IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 13, No. 4, April, 2026, pp. 652-662
Journal homepage: http://iieta.org/journals/mmep

Random Forest–Based Multilevel Models with and without Lagged Responses in Longitudinal Data

Siti N. Azizah[1], Budi Susetyo[1*], Anwar Fitrianto[1], Irsyad Zamjani[2]

[1] Statistics and Data Science Study Program, School of Data Science, Mathematics, and Informatics, IPB University, Bogor 16680, Indonesia
[2] The Center for Educational Standards and Policy, Ministry of Primary and Secondary Education, Jakarta 12410, Indonesia

Corresponding Author Email: budisu@apps.ipb.ac.id

https://doi.org/10.18280/mmep.130404

Received: 24 February 2026
Revised: 20 April 2026
Accepted: 27 April 2026
Available online: 15 May 2026

Keywords:
lagged response, longitudinal prediction, Mixed-Effects Random Forest, Stochastic Mixed-Effects Random Forest, national assessment, numeracy achievement, school-level heterogeneity

ABSTRACT

Longitudinal educational data are often characterized by temporal dependence and substantial heterogeneity across observational units, posing challenges for conventional parametric modelling. This study compares the predictive performance of Linear Mixed Models (LMMs), Mixed-Effects Random Forest (MERF), and Stochastic Mixed-Effects Random Forest (sMERF), both with and without a lagged response, for forecasting school-level numeracy achievement. The analysis uses Indonesia's National Assessment data from 1,384 senior high schools in West Java Province observed between 2021 and 2024. A total of 63 predictors were derived from the Learning Environment Survey, and prior-year numeracy scores were incorporated as lagged responses to capture temporal persistence in school performance. To respect the chronological structure of the data, models were assessed using a temporal hold-out design, with observations from 2024 reserved for testing. The results show that the two Random Forest–based mixed-effects approaches outperform the parametric LMM. Among all specifications, MERF with a lagged response achieves the lowest prediction error, with a mean absolute percentage error of 5.11%. Variable importance results indicate that prior-year numeracy is the strongest predictor, confirming the persistence of school-level achievement over time. Several learning-environment variables, including cultural equality, inclusivity, and diversity climate, also contribute meaningfully to prediction. These findings suggest that combining historical achievement information with school-contextual indicators can improve prediction-oriented longitudinal modelling in educational assessment.

1. INTRODUCTION

In the era of big data, analytical challenges arise not only from data volume but also from structural complexity. When researchers measure observational units repeatedly over multiple time periods, the resulting observations become statistically dependent. This structure defines longitudinal data, which records repeated measurements over time within consistent units such as individuals, institutions, or regions [1].

Longitudinal data exhibit a two-level hierarchical structure in which time-specific measurements (Level 1) are nested within observational units (Level 2) [2]. Although this configuration resembles hierarchical clustering (e.g., students nested within teachers or schools) [3], longitudinal data differ in that they explicitly incorporate a temporal dimension. Their defining characteristic is the repeated observation of the same unit across time.

Longitudinal data analysis enables researchers to examine within-unit change while accounting for heterogeneity across units [4, 5]. Consequently, appropriate analytical frameworks must address both within-unit dependence and between-unit heterogeneity. Multilevel or mixed-effects models provide a structured approach to accommodate these features [6]. Linear Mixed Models (LMMs), also referred to as multilevel or hierarchical models, are widely used to analyze clustered and longitudinal data. Within this framework, fixed and random effects are integrated in a unified parametric formulation [7].

Building on the LMM framework, generalized LMMs and their semiparametric extensions enable more flexible modeling of hierarchical dependence and nonlinear trajectories [8]. Dynamic extensions and state-space approaches have also been explored for temporal prediction in complex time series [9]. However, these parametric formulations remain sensitive to model misspecification and may struggle to capture complex nonlinear relationships [10].

Parallel to developments in mixed modeling, time-series forecasting methods have evolved substantially. Traditional statistical models, such as Autoregressive Integrated Moving Average (ARIMA) and its extension with exogenous variables (ARIMAX), are effective for capturing temporal dependence in relatively low-dimensional settings, particularly under approximate linearity assumptions. More recently, deep learning architectures, including recurrent neural networks and long short-term memory networks, have demonstrated

strong capabilities for learning complex nonlinear temporal patterns without explicit parametric assumptions [11].

Despite these methodological advances, significant challenges remain for longitudinal educational data, which often combine hierarchical dependence, temporal correlation, and complex nonlinear relationships. Classical time-series models are not explicitly designed to accommodate unit-specific random effects in hierarchical structures. In contrast, deep learning approaches typically require large training samples and extensive hyperparameter tuning. These considerations are particularly relevant in educational forecasting contexts, where data often exhibit multilevel structure and moderate sample sizes.

To address these challenges, mixed-effects machine learning approaches have emerged that combine the strengths of tree-based machine learning with random-effects modeling. Such hybrid frameworks are particularly well-suited for hierarchical data with repeated measurements and complex covariate interactions.

Within this methodological context, Mixed Effects Random Forest (MERF) introduces a hybrid framework that integrates the hierarchical structure of mixed-effects models with the nonlinear predictive capacity of tree-based machine learning [12]. The standard formulation was originally developed for clustered data settings. Empirical studies indicate that MERF improves predictive performance relative to standard random forests when random effects contribute substantially to outcome variation [13, 14].

Despite its strong predictive performance, the standard MERF formulation has several limitations when applied to longitudinal data. Although initially developed for clustered data, the hierarchical structure of MERF aligns conceptually with longitudinal data, in which repeated measurements over time are nested within the same observational units. This structural compatibility supports the extension of MERF to longitudinal prediction settings.

Lewis et al. [15] applied MERF to predict patient depression scores using repeated measurements and reported improved predictive accuracy relative to standard random forests. Their specification followed the standard MERF formulation, which assumes independence of residual errors across time points and attributes temporal dependence primarily to unit-specific random effects.

When random effects adequately capture within-unit heterogeneity, this assumption may be sufficient. However, the framework does not explicitly model serial correlation across repeated observations. In longitudinal settings where residual temporal dependence persists beyond random effects, the standard MERF formulation may therefore provide an incomplete representation of temporal dynamics.

To address this limitation, Capitaine et al. [16] proposed Stochastic MERF (sMERF), which extends MERF by incorporating a structured stochastic process into the residual component. This modification enables explicit modeling of serial correlation within observational units and is particularly suitable for longitudinal data exhibiting residual temporal dependence. Simulation and empirical studies indicate that sMERF improves predictive performance when temporal correlation is substantial, whereas MERF performs comparably when such correlation is weak or absent [17, 18].

Beyond residual correlation, temporal dependence in longitudinal data may also manifest directly in the response trajectory. In many applied settings, earlier response values may contain information about the underlying dynamic process. Consequently, lagged responses are commonly incorporated to capture temporal dependence across time periods [1, 19]. In parametric dynamic models, the lagged response serves as a central mechanism for representing temporal dynamics.

Within machine learning frameworks, lagged responses are typically introduced through feature engineering as additional predictors rather than as structural autoregressive parameters. This strategy enables models to exploit historical information while avoiding strong parametric assumptions about the temporal process.

Stochastic residual modeling and lagged responses both address temporal dependence, but they operate through distinct mechanisms. Lagged responses represent persistence at the response level, whereas sMERF captures temporal dependence at the error level through a structured stochastic process. Empirical research examining the joint integration of lagged responses within the MERF and sMERF frameworks remains limited, and direct comparisons between these two temporal mechanisms in longitudinal prediction contexts are scarce.

This study uses educational data as a case study to illustrate the proposed longitudinal prediction framework. Specifically, the analysis employs National Assessment data that provide annual school-level numeracy measurements. The response variable is continuous and follows a hierarchical longitudinal structure. The educational context serves primarily as an applied setting for methodological comparison, while selected substantive interpretations are provided to illustrate the practical relevance of the predictors. This setting offers real-world longitudinal data with observable temporal dependence and heterogeneity across units.

Based on this background, the study evaluates and compares the predictive performance of LMM, MERF, and sMERF under specifications that include and exclude lagged responses as additional features. The analysis focuses on prediction rather than causal inference and examines how historical information and distinct temporal dependence mechanisms contribute to predictive accuracy in hierarchical longitudinal data.

2. MATERIALS AND METHOD

2.1 Data description

This study used longitudinal data from the West Java Provincial National Assessment for senior high schools covering the period 2021–2024. The dataset was provided by the Ministry of Primary and Secondary Education. The final analytical sample comprised 1,384 schools observed consistently over four consecutive years.

The response variable was the school-level numeracy score. Numeracy is defined as students' ability to apply mathematical concepts, procedures, facts, and tools to solve everyday problems in contexts relevant to individuals as citizens in national and global settings. A total of 63 predictors were derived from the Learning Environment Survey conducted by the Indonesian Ministry of Primary and Secondary Education, covering six educational-quality dimensions: learning quality, learning reflection and improvement, instructional leadership, safety climate, diversity climate, and inclusiveness climate.

Learning quality captures the quality of instructional interactions among teachers, students, and learning materials

during the teaching–learning process. Learning reflection and improvement refers to teachers' reflective practices and ongoing efforts to enhance instruction. Instructional leadership reflects principals' roles in supporting improvements in instructional quality. Safety climate describes school conditions that promote students' physical protection and psychological safety. Diversity climate reflects the attitudes and behaviors of principals, teachers, and students in fostering religious and cultural tolerance and national commitment. Inclusiveness climate concerns how the school environment responds to student diversity, including differences in individual characteristics, identity, and socio-cultural background.

The 2021 numeracy score was used solely as a lagged response variable in model development and was not treated as a contemporaneous outcome in subsequent years. The study included schools that met the following criteria: (i) public or private senior high school status, (ii) a minimum participation rate of at least 85%, and (iii) complete data for all predictors and response variables throughout the observation period.

2.2 Multilevel models and Random Forest

Multilevel models, such as LMMs, extend conventional linear regression by incorporating random effects to represent dependence among repeated observations within the same unit [10]. The LMM formulation is given by:

$$y_{it} = X_{it}\beta + Z_{it}^{T}b_i + \varepsilon_{it} \quad (1)$$

where, y_{it} denotes the response for unit i at time t, X_{it} is the fixed-effects covariate vector, β is the fixed-effects parameter vector, Z_{it} is the random-effects covariate vector, b_i represents the random effects for unit i, and ε_{it} is the residual error term. The random effects and residuals are assumed to follow $b_i \sim N(0, D)$ and $\varepsilon_{it} \sim N(0, \sigma^2)$, respectively. Let n denote the total number of observations. Matrix D represents the covariance of the random effects, and R_i denotes the covariance matrix of the residual errors.

The LMM also provides the intraclass correlation coefficient (ICC), which helps assess the suitability of multilevel modeling. However, parametric approaches such as LMM may have limited flexibility in capturing nonlinear relationships and complex interactions among predictors.

Random Forests, introduced by Breiman [20], learn nonlinear patterns without requiring explicit distributional assumptions. The method is robust to noise and multicollinearity and provides measures of variable importance [21]. However, standard Random Forest does not explicitly accommodate hierarchical structure or residual temporal dependence.

2.3 Model specifications

2.3.1 Mixed Effects Random Forest

The MERF model proposed by Hajjem et al. [12] for continuous responses integrates Random Forest with mixed-effects modeling to simultaneously capture nonlinear predictor–response relationships and hierarchical unit heterogeneity. In this framework, a Random Forest ensemble models the nonlinear fixed-effects component, whereas the random effects are specified parametrically following an LMM structure. The standard MERF formulation for clustered data is

$$y_i = f(X_i) + Z_i^{T}b_i + \varepsilon_i \quad (2)$$

where, $f(X_i)$ denotes the nonlinear function estimated by Random Forest, and the random-effects component retains a parametric specification analogous to that of LMM. The MERF algorithm proceeds through an iterative estimation scheme similar to the Expectation–Maximization procedure commonly used in mixed-effects modeling.

Although the original MERF formulation targets clustered data without explicit temporal indexing, the framework extends naturally to longitudinal settings in which repeated observations are indexed over time within the same hierarchical structure. Under this notation, the model becomes

$$y_{it} = f(X_{it}) + Z_{it}^{T}b_i + \varepsilon_{it} \quad (3)$$

where, t indexes repeated measurements for unit i. This representation preserves the fundamental MERF structure while enabling its application to repeated-measures data.

2.3.2 Stochastic Mixed Effects Random Forest

The sMERF model proposed by Capitaine et al. [16] extends the MERF framework by incorporating an explicit stochastic process component to capture temporal dependence in longitudinal data. Although MERF accommodates hierarchical heterogeneity through random effects, it assumes that residual errors are independent across time. The sMERF formulation relaxes this assumption by introducing a temporally correlated stochastic component. The general sMERF formulation for longitudinal data is

$$y_{it} = f(X_{it}) + Z_{it}^{T}b_i + \omega_{it} + \varepsilon_{it} \quad (4)$$

In this model, b_i, ω_{it}, and ε_{it} are assumed to be mutually independent. The residual term follows $\varepsilon_{it} \sim N(0, \sigma^2)$, and the random effects follow $b_i \sim N(0, D)$. The stochastic component ω_{it} is modeled as a Gaussian process with covariance kernel K_i to capture temporally correlated variability within observational units. For each unit i, the stochastic vector

$$\omega_i = (\omega_{i1}, \omega_{i2}, \ldots, \omega_{iT})$$

follows the multivariate normal distribution

$$\omega_i \sim N(0, \gamma^2 K_i)$$

where, γ^2 denotes the stochastic variance parameter and K_i is a positive-definite kernel matrix defining temporal dependence across repeated observations. The kernel matrix encodes the temporal correlation structure and can be specified using stochastic processes such as Brownian motion (BM) or Ornstein–Uhlenbeck (OU) dynamics, allowing the model to represent smooth temporal evolution and mean-reverting behavior.

When the stochastic component is absent ($\omega_{it} = 0$), the sMERF formulation reduces to the standard MERF model. Thus, MERF constitutes a special case of sMERF in the absence of temporally structured stochastic variability.

2.3.3 Lagged responses as predictor features

Longitudinal data typically exhibit temporal dependence because the same units are observed repeatedly over time. Consequently, past response values often contain information that is useful for predicting future outcomes.

Temporal dependence in longitudinal modeling can be accommodated through several approaches [2]. The first approach models autocorrelation in the error structure by specifying residual correlation, for example, through autoregressive error processes. This strategy focuses on residual temporal dependence after accounting for covariates and random effects.

The second approach uses dynamic or autoregressive longitudinal models that incorporate previous responses as structural components. A general autoregressive LMM formulation is [2]

$$y_{it} = \rho y_{i,t-1} + X_{it}\beta + Z_{it}^{T} b_i + \varepsilon_{it} \quad (5)$$

where, ρ denotes the autoregressive coefficient governing linear temporal dependence.

In supervised machine learning frameworks, temporal information is commonly incorporated through feature engineering. This strategy constructs derived variables that encode historical information as additional predictors, without explicitly specifying an autoregressive structure in either the mean function or the error term. Transforming temporal information into features, such as lagged responses, allows nonlinear predictive algorithms to exploit historical context while avoiding strong parametric assumptions about the temporal process.

Prediction-oriented studies support the inclusion of lagged responses as informative predictors. Moges et al. [22] showed that Random Forest models incorporating lagged discharge variables achieved higher predictive accuracy than models based solely on contemporaneous predictors.

Building on this evidence, the present study integrates the lagged response into the MERF and sMERF frameworks. Operationally, the lagged response is constructed by aligning each observation with the response from the previous period within the same unit. For the first observation in each unit, the lagged value is undefined; these observations were therefore excluded or handled according to the modeling requirements (Table 1).

Table 1. Illustration of lag construction within each observational unit

Unit (*i*)	Time (*t*)	y_{it}	$y_{i,t-1}$
A	2021	50	-
A	2022	45	50
A	2023	67	45
B	2021	89	-
B	2022	34	89
B	2023	77	34
...	...	...	...
...	...	...	...

The study focuses on predictive modeling rather than parametric inference. Under this feature-based formulation, the Random Forest component directly incorporates the lagged response into the nonlinear predictor function. The MERF model with the lagged response is given by

$$y_{it} = f\left(X_{it}, Y_{i,t-1}\right) + Z_{it}^{T} b_i + \varepsilon_{it} \quad (6)$$

For the stochastic extension, the sMERF formulation becomes

$$y_{it} = f\left(X_{it}, Y_{i,t-1}\right) + Z_{it}^{T} b_i + \omega_{it} + \varepsilon_{it} \quad (7)$$

In this specification, the model uses the lagged response as an input variable to enhance predictive performance by incorporating historical context. The stochastic component ω_{it} and residual term ε_{it} retain their respective roles in representing temporally structured stochastic dependence and idiosyncratic noise. This formulation enables the MERF and sMERF frameworks to jointly capture nonlinear predictor–response relationships, unit-level heterogeneity, and temporal dynamics in longitudinal data.

2.4 Procedure of data analysis

All preprocessing and modeling procedures were conducted in R version 4.5.1 using the RStudio 2025.09.1+40 interface. The analysis primarily relied on the *longituRF* package developed by Capitaine et al. [16]. The following subsections describe each stage of the analytical workflow.

2.4.1 Data preparation

(a) Integration of annual data (2021–2024). Annual data from 2021 to 2024 were integrated into a longitudinal structure. All records were merged into a single long-format dataset, in which the school identifier served as the Level-2 unit and the observation year as the Level-1 unit. Each school–year combination corresponded to one row of data. Because four years of observations were available, each school contributed four rows reflecting changes in characteristics and numeracy scores over time. This long-format structure was required for multilevel modeling and longitudinal random forests to capture temporal dynamics at the school level.

(b) Construction of lagged response. To represent temporal dynamics in numeracy outcomes, each observation was assigned a lagged response equal to the school's numeracy score from the previous year. Table 1 illustrates this construction. For a given year (e.g., 2022), the dataset included an additional column containing the prior-year score (2021). This procedure was applied consistently across all years. The lag-based specification utilized the full 2021–2024 period for model training to fully exploit longitudinal information.

(c) Compilation of analysis datasets. The preprocessing stage produced two datasets: one including the lagged response (64 predictors) and one excluding it (63 predictors).

(d) Assessment of data completeness. All predictor and response variables were screened for missing values to ensure data completeness.

(e) Exploratory data analysis. Descriptive analyses were conducted to examine the distributional characteristics of predictors and outcomes and to explore preliminary relationships among variables.

(f) Multicollinearity assessment. As part of predictor screening, multicollinearity was evaluated using the variance inflation factor (VIF) computed from the pooled longitudinal dataset. Predictors exceeding the conventional threshold (VIF > 10) were flagged as potentially redundant. These diagnostics were used as screening tools rather than as strict exclusion criteria.

(g) Intraclass correlation analysis. The ICC was computed to justify the use of multilevel modeling before incorporating random effects. The ICC quantifies the proportion of total variance attributable to between-group differences, such as variation across schools [23]. For a random-intercept model, the ICC is defined as

$$ICC = \frac{\sigma^2_{between}}{\sigma^2_{between} + \sigma^2_{within}}$$

where, $\sigma^2_{between}$ denotes the between-school variance, and σ^2_{within} denotes the within-school variance.

2.4.2 Modeling and evaluation

(a) Model training. The LMM, MERF, and sMERF models were trained on two versions of the dataset: with and without the lagged response. This design enabled the evaluation of the extent to which historical numeracy information improves predictive accuracy within the multilevel Random Forest framework.

(b) Model structure specification. The fixed-effects component was modeled nonparametrically using Random Forest to accommodate nonlinear and high-dimensional predictor–response relationships. Random effects were specified as school-level random intercepts to represent Level-2 variability across schools. Before iterative estimation, a hyperparameter grid was defined to explore model performance. The search space included: (a) mtry: $p, p/3, p/5$ with $p = 63 - 64$ predictors; (b) *ntree*: 100, 200, 300, 500; (c) *iter*: 20, 50, 100; and (d) *delta*: 1×10^{-4}.

(c) Stochastic process specification via the *sto* argument. The primary distinction between MERF and sMERF lies in the stochastic component specified for the residual structure. The *sto* argument determines the type of stochastic process, with the following options: (a) "none", corresponding to MERF with no stochastic process; (b) "BM", representing Brownian motion with persistent temporal correlation and no mean reversion; and (c) "OrnUhl", representing Ornstein–Uhlenbeck dynamics with mean-reverting temporal dependence. This selection was intended to capture alternative temporal dependence patterns while maintaining interpretability and computational feasibility.

(d) Iterative model estimation. Model estimation followed an iterative algorithm until convergence. At each iteration, a Random Forest was fitted using the specified values of *ntree* and *mtry* to estimate the fixed-effects component. The algorithm then updated the random effects and stochastic components using estimates from the previous iteration. Convergence was declared when parameter changes fell below the tolerance threshold (*delta*). Predicted numeracy scores were obtained by summing three components: (i) the fixed-effects contribution from the Random Forest, (ii) the school-level random intercept, and (iii) the stochastic process component capturing serial correlation among repeated measurements.

(e) Temporal hold-out validation. Temporal hold-out validation was applied to preserve chronological ordering and prevent the use of future observations during model training. Models were trained on observations from 2022–2023 and evaluated on 2024 data to assess predictive performance on unseen future observations. This strategy preserves temporal dependence, unlike random resampling approaches such as k-fold cross-validation, which may introduce bias when applied to temporally structured data [24].

(f) Performance evaluation and uncertainty quantification. Model performance on the 2024 test set was evaluated using multiple accuracy metrics, including mean absolute percentage error (MAPE), root mean squared error (RMSE), and the coefficient of determination (R^2), to provide complementary perspectives on predictive performance, as commonly recommended in recent forecasting studies [25]. To quantify uncertainty, a cluster-based bootstrap procedure was implemented. Resampling was performed at the observational-unit level with replacement to preserve the hierarchical and longitudinal dependence structure. For each bootstrap replicate, model predictions and performance metrics were recomputed. Percentile-based 95% confidence intervals were derived from 1,000 bootstrap samples.

(g) Variable importance analysis**.** Variable importance measures from the Random Forest component were used to identify the most influential predictors. Importance scores were computed based on the increase in mean squared error (MSE) following permutation of each predictor.

The methodological workflow is illustrated in Figure 1.

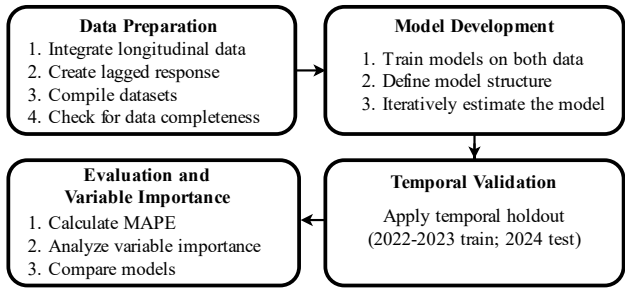


Figure 1. Workflow of the proposed method

3. RESULTS AND DISCUSSION

3.1 Descriptive statistics of response variables

Before model estimation, an exploratory analysis of the response variable was conducted to examine its temporal structure and between-unit variability. This step aimed to characterize the longitudinal patterns in numeracy outcomes that motivate the use of a longitudinal predictive framework.

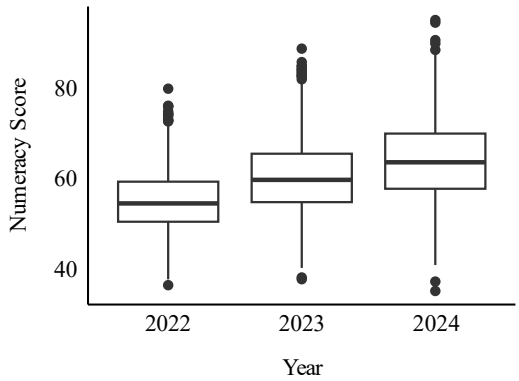


Figure 2. Distribution of school-level numeracy scores across 2022–2024, illustrating overall score dispersion and year-to-year variability

Apparent temporal variation in school-level numeracy scores is observed across the study period (Figure 2). The distribution shifts over time, as reflected in changes in median values and interquartile ranges. The median numeracy score

exhibits a gradual increase across successive years, suggesting a systematic temporal pattern.

The spread of the box plots across schools also reflects substantial between-unit variability, indicating heterogeneity in response levels. This pattern suggests that response values vary meaningfully across units and supports treating the data within a longitudinal framework rather than as independent cross-sectional observations. The observed variability further supports the inclusion of school-level random effects in the mixed-effects predictive modeling framework.

Substantial heterogeneity is evident in the longitudinal trajectories across schools. While some schools exhibit relatively stable trajectories, others show noticeable increases or fluctuations over time. These descriptive findings provide empirical support for the use of multilevel modeling, as the response exhibits meaningful between-school variation alongside within-school temporal dependence (Figure 3). The observed heterogeneity supports the inclusion of school-level random effects to represent differences in baseline performance.

Moreover, the presence of distinct temporal trajectories highlights the value of incorporating longitudinal information into the predictive modeling framework, particularly through stochastic components or lagged-response features that capture within-unit temporal dynamics.

A strong positive association is observed between current numeracy scores and their one-year lagged values (Figure 4). This pattern suggests that schools with higher numeracy performance in the previous year tend to maintain relatively higher performance levels in subsequent years, indicating persistence in school-level achievement.

This relationship is further quantified through year-specific correlations between numeracy scores at time t and $t-1$. The results indicate consistently strong positive associations, with correlation coefficients of 0.86 in 2022, 0.88 in 2023, and 0.88 in 2024. The similarity and magnitude of these coefficients suggest that the observed dependence is not driven by a single transition period but instead reflects a persistent temporal pattern. Taken together, the graphical and statistical evidence support the inclusion of the lagged response in the predictive modeling framework.

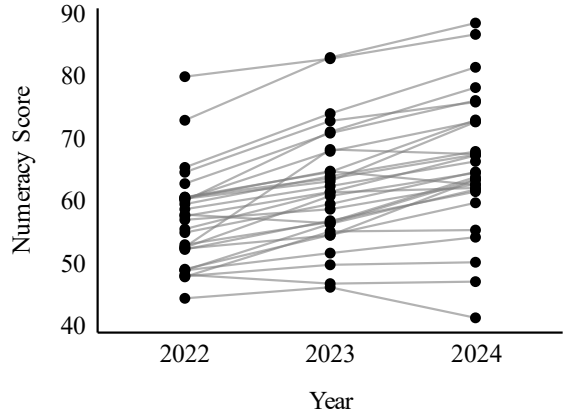


Figure 3. Longitudinal trajectories of numeracy scores for 30 schools (2022–2024), illustrating heterogeneous performance trends

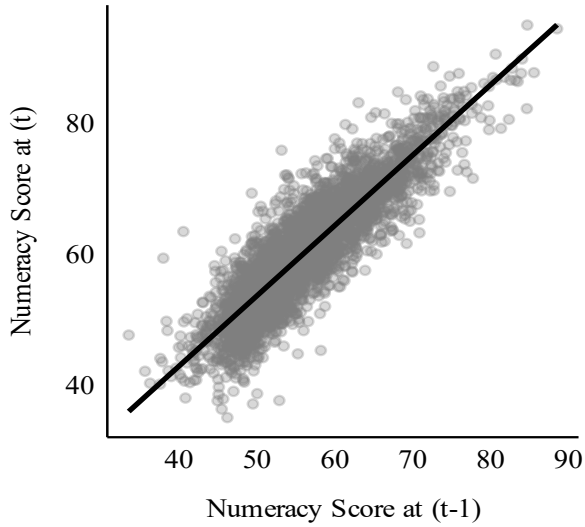


Figure 4. Relationship between current numeracy scores at time (t) and one-year lagged values $(t-1)$, illustrating strong temporal persistence in school-level achievement

Table 2. Distributional summary of school-level indicators by dimension

Dimension	Number of Predictors	Mean Score Range	Standard Deviation Range	Min Score Range	Max Score Range	Skewness Range
Learning quality	21	(51.94–79.39)	(4.56–12.79)	(21.05–52.60)	(85.56–100.00)	(−0.30–1.33)
Learning reflection and improvement	3	(57.44–64.21)	(6.83–7.61)	(33.08–37.26)	(96.31–100.00)	(0.20–0.88)
Instructional leadership	3	(54.62–59.66)	(7.30–10.21)	(31.51–38.75)	(94.42–94.58)	(0.30–0.73)
Safety climate	17	(60.29–87.57)	(6.28–15.32)	(2.26–49.00)	(92.78–100.00)	(−1.24–0.78)
Diversity climate	9	(53.45–88.46)	(4.72–10.04)	(22.32–48.24)	(91.00–100.00)	(−0.77–0.88)
Inclusiveness climate	10	(49.85–75.35)	(5.97–11.32)	(11.00–39.18)	(85.00–100.00)	(−0.58–0.91)

Note: All indicators are measured at the school level and scaled from 0 to 100.

3.2 Descriptive statistics of predictor variables

The 63 predictors used in this study were grouped into six main dimensions, as presented in Table 2. Overall, school-level predictors exhibit moderate to high average levels, although the degree of variation differs across dimensions.

In the learning quality dimension, which comprises 21 predictors, the relatively wide ranges of means and standard deviations compared with several other dimensions indicate substantial variability in learning quality across schools. The minimum value of 21.05 suggests that some schools perform relatively poorly on specific indicators, whereas the maximum value of 100 reflects the presence of very high-performing schools. Skewness values ranging from negative to positive indicate that no single direction of asymmetry dominates the distribution of learning quality indicators.

The learning reflection and improvement and instructional leadership dimensions exhibit comparatively more homogeneous patterns, as reflected in their narrower ranges of means and standard deviations. These predictors tend to

cluster around moderate levels, although some between-school variation remains. The maximum instructional leadership value (approximately 94.58) further indicates the presence of high-performing schools in this dimension.

In contrast, the safety climate dimension, which includes 17 predictors, displays greater heterogeneity. The relatively high mean range suggests that school safety conditions are generally favorable. However, the wider dispersion and the very low minimum value (approximately 2.26) indicate that safety and preparedness are not evenly distributed across schools. The presence of negative skewness values further suggests that several indicators are concentrated at higher levels.

For the diversity climate dimension (nine predictors), the overall pattern also indicates relatively conducive conditions. The high mean range, accompanied by skewness values ranging from near zero to positive, suggests that most schools demonstrate moderate to high levels of tolerance and national commitment, although inter-school variation remains evident.

Finally, the inclusiveness climate dimension, comprising 10 predictors, exhibits notable internal variability. This is reflected in the relatively wide ranges of means and standard deviations, as well as skewness values spanning from negative to positive. These patterns indicate that inclusivity-related indicators have improved in some schools, although the degree of strengthening across aspects remains uneven.

Taken together, the observed dispersion across schools, particularly in the safety and diversity climate dimensions, suggests meaningful between-school heterogeneity, providing preliminary support for modeling approaches that explicitly account for the hierarchical structure of the data. In addition to between-school variability, relationships among predictors were examined to assess potential redundancy.

VIF diagnostics indicated that severe multicollinearity was not widespread, although three predictors exceeded the conventional threshold of 10. Only three of the 63 variables exceeded the conventional threshold of 10: teaching enjoyment (teacher-reported; VIF = 14.33), classroom climate conducive to learning (student-reported; VIF = 11.81), and constructive feedback (teacher-reported; VIF = 10.38). Given the known robustness of Random Forest–based models to correlated predictors, all variables were retained for subsequent modeling.

3.3 Intraclass correlation coefficient

The ICC values were estimated using a random-intercept LMM under two specifications: (i) an unadjusted (null) model without covariates to represent baseline between-school variability and (ii) an adjusted model including predictors to assess whether covariate adjustment altered the variance decomposition. In practice, both specifications yielded identical ICC estimates, indicating that between-school heterogeneity remained substantial after accounting for the observed predictors.

For the full dataset, the ICC was 0.613, indicating that more than 60% of the total variability in numeracy scores was attributable to differences between schools. After partitioning the data into training and test subsets, the ICC was recalculated using the training data to verify that the hierarchical structure was preserved. The resulting ICC of 0.661 suggests that the training subset retained a similar clustering pattern.

These ICC findings are consistent with the descriptive evidence of substantial between-school dispersion observed across several predictor dimensions, particularly safety climate and diversity climate.

Table 3. Predictive performance comparison of LMM, MERF, and sMERF models with 95% confidence intervals (CI)

Model	Stochastic Type	Without a Lagged Response			With a Lagged Response		
		RMSE (95% CI)	MAPE (95% CI)	R-Square	RMSE (95% CI)	MAPE (95% CI)	R-Square
LMM	-	7.63 (7.39–7.88)	9.56 (9.30–9.90)	28.13	4.75 (4.55–4.94)	5.80 (5.58–6.03)	72.17
MERF	none	6.88 (6.65–7.13)	8.28 (8.02–8.54)	41.55	4.42 (4.22–4.64)	5.11 (4.90–5.32)	75.86
sMERF	Brownian motion (BM)	7.01 (6.78–7.24)	8.53 (8.26–8.80)	39.24	4.55 (4.34–4.78)	5.23 (5.03–5.47)	74.43
sMERF	OrnUhl	7.01(6.76–7.26)	8.39 (8.11–8.67)	39.30	4.53 (4.31–4.76)	5.24 (4.99–5.48)	74.61

Notes: 1. MERF (without lag): mtry = 21, ntree = 300, iter = 20; 2. sMERF-BM (without lag): mtry = 21, ntree = 100, iter = 50; 3. sMERF-OrnUhl (without lag): mtry = 21, ntree = 100, iter = 100; 4. MERF (with lag): mtry = 21, ntree = 500, iter = 50; 5. sMERF-BM (with lag): mtry = 21, ntree = 100, iter = 100; 6. sMERF-OrnUhl (with lag): mtry = 21, ntree = 200, iter = 100; 7. MAPE and R^2 are reported in percentage (%). Bootstrap CI were computed for RMSE and MAPE; R^2 is reported as a point estimate.

LMMs: Linear Mixed Models; MERF: Mixed-Effects Random Forest; sMERF: Stochastic Mixed-Effects Random Forest; MAPE: mean absolute percentage error, RMSE: root mean squared error, and the R^2: coefficient of determination.

3.4 Performance evaluation

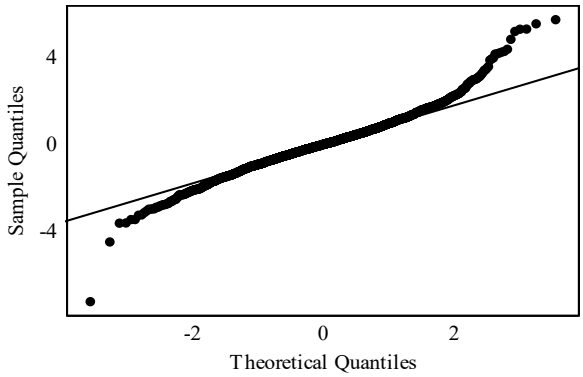
Model performance was evaluated under two specifications: without the lagged response and with a first-order lag. Detailed results for both specifications are presented in Table 3. In the baseline specification, the LMM exhibited the weakest predictive accuracy across all metrics, with the highest RMSE and MAPE and the lowest R^2. MERF substantially improved performance, reducing RMSE from 7.63 to 6.88 and increasing R^2 from 28.13% to 41.55%. The largely separated confidence intervals suggest that the performance improvement is unlikely to be due only to sampling variability. The sMERF variants with BM and Ornstein–Uhlenbeck (OrnUhl) stochastic processes also outperformed LMM but remained comparable to MERF and slightly below it.

Incorporating the lagged response markedly improved predictive accuracy across all models. For example, MAPE for LMM decreased from 9.56% to 5.80%, while MERF achieved the lowest error (5.11%), followed closely by the sMERF variants (approximately 5.23%–5.24%). Similar improvements were observed for RMSE and R^2.

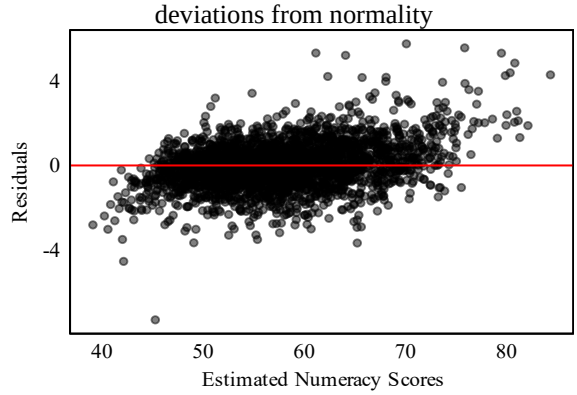
Despite the inclusion of stochastic residual structures, sMERF did not outperform MERF under the lagged specification. Although MERF achieved the lowest mean errors, the overlapping confidence intervals indicate that performance differences among the hybrid models are modest and should be interpreted with caution.

Residual diagnostics were conducted to further assess model adequacy. Overall, the diagnostics do not indicate substantial model misspecification for predictive purposes. The diagnostic plots were used as descriptive checks of residual behavior rather than as formal tests of distributional

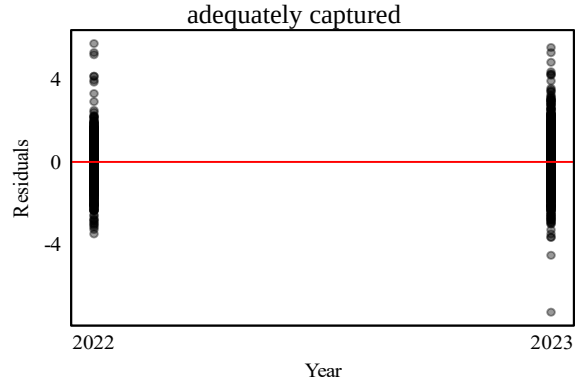
assumptions. (Figure 5(a)). However, the residuals plotted against the fitted values show no clear systematic pattern (Figure 5(b)), suggesting that the model adequately captures the nonlinear mean structure. Similarly, residuals over time remain centered around zero with no visible temporal trend (Figure 5(c)), indicating no strong evidence of remaining unmodeled longitudinal dependence.



(a) Normal Q–Q plot of residuals showing minor tail deviations from normality



(b) Residuals versus fitted values showing no clear systematic pattern, suggesting that the functional form is adequately captured



(c) Residuals over time remaining centered around zero with no visible temporal trend, indicating no substantial unmodeled longitudinal dependence

Figure 5. Residual diagnostics for the best-performing Mixed-Effects Random Forest (MERF) model with the lagged response

Overall, the MERF model with the lagged response achieved the lowest prediction error among the evaluated models. The results highlight three main patterns: (i) MERF and sMERF consistently outperformed LMM in longitudinal prediction; (ii) incorporating the lagged response substantially improved predictive accuracy across model classes; and (iii) once response-level temporal dependence was captured through the lagged response, MERF performed comparably to or slightly better than the stochastic extensions.

3.5 Variable importance analysis

Following the performance evaluation, a variable-importance analysis was conducted to examine how predictors contribute to longitudinal numeracy prediction within the MERF and sMERF frameworks.

Importantly, variable-importance patterns remained qualitatively similar across sMERF configurations with different stochastic residual specifications. Because the Random Forest component is identical across these models, only the MERF results are presented for clarity. This consistency suggests that the ranking of influential predictors is driven primarily by the fixed-effects component rather than by differences in the stochastic residual specification.

Without the lagged response, several non-cognitive school environment indicators consistently ranked among the most influential predictors, suggesting that contemporaneous covariates provided the primary predictive signal for numeracy outcomes (Figure 6).

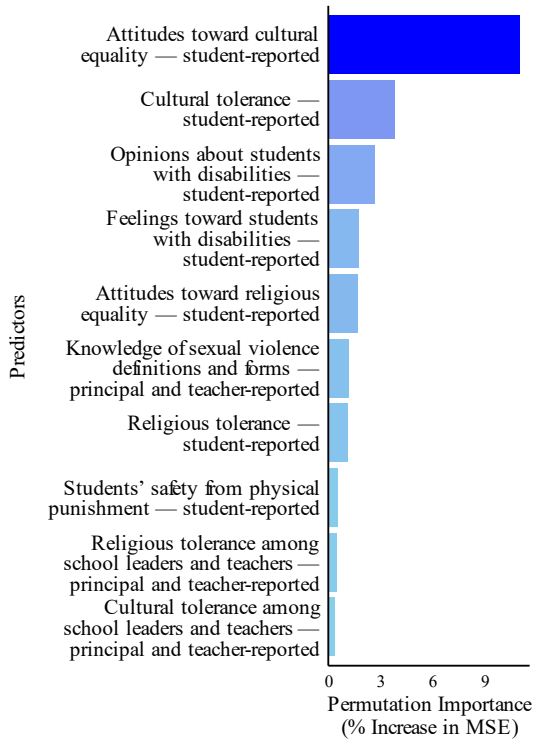


Figure 6. Top 10 predictor importance rankings in the Mixed-Effects Random Forest (MERF) without lag, highlighting the relative contributions of contemporaneous school-environment indicators to predictive performance

After incorporating the lagged response, the previous year's numeracy score became the dominant predictor, substantially exceeding the contributions of other covariates (Figure 7). This shift is consistent with strong temporal persistence, whereby historical achievement accounts for a large share of the predictive information for subsequent performance.

Overall, the variable-importance analysis complements the performance evaluation by indicating that improvements in predictive accuracy are largely associated with the explicit incorporation of response history. This finding reinforces the role of the lagged response as a key feature for longitudinal

prediction within the MERF and sMERF frameworks, particularly in short-panel settings in which temporal dependence is primarily reflected in historical outcomes rather than in residual correlation structures.

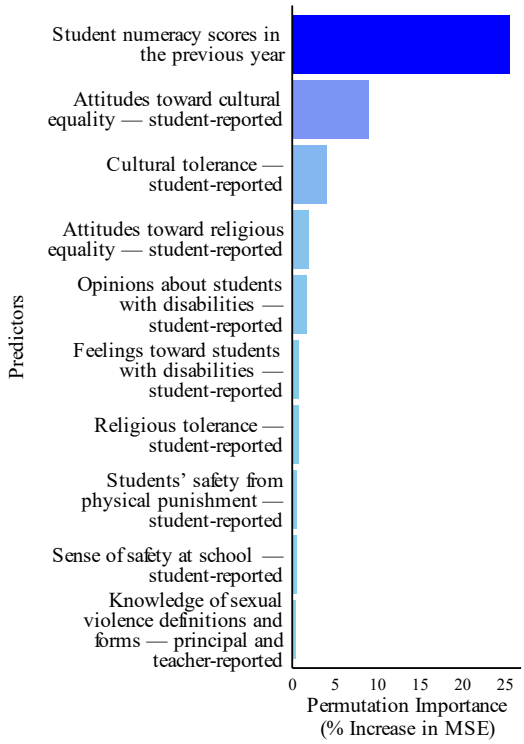


Figure 7. Top 10 predictor importance rankings in the Mixed-Effects Random Forest (MERF) with lag demonstrating the dominant predictive contribution of prior numeracy achievement

Although Random Forest–based models are generally tolerant of correlated inputs, the potential influence of multicollinearity was examined. Predictors flagged as multicollinear during the screening stage did not appear among the top-ranked variables, and their importance scores were substantially lower than those of the leading predictors, indicating limited empirical influence on the overall importance structure.

3.6 Discussion

This study examined longitudinal prediction within the MERF and sMERF frameworks by comparing two mechanisms for capturing temporal dependence: lagged responses and stochastic residual structures. The findings should be interpreted in light of the temporal validation design adopted to assess predictive performance.

Because temporal ordering must be preserved in longitudinal prediction, this study employed temporal hold-out validation rather than random cross-validation. Random cross-validation, although common in machine learning, can introduce information leakage by allowing observations from future time periods to appear in the training set when predicting earlier periods [24]. The temporal split ensures that models are trained only on past observations to predict future outcomes, yielding performance estimates that better reflect real-world deployment. This validation strategy is therefore more appropriate for longitudinal forecasting settings in which predictions are made for future time points.

The earlier ICC analysis indicated substantial between-school variability in numeracy outcomes, confirming the presence of a strong hierarchical data structure. This structural characteristic underscores the importance of modeling approaches that explicitly accommodate clustered longitudinal data.

Across all scenarios, MERF and sMERF consistently outperformed LMM, supporting prior evidence that tree-based ensembles can better capture nonlinear relationships and complex interactions. As a parametric baseline, LMM exhibited lower predictive performance, highlighting the limitations of purely parametric approaches for longitudinal prediction. It is important to note that the present evaluation focuses on predictive performance rather than causal inference, and the relative advantages observed here should be interpreted within a prediction-oriented framework.

Incorporating the lagged response substantially improved predictive accuracy across all models, with MAPE decreasing from approximately 8% to about 5%. This pattern indicates strong temporal persistence in school-level numeracy achievement and suggests that historical response information carries a dominant share of the predictive signal, consistent with principles from dynamic and autoregressive modeling. This interpretation is further supported by the descriptive correlation patterns and graphical diagnostics (Figure 4), which indicate strong year-to-year persistence in numeracy outcomes.

The empirical results indicate that sMERF with a lagged response did not consistently outperform MERF with a lagged response, likely due to the short temporal span of the panel. With only a few time points, the effective memory of serial dependence is limited, reducing the amount of residual temporal structure available for stochastic modeling. The substantial performance improvements observed after incorporating the lagged response suggest that the dominant temporal signal is captured directly through prior outcomes rather than through residual autocorrelation.

Importantly, the overlapping bootstrap confidence intervals indicate that MERF with a lagged response does not uniformly dominate sMERF; rather, both models exhibit broadly comparable predictive performance in this dataset. From a bias–variance perspective, the additional stochastic flexibility in sMERF may introduce extra estimation variability without providing sufficiently large additional predictive signal once the lagged response has already captured most of the temporal persistence.

This interpretation is consistent with the findings of Capitaine et al. [16], who reported that when serial correlation is weak or absent, MERF performs comparably to or better than sMERF. Further evaluation using longer panels would be valuable to determine whether the relative advantage of sMERF becomes more pronounced as the temporal dimension increases.

Consistent with this pattern of temporal persistence, the variable-importance results further reinforce the interpretation. Without the lagged response, several non-cognitive school environment indicators emerged as influential predictors, indicating that the Random Forest component captured meaningful nonlinear associations with numeracy outcomes. After the inclusion of the lagged response, prior numeracy achievement became the dominant predictor, consistent with strong temporal persistence and prior evidence on the cumulative nature of academic performance [26]. Substantively, this pattern highlights the

importance of early learning trajectories, suggesting that timely identification and targeted support for students with low prior numeracy may be critical for improving later achievement.

Beyond prior achievement, socio-psychological indicators related to cultural equality, tolerance, religious equality, and perceptions of students with disabilities remained relevant. This relevance is consistent with earlier descriptive evidence showing notable between-school dispersion in several school-climate-related predictors. These variables reflect school inclusivity and social climate, which have been linked to academic outcomes through student engagement and well-being [27, 28]. From a social–emotional learning perspective, supportive environments may function as enabling contexts for cognitive development [29]. From a policy standpoint, the results suggest that strengthening inclusive school climates may indirectly support numeracy development by fostering more supportive and equitable learning environments.

The limited presence of multicollinearity and its minimal impact on variable importance further support the robustness of the modeling framework. Although a small number of predictors exhibited redundancy, they did not rank among the dominant features, consistent with the known tolerance of tree-based ensembles to correlated inputs. Nevertheless, importance rankings should be interpreted with appropriate caution when correlated predictors are present.

These findings are context-specific and should not be generalized to all longitudinal settings. The relative advantage of stochastic extensions likely depends on the strength of residual serial correlation and the length of the time series. In short-panel settings such as the four-year structure examined here, residual stochastic processes may be less pronounced. Future work using simulation studies or datasets with different temporal properties is needed to clarify when stochastic residual modeling provides additional predictive value.

Finally, the empirical evaluation relied on internal validation within the West Java dataset due to restrictions on access to administrative data. External validation across regions was therefore not conducted. Broader multi-region studies are needed to further assess the generalizability and transferability of the proposed framework.

4. CONCLUSION

Longitudinal modeling of the 2021–2024 National Assessment data in West Java indicates that both MERF and sMERF outperform the baseline LMM, with MERF incorporating the lagged response achieving the lowest prediction error (MAPE = 5.11%) among the evaluated configurations. This result highlights the importance of explicitly incorporating historical response information to improve predictive accuracy within a nonparametric mixed-effects framework.

The findings further suggest that the lagged response functions as an effective temporal memory mechanism in hierarchical longitudinal data, capturing most of the relevant temporal dependence without requiring additional stochastic complexity. Variable-importance analysis also indicates that, beyond historical performance, several non-cognitive predictors contribute meaningfully to prediction, although their substantive interpretation lies beyond the methodological scope of this study.

Overall, integrating a lagged response into the MERF framework provides a practical and computationally efficient strategy for prediction-oriented longitudinal modeling. These results are context-specific and should not be generalized broadly. Future research should examine conditions under which stochastic residual modeling offers additional predictive benefits, particularly through simulation studies or applications to data with alternative temporal structures.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Primary and Secondary Education of the Republic of Indonesia, which provided access to the National Assessment data used in this study.

REFERENCES

[1] Wang, L., Fang, Y., Bergeman, C.S. (2026). Dynamic modeling with intensive longitudinal data: One-step and two-step DSEM approaches. Structural Equation Modeling: A Multidisciplinary Journal, 33(2): 235-251. https://doi.org/10.1080/10705511.2025.2597284

[2] Fu, L., Wang, Y.G., Wu, J. (2024). Recent advances in longitudinal data analysis. In Handbook of Statistics, pp. 173-221. https://doi.org/10.1016/bs.host.2023.10.007

[3] Zhang, L. (2025). Estimating school and teacher effects on students' academic performance using multilevel models with three or more levels: A literature review. Discover Education, 4(1): 4. https://doi.org/10.1007/s44217-024-00392-4

[4] Hu, J., Szymczak, S. (2023). A review on longitudinal data analysis with random forest. Briefings in Bioinformatics, 24(2): 1-11. https://doi.org/10.1093/bib/bbad002

[5] Speiser, J.L. (2021). A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data. Journal of Biomedical Informatics, 117: 103763. https://doi.org/10.1016/j.jbi.2021.103763

[6] Asrirawan, Notodiputro, K.A., Susetyo, B., Oktarina, S.D. (2025). Modified tree-based selection in hierarchical mixed-effect models with trees: A simulation study and real-data application. MethodsX, 14: 103312. https://doi.org/10.1016/j.mex.2025.103312

[7] Jahangiri, M., Kazemnejad, A., Goldfeld, K.S., Daneshpour, M.S., Momen, M., Mostafaei, S., Khalili, D., Akbarzadeh, M. (2025). Leveraging mixed-effects regression trees for the analysis of high-dimensional longitudinal data to identify the low and high-risk subgroups: Simulation study with application to genetic study. BioData Mining, 18: 22. https://doi.org/10.1186/s13040-025-00437-w

[8] Sørensen, Ø., Fjell, A.M., Walhovd, K.B. (2023). Longitudinal modeling of age-dependent latent traits with generalized additive latent and mixed models. Psychometrika, 88(2): 456-486. https://doi.org/10.1007/s11336-023-09910-z

[9] Patharkar, A., Cai, F., Al-Hindawi, F., Wu, T. (2024). Predictive modeling of biomedical temporal data in healthcare applications: Review and future directions. Frontiers in Physiology, 15: 1386760. https://doi.org/10.3389/fphys.2024.1386760

[10] Hu, S., Wang, Y.G., Drovandi, C., Cao, T. (2023). Correction: Predictions of machine learning with mixed-effects in analyzing longitudinal data under model misspecification. Statistical Methods & Applications, 32(2): 713-714. https://doi.org/10.1007/s10260-022-00662-1

[11] Lara-Benítez, P., Carranza-García, M., Riquelme, J.C. (2021). An experimental review on deep learning architectures for time series forecasting. International Journal of Neural Systems, 31(3): 2130001. https://doi.org/10.1142/S0129065721300011

[12] Hajjem, A., Bellavance, F., Larocque, D. (2014). Mixed effects random forest for clustered data. Journal of Statistical Computation and Simulation, 84(6): 1313-1328. https://doi.org/10.1080/00949655.2012.741599

[13] Mayapada, R., Susetyo, B., Sartono, B., Rahmawati. (2021). A comparison between random forest and mixed effects random forest to predict students' math performance in Indonesia. International Journal of Sciences: Basic and Applied Research, 57(1): 1-8.

[14] Boufarh, R., Boursas, F., Bakri, M. (2026). Factor of safety prediction for high road embankments using mixed effects random forest and bee colony optimization. Scientific Reports, 16(1): 6003. https://doi.org/10.1038/s41598-026-35431-7

[15] Lewis, R.A., Ghandeharioun, A., Fedor, S., Pedrelli, P., Picard, R., Mischoulon, D. (2023). Mixed effects random forests for personalised predictions of clinical depression severity. arXiv Preprint arXiv.2301.09815. https://doi.org/10.48550/arXiv.2301.09815

[16] Capitaine, L., Genuer, R., Thiébaut, R. (2021). Random forests for high-dimensional longitudinal data. Statistical Methods in Medical Research, 30(1): 166-184. https://doi.org/10.1177/0962280220946080

[17] Koçoğlu, E.Y. (2024). Longitudinal emotion analysis with Globem dataset. M.S. Thesis. Department of Statistics, Middle East Technical University, Ankara, Turkey.

[18] Locker, K.C.S., Bacon, R.A., Caterino, T.L., Breyfogle, L., Alperet, D.J., Sarkar, P., Piliang, M., Davis, M.G. (2025). Understanding the dandruff flare-up: A cascade of measurable and perceptible changes to scalp health. International Journal of Cosmetic Science, 47(4): 703-717. https://doi.org/10.1111/ics.13067

[19] Pooseh, S., Kalisch, R., Köber, G., Binder, H., Timmer, J. (2024). Intraindividual time-varying dynamic network of affects: Linear autoregressive mixed-effects models for ecological momentary assessment. Frontiers in Psychiatry, 15: 1213863. https://doi.org/10.3389/fpsyt.2024.1213863

[20] Breiman, L. (2001). Random forests. Machine Learning, 45: 5-32. https://doi.org/10.1023/A:1010933404324

[21] Loef, B., Wong, A., Janssen, N.A.H., Strak, M., Hoekstra, J., Picavet, H.S.J., Boshuizen, H.C.H., Verschuren, W.M.M., Herber, G.C.M. (2022). Using random forest to identify longitudinal predictors of health in a 30-year cohort study. Scientific Reports, 12(1): 10372. https://doi.org/10.1038/s41598-022-14632-w

[22] Moges, D.M., Virro, H., Kmoch, A., Cibin, R., Rohith, R.A.N., Martínez-Salvador, A., Conesa-García, C., Uuemaa, E. (2024). Streamflow prediction with time-lag-informed random forest and its performance compared to SWAT in diverse catchments. Water, 16(19): 2805. https://doi.org/10.3390/w16192805

[23] Nurfadilah, K., Aidi, M., Notodiputro, K., Susetyo, B. (2024). Multilevel regressions for modeling mean scores of national examinations. BAREKENG: Jurnal Ilmu Matematika Dan Terapan, 18(1): 323-332. https://doi.org/10.30598/barekengvol18iss1pp0323-0332

[24] Hewamalage, H., Ackermann, K., Bergmeir, C. (2023). Forecast evaluation for data scientists: Common pitfalls and best practices. Data Mining and Knowledge Discovery, 37(2): 788-832. https://doi.org/10.1007/s10618-022-00894-5

[25] Ghannam, S. (2026). An explainable comparative study of statistical, machine learning, deep learning, and hybrid models for CO_2 emissions forecasting in Australia. Array, 29: 100639. https://doi.org/10.1016/j.array.2025.100639

[26] Getenet, S. (2024). The influence of students' prior numeracy achievement on later numeracy achievement as a function of gender and year levels. Mathematics Education Research Journal, 36(4): 745-766. https://doi.org/10.1007/s13394-023-00469-7

[27] Erdem, C., Kaya, M. (2024). The relationship between school and classroom climate, and academic achievement: A meta-analysis. School Psychology International, 45(4): 380-408. https://doi.org/10.1177/01430343231202923

[28] Zhao, H., Han, M., Wang, Z., Liu, B. (2024). School connectedness and academic burnout in middle school students: A multiple serial mediation model. Behavioral Sciences, 14(11): 1077. https://doi.org/10.3390/bs14111077

[29] Cipriano, C., Ha, M., Wood, M., Sehgal, K., Ahmad, E., McCarthy, M.F. (2024). A systematic review and meta-analysis of the effects of universal school-based SEL programs in the United States: Considerations for marginalized students. Social and Emotional Learning: Research, Practice, and Policy, 3: 100029. https://doi.org/10.1016/j.sel.2024.100029