

# Prediction With Mixed Effects Models: A Monte Carlo Simulation Study

Educational and Psychological  
Measurement  
1–25

© The Author(s) 2021

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0013164421992818

journals.sagepub.com/home/epm



Anthony A. Mangino<sup>1</sup>  and W. Holmes Finch<sup>1</sup> 

## Abstract

Oftentimes in many fields of the social and natural sciences, data are obtained within a nested structure (e.g., students within schools). To effectively analyze data with such a structure, multilevel models are frequently employed. The present study utilizes a Monte Carlo simulation to compare several novel multilevel classification algorithms across several varied data conditions for the purpose of prediction. Among these models, the panel neural network and Bayesian generalized mixed effects model (multilevel Bayes) consistently yielded the highest prediction accuracy in test data across nearly all data conditions.

## Keywords

multilevel modeling, predictive modeling, classification

Statistical classification analyses are a commonly used family of modeling techniques designed to predict into which of  $k$  groups an individual or case belongs based on the set of  $p$  predictors (Hastie et al., 2009). These analyses can be used for the purpose of explanation—enhanced understanding of the relative roles and efficacy of a variety of predictors in identifying the group to which a case belongs—and/or prediction—the estimation of a model that could correctly classify new cases (Steyerberg, 2019). In considering analyses for the latter purpose, the intended outcome of the present study, two data sets are typically used: The training set, on which a model is initially fit or trained, and the test or cross-validation set, which consists of new, previously unused cases. The model obtained using the training set is then applied to the test data

---

<sup>1</sup>Ball State University, Teachers College, Muncie, IN, USA

## Corresponding Author:

Anthony A. Mangino, Ball State University, Teachers College, 1950 West Riverside Avenue, Room 505, Muncie, IN 47306, USA.

Email: [aamangino@bsu.edu](mailto:aamangino@bsu.edu)

from which predictions are obtained. The accuracy of these cross-validated predictions is used to ascertain the performance of the model.

Predictive classification models have been employed in a wide array of situations, including for academic learning disability and intervention determinations (Stuebing et al., 2012; VanDerHeyden, 2013), suicide attempts (Hedegaard et al., 2018; Mann et al., 2008; Ribeiro et al., 2016), school dropout (Morris et al., 2005), and psychiatric diagnosis (Zigler & Phillips, 1961). When considering the factors affecting predictive accuracy in such contexts as those described above, it is imperative to consider the conditions existing within the data in order to select the most optimal classifier. Unbalanced outcome group sizes—such as those found in situations of predicting student dropout (Morris et al., 2005; U.S. Department of Education, National Center for Education Statistics, 2018), severe mental illness (National Institute of Mental Health, 2019), and suicide attempts (Hedegaard et al., 2018; Ribeiro et al., 2016)—as well as poorly separated groups, have the effect of reducing model prediction accuracy (Bolin & Finch, 2014; Holden et al., 2011; Lei & Koehly, 2003). Under the condition of unbalanced groups, classifiers tend to yield high overall accuracy and large-group recovery (LGR) rates but with a concomitant diminution of small-group recovery (SGR) rates (Kessler et al., 2003; Lei & Koehly, 2003; Mann et al., 2008; Morris et al., 2005). Similarly, when groups are poorly separated (as determined by the effect size, Cohen's  $d$  [Cohen, 1988]; or Mahalanobis' distance, [Mahalanobis, 1936]), more cases are likely to be misclassified into the larger group (Fletcher et al., 2014; Ho & Basu, 2002; Luengo & Herrera, 2012). This misclassification is particularly evident for “border” cases residing close to the decision boundary—the point, plane, or probability at which the model determines classification into one group or the other.

Given the factors salient within predictive modeling contexts, the present study has two distinct goals:

1. Compare the relative predictive capability of six different classifiers—five of which being relatively new mixed effects models—in the presence of multilevel data structures under various conditions; and
2. Assess the effects of varied conditions within nested data structures on the classifiers through the systematic manipulation of these conditions within a Monte Carlo simulation.

Our results will indicate the methods with the greatest predictive efficacy across differing conditions within multilevel data and will serve to provide recommendations for model selection to practitioners and researchers within various areas of the social sciences.

### *Multilevel Data Structure*

Much of the data obtained in the social sciences, business, and health care fields, particularly those collected in educational contexts, are structured such that individual

participants are nested within higher order clusters (e.g., students nested within schools). These situations result in cases within the same cluster correlating with one another (intraclass or intracluster correlation; ICC), thereby violating the required assumption of independence of cases inherent in many statistical models. Consequently, multilevel, or mixed effects, models must be employed to account for the high ICC resulting from the nesting of individual cases within higher order clusters (Raudenbush & Bryk, 2002). Despite the relative frequency with which this data structure is seen, only the most typical multilevel model—the generalized linear mixed model (GLMML)—has been regularly employed as a classifier.

A primary issue for researchers working with multilevel models concerns the number of higher order clusters that are necessary for models to fit the data appropriately: Maas and Hox (2005) suggest between 60 and 100 clusters; Meuleman and Billiet (2009) suggest 40 to 60 clusters; and McNeish and Kelley (2019) propose 30 to 50 clusters. However, no general consensus has been reached among researchers in this area and, consequently, many studies employing multilevel methodologies may do so with minimal theoretical, empirical, or conventional guidance. Additionally, the impact of the number of clusters has yet to be fully determined in the context of classification analyses. Studies considering single-level classifiers have found that increased sample sizes tend to yield more stable results across iterations within simulation studies (Bolin & Finch, 2014; Pohar et al., 2004), but no analogous results have been found for multilevel classifiers. Also uniquely pertaining to multilevel contexts is the ICC and between-cluster correlations. To date, no research has been conducted considering the effects of various correlation structures on the accuracy of multilevel models, neither for regression nor classification. While ICCs as low as or 0.05 may be considered sufficiently large to employ multilevel models (LeBreton & Senter, 2008), no simulation or archival analysis studies have considered varied ICCs in their comparisons of model accuracy rates. Consequently, a number of factors inherent in classification models have been previously studied, but many multilevel classifiers have not been subject to the same rigorous empirical treatment. A primary purpose of the current study was to investigate the impact of sample size, at both Levels 1 and 2, on the prediction accuracy of several commonly used multilevel classification models.

### *Mixed Effects Models*

While a regression-based general linear model is often implicated when researchers discuss multilevel models, a number of other techniques have since become tractable with commonly used statistical software packages, including R and Python. These models utilize varied estimation methods and statistical paradigms in order to calculate parameters and predict outcomes. This study considers six classifiers that have been implemented and include estimates of random effects, thereby allowing their use in multilevel contexts. Each algorithm is briefly described below. A generalized representation of the mixed effects model for each of the following algorithms can be represented as

$$Y_i = X_i\beta + Z_i\gamma_i + \varepsilon_i, \quad (1)$$

where  $Y_i$  = response for case  $i$ ,  $X_i$  = matrix of fixed effects for case  $i$ ,  $\beta$  = vector of coefficients for fixed effects,  $Z_i$  = matrix of random effects for case  $i$ ,  $\gamma_i$  = vector of coefficients for random effects, and  $\varepsilon_i$  = random error for case  $i$  (Bagiella et al., 2000).

The fundamental differences among the algorithms currently utilized exists in the estimation method for  $\beta$  and  $\gamma_i$ , thus allowing this model to serve as a basic mathematical representation of the currently employed algorithms. It is in these novel methodologies that the present inquiry is focused and, as such, five mixed effects models relatively novel to the social sciences (i.e., those that are not yet widely employed) are considered within the context of a Monte Carlo simulation study. The models were selected based on their novelty and paucity of empirical examination, or their superiority to the more traditional generalized linear mixed effects model (multilevel logistic regression; GLMML) commonly employed in the social sciences. A more extensive discussion comparing these methods and illustrating the rationale for their selection in this study follows.

### *Previous Comparisons*

Few existing studies have compared the above-discussed methods with one another in terms of classification accuracy with a considerable number more focusing on continuous outcome variables. To our knowledge, no research has been published, as yet, comparing these classification approaches to one another using a Monte Carlo simulation study design across a wide variety of data conditions and parameters. Initially, Sela and Simonoff (2012) found that recursive partitioning expectation minimization trees (REEMTree) proved to be comparable to GLMML in both simulated and archival regression data regardless of the underlying data structure but did not demonstrate substantively improved model accuracy. Additionally, Finch (2015) illustrated the utility of REEMTree for regression in the multilevel context, using an extant data set. In the 2014, introduction of mixed effects random forests (MERF), Hajjem et al. (2014) found that MERF outperformed standard linear models, GLMML, single-level classification or regression tree (CART), and random forests in a comparison on box office revenue data. Hajjem et al. (2017) built on the prior study by conducting a simulation study incorporating variants of GLMML and a generalized mixed effects regression tree (GMERT) approach across conditions of large and small fixed effects, and no random effects, random intercepts only, and random coefficients. The GMERT models consistently outperformed the GLMML variants on measures of predictive mean absolute deviation and predictive misclassification rates; to date, this represents the most comprehensive simulation study on multilevel classifiers. Expanding on this finding, Kilham et al. (2019) compared REEMTree, MERF, and GLMML, with MERF performing best on measures of both variance explained and root mean squared error in harvest prediction data. Capitaine et al. (2019) also found substantially reduced bias in estimates for MERF above REEMTree and GLMML in

the context of both simulated and archival low- and high-dimensional genetic data. Collectively, the findings of these studies suggest that MERF may demonstrate a high degree of efficacy when compared with a number of other mixed effects models.

Additional work examining the performance of these multilevel models has been conducted in the longitudinal data context, where individuals are measured at multiple points in time. The individual is then at Level 2 (L2), and the individual measurements are at Level 1 (L1). Ngufor et al. (2019) found that all mixed effects machine learning algorithms—MERF, Megbm, REEMTree, and GLMML—provided improvements on their single-level counterparts in predicting patient hemoglobin A1c levels, change in glycemic control, and lung cancer remission, among other outcomes. Ngufor et al. (2019) also found that as the number of individuals (i.e., L2 units) included in the model increased, so too did all accuracy metrics (i.e., overall accuracy, sensitivity, specificity, area under the curve, and positive predictive value). Crane-Droesch (2017) compared panel neural networks (PNNET) with single-level models—including standard regression, variable selection (i.e., LASSO), and random forest models—in the prediction of agricultural yield with results in favor of PNNET. Xiong et al. (2019) also found a substantive improvement in the accuracy of PNNET over and above single-level models and GLMML in the context of high-dimensional video image data, thereby bolstering Crane-Droesch's (2017) findings. Presently, no studies have been found to have incorporated Hadfield's (2010) Bayesian multilevel framework in comparisons with other multilevel models on accuracy rates in classification or regression contexts.

A more simplistic classifier—in the form of the naive Bayes (NB) classifier—is presently included as a de facto baseline method within the Bayesian statistics paradigm and a more parsimonious, yet efficacious, method against which the multilevel Bayes framework could be compared. Studies by Demichelis et al. (2006) and Zhang et al. (2018) have indicated that multilevel Bayesian frameworks may only marginally outperform NB under some conditions. Presently, this method is considered to illustrate whether a more parsimonious model may be used in lieu of the more complex multilevel Bayes (in a manner analogous to that of single-level random forests when compared with MERF in Kilham et al.'s, 2019 study).

The literature reviewed presently has illustrated that the standard GLMML has been broadly outperformed by many of the neoteric methods discussed above and, thus, has sufficient evidence as to warrant exclusion from the present study. Consequently, the present study focuses on a comparison of REEMTree, MERF, Megbm, PNNET, and Bayes in addition to the simpler NB method as a more parsimonious comparative baseline method. The mathematical underpinnings of these algorithms are detailed in the following section.

### *Random Effects Expectation Minimization Trees*

Random effects expectation minimization trees (REEMTree) were proposed by Sela and Simonoff (2012) as a method for fitting a recursive partitioning model (CART)

while accounting for the L2 variance (random effects) due to the clusters. In this method, CART is first fit to the data by identifying increasingly more homogeneous groups within the data such that a classification decision can be made, then the random effects are estimated using the CART model parameters. This process is repeated until the model yields maximally homogeneous terminal groups (nodes) and thereby converges (Finch, 2015). The resulting model could be described as

$$y_{ij} = Z_{ij}\mathbf{b}_i + f(x_{ij1}, \dots, x_{ijK}) + \varepsilon_{ij}, \quad (2)$$

where  $Z_{ij}$  = design matrix for random effects of dimensions  $i * j$ ,  $\mathbf{b}_i$  = vector of L2 effects,  $f$  = known linear function for L1 effects,  $x_{ij1}, \dots, x_{ijK}$  = L1 inputs for variables 1 to K, and  $\varepsilon_{ij}$  = error for case  $i$  in cluster  $j$  (Sela & Simonoff, 2012).

### Mixed Effects Random Forests

Random forests are an ensemble method in which multiple iterations of a CART model are estimated using bootstrapped samples of both the individuals and predictor variables (Hastie et al., 2009). Once a set of  $m$  (e.g., 500) such trees has been obtained, predictions for members of the test sample are obtained for each tree, results of which are averaged, and from which classification accuracy can be calculated. MERF take this concept further by estimating the random effects via the parameters of their associated singular trees within the ensemble (Hajjem et al., 2014). This model then makes classification decisions based on the population-averaged single-level random forest prediction and its corresponding random effects. The resulting model largely resembles that of REEMTree such that

$$y_i = f(X_i) + Z_i b_i + \varepsilon_i, \quad (3)$$

where  $f(X_i)$  = fixed effects for input  $X$  for case  $i$ ,  $Z_i$  = Matrix of L2 inputs,  $b_i$  = vector of random effects where  $b_i \sim N(0, D)$ , and  $\varepsilon_i$  = error for case  $i$ , where  $\varepsilon_i \sim N(0, R_i)$  (Hajjem et al., 2014).

As was also the case with both GLMML and REEMTree, MERF assumes that the composite random effects  $Z_i b_i$  are linear in nature, thus leading to potential limitations in the relationships that can be modeled in the random effects.

### Mixed Effects Gradient Boosting Machines

Ngufor et al. (2019) incorporated random effects estimation into a gradient boosting framework (Megbm) in which a sequence of weak learners in an ensemble progressively became increasingly more accurate in their estimation via learning through the loss function; this occurs for both the fixed and random effects. By estimating weak learners sequentially, the negative gradient of the loss function can be learned, thereby minimizing the loss function when aggregated across the ensemble. This method follows Friedman's (2001) standard gradient boosting algorithm, which

learns progressively through the previous model's residuals. The base boosting model can be represented as

$$f(x) = \sum_{b=1}^B \lambda \hat{f}^b(x), \quad (4)$$

where  $\lambda$  = shrinkage parameter for model learning and  $\hat{f}^b$  = weak learners 1, . . . ,  $B$  (James et al., 2013).

The mixed effects expansion of boosting can then be considered as

$$f(x) = \sum_{v=1}^V c_v I(x \in R_v), \quad (5)$$

where  $c_v$  = constant for mean  $v$  of responses from observations  $x \in R_v$ ,  $R_v$  = set of disjointed regions in the feature space, and  $I$  = Indicator function mapping  $x$  to regions in  $R_v$ , (Ngufor et al., 2019).

### Panel Neural Networks

Xiong et al. (2019) also proposed an expansion of the neural network architecture—based on the premise of Crane-Droesch's (2017) PNNET—to include the estimation of random effects while accommodating nonlinearity in the model structure. The resulting PNNET utilizes a nonlinear transformation to create derived features before estimating the combined fixed and random effects from those derived features. A mixed effects linear model is then estimated from the derived features and parameters with the goal of minimizing the loss function. Parameters for PNNET are estimated via the expectation maximization algorithm with embedded optimization via stochastic gradient descent. It should be noted that PNNET allows for the inclusion of linear fixed effects at L1 and nonlinear random effects at L2, differentiating it from the previously discussed models (Xiong et al., 2019). The resulting PNNET is derived from nonlinear mixed effects model.

$$y_i = v(\Phi_i) + \varepsilon_i, \quad (6)$$

$$\Phi_i = X_i \beta + Z_i u_i,$$

where  $\beta$  = coefficient for fixed effects,  $u_i$  = coefficient for random effects,  $v$  = non-linear function for composite predictor effects  $\Phi_i$ , and  $\Phi_i$  = effect for case  $i$  in terms of fixed effect  $\beta$  and random effects  $u_i$  (Xiong et al., 2019).

The resulting PNNET, then incorporates the nonlinear structure in Equation (6) to form the resulting model

$$y_i = \Gamma(X_i) \beta + \Gamma(X_i) u_i, \quad (7)$$

where  $\Gamma$  = nonlinear transformation for fixed and random effects,  $X_i$  = assumed output from neural network  $\Gamma$ ,  $\beta$  = fixed effects coefficient for final layer, and  $u_i$  = random effects coefficient for final layer. The model in Equation (7) allows for

separate estimates of the fixed  $\Gamma(X_i)\beta$  and random effects  $\Gamma(X_i)u_i$  (Xiong et al., 2019).

### *Bayesian Markov-Chain Monte Carlo Model*

The Bayesian Markov chain Monte Carlo method (Bayes) utilizes the Bayesian statistical paradigm in which a prior distribution is hypothesized based on existing data or theory and is incorporated into the model parameters such that the posterior distribution can be estimated (Hadfield, 2010). Fundamentally, Bayesian methods use a hypothesized probability distribution (the prior) to estimate a probability of a given outcome (the posterior), which can be represented as

$$f_i(y_i|l_i), \quad (8)$$

Equation 8 implies the estimation of a probability density function for the observed data  $y$ , given a latent variable  $l$  for each case  $i$  (see Gelman et al., 2013 or Kruschke, 2014 for more comprehensive treatments of Bayesian methods). In order to estimate the posterior distribution for the outcome variable, the model proposed by Hadfield (2010) then takes the form:

$$f_P(y_i|\lambda = \exp(l_i)), \quad (9)$$

with canonical parameter  $\lambda$  for assumed Poisson density function  $f_P$ , and vector of latent variables  $l$  predicted by the equation:

$$l = X\beta + Zu + e, \quad (10)$$

where  $X$  = design matrix for fixed effects  $\beta$ ,  $Z$  = design matrix for random effects  $u$ , and  $e$  = model residual vector.

The model described in Equations 9 and 10 can be estimated via Markov chain Monte Carlo (MCMC), with the latent variable component being estimated via the Metropolis–Hastings algorithm. Parameters  $\beta$  and  $u$  are then Gibbs sampled and follow an assumed multivariate normal distribution (Hadfield, 2010).

### *Naive Bayes Classifier*

While the Bayesian MCMC generalized linear mixed model functions in a manner analogous to the GLMML, it also carries with it an element of complexity inherent in any multilevel analytical framework. A more simplistic, yet frequently employed, Bayesian framework is the NB classifier. This classifier operates by estimating the marginal densities for each of the outcome groups given the predictors of the model resulting in the form

$$f_k(X) = \prod_{k=1}^p f_{jk}(X_k), \quad (11)$$



where  $f_k$  = unconditional class density function and  $f_{jk}$  = class-conditional marginal density function (Hastie et al., 2009).

The NB model assumes that the predictors  $X_k$  are independent—an assumption violated by multilevel data—yet despite the potential bias within estimates for each class, the model often outperforms many more complex classifiers. In studies comparing standard and multilevel NB models, the standard model NB was often only marginally outperformed by the hierarchical model (e.g., Demichelis et al., 2006; Zhang et al., 2018).

### **Current Study**

Of the literature described above, only three studies focused specifically on classification situations—namely Hajjem et al. (2017), Kilham et al. (2019), and Ngufor et al. (2019)—rather than regression situations. Given the sparsity of comparisons among the current algorithms in the classification context, this work stands as a novel contribution via the relatively comprehensive set of conditions on which the data were simulated, and the breadth and relative novelty of algorithms employed. These conditions allowed for a thorough comparison of the present classifiers in order to provide researchers and practitioner-researchers with guidelines on the optimal methods to employ under varied data conditions.

Given the exploratory nature of this study, the principal research questions pertain to which of the present multilevel classifiers will prove most efficacious in predictive accuracy on test data under various conditions. This study was also designed to identify the relative effects of the varied data conditions on the accuracy rates of the classifiers. Specifically, the extent to which ICCs, between-cluster correlations, group size ratios, cluster size, and sample sizes affected the predictive capability of the classifiers was examined.

### **Method**

To address the goals of this study, a Monte Carlo simulation was employed to facilitate specific and controlled manipulation of the parameters of interest. The existing literature, as described above, considered both archival and simulated data sets but in somewhat limited cases. The purpose of the current study was to expand on this earlier work by including a wider array of methods and study conditions than have been used previously. The conditions currently employed were based on principle data features in either multilevel and classification contexts. Sela and Simonoff (2012) utilized varied L1 and L2 sample sizes with the former featuring values from 50 to 2,000 and the latter featuring values from 10 to 100; Ngufor et al. (2019) also incorporated varied numbers of clusters and time points from 10 to 60 clusters and 1 to 4 time points (within-cluster sample size). Additionally, Bolin and Finch (2014), Holden et al. (2011), and Lei and Koehly (2003) each considered varied parameters in single-level classification situations including varied group size ratios and group

separation, among other factors, in simulated data. The salient effects of the ICC must also be considered in multilevel contexts, as this is precisely the factor that necessitates use of mixed effects models (with LeBreton & Senter's [2008] recommendation being any  $ICC > 0.05$  necessitating multilevel models; Raudenbush & Bryk 2002). Consequently, the present study includes a total of 270 data conditions with varied parameters for group size ratio (2 conditions), number of L1 cases per cluster (3), number of L2 clusters (5), ICC (3), and between-group correlation (3) each compared across seven classification methods (see Appendix Table A1 for full list of manipulated study parameters).

The present study incorporated a fully factorial design in which all the conditions are sequentially and systematically crossed with one another to yield the full set of conditions. That is, the present simulation conditions result in a  $2 \times 3 \times 5 \times 3 \times 3$  factorial design across the aforementioned conditions. The resulting Monte Carlo simulation seeks to assess the relative effects of each parameter to be manipulated within the data (Hammersley, 2013). Furthermore, to maintain simplicity, the data simulated will feature only a single normally distributed L1 predictor variable and each model will be fit with random intercepts only. The simulation follows the Monte Carlo design in which data are simulated, each of the classifiers are fit to a training data set, and cross-validated with a test data test. Each condition was replicated 100 times with raw and mean outcome metrics obtained. Four outcome measures are used in order to obtain results for each aspect of the classification outcomes: Overall accuracy, sensitivity (true positive identification rate or SGR), specificity (true negative identification rate or LGR), and binary cross-entropy (CE). The former three metrics align with Ngufor et al.'s (2019) method and serve to identify the key aspects of classification accuracy. That is, in the unequal group size condition, it is likely for each model to have high overall accuracy and LGR, but SGR will likely be lower as the group of interest will be substantially smaller in size than the larger group. Consequently, SGR should be considered relative to overall accuracy and LGR.

- Overall accuracy acts as the total number of cases correctly classified and is calculated by the following equation:  $\frac{TP + TN}{TP + TN + FP + FN}$  and is represented on a 0 to 1 scale corresponding to a percentage correctly classified.
- Sensitivity or SGR is calculated as:  $\frac{TP}{TP + FN}$  and also results in a value from 0 to 1 representing a percentage correctly classified into the smaller/positive group.
- Specificity or LGR is calculated as:  $\frac{TN}{TN + FP}$  and also results in a 0 to 1 value interpreted as a percentage correctly classified into the larger/negative group.
- CE serves as a flexible measure of shared disorder between two distributions and serves as a measure of overall model uncertainty (Richard & Lippmann,

1991). CE is interpreted such that larger values indicate more disorder and higher error rates.

It should be noted that the overall accuracy, sensitivity, and specificity outcomes used in this study are identical to those used with the receiver operating curve (Fawcett, 2006).

As noted above, the models were initially fit to the training set with all accuracy measures obtained, then they are applied to the test set in accordance with cross-validation methods (Steyerberg, 2019). The outcome metrics were then obtained for each model on the test data. The loss for each of the outcome metrics was then obtained in order to determine the degree to which the model loses accuracy from training to test settings (e.g.,  $CE_{\text{Train}} - CE_{\text{Test}} = CE_{\text{Loss}}$ ). Both the loss and test set accuracy measures were considered as key outcomes for the present analysis. To determine the differences between conditions and classifiers, a series of factorial analyses of variance (ANOVAs; with corrected  $p$  values) were employed to assess for statistically significant differences between conditions on both main effects and interactions using the average outcome values for each condition; the highest order interactions with omega-squared values greater than 0.1 ( $\omega^2 > 0.1$ ) were discussed, as this is the de facto benchmark for practical significance due to its status as a moderate-to-large effect (Finch & French, 2012; Okada, 2013; Rosen & DeMaria, 2012). Furthermore, visual analysis was conducted on a graphical representation of the results to provide for ease of interpretation.

The simulation was implemented in the R programming language (R Core Team, 2019) and utilized a number of packages to encompass the variety of methods presently employed. REEMTree is implemented using the REEMtree package (Sela & Simonoff, 2012); MERF and Megbm utilize the Vira package (Ngufor, 2019); PNNET uses the panelNNET package (Crane-Droesch, 2017); multilevel Bayes employs the MCMCglmm package (Hadfield, 2010), and NB was fit using the naiveBayes function from the e1071 library (Meyer et al., 2020). Multilevel data were simulated using the sim.multi function from the psych package in R (Revelle, 2017).

## Results

### Cross-Entropy

ANOVA results (Table 1) indicated that the interactions of estimation method by ICC, method by number of cases per cluster, and method by number of clusters were each statistically significantly related to the CE value. Figure 1 includes the CE values by method and ICC. Across estimators, CE increased in value concomitantly with increases in the ICC. This increase was steepest for MERF, MEGBM, and REEMTree. Finally, CE was smallest for the Bayesian and PNNET approaches across ICC levels, and largest for MERF, REEMTree, and Megbm when the ICC was 0.50 or greater. The NB classifier yielded the worst performance when the ICC

**Table 1.** Analysis of Variance Results for Statistically Significant Terms.

Term	<i>F</i>	Degrees of Freedom	<i>p</i>	$\omega^2$
Cross-entropy				
Method $\times$ ICC	2071.21	30, 163	<.0001	0.95
Method $\times$ cases	15.10	18, 613	<.0001	0.14
Method $\times$ clusters	17.75	24, 613	<.0001	0.22
Sensitivity				
Method $\times$ proportion in group 1	1906.99	30, 613	<.0001	0.95
Method $\times$ ICC	49.11	18, 613	<.0001	0.44
Method $\times$ cases	15.10	24, 613	<.0001	0.12
Specificity				
Method $\times$ proportion in group 1	450.60	30, 613	<.0001	0.83
Method $\times$ ICC	7.68	18, 613	<.0001	0.40
Method $\times$ cases	2.67	24, 613	<.0001	0.10

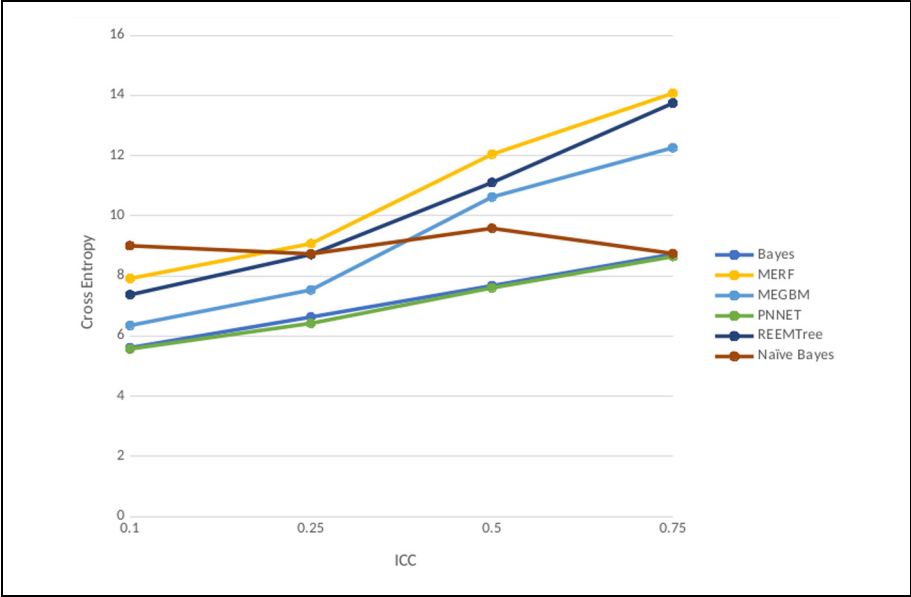
Note. ICC = intraclass or intracluster correlation.

was 0.1 but retained its performance as the ICC increased, performing comparably to multilevel Bayes and PNNET when the ICC was 0.75.

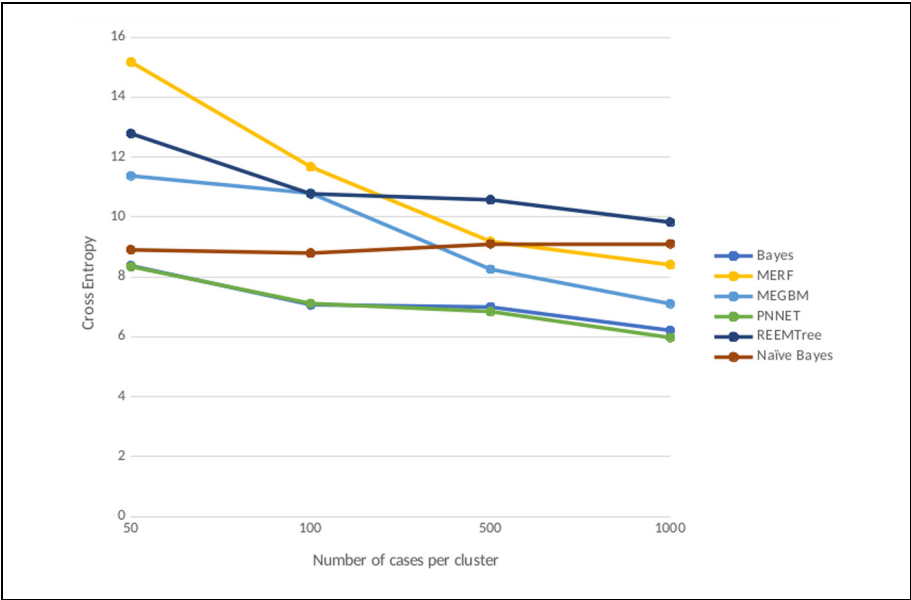
CE by the number of cases per cluster appears in Figure 2. The value of CE was smaller for larger numbers of cases per cluster, with the multilevel Bayesian estimator and PNNET having the lowest values across all number of cases, and MERF and REEMTree the largest for 50 cases. The steepest decline in CE across the number of cases occurred for MERF and Megbm, indicating that the within cluster sample size was more strongly related to performance of these algorithms than was the case for REEMTree, multilevel Bayes, NB, or PNNET. CE by the number of clusters and estimator appears in Figure 3. As was the case for the number of cases per cluster, CE was lower for larger number of L2 clusters. In addition, CE values were smallest for the multilevel Bayesian and PNNET approaches to estimation across all conditions. MERF exhibited the second highest CE for 10 and 20 clusters but performed in the midrange of methods when the number of clusters was 30 or greater.

### Sensitivity Rates

The ANOVA results (Table 1) revealed that the interaction of the proportion of cases in Group 1 with estimation method, ICC by method, and number of cases per cluster by method were all significantly related to the sensitivity rates for successfully classifying members of Group 1. The sensitivity rates by estimation method and ICC appear in Figure 4. Across all levels of ICC, the NB classifier had the highest sensitivity rates among the methods. In addition, when the ICC was 0.1, all other methods were able to accurately classify members of the target group (Group 1) at rates in excess of 0.8. Outside NB, the multilevel Bayesian and PNNET approaches performed the best in this case, with the other methods all being within 0.03 of one



**Figure 1.** Cross-entropy (CE) by estimation method and intraclass or intracluster correlation (ICC).



**Figure 2.** Cross-entropy (CE) by estimation method and number of cases per cluster.

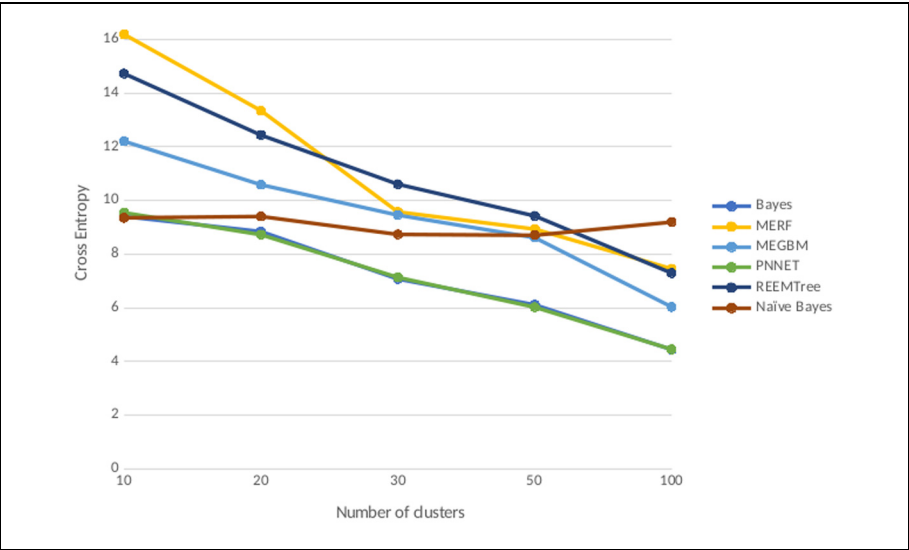


Figure 3. Cross-entropy (CE) rates by estimation method and number of clusters.

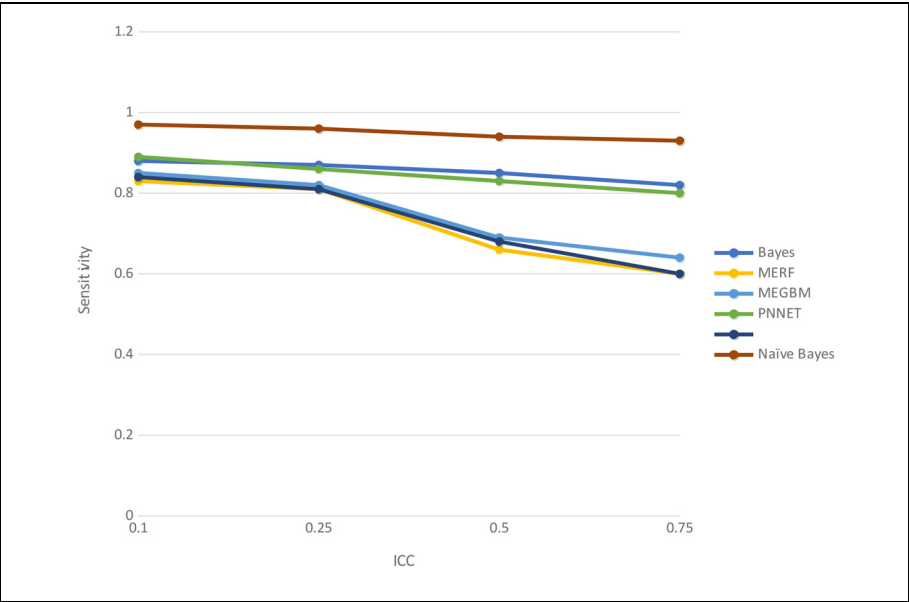
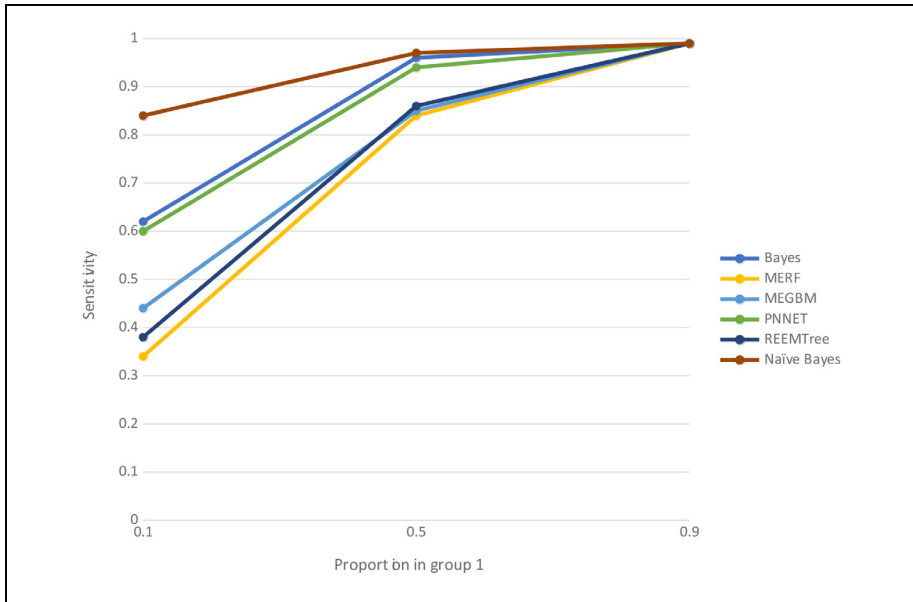


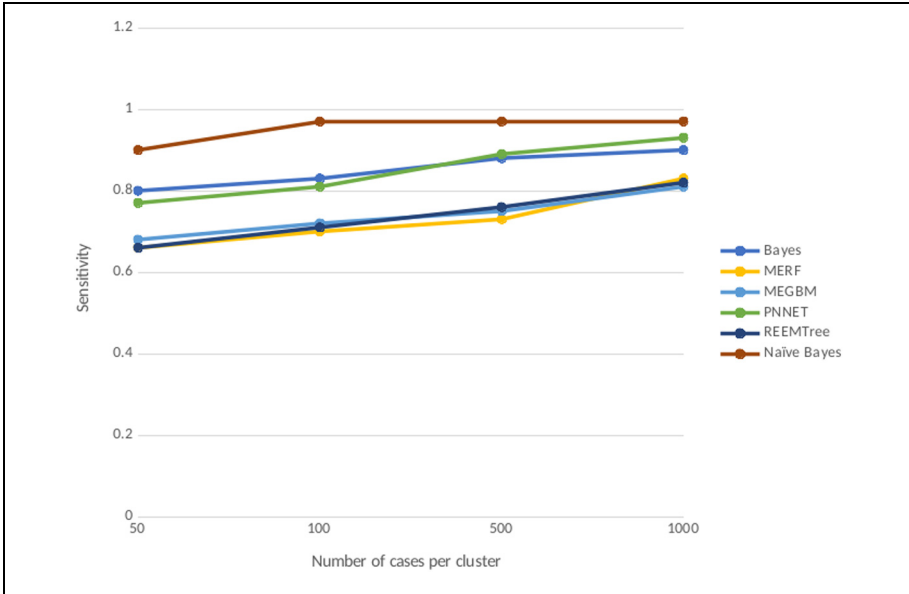
Figure 4. Sensitivity rate by estimation method and intraclass or intracluster correlation (ICC).



**Figure 5.** Sensitivity rates by estimation method and proportion of cases in Group 1.

another, and approximately 0.05 below that of multilevel Bayes and PNNET. Increased values of the ICC were associated with a decline in the sensitivity rates of all methods except for NB, with the sharpest such decline associated with MERF, Megbm, and REEMTree. Multilevel Bayes, NB, and PNNET demonstrated the least amount of decline in sensitivity rates across levels of the ICC. The worst performing methods studied here yielded sensitivity rates between 0.59 and 0.64 at this highest ICC value, with multilevel Bayes and PNNET having sensitivity rates of approximately 0.82, and NB of 0.93.

Sensitivity by the proportion of individuals in group 1 appear in Figure 5. Across methods, sensitivity increased in conjunction with an increased proportion in Group 1. When 90% of cases were simulated to be in Group 1, sensitivity rates for all of the methods studied here were 0.99. On the other hand, when only 10% of the simulated cases were in Group 1, sensitivity differed across the methods. The NB estimator had the highest sensitivity rate, followed by the multilevel Bayesian and PNNET techniques. MERF, GLMML, and REEMTree had comparable sensitivity for the 0.1 Group 1 proportion condition with values between 0.35 and 0.38, with Megbm exhibiting a rate of 0.45. When half of the observations were simulated to belong to Group 1, NB, multilevel Bayes, and PNNET had the highest sensitivity values (between 0.94 and 0.96), whereas the other approaches all performed comparably with sensitivity between those of the top three performers (0.84 to 0.86). Figure 6 contains sensitivity by the number of cases. Sensitivity rates were uniformly highest



**Figure 6.** Sensitivity rates by estimation method and number of cases per cluster.

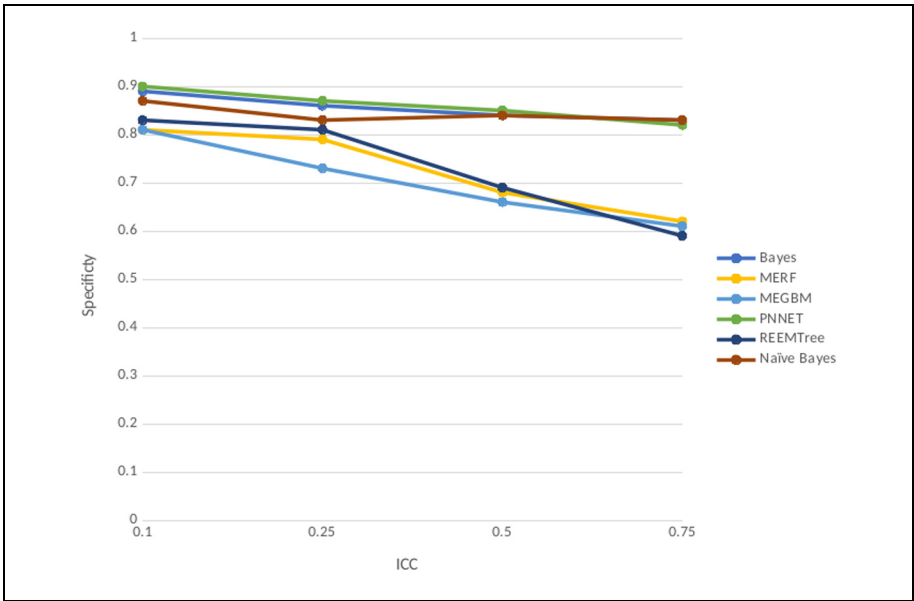
for the NB estimator, followed by the multilevel Bayes and PNNET approaches across conditions. MERF, Megbm, and REEMTree had the lowest sensitivity rates across all conditions of number of cases per cluster. Finally, sensitivity rates increased concomitantly with increases in the number of cases per cluster, with these increases being comparable across methods.

### Specificity Rates

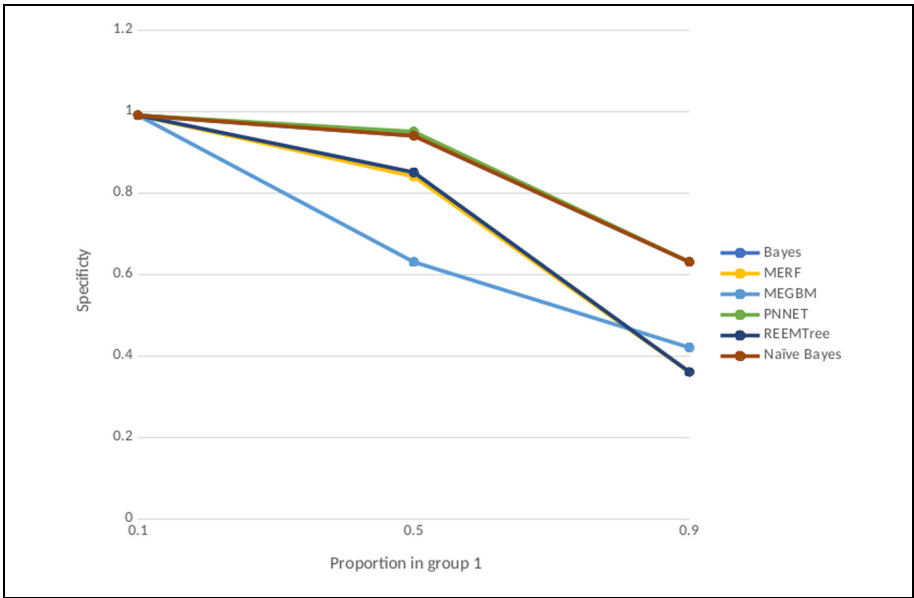
The specificity rates for Group 2 were found to be associated with the interaction of the proportion of cases simulated to be in Group 1 by estimator, ICC by estimator, and number of cases per cluster by estimator. The specificity rates by ICC and estimation method, appearing in Figure 7, reveal that as ICC increased accuracy for correctly classifying Group 2 decreased. Multilevel Bayes, NB, and PNNET consistently had the highest specificity rates across conditions. When the ICC was 0.1, the other methods all exhibited comparable specificity values, between 0.8 and 0.82. However, as ICC increased in value, specificity for REEMTree and MERF declined more quickly than was the case for the other methods.

The specificity rates for Group 2 by method and proportion of cases in Group 1 are displayed in Figure 8. When the proportion of cases in Group 1 was 0.1, specificity for correctly classifying members of Group 2 were 0.99. As the proportion in Group 1 increased, the specificity rates declined for all methods, with those of NB

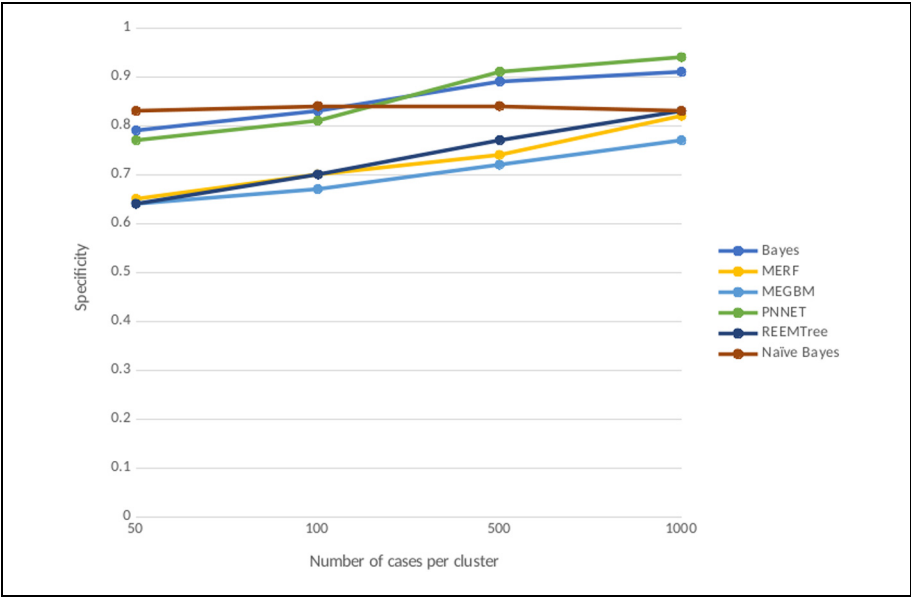




**Figure 7.** Specificity rates by estimation method and intraclass or intracluster correlation (ICC).



**Figure 8.** Specificity rates by estimation method and proportion of cases in Group I.



**Figure 9.** Specificity rates by estimation method and number of cases per cluster.

and PNNET declining the least. When 90% of observations were simulated to be in Group 1, the proportion of Group 2 cases correctly classified was approximately 0.36 for REEMTree and MERF, 0.62 for NB and PNNET, and 0.41 for Megbm. Specificity rates for all estimation methods were higher for a larger number of cases per cluster (Figure 9). Overall, Multilevel Bayes and PNNET performed better than the other approaches across the number of cases. The NB classifier had specificity rates comparable to multilevel Bayes and PNNET for 50 and 100 clusters, whereas for 500 and 1000 clusters its performance did not improve in contrast to the pattern for multilvel Bayes and PNNET. Megbm displayed the lowest specificity rates once the number of cases per cluster increased to 100 and greater.

**Discussion**

The present study was designed to investigate and to compare the relative predictive efficacy of several mixed effects classifiers across a number of conditions relevant to both single- and multilevel data. The results of the simulation illustrated a number of differences between the six methods employed with Megbm, MERF, and REEMTree frequently performing similarly to one another, and PNNET and multilevel Bayes performing similarly. The NB classifier, in contrast, appeared relatively robust to many of the conditions manipulated, most notably the differing ICCs. While this accuracy differential between multilevel models has been demonstrated universally in regression contexts, it is presently shown to remain relevant in classification

contexts. Furthermore, while results by Speiser et al. (2019, 2020) and Kilham et al. (2019) have demonstrated that conditions may exist in which single-level classification trees and random forests may outperform their multilevel analogues, this study adds to the debate of the necessity of multilevel classifiers through a demonstration of NB's efficacy under multilevel data conditions. These results bolster the findings of Demichelis et al. (2006) and Zhang et al. (2018) illustrating that NB does perform comparably to or greater than the multilevel Bayes classifier (among others) under the same conditions.

Across all classifiers, a consistent decrease in CE and an increase in sensitivity was noted as the number of cases per cluster increased with a concomitant decrease in CE and increase in sensitivity when the number of clusters increased. While no unitary sample size conditions were employed, these effects illustrate a pattern somewhat different from single-level classifiers. That is, previous studies comparing single-level classifiers (e.g., Bolin & Finch, 2014; Holden et al., 2011; Lei & Koehly, 2003) have found that as the sample size increases, estimates tend to become more stable but do not necessarily increase appreciably in accuracy. However, in the multilevel context, it is evident that a higher number of clusters and cases per cluster is associated with noticeable increases in accuracy. Additionally, increased sensitivity and decreased specificity rates across all methods were observed as group size ratios became increasingly more discrepant. That is, when the proportion of cases in group 1 was 0.9, all methods decreased substantially in their capability to predict cases belonging in group 0 despite an increase in accuracy of predictions for Group 1. This differential in group-specific accuracy rates is a well-documented phenomenon in classification literature and remains consistent in multilevel contexts.

One finding of particular note pertains to the differing ICCs, which is a fundamental characteristic of multilevel data for which guidelines have not been established, particularly in the classification context. The present results demonstrate a consistent decrease in sensitivity and specificity, with an increase in CE as the ICC becomes stronger with methods losing ~8% to 30% accuracy when the ICC went from 0.1 to 0.75. This finding illustrates that even in the context of multilevel models designed to capture the variance due to case nonindependence, highly correlated intracluster cases are likely to decrease predictive classification accuracy.

Across the multilevel classifiers presently employed, PNNET and Bayes demonstrated the lowest CE and highest accuracy rates of all methods across all conditions. Furthermore, in cases of increased ICC and group size discrepancy, and lower sample sizes and numbers of clusters, PNNET and Bayes demonstrated comparable accuracy rates to one another, both notably higher than all other methods. Therefore, it is likely that prediction tasks in the context of classification with multilevel data are best conducted and assessed using these methods across many conditions. Furthermore, given the practical limitations of the implementation of either method, it is likely Bayes would provide the most useful output for practitioner-researchers due to the ability of the software implementation yielding coefficient estimates and MCMC  $p$  values largely resembling traditional significance tests.

As discussed of single-level random forests compared with their multilevel analogues in Speiser et al. (2019), the present study illustrates that the NB classifier, while not accounting for the nesting structure of the data, is likely to perform well under many conditions relevant to multilevel data. Of particular note is NB's consistent performance across nearly all conditions: While all other methods demonstrated some amount of change in performance as ICC, group size ratio, and disaggregated sample size conditions changed, NB remained relatively robust to these changing conditions. For example, while yielding the highest CE of all methods when the ICC was 0.1 ( $CE \approx 9$ ), NB's performance remained at this level while all other methods' CE increased appreciably; at the highest ICC condition of 0.75, NB performed comparably to PNNET and multilevel Bayes. Such behavior and consistently high performance calls into question whether the more complex models truly add interpretive and predictive potential above and beyond a simpler classifier.

### Limitations

While this study featured a more comprehensive and controlled assessment of the available mixed effects classifiers and data conditions relevant in both classification and multilevel data contexts, several limitations should be noted. One limitation is that the simulation design did not allow for the unique manipulation of some clusters. That is, when simulated according to different cluster size, all clusters in the data set were created with such parameters. This is often not the case in most situations using nested data, as clusters may differ in their membership (with the exception being in the case of dyadic data; Knight & Humphrey, 2019). However, the effects of these equal and unequal cluster membership conditions were assessed by Milliren et al. (2018) in the context of three-level cross-classified data. However, the effects of unequal cluster sizes in two-level and uncrossed models remain untested.

An additional limitation pertains to the simplicity of the models constructed. That is, only a random intercept with a single L1 predictor were used to simulate the data. Additional considerations of multiple predictors, interacting and nonlinear predictors, and random coefficient models would allow for greater insight into the efficacy of the six multilevel models currently employed. Furthermore, as was noted above, GLMML, REEMTree, and MERF all assume a linear function to be estimated whereas models such as PNNET make no such assumption. This adherence to an assumption of linearity may be prohibitive, particularly when an oft-cited rationale for employing complex machine learning models (e.g., random forests, classification and regression trees, gradient boosting machines) being the absence of requirement for *a priori* specification of nonlinear and interaction terms.

### Conclusion

While a number of factors have not yet been assessed in the presently employed models—particularly MERF, Megbm, Bayes, and PNNET—the present study has demonstrated appreciable differences between these algorithms in their predictive

capability in classification contexts. The current results illustrate both Bayes and PNNET had not only in the highest rates of classification accuracy across multilevel methods but also were robust to changing data conditions when compared with the other methods studied here. However, in many conditions, NB performed as well as or better than both multilevel Bayes and PNNET. Therefore, while many multilevel analyses utilize GLMML due to its comparability to single-level logistic regression, the fundamental conditions of the data should be considered such that the most appropriate model could be effectively considered. In many cases, Bayes may prove to yield not only more accurate classification decisions but also software output largely resembling that from GLMML (insofar as the R statistical software package implementations of both). This recommendation is further bolstered by the multilevel Bayesian model's ability to provide conceptually interpretable parameter estimates. However, in the case of the construction of classification models for the purpose solely of prediction, conditions exist in which NB may prove to be more efficacious. While the nesting structure of the data may be ignored in NB (and its fundamental assumption of independence being violated), this model yielded highly accurate predictions for both Group 0 and Group 1, particularly in cases of smaller sample sizes (number of cases per cluster and number of clusters), larger ICCs, and unequal group sizes. Additional investigations should be considered using both simulated and real data in order to ascertain an augmented understanding of the efficacy of the present classifiers in varied data conditions and with additional model complexity.

Appendix

Table A1. Data Simulation Conditions.

Simulation variable	Conditions
Outcome variable group size ratio	0.5; 0.9
Number of Level-1 cases per Level-2 cluster	100; 500; 1000
Number of Level-2 clusters	10; 20; 30; 50; 100
Correlation within Level-2 clusters (intraclass correlation)	0.1; 0.3; 0.8
Correlation between Level-2 clusters	0.1; 0.3; 0.8
Method	GLMML; REEMTree; MERF; Megbm; PNNET; Bayes; Naive Bayes

Note. GLMML = generalized linear mixed model; REEMTree = recursive partitioning expectation minimization trees; MERF = mixed effects random forests; PNNET = panel neural network; Megbm = mixed effects gradient boosting machine; Bayes = Multilevel Bayesian classifier.


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 received no financial support for the research, authorship, and/or publication of this article.

## ORCID iDs

Anthony A. Mangino  <https://orcid.org/0000-0002-2699-9989>

W. Holmes Finch  <https://orcid.org/0000-0003-0393-2906>

## References

- Bagiella, E., Sloan, R. P., & Heitjan, D. F. (2000). Mixed-effects models in psychophysiology. *Psychophysiology*, 37(1), 13-20. <https://doi.org/10.1111/1469-8986.3710013>
- Bolin, J. E., & Finch, H. (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://doi.org/10.3389/fpsyg.2014.00118>
- Capitaine, L., Genuer, R., & Thiébaud, R. (2019). *Random forests for high-dimensional longitudinal data*. arXiv. <https://arxiv.org/abs/1901.11279>
- Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Academic Press.
- Crane-Droesch, A. (2017). *Semiparametric panel data models using neural networks*. arXiv. <https://arxiv.org/pdf/1702.06512.pdf>
- Demichelis, F., Magni, P., Piergiorgi, P., Rubin, M. A., & Bellazzi, R. (2006). A hierarchical naive bayes model for handling sample heterogeneity in classification problems: An application to tissue microarrays. *BMC Bioinformatics*, 7(1), 514-514. <https://doi.org/10.1186/1471-2105-7-514>
- 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>
- Finch, H. (2015). Recursive partitioning in the presence of multilevel data. *General Linear Model Journal*, 41(2), 30-44.
- Finch, W. H., & French, B. F. (2012). A comparison of methods for estimating confidence intervals for omega-squared effect size. *Educational and Psychological Measurement*, 72(1), 68-77. <https://doi.org/10.1177/0013164411406533>
- Fletcher, J. M., Stuebing, K. K., Barth, A. E., Miciak, J., Francis, D. J., & Denton, C. A. (2014). Agreement and coverage of indicators of response to intervention: A multi-method comparison and simulation. *Topics in Language Disorders*, 34(1), 74-89. <https://doi.org/10.1097/TLD.0000000000000004>
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189-1232. <https://doi.org/10.1214/aos/1013203451>
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman & Hall/CRC Press.
- Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, 33(2), 1-22. <https://doi.org/10.18637/jss.v033.i02>
- 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>

- Hajjem, A., Larocque, D., & Bellavance, F. (2017). Generalized mixed effects regression trees. *Statistics & Probability Letters*, 126, 114-118. <https://doi.org/10.1016/j.spl.2017.02.033>
- Hammersley, J. (2013). *Monte Carlo methods*. Springer Science & Business Media.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.
- Hedegaard, H., Curtin, S. C., & Warner, M. (2018). *Suicide mortality in the United States, 1999–2017* (NCHS Data Brief, 330). National Center for Health Statistics. <https://www.cdc.gov/nchs/products/databriefs/db330.htm>
- Ho, T. K., & Basu, M. (2002). Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 24(3), 289-300. <https://doi.org/10.1109/34.990132>
- Holden, J. E., Finch, W. H., & Kelley, K. (2011). A comparison of two-group classification methods. *Educational and Psychological Measurement*, 71(5), 870-901. <https://doi.org/10.1177/0013164411398357>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning (Vol. 112)*. Springer.
- Kessler, R., Barker, P., Colpe, L., Epstein, J., Gfroerer, J., Hiripi, E., Howes, M., Normand, S., Manderscheid, R., Walters, E., & Zaslavsky, A. (2003). Screening for serious mental illness in the general population. *Archives of General Psychiatry*, 60(2), 184-189. <https://doi.org/10.1001/archpsyc.60.2.184>
- Kilham, P., Hartebrodt, C., & Kändler, G. (2019). Generating tree-level harvest predictions from forest inventories with random forests. *Forests*, 10(1), 20-45. <https://doi.org/10.3390/f10010020>
- Knight, A. P., & Humphrey, S. E. (2019). *Dyadic data analysis*. In S. E. Humphrey & J. M. LeBreton (Eds.), *The handbook of multilevel theory, measurement, and analysis* (pp. 423-447). American Psychological Association. <https://doi.org/10.1037/0000115-019>
- Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.
- LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, 11(4), 815-852. <https://doi.org/10.1177/1094428106296642>
- Lei, P. W., & Koehly, L. M. (2003). Linear discriminant analysis versus logistic regression: A comparison of classification errors in the two-group case. *Journal of Experimental Education*, 72(1), 25-49. <https://doi.org/10.1080/00220970309600878>
- Luengo, J., & Herrera, F. (2012). Shared domains of competence of approximate learning models using measures of separability of classes. *Information Sciences*, 185(1), 43-65. <https://doi.org/10.1016/j.ins.2011.09.022>
- Maas, C. J., & 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>
- Mahalanobis, P. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Science of India*, 2(1), 49-55.
- Mann, J. J., Ellis, S. P., Waternaux, C. M., Liu, X., Oquendo, M. A., Malone, K. M., Brodsky, B. S., Haas, G. L., & Currier, D. (2008). Classification trees distinguish suicide attempters in major psychiatric disorders: A model of clinical decision making. *Journal of Clinical Psychiatry*, 69(1), 23-31. <https://doi.org/10.4088/JCP.v69n0104>
- 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>
- Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Survey Research Methods*, 3(1), pp. 45-58. <https://doi.org/10.18148/srm/2009.v3i1.666>
- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., & Lin, C.-C. (2020). *Package e1071* (R Package Version 1.7-4). R Project for Statistical Computing. <https://cran.r-project.org/web/packages/e1071/e1071.pdf>
- Milliren, C. E., Evans, C. R., Richmond, T. K., & Dunn, E. C. (2018). Does an uneven sample size distribution across settings matter in cross-classified multilevel modeling? Results of a simulation study. *Health & Place*, 52, 121-126. <https://doi.org/10.1016/j.healthplace.2018.05.009>
- Morris, L. V., Wu, S. S., & Finnegan, C. L. (2005). Predicting retention in online general education courses. *American Journal of Distance Education*, 19(1), 23-36. [https://doi.org/10.1207/s15389286ajde1901\\_3](https://doi.org/10.1207/s15389286ajde1901_3)
- National Institute of Mental Health. (2019). *Mental illness*. <https://www.nimh.nih.gov/health/statistics/mental-illness.shtml#:~:text=Mental%20illnesses%20are%20common%20in,mild%20to%20moderate%20to%20severe.>
- Ngufor, C. (2019). *Vira: Virtual intelligent robot assistant* (R Package Version 0.1). rddr.io. <https://rddr.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>
- Okada, K. (2013). Is omega squared less biased? A comparison of three major effect size indices in one-way ANOVA. *Behaviormetrika*, 40(2), 129-147. <https://doi.org/10.2333/bhmk.40.129>
- Pohar, M., Blas, M., & Turk, S. (2004). Comparison of logistic regression and linear discriminant analysis: A simulation study. *Metodoloski Zvezki*, 1(1), 143-161.
- R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.
- Revelle, W. R. (2017). *psych: Procedures for personality and psychological research* (R Package Version 1.8.4). R Foundation for Statistical Computing. <https://CRAN.R-project.org/package=psych>
- Ribeiro, J. D., Franklin, J. C., Fox, K. R., Bentley, K. H., Kleiman, E. M., Chang, B. P., & Nock, M. K. (2016). Suicide as a complex classification problem: Machine learning and related techniques can advance suicide prediction-a reply to Roaldset (2016). *Psychological medicine*, 46(9), 2009-2010. <https://doi.org/10.1017/S0033291716000611>
- Richard, M. D., & Lippmann, R. P. (1991). Neural network classifiers estimate Bayesian a posteriori probabilities. *Neural Computation*, 3(4), 461-483. <https://doi.org/10.1162/neco.1991.3.4.461>
- Rosen, B. L., & DeMaria, A. L. (2012). Statistical significance vs. practical significance: An exploration through health education. *American Journal of Health Education*, 43(4), 235-241. <https://doi.org/10.1080/19325037.2012.10599241>



- Sela, R. J., & Simonoff, J. S. (2012). RE-EM trees: A data mining approach for longitudinal and clustered data. *Machine Learning*, 86(2), 169-207. <https://doi.org/10.1007/s10994-011-5258-3>
- 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 International.
- Stuebing, K. K., Fletcher, J. M., Branum-Martin, L., & Francis, D. J. (2012). Evaluation of the technical adequacy of three methods for identifying specific learning disabilities based on cognitive discrepancies. *School Psychology Review*, 41(1), 3-22. <https://doi.org/10.1080/02796015.2012.12087373>
- U.S. Department of Education, National Center for Education Statistics. (2018). *The condition of education 2018* (NCES 2018-144). <https://nces.ed.gov/pubs2018/2018144.pdf>
- 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://doi.org/10.1080/02796015.2013.12087462>
- Xiong, Y., Kim, H. J., & Singh, V. (2019). Mixed Effects Neural Networks (MeNets) with applications to gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7743-7752). IEEE.
- Zhang, N., Wu, L., Yang, J., & Guan, Y. (2018). Naive Bayes bearing fault diagnosis based on enhanced independence of data. *Sensors (Basel, Switzerland)*, 18(2), 463. <https://doi.org/10.3390/s18020463>
- 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>