Original Research Article

Educational and Psychological
Measurement
1–30
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00131644221108180
journals.sagepub.com/home/epm
SAGE

Fixed Effects or Mixed Effects Classifiers? Evidence From Simulated and Archival Data

Anthony A. Mangino[1,2], Jocelyn H. Bolin[1] and W. Holmes Finch[1]

Abstract

This study seeks to compare fixed and mixed effects models for the purposes of predictive classification in the presence of multilevel data. The first part of the study utilizes a Monte Carlo simulation to compare fixed and mixed effects logistic regression and random forests. An applied examination of the prediction of student retention in the public-use U.S. PISA data set was considered to verify the simulation findings. Results of this study indicate fixed effects models performed comparably with mixed effects models across both the simulation and PISA examinations. Results broadly suggest that researchers should be cognizant of the type of predictors and data structure being used, as these factors carried more weight than did the model type.

Keywords

predictive classification, mixed effects models, Monte Carlo simulation, Program for International Student Assessment

Nearly uniform among practicing statisticians and methodological researchers is the desire to pursue ever more informative methodologies to permit increasingly more accurate inquiry into complex phenomena. Simultaneously, researchers in the social

[1]Ball State University, Muncie, IN, USA
[2]University of Kentucky, Lexington, USA

Corresponding Author:
Anthony A. Mangino, Biostatistics Consulting and Interdisciplinary Research Collaboration Lab, Department of Biostatistics, University of Kentucky, 725 Rose Street, Room 205, Lexington, KY 40506, USA.
Email: Anthony.Mangino@uky.edu

sciences—and the methodologists in these various domains—often seek to utilize only those methods as complex as necessary to facilitate maximally effective investigations into social, psychological, and educational phenomena (Little, 2013; Murtaugh, 2007; Zellner et al., 2001). This study seeks to establish guidelines for facilitating parsimony from analysis of complex data structures favoring simplicity and recommendations for effective practice on the part of social science researchers in the area of predictive classification. That is, the intersection of classification methodology and multilevel data structures is as yet sparsely studied and, thus, recommendations for practice are not well established; this study then serves as an examination of this intersection through both simulated and archival data analyses to identify such optimal recommendations. This study compares fixed effects logistic regression (LR) and random forest (RF) models to that of generalized linear mixed model (GLMM) and Ngufor et al.'s (2019) mixed effects random forest (MERF) algorithm in both simulated and the publicly available version of the United States Program in International Student Assessment (PISA) data set to predict whether students will be held back in school based on a number of predictors.

Supervised classification analysis is a family of methods designed to assign cases into one of two or more known outcome groups (Hastie et al., 2009). The purpose germane to this study is the use of classifiers to predict group membership of cases not previously seen in the data set following an initial training of the model using observations for which group membership is known (Steyerberg, 2019). This prediction capacity follows a method as such: A classifier is trained on data consisting of a set of cases for which group membership is known (identified as the training set); this trained classifier is then applied to a new data set with cases previously unseen and for which group membership is not known (the test set). These predictions provide researchers and practitioners with the information necessary to engage in maximally effective practices such as implementing interventions in the classroom (VanDerHeyden, 2013) or for psychiatric treatment (Zigler & Phillips, 1961), among others. The estimation of classifiers is plagued by a number of methodological issues due to the number of unknown conditions and results of cases within the test set and, as such, requires additional consideration to identify the optimal methods when used to make optimally accurate predictions such that early intervention may be employed.

A secondary factor germane to the present investigation considers a feature of many contexts in the psychoeducational domains: Individuals are situated within and affected by their context. Presently considered is the multilevel data structure in which cases are nested within higher-level units or clusters (Luke, 2019). This condition does present somewhat frequently in the social sciences—in both cross-sectional and longitudinal paradigms—and carries with it numerous considerations beyond those in single-level data that can be problematic for many simpler analytical frameworks. A classic exemplar of this structure is Raudenbush and Bryk's (2002) motivating example of students being situated within classrooms and those classrooms being nested within schools (thus resulting in a three-level data structure).

The present investigation examines the intersection between these two analytical frameworks—classification and multilevel modeling—with respect to several factors relevant to both frameworks, specifically focusing on multilevel random-intercepts models to maintain simplicity. While traditional classifiers in the fixed and mixed effects frameworks—LR in the single-level and the GLMM in the multilevel—have been frequently used in classification settings both explanatory and predictive within the social sciences, machine learning research has yielded substantive advances in model construction for both regression and classification. However, while myriad studies have considered many of these novel frameworks for both classification and regression, few have fully compared and examined the relative efficacy of fixed and mixed effects models under various conditions unique to classification with nested data structures. Furthermore, fixed effects models may perform comparably with or even better than mixed effects classifiers in predictive contexts under certain conditions (Kilham et al., 2019; Speiser et al., 2019). Specifically, Speiser et al. (2019) found that when comparing a standard RF classifier to their novel binary mixed model random forest (BiMM forest) model, the RF model performed at least as well as the BiMM forest and was only marginally outperformed under some conditions. Similarly, Kilham et al. (2019) found that when including both Level-1 and Level-2 predictors, a standard RF model could paradoxically outperform GLMM despite functionally ignoring the nesting structure of the data in prediction contexts. The conditions dictating this choice, however, are paramount as mixed effects models reduce to fixed effects only when employed with cases from new clusters in contrast to new cases within existing clusters in which models can retain random effect components. Consequently, it is key to understand the conditions both within the data and the research design.

The architectures and estimation methods of the principal four are detailed further, followed by a review of the existing literature comparing classifiers.

Current Prediction Methods

Despite the plethora of fixed effects classifiers available to researchers, each with their unique benefits and drawbacks, only RF and its mixed effects analogue MERF have been compared directly with LR and its mixed effect analogue GLMM (e.g., Kilham et al., 2019; Speiser et al., 2019). Therefore, to remain consistent with the sparse existing literature comparing fixed and mixed effects classifiers—as well as its consistent noteworthy performance in classification tasks—LR and RF are considered in both their fixed and mixed effects forms.

Logistic Regression

Among others, LR has long been a commonly used method due to its broad utility, relative simplicity, interpretability, and commonality across various software packages. Functionally, LR operates by calculating the probability of group membership {0,1} based on estimation of a linear function for each input. Following the form of a linear regression model, LR instead estimates these probabilities via the logistic function:

$$p(X) = \frac{e^{\beta_0 + \Sigma \beta_1 X_1 \ldots \beta_p X_i}}{1 + e^{\beta_0 + \Sigma \beta_1 X_1 \ldots \beta_p X_i}}, \tag{1}$$

where β_0 = Intercept coefficient; β_p = Coefficient estimate for predictor X_i; e = Euler's constant, and base natural logarithm ≈ 2.718 (Hastie et al., 2009; James et al., 2013).

The resulting equation could then be interpreted with respect to the estimation of the log-odds (logit) of group membership whereby the above equation can be simplified to:

$$\log\left(\frac{p(X)}{1 - p(x)}\right) = \beta_0 + \sum \beta_1 X_1 \ldots \beta_p X_i, \tag{2}$$

to approximate more directly approximate $\Pr(Y = 1|X)$ (James et al., 2013). Logistic regression parameters are estimated via maximum likelihood.

Random Forests

In contrast to LR, RF operates on a substantially different paradigm based on two principle foundations: A nonparametric recursive partitioning algorithm and an ensemble classifier framework. Breiman (2001) then proposed an alternative method in the form of RF based on the notion of ensembles: a series of models consisting of multiple singular classification tree classifiers, each making their own decisions before being aggregated through various algorithms to improve model accuracy and diversity (Bauer & Kohavi, 1999; Dietterich, 2000; Strobl et al., 2009). Random forests are trained as such:

- A random bootstrapped sample (sampling with replacement) of size N is drawn from the training set.
- A singular decision tree T_b is fit to the bootstrapped sample.
- At each split in trees $T_1, \ldots, T_b$, a predictor is selected at random from a random subset of all predictors m, such that $m \approx \sqrt{p}$; where p = total number of predictors.
- The ensemble of trees $\overline{\{T_b\}_1^B}$ is output.
- Group membership predictions $\widehat{C_b}(x)$ are yielded from each of the trees, then aggregated such that $\widehat{C_{rf}^B} = majorityvote\{\widehat{C_b}(x)\}_1^B$ (Hastie et al., 2009).

In contrast to the a priori specification of more conventional models like LR or GLMM, RF models are trained algorithmically, thus allowing for a more flexible architecture.

Generalized Linear Mixed Effects Models

The most commonly employed mixed effects model is the GLMM, which acts as the multilevel analogue to the fixed effects LR model discussed earlier. As an extension of the standard hierarchical linear mixed effects model, GLMM makes use of a similar parameter estimation method to estimate the probability of group membership, as was the case with LR. The basic structure of GLMM makes use of a binomial sampling model with a logit link function at Level-1. The overall model estimates the number of cases belonging to a particular group with the equation:

$$Y_{ij}|\varphi_{ij} \sim B(m_{ij}, \varphi_{ij}), \tag{3}$$

where Y_{ij} = Number of cases identified as 1 (assuming group labels of $\{0, 1\}$) in m_{ij} trials, distributed as binomial with φ_{ij} probability of success per trial across m_{ij} trials; and φ_{ij} = Probability of identification in Group 1 on each trial with a structural model resembling a traditional regression model:

$$\eta_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \cdots + \beta_{pj}X_{pij}, \tag{4}$$

where β_{0j} = overall intercept, grand mean; β_{pj} = coefficient estimate for cluster j for predictor p; and X_{pij} = Predictor p value for case i in cluster j; and a link function represented as:

$$\eta_{ij} = \log\left(\frac{\phi_{ij}}{1 - \phi_{ij}}\right), \tag{5}$$

where η_{ij} = Probability of membership in Group 1 (Raudenbush & Bryk, 2002).

The Level-2 model then seeks to estimate the cluster-level effects on Level-1 parameter estimates as represented by:

$$\beta_{qj} = \gamma_{q0} + \sum_{s=1}^{S_q} \gamma_{qs} W_{sj} + u_{qj}, \tag{6}$$

where β_{qj} = Level-1 coefficient estimate q for cluster $j;$ γ_{q0} = Intercept for random effect $q;$ γ_{qs} = Level-2 coefficient estimate s for random effect $q;$ W_{sj} = Level-2 predictor effect s for cluster j; and u_{qj} = Random error for random effect q in cluster j (Raudenbush & Bryk, 2002).

Mixed Effects Random Forests

Expanding upon the baseline GLMM, Hajjem et al. (2014) proposed a framework for developing the MERF algorithm as one of several multilevel extensions of RF including, among others, Speiser et al.'s (2019) binary mixed model random forest (BiMM forest) and Capitaine et al.'s (2021) stochastic mixed effects random forest (SMERF), none of which are as yet widely available in commonly used statistical software packages. Among these numerous frameworks, the only one presently available for implementation in the R Statistical Software package (R Core Team, 2020) is Ngufor et al.'s (2019) expansion of Hajjem et al.'s (2014) MERF algorithm via Ngufor's (2019) vira package in R, hence its present use. Consequently, while previous comparisons have considered alternative frameworks, it is Ngufor et al.'s (2019) MERF framework that will be presently utilized due to its availability in the R statistical software package. Although it should be noted that Hajjem et al.'s original 2014 algorithm is available in Python, this study focused on methods readily available in R (thus the use of Ngufor et al.'s implementation).

The MERF algorithm follows a function estimating response variable y^* with an RF model estimating fixed effects parameters and random effects assumed to be linear. Fixed and random effects within the model are estimated iteratively with each successive iteration updating the residuals of the previous until the algorithm converges and population-averaged predictions are yielded for the training set. This model is represented as follows:

$$y_i = f(X_i) + Z_i b_i + \varepsilon_i, \tag{7}$$

where $f(X_i)$ = RF function for fixed effects X_i; Z_i = random effects matrix of covariates with dimensions $n_i{*}q$; b_i = unknown random effects vector for cluster i with distribution $b_i \sim N(0, D)$; and ε_i = Vector of errors of dimensions $n_i{*}$ 1 with distribution $\varepsilon_i \sim N(0, R_i)$ (Hajjem et al., 2014).

Each iteration begins by estimating $f(X_i)$ before updating the variance components for the model error ε_i and for the random effects b_i (Hajjem et al., 2014). A generalized log-likelihood criterion is calculated at each iteration until the change is sufficiently minimal as to characterize convergence. Predictions from this model are estimated using population-averaged parameters $f(x_{ij})$ and $Z_i b_i$ in cases where new data belong to existing clusters; predictions for new cases in new clusters use $f(x_{ij})$ only.

Previous Comparisons Between Fixed and Mixed Effects Classifiers

Regarding the present argument leveraged from Speiser et al.'s (2019) and Kilham et al.'s (2019) findings favoring RF over multilevel classifiers, the question of how these models compare becomes relevant. Coupled with Ngufor et al.'s (2019) findings that mixed effects models (including Hajjem et al.'s, 2014, mixed effects random forest [MERF] and GLMM) generally outperform fixed effects classifiers (including RF), the evidence is as yet inconclusive. However, when considering classification contexts in the machine learning domain, seldom are mixed effects models actively used, often eschewed in favor of the ''black box'' methods noted earlier (Abu-Nimeh et al., 2007; Palvanov & Cho, 2018; Wu & Zhang, 2010; Zhang & Haerdle, 2010). Furthermore, researchers such as McNeish and colleagues (McNeish et al., 2017; McNeish & Kelley, 2019; McNeish & Stapleton, 2016) argue in favor of more parsimonious models such as population-averaged or fixed effects models. Fixed effects models are often used when the number of higher-level clusters is small (typical guidelines of 30 clusters were derived from Kreft's unpublished, but often-cited, 1996 study) and simply include the cluster identifier variable as dummy coded predictors and serve to account for the variance due to cluster without explicitly estimating the variance decomposition (as in the case of GLMM). Fixed effects models, then, act as a method for controlling for the variability due to cluster membership while retaining the simplicity of interpretation social science researchers and methodologists often seek. However, no studies to date have explicitly compared fixed effects RF models with fixed effects LR or any mixed effects models.

In many studies comparing multilevel models, it has been common to utilize a fixed or mixed effects model as a comparison against which more complex multilevel models can be assessed. The present case, however, employs a novel method in the form of MERF and, consequently, has fewer studies considering its efficacy relative to fixed effects models and simpler multilevel classifiers. Hajjem et al. (2014) instantiated MERF for the purposes of regression and have not since assessed it for the purposes of classification. Mangino and Finch (2021) employed Hajjem et al.'s MERF framework for the purposes of predictive classification and found it outperformed GLMM (and several other more complex mixed effects models) under many different conditions including those of differing sample sizes and intraclass correlations (ICCs; ratio of between-cluster variance to the total outcome variance, $\sigma_b^2/(\sigma_b^2+\sigma_r^2)$). Speiser et al. (2019) postulated an alternative multilevel RF framework in the form of the Binary Mixed Model forest (BiMM forest; a binary classification extension of a multilevel RF framework). However, the BiMM forest algorithm is not yet available in commonly used statistical software packages, whereas MERF is via Ngufor's (2019) Vira R package. Hajjem et al. (2014) demonstrated that in the regression context, MERF appreciably outperformed RF and GLMM. Furthermore, Kilham et al. (2019), in the context of tree-level harvest predictions, found that when simultaneously accounting for tree-level (Level-1) and plot-level (Level-2) predictors, RF outperformed GLMM with substantially less requirement for proper model effect specification and while preserving model interpretability. That is, Kilham et al. (2019) concluded that while RF did not provide a detailed consideration of plot-level effects, predictions were equal to or more accurate than those obtained from GLMM.

Considering fixed effects RF compared with LR, nearly all comparisons of classifiers—particularly in cases of unbalanced data and in predictive contexts—RF broadly outperforms LR. For example, when predicting civil war onset in an unbalanced data set (the ratio of peaceful to bellicose years was approximately 100:1, even more extreme than the retention grouping variable in the PISA data set at approximately 8.5:1), Muchlinski et al. (2016) found that RF yielded substantially more accurate predictions with less model specification required. It was hypothesized that because of RF's (and tree-based models, broadly) flexibility and ability to model complex interactions and nonlinear relationships, it was able to account for the variety of variable types and relationships that would otherwise need to be specified a priori in an LR model. Similarly, a thesis by Yan (2019) illustrated that RF outperformed other tree-based classifiers when data were unbalanced at an approximately 29:1 ratio, thus making it a more prominent candidate for consideration than other tree-based methods. In addition, in the context of calculating propensity scores, Westreich et al. (2010) noted key limitations of LR, namely, the requirement of proper model specification and assumptions of linear relationships among variables. Alternatively, Westreich et al. (2010) discuss the benefits of tree-based methods while specifically favoring RF as a preferred alternative. Furthermore, when predicting dementia diagnoses using a number of classical and novel statistical and machine learning models—including RF, LR, neural networks, support vector machines, and classification trees, among others—Maroco et al. (2011) found that RF performed the most consistently across a number of key accuracy metrics (including large- and small-group recovery, as in the present case) compared with more complex machine learning classifiers as well as the classical LR.

Thus far, few studies have compared Ngufor et al.'s (2019) MERF algorithm with other fixed or mixed effects models in either regression or classification contexts (e.g., Kilham et al., 2019; Mangino & Finch, 2021; Speiser et al., 2019). Hajjem et al.'s (2014) original study and MERF framework found that MERF outperformed LR, RF, and GLMM across all conditions in a regression context. However, in many data sets, the difference between RF and MERF only gave marginal favor to MERF over RF and similar performance of RF to GLMM. In addition, a study by Ngufor et al. (2019) featured perhaps the most comprehensive applied classification comparison to date, comparing a number of fixed and mixed effects classifiers on several data sets predicting longitudinal hemoglobin A1c change. In this study, MERF consistently performed as well as or better than GLMM and outperformed RF in three of the four data sets considered, often only in situations with larger cluster sizes; in many cases, this performance differential was marginal (Ngufor et al., 2019). Furthermore, MERF and GLMM tended to perform similarly to one another in nearly all of the data sets and cluster sizes Ngufor et al. (2019) considered; GLMM also tended to marginally outperform RF in many of the data sets. Alternatively, Capitaine et al. (2021) found that MERF consistently yielded substantially lower error than GLMM in simulations using both deterministic and stochastic model construction in the context of high-dimensional data (more predictors than cases). However, similar to findings by Speiser et al. (2019) and Kilham et al. (2019), when applied to HIV vaccine trial data, MERF and RF performed comparably with one another with RF holding only marginally larger standard errors (Capitaine et al., 2021). This similarity in performance was also found in Karpievitch et al.'s (2009) comparison of RF with subject-level averaged and bootstrapped RF methods to account for clustering revealed that the standard RF method performed nearly identically to the cluster-adjusted methods.

Present Study

Given the previous literature, the present investigation encompasses both a simulation study and archival data analysis to determine the comparative predictive accuracy of fixed and mixed effects classifiers under varied data conditions. The principle purpose of this study is to ascertain the predictive capability of various fixed and mixed effects classifiers in binary classification settings. An initial simulation study serves to provide a preliminary consideration of fixed effects classifiers compared with mixed effects models such that the conditions inherent in data featuring a nested structure (discussed later) are thoroughly considered. Results from the simulation study then serve to inform the methods to be employed in the archival analyses, given the conditions within the PISA data set presently utilized. Throughout both investigations, the R Statistical Software Package (R Core Team, 2020) was employed with several additional packages utilized to implement the various analytical techniques currently considered (to be detailed later).

Given the literature reviewed and the methodology utilized, the principal guiding research question could be identified as follows: Under what conditions could fixed effects—rather than mixed effects—classifiers be utilized for predictive classification with multilevel data structures?

Methods

Simulation

To determine the comparative efficacy of fixed and mixed effects models in the presence of multilevel data, a Monte Carlo simulation provided preliminary evidence to inform the conditions under which fixed or mixed effects classifiers should be employed. Within the Monte Carlo simulation framework, six data characteristics were sequentially permuted (243 total conditions); each iterated 500 times (due to the computational capacity required to estimate the current models) to approximate the likely outcomes to arise while accounting for the stochasticity of any single data generation process (Harrison, 2010). Permissions were obtained to utilize Ball State University's Beowulf Computing Cluster to run all simulation conditions. The conditions are as follows: Number of cases per Level-2 cluster, number of Level-2 clusters, ICC, group size ratio (ratio of outcome Group 0 to outcome Group 1), and the type of predictors used (all three at Level-1; one at Level-2 with two at Level-1, and two at Level-2 with one at Level-1). The conditions manipulated and classifiers employed are illustrated in Table 1. Data were simulated to the conditions in Table 1 and split into two equal-sized partitions to serve as training and test data sets; all results presented are drawn from the test set.

Considering the number of Level-1 cases, Hajjem et al.'s (2014) use of cluster sizes of 10 and 50, and Mangino and Finch's (2021) use of cluster sizes up to 100 led to the decision to use these cluster sizes. The number of clusters has been previously examined with larger size conditions (e.g., Crane-Droesch [2017] used a

Table 1. Data Simulation Conditions.

Simulation variable	Conditions
Number of Level-1 cases per Level-2 cluster	10; 50; 100
Number of Level-2 clusters	10; 50; 100
Correlation within Level-2 clusters (intraclass correlation)	0.1; 0.3; 0.8

Number of predictors	3 at Level-1;
	2 at Level-1 and 1 at Level-2;
	1 at Level-1 and 2 at Level-2

Outcome group size ratio	50:50; 75:25; 90:10
Method	LR; GLMML RF; MERF
Outcome metrics	Large-group recovery Small-group recovery Area under the curve Binary cross-entropy

Note. LR = logistic regression; GLMML = generalized linear mixed model; RF = random forest; MERF = mixed effects random forest.

number of clusters from 900 to 2,700). However, Paccagnella's (2011) finding that 50 or more clusters provides robust standard errors led to the use of these conditions. McNeish and Stapleton's (2016) results indicated that multilevel models can provide reliable parameter estimates with as few as 10 clusters, though the models are underpowered. Furthermore, Mangino and Finch (2021) employed conditions with 100 clusters, thus encompassing a greater number of clusters. Consequently, conditions of 10, 50, and 100 clusters are presently considered. The total number of cases, thus, ranges from $N = 100$ (10 cases to each of 10 clusters) to N = 10000 (10 cases to each of 100 clusters).

The ICC conditions were determined based on the only study in the presently reviewed literature base to specify exact ICC values rather than the oblique ''large'' and ''small'' random effects language. While Speiser et al. (2019, 2020) specified cluster standard deviations of 0.1 and 0.5 to mean small and large random effects, respectively, the actual calculated ICC is unknown for these studies. Mangino and Finch (2021) specified their ICC conditions to be 0.1, 0.3, and 0.8 as they resemble a wide variety of cluster heterogeneity. The outcome group size ratios (ratio of case membership in Group 0 to Group 1) were determined by previous studies in the classification literature base that utilized equal (50:50) and various unequal (75:25 and 90:10, among others) group size ratios (Bolin & Finch, 2014; Lei & Koehly, 2003). Furthermore, given the conditions of the PISA data set—with a 5,108/604 split of individuals who were not retained to those who were (an approximately 8.5:1 ratio)—it is evident that a comparison between equal and various levels of unequal group size ratios should be considered. Therefore, group size ratios of 50:50, 75:25, and 90:10 were selected.

Many studies reviewed in earlier featured a number of predictors ranging from two (Lavery et al., 2019; Maas & Hox, 2005) to 8000 (Capitaine et al., 2021). However, the number and type of predictors is a set of conditions as yet minimally examined, particularly in the domain of multilevel classifiers. Few studies have explicitly examined the effects of Level-1 v. Level-2 predictors on model accuracy in either the classification or regression contexts (Kilham et al., 2019, and Downes and Carlin, 2020, have done so in applied contexts with real data), but recommendations for practice have yet to be established based on the intersection of tightly controlled simulation and real data settings. Consequently, the consideration of the number and type of predictors is a highly exploratory area. Therefore, the present study sought to examine a relatively simple constellation of predictors with some resemblance to the behavior of the predictor and predictor–outcome relationships in the PISA data. The predictor–outcome relationship was held constant at 0.13 while predictor intercorrelations were held constant at 0.31. These values represent a commonly identified low effect size, but largely resemble the median absolute correlations between predictors and the grade repetition outcome in the PISA data.

In addition, given the paucity of literature on the type of predictors utilized in situations with nested data, it was determined that the level of predictor be considered while holding constant the number of predictors. Therefore, a total of three continuous, normally distributed, and significant predictors were simulated across all conditions with one condition featuring all predictors at Level-1, one condition featuring two at Level-1 and one at Level-2, and a third condition featuring one at Level-1 and two at Level-2. Previous studies in this area have employed simulated data with conditions ranging from eight continuous and categorical predictors resembling clinical measurements (Speiser et al., 2019, 2020) to those resembling genomics data with 8000 predictors across multiple time points (Capitaine et al., 2021). While the presently simulated data are decidedly simpler in their data generation process than is the case in previous literature, many previous studies have focused on data sets in niche settings (e.g., genomics) in contrast to the presently more general framework of continuous, normally distributed predictors employed across a wide variety of other possible conditions.

Each of the four models discussed earlier—LR, GLMM, RF, and MERF—were fit to the data at each iteration with average outcome metrics obtained across each condition's 500 iterations.

Empirical Example

The publicly available U.S. PISA data set includes 5712 unique students in 177 schools with 947 variables, many of which are multicategorical coded representations of variables from the restricted data set. The outcome of retention was chosen due to both its status as a binary outcome (either students have or have not been retained) and the fact that it is a powerful predictor of later school dropout and poor academic performance (Glick & Sahn, 2010). Several predictors were used aligning with the literature base associated with student dropout. More positive attitudes toward school (Ikeda & García, 2014), greater sense of belonging (McMahon et al., 2008), and a higher expected educational attainment (Lee & Stankov, 2018) were associated with a reduced likelihood of being retained. In addition, parent factors including improved parent education (Corman, 2003), and greater home resources (e.g., books), wealth, and income were identified as resulting in a reduced likelihood of students being retained (Choi et al., 2018; Corman, 2003; Eisemon et al., 1997; Wößmann, 2003). Given that missing data were present (range: 0%–20%; Total Missing: 5.32%), for the purposes of the present investigation, all missing data were imputed using multivariate imputation by chained equations (MICE) using an RF estimator as per recommendations by Shah et al. (2014) and Waljee et al. (2013). The above-described variables were all measured at the student level (Level-1) and, thus, do not fully represent the conditions considered in the simulation. Therefore, school free and reduced lunch participation rate and the estimated percentage of students from socioeconomically disadvantaged homes were included. Two sets of analyses were thus conducted: The first including only student-level predictors and the second including all student-level predictors and two school-level predictors. Table 2 includes descriptive statistics for all predictors and the outcome used.

The full U.S. PISA data set was split into two equally sized partitions to serve as training and test data sets; like the simulation study, all results presented are from the test set.

Outcome Measures

To achieve a holistic comparison across fixed and mixed effects classifiers, both raw classification metrics (e.g., large-group recovery) should be considered in conjunction with computational accuracy metrics (e.g., area under the curve [AUC]). Consequently, four prime outcome metrics were considered and are briefly described:

- *Large group recovery* is the proportion of cases correctly classified into the larger of the two groups. When group sizes are equal, this is Group 0.
- *Small group recovery* is the proportion of cases correctly classified into the smaller of the two groups. When group sizes are equal, this is Group 1.
- *AUC* is interpreted as the likelihood a classifier will identify a higher probability of a random Group 1 case belonging to its correct group compared with the same likelihood for a random Group 0 case (Fawcett, 2006). Higher values indicate better case discrimination.
- *Cross-entropy*, a metric drawn from information theory, is interpreted as the amount of uncertainty in a model given the data with which it is provided (Ramos et al., 2018). Higher values indicate more uncertainty.

Table 2. Descriptive Statistics for Raw and Imputed PISA Data Sets.

Variable (total missing: 5.32%)	*M* (*SD*)	
	PISA original	PISA imputed
FRPL (missing = 0; 0%) NAT/percentage free/reduced price lunch	3.512 (1.156)	3.512 (1.156)
SC048Q03NA (missing = 299; 5.23%) Est. percent. the 10th grade. Students from socioeconomic disadvantaged homes	50.852 (26.187)	50.818 (26.179)
MOTIVAT (missing = 122; 2.14%) Student attitudes, preferences, and self-related beliefs: achieving motivation	0.661 (0.943)	0.675 (0.946)
HEDRES (missing = 97; 1.52%) Home educational resources	−0.115 (1.142)	−0.1053 (1.145)
WEALTH (missing = 55; 0.96%) Family wealth	0.473 (1.086)	0.476 (1.085)
ESCS (missing = 74; 1.29%) Index of economic, social and cultural status	0.079 (1.003)	0.083 (1.002)
AGE (missing = 0; 0%) Age	15.806 (0.287)	15.806 (0.287)
BMMJ1 (missing = 1,093; 19.13%) International socioeconomic index of occupational status of mother	49.873 (21.777)	47.976 (22.046)
BFMJ2 (missing = 1,145; 20.04%) International socioeconomic index of occupational status of father	43.317 (22.309)	41.334 (22.063)
PARED (missing = 99; 1.73%) Index highest parental education in years of schooling	13.621 (2.802)	13.632 (2.791)
BELONG (missing = 132; 2.31%) Subjective well-being: Sense of belonging to school	−0.085 (1.013)	−0.074 (1.024)
Hisei (missing = 453; 7.93%) Index highest parental occupational status	53.643 (21.713)	53.667 (21.593)
HOMEPOS (missing = 49; 0.86%) Home possessions	0.022 (1.112)	0.227 (1.110)
BSMJ (missing = 760; 1.33%) Students' expected occupational status	62.419 (17.045)	63.247 (17.309)
	Repeated/Not (%)	
REPEAT (missing = 179; 3.13%) Grade repetition	4941 (89.30)/592 (10.70)	5108 (89.43)/604 (10.57)

Note. PISA = United States Program in International Student Assessment.

While LGR and SGR are simplistic metrics of raw accuracy, AUC and CE both act as computational metrics to indicate the degree to which the estimation of the model is sufficiently accurate. Only the practically significant ($\eta_p^2 > 0.1$; Bolin & Finch, 2014; Lei & Koehly, 2003) analysis of variance (ANOVA) interactions (maximum interaction depth = 4) across all four metrics were considered to compare the raw prediction metrics and computational accuracy metrics.

Results

Three principal four-way interactions were found to be practically significant across all four metrics: The method by group size ratio by predictor type by ICC interaction; the method by group size ratio by predictor type by cluster size interaction; and the group size ratio by ICC by predictor type by cluster size interaction.

Method by Group Size Ratio by Predictor Type by ICC

This interaction was characterized by a concomitant increase in LGR and decrease in SGR as the group size ratio became increasingly more unequal irrespective of the predictor type and ICC. However, the degree to which these trends occurred was not as pronounced when the ICC increased to 0.8 and at least one predictor was included at Level-2. This effect was consistent across all methods, sans MERF in the case of unequal group sizes with two Level-2 predictors. In addition to raw prediction metrics, AUC indicated near-uniform performance across all methods with MERF yielding the least consistency. Broadly, AUC increased slightly as group sizes became increasingly more unequal except in the case of an ICC of 0.1 in which the 75:25 group size ratio yielded the highest AUC values. Conversely, CE increased appreciably as group sizes became increasingly more unequal and the ICC was increased; this effect was more pronounced as more Level-2 predictors were added. Among the classifiers, MERF yielded the highest CE values, thus corroborating its marginally lower performance compared with the other three methods. Figure 1 illustrates all interactions across all outcome metrics.

Method by Group Size Ratio by Predictor Type by Cluster Size

This interaction illustrated similar trends to those described earlier with LR, GLMM, and RF largely performing comparably to one another under most conditions; MERF consistently yielded the lowest AUC and SGR while achieving the highest CE of all methods. In this interaction, AUC, LGR, and SGR tended to increase as more predictors were simulated at Level-2 with only a marginal accompanying increase in CE for all methods except MERF. In addition, MERF yielded the lowest SGR of all methods when group sizes were unequal; under these same conditions, MERF also yielded CE values more than twice as large as RF, the method with the second highest CE values. A notable pattern is the slight decrease in LGR and SGR when group sizes are equal,

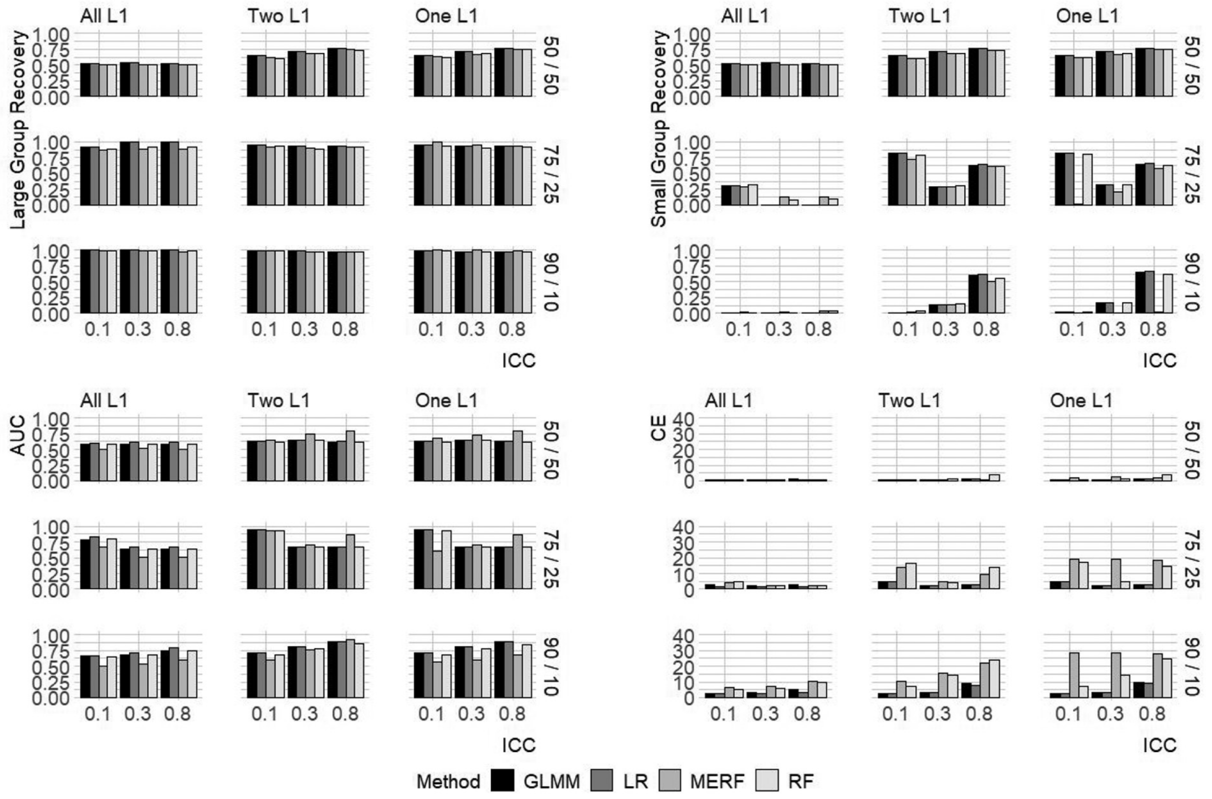

Figure 1. Method by Group Size Ratio by Predictor Type by ICC; All Outcomes

Note. AUC = area under the curve; CE = Cross-entropy; ICC = intraclass correlations; GLMM = generalized linear mixed model; LR = logistic regression; MERF = mixed effects random forest; RF = random forest.

and the accompanying increase in AUC and CE across all group size ratios as the cluster size increases. Graphs for all outcome metrics are shown in Figure 2.

Group Size Ratio by ICC by Predictor Type by Cluster Size

The final key interaction operated across all four classifiers and demonstrated an appreciable shift in all outcome metrics as the cluster size increased. When the group sizes were equal and the ICC was 0.3 or higher, both LGR and SGR decreased as the cluster size was increased from 10 to 50 or 100; this was not evident when group sizes were unequal. Under these same conditions, however, AUC and CE increased as the cluster size increased across all group size ratio and ICC conditions. A notable diminution of SGR and CE was observed when the group size ratio was 75:25 and the ICC was 0.3 before increasing again with an ICC of 0.8. Plots for all metrics in this interaction are shown in Figure 3.

PISA Examination

In the prediction of whether students were held back in the public version of the PISA data set, all outcome metrics were obtained for all methods. The conditions of the imputed PISA data set most closely resembled those seen in the simulation conditions with 50 clusters, 10 cases per cluster, an ICC of 0.1, a group size ratio of 90:10, and varying numbers of Level-1 and Level-2 predictors.

Overall, the pattern across all methods was largely consistent with each classifier yielding approximately 1.0 LGR and near-zero LGR; only when all predictors were situated at Level-1 did MERF yield SGR of approximately 0.007, the highest of all methods on this metric. Similarly, all methods performed comparably on AUC with MERF yielding a slight advantage, particularly with two Level-2 predictors. However, despite its equivalent performance on SGR and LGR, RF yielded the lowest CE values of all methods. When all predictors were situated at Level-1, the CE value for MERF decreased compared with when Level-2 predictors were included, opposite the pattern observed in the simulation. All accuracy metrics for all methods with both sets of predictors are shown in Figure 4.

Balanced Accuracy

To further illustrate the comparison between the classifiers across all significant simulation conditions described earlier—as well as the PISA data—the balanced accuracy metric was employed. Balanced accuracy is calculated as the average of LGR and SGR, thus accounting for both true positive and true negative classification rates; these results are shown in Figure 5.

Across the conditions compared, it was evident that the increasing ICC facilitated higher balanced accuracy, particularly when coupled with a greater number of Level-

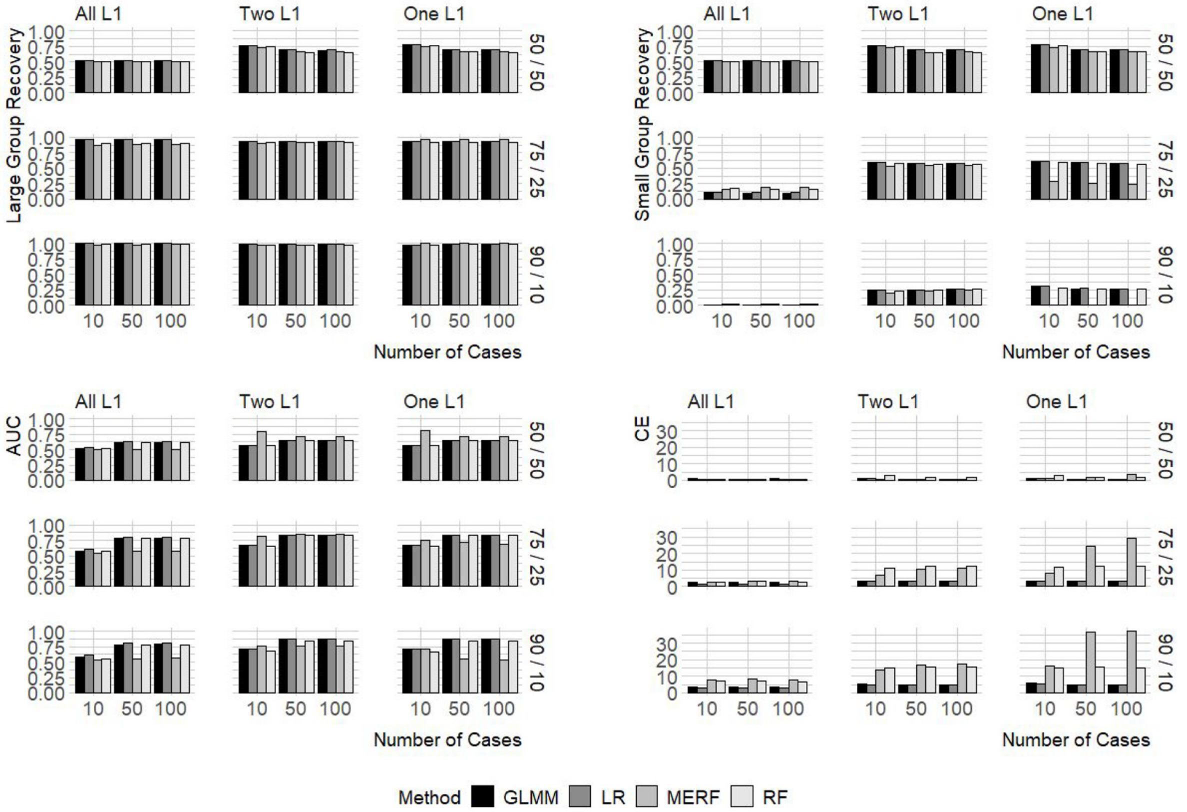

Figure 2. Method by Group Size Ratio by Predictor Type by Cluster Size; All Outcomes

Note. AUC = area under the curve; CE = Cross-entropy; GLMM = generalized linear mixed model; LR = logistic regression; MERF = mixed effects random forest; RF = random forest.

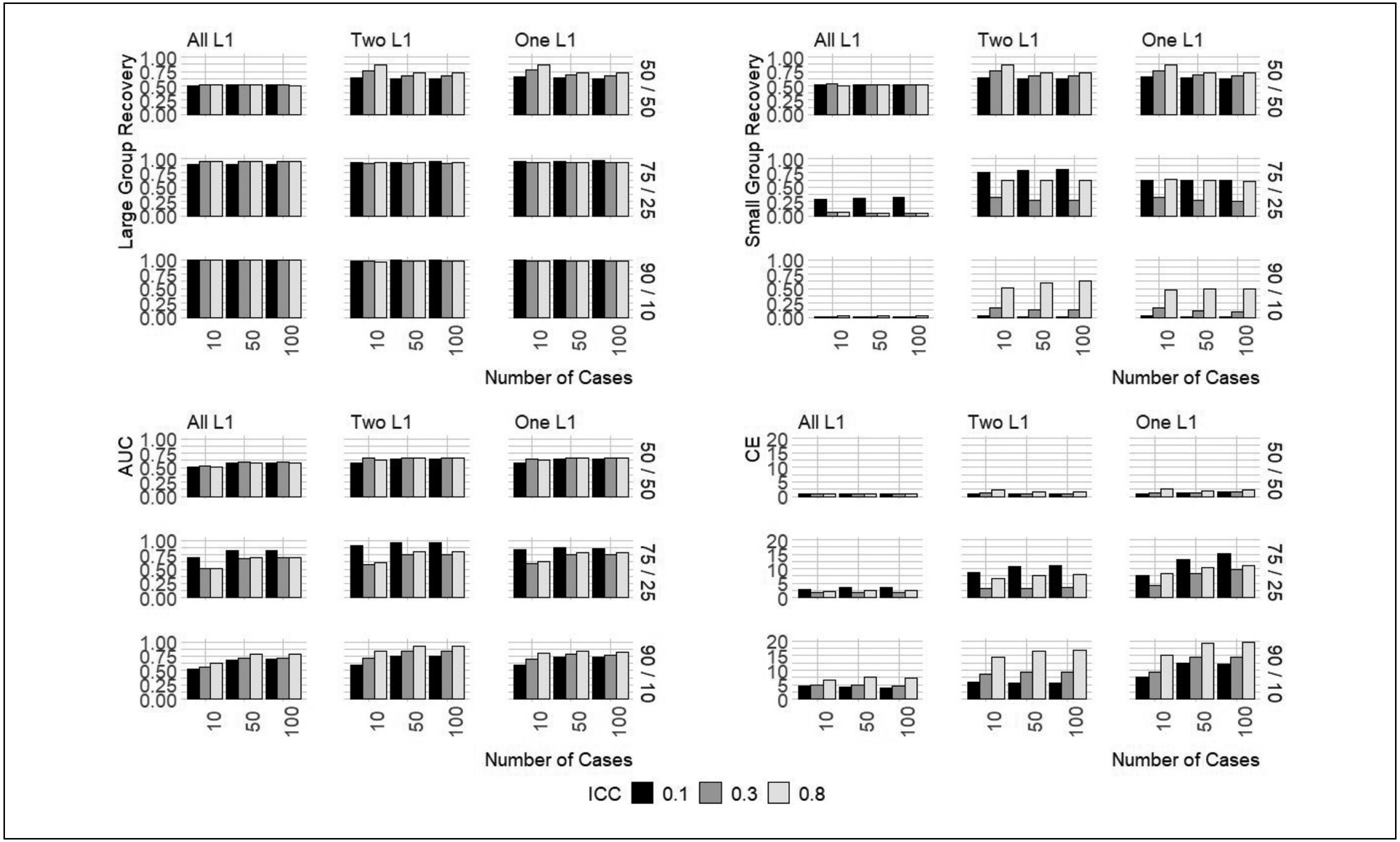

Figure 3. Group Size Ratio by ICC by Predictor Type by Cluster Size; All Outcomes

Note. AUC = area under the curve; CE = Cross-entropy; ICC = intraclass correlations.

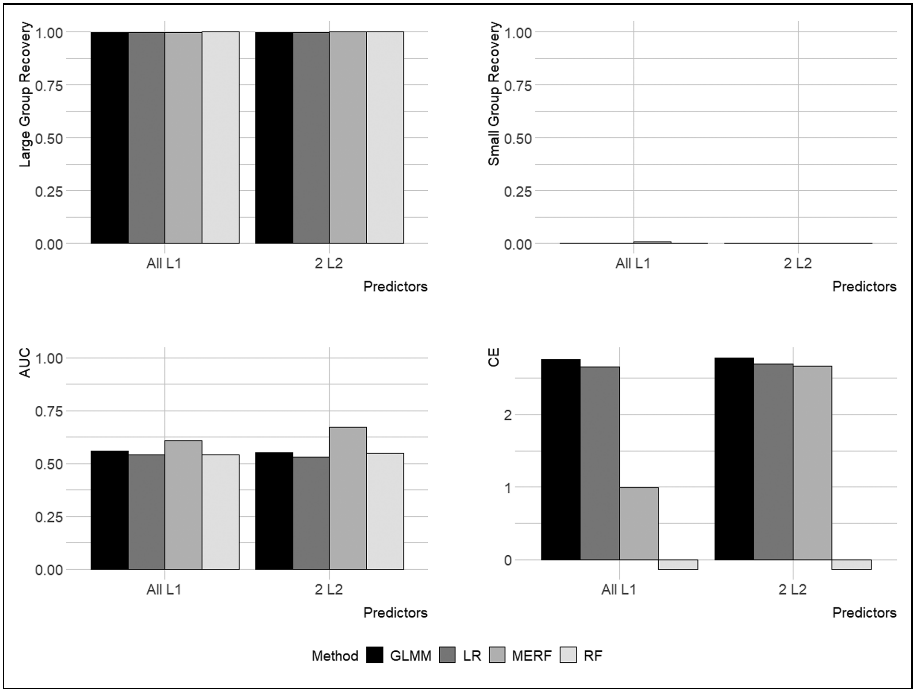

Figure 4. Accuracy Metrics for All Methods; PISA Data Examination.
Note. GLMM = generalized linear mixed model; LR = logistic regression; MERF = mixed effects random forest; RF = random forest; PISA = United States Program in International Student Assessment.

1 cases. In addition, as the number of Level-2 predictors and the ICC increased, so too did balanced accuracy. This finding was not evident in the PISA data.

Discussion

Taken holistically, the results yield limited evidence indicating any appreciable difference between fixed and mixed effects models on prediction accuracy or computational accuracy metrics. These results—coupled with the notable uniform differences across data conditions rather than method—suggest that for the purposes of predictive classification, the implementation of fixed or mixed effects classifiers only has a marginal effect on prediction and computational accuracy.

In particular, the effects of group size ratio, type of predictors, and ICC had considerable impact on outcome metrics with SGR and CE being the most profoundly affected. Given that the smaller group is typically of greater interest when considering binary classification problems, the substantive differential effects of the predictor type and ICC across various group size ratios indicated that classification in a multilevel context may become a simpler endeavor when more features of the higher-level cluster are incorporated. That is, as the ICC increased and more Level-2 predictors

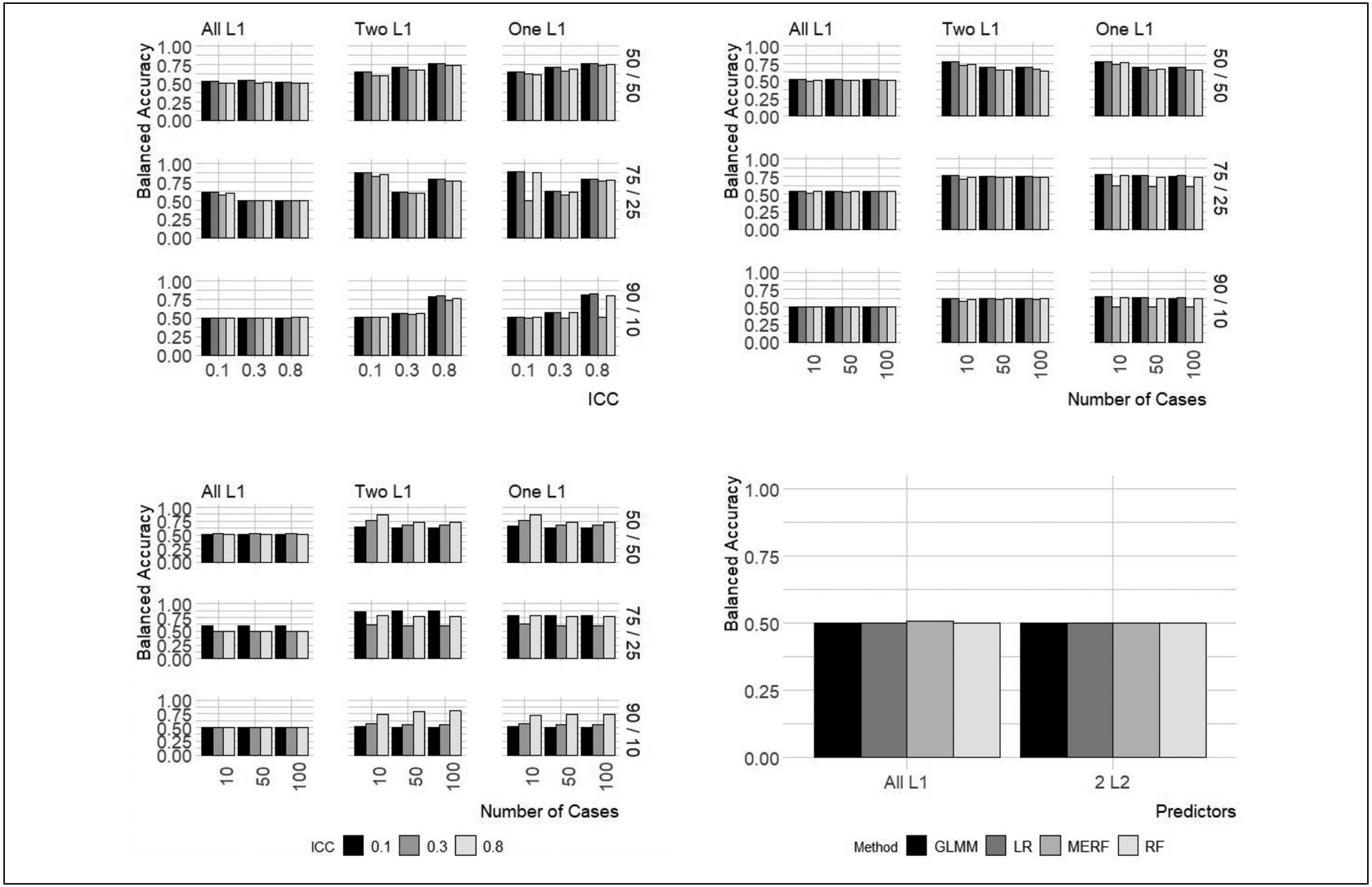

Figure 5. Balanced Accuracy for All Significant Simulation Interactions and PISA Data.

Note. ICC = intraclass correlations; GLMM = generalized linear mixed model; LR = logistic regression; MERF = mixed effects random forest; RF = random forest; PISA = United States Program in International Student Assessment.

were added to the models, LR, GLMM, and RF all improved appreciably in SGR, LGR, and AUC while accompanying a marginal increase in CE. This finding aligns with Kilham et al.'s (2019) finding that including a mix of Level-1 and Level-2 predictors can allow for a greater degree of model accuracy in classification settings; Kilham et al.'s finding was replicated in the present simulation setting.

Of particular note is the appreciable increase in both LGR and SGR metrics as both the ICC increased and as more predictors were simulated at Level-2. As was found in the study by Kilham et al. (2019), the inclusion of Level-2 predictors allowed for a more holistic representation of the in vivo setting of the outcome being predicted. While perhaps statistically (though not conceptually) counterintuitive, this increase in both LGR and SGR (and consequential increase in overall accuracy) is likely due to the inclusion of context as a salient factor within the model being estimated. Two mechanisms appear to be at work in this situation: The greater emphasis on the context (Level-2 cluster membership) as a key factor in accounting for variability in the outcome (increased ICC); and the inclusion of contextual factors (Level-2 variables) as predictors of outcome groups. However, the finding was not replicated in the current PISA examination with all methods except MERF yielding non-zero SGR only when all predictors were situated at Level-1. Consequently, it cannot be fully determined whether it is simply the mechanism of including context-level predictors or the niche situation of the constellation of PISA predictors that affects model accuracy with versus without Level-2 predictors.

Of additional note is the consideration of sample size as a salient factor affecting accuracy metrics. As was noted in the literature review, estimating a mixed effects model can be performed and may yield relatively accurate parameter estimates, but is likely to be underpowered (McNeish & Stapleton, 2016). However, this degradation in statistical power may not result in a concomitant decrease in LGR and SGR metrics. A smaller sample size in classification models with multilevel data—by way of either smaller cluster size, fewer clusters, or both—may be accompanied by a concomitant increase in model computed error metrics, even if not necessarily raw accuracy rates (e.g., LGR and SGR; Beleites et al., 2012; Figueroa et al., 2012; Raudys & Jain, 1991) when models are estimated on smaller samples. The absence of a decrease in LGR, SGR, and AUC with smaller sample sizes was most apparent in LR, RF, and GLMM. While an increase in cluster size slightly decreased SGR while slightly increasing AUC and CE, the predictive capabilities of all models currently employed (sans MERF under some conditions) were largely comparable with one another.

Holistically, these results largely indicate that regardless of the model selected, all of the prediction accuracy metrics were approximately equivalent: Across the conditions assessed presently, LR, GLMM, and RF all performed comparably to one another, usually within a 1%–3% differential. This finding indicates that it is rather the conditions present within the data, not necessarily the classifier itself, that has the most prominent effect on prediction accuracy. However, the MERF algorithm stood out as having the least consistent results, particularly with respect to SGR. While MERF's LGR was consistently high—aligning with that of the other algorithms—its

SGR, AUC, and CE values were either commensurate with those of RF when all predictors were simulated at Level-1 or when one was simulated at Level-2, but became dramatically worse in nearly all conditions when two Level-2 predictors were included. That this reduction was observed most often when group sizes were unequal, it is likely that MERF is considerably more biased toward the larger group compared with the other methods. Furthermore, when no Level-2 predictors were used in the PISA examination, the pattern of metrics across classifiers largely paralleled their values within the simulation for conditions in which all predictors were simulated at Level-1; this finding was not replicated when Level-2 predictors were included. Therefore, while there was little difference between classifiers under most conditions, it is clear that the implementation of MERF requires additional investigation and careful tuning to make the algorithm viable for predictive classification purposes. A further possible implication of the reduction in accuracy when including Level-2 predictors is that less granular context-level predictors may bias the classifiers toward the larger group. Additional research is necessary to fully disentangle this phenomenon in both simulated and archival data contexts.

Nonetheless, it is apparent that the findings of Kilham et al. (2019) and Speiser et al. (2019) demonstrating similar performance of RF, its mixed effects analogue MERF, and GLMM were largely supported within this study.

Limitations

This study sought to assess the question of predictive capability of fixed and mixed effects models in multilevel data with small samples through both simulation and archival data examinations. Despite the robust nature of the study, several limitations are noted with potential corrections and extensions proposed.

One key point noted in both the Kilham et al. (2019) and Speiser et al. (2019) articles is the statement that while RF may perform similarly to mixed effects classification frameworks (including GLMM, MERF, and Speiser et al.'s BiMM Forest), fixed effects models such as RF do not effectively account for the nesting structure of multilevel data and, thus, do not provide the most accurate conceptual or statistical representation of the phenomenon being investigated. While the inclusion of both Level-1 and Level-2 predictors better accounts for this representation, it does not account for the effect of the nesting structure itself. The absence of conceptual and statistical congruence may limit researchers' ability to effectively use models such as RF or MERF as these models include limited conceptually useful information, in contrast to LR and GLMM.

All models were fit as two-level random-intercepts classifiers to account for the nesting structure of the data. The results presently obtained may differ (particularly in archival data examinations, such as the PISA data) if models were specified with random coefficients. The consideration of nesting structure and the degree to which random effects should be included would become notably more difficult in the context of three-level models with a tertiary nesting structure (e.g., time points within students within schools).

An additional limitation is in the verification of results across the simulation and PISA examinations due, in large part, to the conditions simulated and variables selected in the PISA data set. To maintain continuity across both investigations, the constellation of predictors used were all continuous in nature. However, many of the variables that would likely be more eminently related to the outcome of student retention were either coded into categorical representations of continuous variables, or were Likert-type items with a limited range of potential values. Given the exploratory nature of this study, the use of continuous predictors was reasonable, though future investigations would likely make use of categorical or restricted-range numeric (i.e., Likert-type) variables, particularly with various non-normal distributions. Similarly, the use of three normally distributed predictors in the simulation study represents a limited—and unlikely—scenario and should be further expanded upon in subsequent inquiry. It is likely that the expansion of the number and type of predictors used would result in different results than those presently obtained, and should be investigated in future research.

A tertiary limitation could be found in the form of the four-way ANOVA interactions considered. In many cases, five-way interactions were found to be statistically and practically significant, but were functionally uninterpretable and, thus, severely limited in utility. The five-way interactions tend toward a case similar to common critiques of maximalist theoretical models in which all factors are entangled and relevant, but unable to sufficiently explain observed outcomes. Therefore, while the factors identified within the present analyses were shown to have an appreciable impact, the question of true magnitude and explanatory power stands.

Implications for Practice

Given the results of the present investigation, several recommendations for practice may be proposed. When considering model selection—fixed or mixed effects—the research questions, type of data collected, and purpose of model estimation should be carefully and intentionally defined. The findings of this study largely matched those found by Kilham et al. (2019) and Speiser et al. (2019) with respect to fixed effects models, specifically RF, being viable prediction models despite the multilevel structure of data. Therefore, it is likely unnecessary for researchers to use multilevel classifiers for the purpose of *predictive* classification (in which raw accuracy metrics are favored over accurate parameter estimates), as the results largely do not differ substantially from fixed effects classifiers. Conversely, this result does not indicate any recommendation for models estimated for the purpose of *explanation* (in which parameter estimates are more eminently salient) and, instead, would recommend against the use of MERF for this purpose as not only were its predictions less consistent than the other models, but its R implementation does not feature any explanatory information (e.g., variable importance, coefficient estimates). Therefore, the findings of this study indicate that LR, GLMM, and RF are all approximately equal in predictive capability, regardless of the multilevel data structure. However, it should be noted that—due to the method by which the training and test sets were defined, with clusters being independent across the two data sets—that mixed effects models reduce to fixed effects only in the case of prediction. Test sets consisting of new cases within existing clusters would likely yield different results.

A second recommendation is drawn from the near-uniform pattern of increase in both LGR and SGR as more predictors were simulated at Level-2 and the ICC increases as the salience of context increases, it then becomes increasingly more important to consider these factors within the model chosen. That is, it is strongly recommended that variables measured at both the case and cluster levels be included in classification models when data are nested in nature. This is not recommended, however, when employing MERF, as this classifier yielded worse results when including Level-2 predictors. Considering both the simulation and PISA examination results, including case-level predictors resulted in notably lower SGR, particularly when outcome group sizes were unequal. It should be noted that in the simulation context, even when groups were highly unequal in size (90:10 ratio), both LGR and SGR rates were improved (in the case of SGR, substantially so) when even one Level-2 predictor was included, particularly when the ICC was 0.3 or larger. While it is not feasible to know the ICC of a test/cross-validation data set *a priori*, researchers should take care to consider that of the training set when determining the prediction model architecture. Holistically, a representative collection of predictors should be employed in predictive classification settings to ensure classifiers are provided with sufficient information to make reasonable predictions.

Collectively, the results suggest that when accurate predictions are desired and the explanatory power of the predictors is not of concern, LR, GLMM, or RF would all be eminently appropriate in this task provided that a constellation of both Level-1 and Level-2 predictors is employed, particularly if the ICC is 0.3 or higher. This finding suggests a substantial departure from the conventional and statistical wisdom that nested data must be accounted for within analytical frameworks (due to the correlation structure and violation of independence among cases) and, thus, will require much additional study before universal practical recommendations can be offered. However, this finding is promising for situations in which multilevel models may be theoretically appropriate, but infeasible due to issues such as missing data. Furthermore, while the train-test split of the present study contained different clusters, these results were largely consistent with studies in which the split was conducted at the case-level (e.g., Speiser et al., 2019, 2020) with RF performing similarly to GLMM and/or MERF. Therefore, despite this divergence and violation of conventional assumptions, this predictive capability of fixed effects classifiers relative to mixed effects classifiers should be more closely examined.

Directions for Future Research

Given the limitations of the present study and niche focus on *predictive* classification, additional research is needed to determine the degree to which these results would be found for explanatory classification models. It could be hypothesized that, as is the case with regression models, coefficient estimates may be biased while raw classification metrics may be insensitive to this bias (notwithstanding CE's sensitivity to data conditions). Furthermore, it could be hypothesized that SGR and LGR for the training set (for which explanatory power is afforded through the estimation of model parameters) would be higher than the values observed presently for prediction. This is common in prediction settings in which models are initially trained, but may not generalize well to new data (Steyerberg, 2019). Therefore, an investigation into the effects of multilevel data on the computation of explanatory classifiers—considering both parameter estimates and classification accuracy—may be warranted and contrasted with the results of the present study.

A secondary preeminent future direction would be an expansion of the type of predictors to those categorical or range-restricted numeric (Likert-type items) in nature. Given that much of the data in the social sciences are measured on Likert-type scales (or, minimally, seldom truly continuous), it is evident that this type of variable must be considered with respect to their effects on both fixed and mixed effects classifiers. Similarly, the consideration of various distributions of predictors (heavily skewed, bimodal, etc.) may become eminently reasonable. Several predictors used in the PISA data set were heavily left-skewed in nature, in contrast to the normally distributed predictors of the simulation. Although the results of both the simulation and the PISA examination were largely consistent, the effects of the different predictor distributions are not well-documented, particularly when observed in Likert-type items. Therefore, the consideration of both alternative predictor distributions and types should be a notable area of further inquiry.

Conclusion

The task of prediction is neither a simplistic nor lightly approached endeavor, particularly when the complexities of realistic hierarchical structures enter into the realm of statistical representations, the entanglement of person and context is inevitable, thus dictating the importance of considering multilevel data. This study sought to examine predictive classification models in the presence of multilevel data to uncover the simplicity within the phenomenon. Through both simulation and archival data examinations, it was found that three such classifiers used—LR, GLMM, and RF—all performed similar to one another under most data conditions simulated with results verified in the PISA examination. Similarly, it was found that when incorporating factors at both the case level and cluster level in multilevel contexts, most models performed substantially better than when only Level-1 predictors were used. Therefore, the theoretical postulate that both the micro-level units (e.g., person, time point) and context have an impact on outcomes was supported. Consequently, it is important for researchers to consider and represent all reasonably available levels of data when engaging in predictive modeling. Ultimately, with respect to estimating and using prediction models, it is in the model itself that simplicity may be maintained, as little difference was found between fixed and mixed effects classifiers. However, additional research is required to derive comprehensive practical recommendations in such empirical settings as those presently examined. To make irreducible the basic elements of this contention: A simpler model, when used properly despite complex data, may provide a more parsimonious and equally efficacious tool in the complex task of prediction.

Authors' Note

This manuscript is based on Anthony A. Mangino's doctoral dissertation research, published by ProQuest as *Fixed Effects or Mixed Effects?: Predictive Classification in the Presence of Multilevel Data*.

Acknowledgments

The authors thank Dr. Linda Martin in the Department of Elementary Education and Dr. Lisa Rubenstein in the Department of Educational Psychology at Ball State University for all their support and feedback throughout the dissertation process that led to this article. Their mentorship and support has been so greatly appreciated!

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work utilized Ball State University's Beowulf computing cluster, which is supported by the National Science Foundation (MRI-1726017) and Ball State University, Muncie, Indiana.

ORCID iDs

Anthony A. Mangino https://orcid.org/0000-0002-2699-9989
W. Holmes Finch https://orcid.org/0000-0003-0393-2906

References

Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007). A comparison of machine learning techniques for phishing detection. In *Proceedings of the Anti-Phishing Working Groups 2nd Annual eCrime Researchers Summit (eCrime '07)* (pp. 60–69). Association for Computing Machinery. https://doi-org.proxy.bsu.edu/10.1145/1299015.1299021

Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, *36*(1–2), 105–139.

Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., & Popp, J. (2012). Sample size planning for classification models. *Analytica Chimica Acta*, *760*, 25–33. https://doi.org/10.1016/j.aca.2012.11.007

Bolin, J., & Finch, W. (2014). Supervised classification in the presence of misclassified training data: A Monte Carlo simulation study in the three group case. *Frontiers in Psychology*, *5*, Article 118. https://www.doi.org/10.3389/fpsyg.2014.00118

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

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

Choi, A., Gil, M., Mediavilla, M., & Valbuena, J. (2018). Predictors and effects of grade repetition. *Revista De Economía Mundial*, *48*, 21–42.

Corman, H. (2003). The effects of state policies, individual characteristics, family characteristics, and neighbourhood characteristics on grade repetition in the United States. *Economics of Education Review*, *22*(4), 409–420. https://doi.org/10.1016/S0272-7757(02)00070-5

Crane-Droesch, A. (2017). Semiparametric panel data models using neural networks. *arXiv preprint arXiv*, 1702.06512, https://arxiv.org/pdf/1702.06512.pdf

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, *40*(2), 139–157. https://doi.org/10.1023/A:1007607513941

Downes, M., & Carlin, J. B. (2020). Multilevel regression and poststratification as a modeling approach for estimating population quantities in large population health studies: A simulation study. *Biometrical Journal*, *62*(2), 479–491. https://doi.org/10.1002/bimj.201900023

Eisemon, T. O., United Nations Educational, Scientific, and Cultural Organization, Paris (France), & International Institute for Educational Planning. (1997). *Reducing repetition: Issues and strategies* (Fundamentals of Educational Planning Series, Number 55). International Institute for Educational Planning.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, *12*(1), 8. https://doi.org/10.1186/1472-6947-12-8

Glick, P., & Sahn, D. E. (2010). Early academic performance, grade repetition, and school attainment in senegal: A panel data analysis. *The World Bank Economic Review*, *24*(1), 93–120. https://doi.org/10.1093/wber/lhp023

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

Harrison, R. L. (2010, January). Introduction to Monte Carlo simulation. *AIP Conference Proceedings*, *1204*(1), 17–21.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Ikeda, M., & García, E. (2014). Grade repetition: A comparative study of academic and non-academic consequences. *OECD Journal: Economic Studies*, *2013*(1), 269–315. https://doi.org/10.1787/eco_studies-2013-5k3w65mx3hnx

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (*Vol. 112*, p. 18). Springer.

Karpievitch, Y. V., Hill, E. G., Leclerc, A. P., Dabney, A. R., & Almeida, J. S. (2009). An introspective comparison of random forest-based classifiers for the analysis of cluster-correlated data by way of RF. *PLOS ONE*, *4*(9), Article e7087. https://doi.org/10.1371/journal.pone.0007087

Kilham, P., Hartebrodt, C., & Kändler, G. (2019). Generating tree-level harvest predictions from forest inventories with random forests. *Forests*, *10*(1), 20. https://doi.org/10.3390/f10010020

Kreft, I. G. (1996). *Are multilevel techniques necessary? An overview, including simulation studies* [Unpublished manuscript]. California State University, Los Angeles.

Lavery, M. R., Acharya, P., Sivo, S. A., & Xu, L. (2019). Number of predictors and multicollinearity: What are their effects on error and bias in regression? *Communications in Statistics-Simulation and Computation*, *48*(1), 27–38. https://doi.org/10.1080/03610918.2017.1371750

Lee, J., & Stankov, L. (2018). Non-cognitive predictors of academic achievement: Evidence from TIMSS and PISA. *Learning and Individual Differences*, *65*, 50–64. https://doi.org/10.1016/j.lindif.2018.05.009

Lei, P. W., & Koehly, L. M. (2003). Linear discriminant analysis versus logistic regression: A comparison of classification errors in the two-group case. *The Journal of Experimental Education*, *72*(1), 25–49. https://doi.org/10.1080/00220970309600878

Little, R. J. (2013). In praise of simplicity not mathematistry! Ten simple powerful ideas for the statistical scientist. *Journal of the American Statistical Association*, *108*(502), 359–369. https://doi.org/10.1080/01621459.2013.787932

Luke, D. A. (2019). *Multilevel modeling* (*Vol. 143*). SAGE.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Mangino, A. A., & Finch, W. H. (2021). Prediction with mixed effects models: A Monte Carlo simulation study. *Educational and Psychological Measurement*, *81*(6), 1118–1142. https://doi.org/10.1177/0013164421992818

Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., & de Mendonça, A. (2011). Data mining methods in the prediction of dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Research Notes*, *4*(1), 299–313. https://doi.org/10.1186/1756-0500-4-299

McMahon, S. D., Parnes, A. L., Keys, C. B., & Viola, J. J. (2008). School belonging among low-income urban youth with disabilities: Testing a theoretical model. *Psychology in the Schools*, *45*(5), 387–401. https://doi.org/10.1002/pits.20304

McNeish, D., & Kelley, K. (2019). Fixed effects models versus mixed effects models for clustered data: Reviewing the approaches, disentangling the differences, and making recommendations. *Psychological Methods*, *24*(1), 20–35. https://doi.org/10.1037/met0000182

McNeish, D., & Stapleton, L. M. (2016). Modeling clustered data with very few clusters. *Multivariate Behavioral Research*, *51*(4), 495–518. https://doi.org/10.1080/00273171.2016.1167008

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, *22*(1), 114–140. https://doi.org/10.1037/met0000078

Muchlinski, D., Siroky, D., He, J., & Kocher, M. (2016). Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis*, *24*(1), 87–103. https://doi.org/10.1093/pan/mpv024

Murtaugh, P. A. (2007). Simplicity and complexity in ecological data analysis. *Ecology*, *88*(1), 56–62. https://doi.org/10.1890/0012-9658(2007)88[56:SACIED]2.0.CO;2

Ngufor, C. (2019). *Vira: Virtual Intelligent Robot Assistant* (R Package Version 0.1). https://rdrr.io/github/nguforche/Vira/

Ngufor, C., Van Houten, H., Caffo, B. S., Shah, N. D., & McCoy, R. G. (2019). Mixed effect machine learning: A framework for predicting longitudinal change in hemoglobin A1c. *Journal of Biomedical Informatics*, *89*, 56–67. https://doi.org/10.1016/j.jbi.2018.09.001

Paccagnella, O. (2011). Sample size and accuracy of estimates in multilevel models: New simulation results. *Methodology*, *7*(3), 111–120. https://doi.org/10.1027/1614-2241/a000029

Palvanov, A., & Cho, Y. I. (2018). Comparisons of deep learning algorithms for MNIST in real-time environment. *International Journal of Fuzzy Logic and Intelligent Systems*, *18*(2), 126–134. https://doi.org/10.5391/IJFIS.2018.18.2.126

Ramos, D., Franco-Pedroso, J., Lozano-Diez, A., Gonzalez-Rodriguez, J., Ramos, D., Gonzalez-Rodriguez, J., Franco-Pedroso, J., & Lozano-Diez, A. (2018). Deconstructing cross-entropy for probabilistic binary classifiers. *Entropy*, *20*(3), 208. https://doi.org/10.3390/e20030208

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (*Vol. 1*). SAGE.

Raudys, S. J., & Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*(3), 252–264. https://doi.org/10.1109/34.75512

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *American Journal of Epidemiology*, *179*(6), 764–774. https://doi.org/10.1093/aje/kwt312

Speiser, J. L., Wolf, B. J., Chung, D., Karvellas, C. J., Koch, D. G., & Durkalski, V. L. (2019). BiMM forest: A random forest method for modeling clustered and longitudinal binary outcomes. *Chemometrics and Intelligent Laboratory Systems*, *185*, 122–134. https://doi.org/10.1016/j.chemolab.2019.01.002

Speiser, J. L., Wolf, B. J., Chung, D., Karvellas, C. J., Koch, D. G., & Durkalski, V. L. (2020). BiMM tree: A decision tree method for modeling clustered and longitudinal binary outcomes. *Communications in Statistics-Simulation and Computation*, *49*(4), 1004–1023. https://doi.org/10.1080/03610918.2018.1490429

Steyerberg, E. W. (2019). *Clinical prediction models*. Springer.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973

VanDerHeyden, A. M. (2013). Universal screening may not be for everyone: Using a threshold model as a smarter way to determine risk. *School Psychology Review*, *42*(4), 402–414. https://www.doi.org/10.1080/02796015.2013.12087462

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., & Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, *3*(8), Article e002847. https://doi.org/10.1136/bmjopen-2013-002847

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: Machine learning and classification methods as alternatives to logistic regression. *Journal of Clinical Epidemiology*, *63*(8), 826–833. https://doi.org/10.1016/j.jclinepi.2009.11.020

Wößmann, L. (2003). Schooling resources, educational institutions and student performance: The international evidence. *Oxford Bulletin of Economics and Statistics*, *65*(2), 117–170. https://doi.org/10.1111/1468-0084.00045

Wu, M., & Zhang, Z. (2010). *Handwritten digit classification using the MNIST data set* (Course Project CSE802: Pattern Classification & Analysis). https://www.researchgate.net/profile/Ming-Wu-45/publication/228685853_Handwritten_Digit_Classification_using_the_MNIST_Data_Set/links/5409c76b0cf2f2b29a2c57b5/Handwritten-Digit-Classification-using-the-MNIST-Data-Set.pdf

Yan, P. (2019). *Anomaly detection in categorical data with interpretable machine learning: A random forest approach to classify imbalanced data* [Thesis]. http://liu.diva-portal.org/smash/record.jsf?pid=diva2%3A1330907&dswid=6631

Zellner, A., Keuzenkamp, H. A., & McAleer, M. (2001). *Simplicity, inference and modelling: Keeping it sophisticatedly simple*. Cambridge University Press.

Zhang, J. L., & Haerdle, W. K. (2010). The Bayesian additive classification tree applied to credit risk modelling. *Computational Statistics & Data Analysis*, *54*(5), 1197–1205. https://doi.org/10.1016/j.csda.2009.11.022

Zigler, E., & Phillips, L. (1961). Psychiatric diagnosis: A critique. *Journal of Abnormal and Social Psychology*, *63*(3), 607–618. https://doi.org/10.1037/h0040556