Educational and Psychological Measurement
71(5) 870–901

DOI: 10.1177/0013164411398357
http://epm.sagepub.com

A Comparison of Two-Group Classification Methods

Jocelyn E. Holden[1], W. Holmes Finch[1], and Ken Kelley[2]

Abstract

The statistical classification of *N* individuals into *G* mutually exclusive groups when the actual group membership is unknown is common in the social and behavioral sciences. The results of such classification methods often have important consequences. Among the most common methods of statistical classification are linear discriminant analysis, quadratic discriminant analysis, and logistic regression. However, recent developments in the statistics literature have brought new and potentially more flexible classification models to the forefront. Although these new models are increasingly being used in the physical sciences and marketing research, they are still relatively little used in the social and behavioral sciences. The purpose of this article is to provide a comparison of these modern methods with the classical methods widely used in situations that are relevant in the social and behavioral sciences. This study uses a large-scale Monte Carlo simulation study for the comparisons, as analytic comparisons are often not tractable. Results indicate that classification and regression trees generally produced the highest classification accuracy of all techniques tested, though study design characteristics such as sample size and model complexity can greatly influence optimal choice or effectiveness of statistical classification method.

Keywords

discriminant analysis, logistic regression, multivariate adaptive regression splines, classification and regression trees, boosting, generalized additive models, neural networks, mixture discriminant analysis, Monte Carlo simulation study, classification analysis

[1]Ball State University, Muncie, IN, USA
[2]University of Notre Dame, Notre Dame, IN, USA

Corresponding Author:
Jocelyn E. Holden, Department of Educational Psychology, Ball State University, Teacher's College, Room 524, Muncie, IN 47306, USA
Email: jeholden@bsu.edu

Statistical classification of individuals into observed groups is a very common practice throughout the social, behavioral, and physical sciences (Arabie, Hubert, & De Sote, 1996; Keogh, 2005; Zigler & Phillips, 1961). In education and psychology, examples abound in which researchers attempt to find statistical models that can be used to classify individuals into one of several known categories, such as those based on disability status (e.g., Lillvist, 2010; Mammarella, Lucangeli, & Cornoldi, 2010), career choice (Russell, 2008), and student preferences regarding mode of instruction (Clayton, Blumberg, & Auld, 2010), to name but a few. In all of these cases, the group membership is directly observable rather than latent in nature. It should be noted, however, that there is growing interest in a range of techniques designed specifically for use when group membership cannot be directly observed but rather is latent and thus must be inferred using a set of observed measures. The focus of the current study was on the case where group membership is observable and on a set of methods that can be used in that case. Generally speaking, these methods for the observed group context are not applicable to the situation where group membership is latent, and vice versa. Nonetheless, both scenarios are very applicable in the behavioral and social sciences and worthy of study.

Across these fields, perhaps the most common forms of statistical classification for group membership are linear discriminant function analysis (LDA), quadratic discriminant function analysis (QDA), and logistic regression (LR). Recent advances in the statistical literature, however, have introduced a set of alternative classification techniques with computer software that allow their relatively easy implementation in practice. These advances include procedures such as multivariate adaptive regression splines (MARS), generalized additive models (GAM), classification and regression trees (CART), neural networks (NNET), boosting (BOOST), and mixture discriminant analysis (MDA), each of which has been used to one degree or another in a variety of disciplines. These alternative methods of prediction provide a more flexible framework for modeling complex data structures and interactions than do the more traditional methods of LDA, QDA, and LR. Many of them have been used successfully in areas as diverse as business (Do & Grudnitski, 1992; Lee et al., 2006; Nguyen & Cripps, 2001; Smith & Mason, 1997; West, Brockett, & Golden, 1997), ecology (Moisen & Frescino, 2002; Preatoni et al., 2005), and the medical sciences (Grassi, Villani, & Marinoni, 2001; Ture, Kurt, Kurum, & Ozdamar, 2005). Despite their increasing use in the literature, there have been few if any published studies empirically comparing all of these methods with one another and with the more commonly used LR and LDA under a common set of conditions using Monte Carlo simulation techniques. Indeed, as will be reviewed in more detail below, individual methods have been examined or compared with LDA and/or LR typically using only real data, though occasionally with simulations. In addition, some of these approaches that we will refer to as alternative methods (to LDA and LR) have not been systematically studied under known conditions at all, and there has not been a published study (to our knowledge) where all of these techniques were compared with one another simultaneously under the same set of conditions. Therefore, the purpose of the present study is

to use a Monte Carlo simulation to compare the classification accuracy of these alternative methods with one another and with the popular LR and LDA techniques. We believe that this study would represent the first such comprehensive comparison of this entire set of methods and to that end should serve to assist researchers interested in group classification but unsure regarding which methods might be optimal under a given set of conditions. Following is a brief discussion of each of the alternative methods under consideration, so that the reader may have a context for comparison if he or she is not already familiar with them, followed by a review of prior research in the area and a description of the methods used in this study. These are not intended to be thorough reviews of the methods, and the interested reader is provided with relevant references that describe each in greater detail. We will begin our discussion of these methods with the traditionally most common methods of classification analysis: Discriminant Function Analysis and Logistic Regression.

Discrminant Function Analysis

Discriminant function analysis is a classification method that finds the combination of predictor variables so as to maximize the multivariate distance between groups. Based on this combination of predictors and a prior probability for group membership, the posterior probability of group membership is then computed for each individual in the sample, and they are in turn placed in the group for which their posterior probability is highest. When the group variances are equal, LDA is used, whereas when group variances are unequal, the resulting discriminant function will be quadratic (QDA). For further discussion of these methods, the interested reader is encouraged to see Hastie, Tibshirani, and Friedman (2001).

Logistic Regression

As with LDA, LR also bases group classification on a linear combination of the predictor variables. Specifically, LR finds the set of regression coefficients for the predictor variables so as to optimally predict the logit (log odds of being in one group vs. the other). LR differs from LDA in that it does not assume equal group variances, uses the logit as the response in the linear equation, and obtains parameter estimates through maximum likelihood estimation rather than using ordinary least squares (OLS), as is the case for LDA. For further discussion of LR, see Hastie et al. (2001).

Classification and Regression Trees

CART is based on an iterative decision process in which individuals are repeatedly partitioned using a set of predictor variables into ever more homogeneous groups based on the outcome variable, which in this study is categorical in nature (see Breiman, Friedman, Olshen, & Stone, 1984, for a review). The partitioning of subjects continues until a predefined level of homogeneity based on group membership has been attained, at which point the CART algorithm stops with individuals grouped into

what are known as terminal nodes. The overall goal of CART is to group subjects into maximally homogeneous terminal nodes based on the outcome variable (Williams, Lee, Fisher, & Dickerman, 1999).

The tree resulting from a CART analysis can be tested using a separate cross-validation sample of subjects drawn from the same population as the original (training) sample used to create the tree. The quality of the final tree is assessed by how accurately it can group members of the cross-validation sample for the outcome variable. Predicted group membership is based on which terminal node an individual from the cross-validation sample is placed into, based on the predictor splits identified by CART. The interested reader is encouraged to read more detailed descriptions of CART in Berk (2008) and Hastie et al. (2001).

As with any statistical procedure, care must be taken when using CART. Hothorn, Hornik, and Zeileis (2006) note that when selecting predictor variables on which to split, CART has a tendency to favor those with more distinct values over variables with fewer values. In addition, trees produced by CART can sometimes contain terminal nodes with few individuals or terminal nodes that are very heterogeneous, and can overfit the observed data, all of which characterize tree instability (Berk, 2008). In an attempt to address such weaknesses, researchers have developed alternatives to CART based on the principles outlined above but that construct many trees based on bootstrap samples from a set of data and then combine the results of the bootstrapped trees. Two examples of such approaches are bagging (LeBlanc & Tibshirani, 1996) and random forests (Breiman, 2001). Both approaches are based on selecting B samples with replacement from the initial sample, constructing B trees, and then averaging the B results to obtain a final result. The difference in the two approaches is that for random forests, a random sample from the full set of predictors is selected to be used in each split, whereas for bagging, all predictors are used at each split. A set of functions developed by Buhlmann and Hothorn (2007) for carrying out several tree building algorithms is available in the R software package under the PARTY toolbox of functions. To keep the size of the study manageable, and to ensure the comparison of a variety of methodological approaches, these methods are not included in this study. However, future research focusing particular attention on tree-based methodologies should compare the relative predictive accuracy of these variations on the tree building theme.

Neural Networks

Following is a brief description of the NNET approach to group classification. The interested reader is referred to any of several excellent references for a more thorough treatment of the method and its various alternatives, particularly as they pertain to the social and behavioral sciences (Berk, 2008; Garson, 1998). NNET identifies predictive relationships between an outcome variable, *Y*, and a set of predictors. More specifically, a search algorithm examines various subsets of the predictors as well as interactions among them (known as hidden layers), in conjunction with different

weights for these model terms, and selects the combination of main effects, interactions, and weights that minimizes the common least squared criterion (Marshall & English, 2000). The resulting model typically involves a complex combination of main effects and hidden layers coupled with a number of different weight values. These model terms are selected so as to minimize the least squares criterion common in regression models and may involve very complex combinations of interactions, main effects and weights. To reduce the likelihood of finding locally optimal results that will not generalize beyond the original (training) sample well, random changes to the subset of predictors, not based on model fit, are also made. This method of ascertaining fit and adjusting the model weights and included terms based on the difference between actual and predicted values of the outcome variable is known as *feed-forward back-propagation network* and is perhaps the most commonly used NNET algorithm (Garson, 1998).

Although the NNET approach to prediction has the advantage of being able to identify complex interactions of the predictors that might be associated with group membership, it can also produce models that overfit the training data, thus limiting the generalizability of the resulting model (Schumacher, Robner, & Vach, 1996). To combat this problem of overfitting, most NNET models apply what is called decay, which penalizes (i.e., reduces) the largest weights from the original NNET analysis. Such penalties essentially assume that very large weights in a model are at least partially driven by random variation unique to the training data, which must be ameliorated to some extent (Garson, 1998). Overfitting of the training data is typically identified through the use of cross-validation, in much the same way that CART models are tested, as described above.

Multivariate Adaptive Regression Splines

MARS, like NNET, is an extension of standard linear models where nonlinear relationships and interactions involving the predictor variables are modeled automatically through the use of smoothing splines (Simonoff, 1996). The resulting basis function (sometimes known as a hinge function) is piecewise linear and has change points, or knots, at the location where the relationship between a predictor and the outcome variable, Y, changes direction (Hastie et al., 2001). In the case of a dichotomous categorical outcome variable such as that used in this study, Y takes the form of the familiar logit function, $\ln(p_i/[1 - p_i])$, where p_i is the sample estimate of the probability of an individual being in group i.

MARS first estimates a predictive model using a forward stepwise methodology, beginning with the inclusion only of β_0 and then proceeding to add new basis functions to the model at each step, so as to maximize the reduction in the sum of squared residuals. At each step, the newly added term will include a term already in the model multiplied by the new hinge function, which may itself include both main effects and interactions of the predictor variables. When deciding which new basis function to add to the existing model, MARS searches across all existing terms currently in the model,

all the independent variables included in the analysis, and all possible values for each of the independent variables to select the knot for the new basis function. These terms will be selected automatically by the algorithm, based on the greatest reduction in the sum of squared residuals. Model building continues in this stepwise fashion until the change in sum of squared residuals becomes very small when a new term is entered or until a maximum model size (set by the user) is reached. Finally, to deal with the potential problem of the model overfitting the training data, a stepwise backward deletion procedure is used, in which the least important (from a statistical sense) term is removed at each step and the generalized cross-validation criterion is minimized to identify the optimal model.

The primary advantage of the MARS model–building strategy is its ability to work well locally in the function space (Hastie et al., 2001). Specifically, the use of the basis functions described above allows for the modeling of interactions only in the range of data for which two such functions have a nonzero value. Thus, unlike with the more general polynomial terms commonly used in regression, the entire data space is not required to take a common linear functional form.

Generalized Additive Models

The GAM approach to prediction involves the use of models based on smoothing functions, such as cubic splines and kernel smoothers, which are used to link a set of predictor variables to a response, *Y*. In the case of a dichotomous outcome variable, the actual response is the logit. The smoothing functions are selected for each predictor so as to minimize a penalized sum of squares function. The most common smoothing function used with GAM, and the one used in this study, is the cubic spline (Simonoff, 1996). To minimize the potential problem of overfitting the model to the training data, it is recommended that the number of smoothing parameters be kept relatively small and that cross-validation be used to ensure that the resulting model is generalizable to other samples from the target population (Wood, 2006).

Boosting

Boosting is a machine-learning technique based on the principle that combining a large number of weak predictor variables can result in a single very accurate prediction (Freund & Schapire, 1997). The AdaBoost algorithm was designed for classifying individual cases into known groups based on a set of predictor variables uses the following steps (Berk, 2008):

1. A regression equation using the set of predictors is fit to the original response variable.
2. The residuals for this model are calculated.
3. The original set of predictors is used to predict the residuals obtained in Step 2.

4. The fitted residual values obtained in Step 3 are used to update the fitted value of the response variable.
5. Steps 1 through 4 are repeated until the change in the fitted value of the outcome is below a predefined threshold value, in which case convergence to a solution has been reached.

A variety of approaches for determining at which iteration to stop the boosting algorithm have been investigated, with the current recommendation being to select the iteration which minimizes the value for Akaike's information criterion (Buhlmann & Hothorn, 2007). In practice, a researcher may examine a large number of *m* iterations and then review the resultant Akaike's information criterion values, selecting the model that corresponds with the smallest of these, which was the approach used in the current study.

Boosting will typically lead to complex models involving a large number of residual functions as the number of iterations increases. Therefore, it is generally recommended that the original regression model used to predict *Y* be fairly simple, consisting of a relatively small number of predictor terms and few if any third-order or higher interactions (Buhlmann & Yu, 2003). In addition, it should be noted that although linear regression is quite often the model to which the boosting algorithm is applied, it is entirely possible to use smoothing splines or other functions to relate the response variable to the predictors and then apply the boosting algorithm, thus accounting for potential nonlinearity in the data structure without introducing a large number of terms to the model (Buhlmann & Hothorn, 2007).

Mixture Discriminant Analysis

MDA is a variant of discriminant analysis, in which classes are not modeled as Gaussian distributions, as is common practice for LDA and QDA, but rather classes are modeled as a mixture of Gaussian distributions (Hastie & Tibshirani, 1996). This model represents each observed group by its centroid (like LDA and QDA) but also allows latent classes to exist within each group. In other words, existing groups (e.g., men and women) can themselves contain unobserved groups of individuals. Given its use of mixtures, MDA is applicable whether covariances are equal (as in LDA) or unequal (as in QDA; Hastie & Tibshirani, 1996). MDA typically uses the expectation maximization algorithm (Dempster, Laird, & Rubin, 1977) to estimate the model parameters including subgroup means, the common variance, the within-group mixing proportions, and the between-group mixing proportions—all of which are obtained from the training data. As with LDA, group classification in MDA is based on the discriminant function.

Previous Research

Research into optimal performance for categorical classification analyses is not new. In addition to a small number of published simulation studies (Buhlmann & Yu, 2003;

Kuhnert, Mengersen, & Tesar, 2003; Ripley, 1994; Tibshirani & LeBlanc, 1992), a number of comparative studies of categorical classification already exist in areas such as marketing (Curram & Mingers, 1994; Hart, 1992; West et al., 1997; Yoon, Swales, & Margavio, 1993) and medicine (Grassi et al., 2001; Reibnegger, Weiss, Werner-Felmayer, Judmaier, & Wachter, 1991) as well as various applications of the natural sciences (Bailly, Arnaud, & Puech, 2007; Liu & Chun, 2009; Preatoni et al., 2005; Ture et al., 2005).

The vast majority of existing literature examines the overall percentage of correctly classified cases by each statistical method for a variety of existing data sets. Results of these previous studies using real-world data sets demonstrate conflicting results. For example, Ripley (1994) and Dudoit, Fridlyand, and Speed (2002) found that traditional methods such as LDA, QDA, and LR performed comparably or better to some of the newer classification techniques such as CART but not as well as NNET or MARS when used on a variety of real-world and synthetic data sets. Preatoni et al. (2005) also reported that LDA outperformed CART in terms of classification accuracy; however, in contrast to the preceding studies, they also found that LDA provided more accurate classification results than did NNET. In contrast, Ture et al. (2005) and Yoon et al. (1993) reported that NNET produce higher classification accuracy than traditional methods such as LR and LDA. Ture et al. (2005) found that NNET outperformed CART and MARS as well. Still other results indicated that CART provided the highest classification accuracy when compared with LDA and NNET (Grassi et al., 2001).

In addition to these studies involving single data sets, a number of simulation studies investigating the performance of various classification methods have also been undertaken. For instance, LDA (Curram & Mingers, 1994) and LR (West et al., 1997) were found to perform as well as NNET when groups were linearly separable. However, in the presence of nonlinear relationships between predictors and group membership, the classification accuracy of LR and LDA suffered (Curram & Mingers, 1994; West et al., 1997). Other simulations have suggested that NNET (Curram & Mingers, 1994; Reibnegger et al., 1991; West et al., 1997; Yoon et al., 1993) and CART (Grassi et al., 2001) yield higher classification accuracy than LDA.

Prior simulation work has also found that certain data and distribution characteristics have an impact on classification accuracy. Although the majority of these simulation studies were performed on traditional methods of classification (LDA, QDA, LR, or cluster analysis), it is reasonable to hypothesize that these same characteristics may also affect some or all of the newer classification methods as well. In particular, data characteristics such as sample size (Holden & Kelley, 2010), group size ratio (Holden & Kelley, 2010; Lei & Koehly, 2003), and effect size (Finch & Schneider, 2006, 2007; Harrell & Lee, 1985; Holden & Kelley, 2010; Lei & Koehly 2003) can affect classification accuracy of some techniques. In addition, assumption violations (i.e., normality, homogeneity of variances) can lower the classification accuracy of LDA and LR (Blashfield, 1976; deCraen, Commandeur, Frank, & Heiser, 2006; Finch & Schneider, 2006, 2007; Lei & Koehly, 2003; Rausch & Kelley, 2009). Finally, error perturbation

(Baker, 1974; Breckenridge, 2000), number of true classes in the population (Finch & Schneider, 2007), and the number and type of predictors (Finch & Schneider, 2006, 2007; Krzanowski, 1976; Rausch & Kelley, 2009) can each affect classification accuracy of some techniques.

Current Study

As can be seen from the previous section, examination of the literature reveals a failure to reach consensus regarding optimal classification method choice. This is likely because of the lack of comprehensive comparison studies and relative paucity of simulation studies. There are also few studies comparing classification techniques under varying degrees of model complexity, particularly involving nonlinear relationships. Thus, the purpose for the present study was to provide a comprehensive comparison of traditional and alternative categorical classification methods using a Monte Carlo simulation study in order to carefully test these methods under a variety of known conditions that are based on those encountered in actual behavioral and social science data analyses. These conditions included sample characteristics (e.g., sample size and group size ratio) as well as varying complexity of the relationship between the predictors and the outcome variable. The study focused only on supervised classification techniques (classification methods that use training data sets) including LDA, QDA, LR, CART, NNET, GAM, MARS, BOOST, and MDA to provide a fair comparison between techniques.

In addition, it should be noted that although most prior studies have demonstrated that LR and LDA perform similarly (though not identically) in terms of classification accuracy in the two groups case, both were included in the current study. The reason for including both methods was that they have not been compared extensively when the underlying model is nonlinear, which is the focus of this research. Thus, one question of interest was whether there are any differences in the relative accuracy of these commonly used methods of classification when there are nonlinear terms in the model. Because both LDA and LR are based on linear models unless interactions are purposely introduced into them, we do not hypothesize great differences in their performance. Nonetheless, whether this hypothesis is warranted remains an open question, given the relative paucity of studies comparing them in the nonlinear case. Finally, after the simulation study results, an analysis of a real-world data using these techniques will be presented to demonstrate the utility of these methods (especially the nontraditional techniques) for real-world applications.

Method

Data Generation

Generation of simulated data and classification analyses were conducted with the R statistical software program (R Development Core Team, 2010). Data were generated

Table 1. Simulation Conditions

Variable	Levels
Sample size	100, 200, 500, 1,000
Effect size	0.2, 0.5, 0.8, 1.6
Group size ratio	50:50, 75:25, 90:10
Model complexity	Linear, simple, complex
Prediction method	LDA, QDA, LR, CART, MARS, GAM, NNET, BOOST, MDA

Correlation Matrix for Simulated Predictor Variables

	X_1	X_2	X_3	X_4
X_1	1	.76	.58	.43
X_2	.76	1	.57	.36
X_3	.58	.57	1	.45
X_4	.43	.36	.45	1

Note. LDA = linear discriminant analysis; QDA = quadratic discriminant analysis; LR = logistic regression; CART = classification and regression trees; MARS = multivariate adaptive regression splines; GAM = generalized additive models; NNET = neural networks; BOOST = boosting; MDA = mixture discriminant analysis.

to meet the conditions listed in Table 1. Simulated data consisted of a single two-group classification (outcome) variable and four continuous predictor variables with the correlation matrix among the variables appearing in Table 1. These values are based on those reported in Waller and Jones (2009) for the Wechsler Adult Intelligence Scale—Third Edition subscales, and all the manipulated variables (described below) were fully crossed for a total of 1,296 simulation conditions. A total of 1,000 iterations for each simulation condition were performed. For each method, tuning parameters used were those recommended in the literature discussed above.

Manipulated Study Variables

To address the research questions for the study, a number of variables were manipulated. Group membership was predicted using each of the following methods: LR, LDA, QDA, CART, NNET, MARS, GAM, BOOST, and MDA. In addition to type of classification model used, the impact of four other variables was examined: sample size, effect size, ratio of group sizes, and model complexity. Effect size was defined as the standardized mean difference (commonly referred to as Cohen's *d*) between the groups for each predictor variable. These values were set at 0.2, 0.5, 0.8, and 1.6, with the same values for each predictor. Cohen (1988) has termed effect size values of 0.2, 0.5, and 0.8 as *small*, *medium*, and *large*. Group size ratio had three conditions: (1) equal group sizes (50:50), (2) 75:25, and (3) 90:10. Model complexity refers to the nature of the relationship between the predictor variables and the outcome. Three levels of model complexity (appearing below) were tested, which we termed *linear*,

simple, and *complex*. In the linear condition (1), the four predictor variables were linearly related to the dichotomous outcome variable without any interactions. In the simple interaction condition (2), the predictor variables were related to the dichotomous outcome variable via a simple interaction. In the complex condition (3), the predictor variables were related to the dichotomous variable via several simple and complex interactions.

1. $\text{logit}(y) = X_1 + X_2 + X_3 + X_4$.
2. $\text{logit}(y) = X_1 + X_2 + X_3 + X_4 + X_1 * X_2$.
3. $\text{logit}(y) = X_1 + X_2 + X_3 + X_4 + X_1 * X_2 + X_3^2 + X_4^3$.

Analyses

The outcome variables of interest in this study were the overall percentage of misclassified cases as well as the percentage of misclassified cases from the smaller and larger groups, respectively, for a cross-validation sample drawn from the same population as the training data. To ascertain which of the manipulated variables had a significant impact on the misclassification rates, a factorial analysis of variance (ANOVA) was used. Average results across the 1,000 iterations are reported. Rates of nonconvergence for the prediction methods were very low (less than five replications). When nonconvergence did occur, another replication was run, so that for each combination of conditions, a total of 1,000 replications were obtained.

Results

Overall Misclassification

To identify the significant interactions and main effects for the simulation conditions, a factorial ANOVA was performed that tested all four-way, three-way, and two-way interactions and the main effects. Two four-way interactions were found to be statistically significant: Classification method × Model complexity × Effect size × Group ratio ($p < .01$, $\eta^2 = .516$) and Classification method × Effect size × Group size ratio × Sample size ($p < .01$, $\eta^2 = .474$). As these were the highest level interactions and all other interactions and main effects were subsumed in them, no further interactions or main effects will be discussed. Also, it should be mentioned that the $N = 200$ condition is omitted from all the tables. For all four levels of sample size, the same sample size patterns were observed, and we decided that to keep the tables more readily manageable, the $N = 200$ condition will not be included in the results. The full set of results are available from the first author on request. The technique demonstrating the most accurate classification for each condition is highlighted in boldface in each table.

Table 2 contains the misclassification percentages for the classification techniques by model complexity, group ratio, and effect size. Several immediate observations can be made. Consistent with previous findings, as effect size and group size ratio

Table 2. Misclassification Percentages Broken down by Method, Model Complexity Effect Size, and Group Size Ratio

Model	Group Ratio	Effect Size	Method								
			LDA	QDA	LR	CART	MARS	GAM	NNET	BOOST	MDA
Linear	50:50	0.2	49.5	53.0	45.0	**33.3**	47.3	39.5	40.0	45.0	44.8
		0.5	39.5	43.5	37.8	**28.0**	37.5	34.5	34.0	37.8	37.8
		0.8	31.5	33.5	30.5	**23.8**	30.0	29.3	27.5	30.8	31.0
		1.6	16.5	17.0	16.3	**12.5**	15.5	14.0	14.8	16.0	16.0
	75:25	0.2	25.0	25.5	25.0	**20.3**	25.0	24.5	23.5	25.0	39.8
		0.5	26.0	27.5	26.0	**22.5**	25.8	24.8	24.5	26.0	37.5
		0.8	23.8	26.0	22.8	**17.8**	22.8	21.5	20.5	22.8	27.8
		1.6	13.0	14.0	13.0	**9.5**	12.0	11.0	11.0	14.5	14.5
	90:10	0.2	10.0	10.3	10.0	**9.0**	10.0	10.0	9.5	10.0	45.0
		0.5	10.0	10.3	10.0	**9.0**	10.0	9.8	9.5	10.0	35.8
		0.8	10.3	10.3	10.3	**8.3**	10.0	9.3	9.3	10.0	31.3
		1.6	7.0	7.8	7.0	**5.5**	7.0	6.3	6.3	10.0	14.0
Simple	50:50	0.2	48.5	46.8	45.0	**31.3**	44.0	37.8	38.3	44.8	37.8
		0.5	35.8	31.0	34.0	**14.5**	31.5	29.0	25.0	33.8	18.8
		0.8	28.8	25.3	25.5	**8.3**	23.0	19.5	17.3	27.0	9.0
		1.6	12.5	13.5	11.8	1.8	2.8	1.3	6.8	12.0	**0.0**
	75:25	0.2	24.8	25.8	24.3	**19.0**	21.8	20.5	22.0	24.3	38.8
		0.5	22.8	21.0	22.0	**13.5**	22.5	17.8	18.3	22.0	19.3
		0.8	18.3	14.3	17.8	**7.5**	13.3	13.0	12.0	18.8	8.3
		1.6	9.0	7.5	8.0	1.8	2.0	1.3	4.0	13.0	**0.0**
	90:10	0.2	10.0	10.3	10.0	**9.0**	10.0	9.8	9.5	10.0	39.5
		0.5	10.0	10.3	10.0	**7.3**	10.0	9.0	8.5	10.0	21.3
		0.8	9.3	7.8	8.5	**4.3**	9.0	6.5	5.5	10.0	8.0

(continued)

Table 2. (continued)

Model	Group Ratio	Effect Size	Method								
			LDA	QDA	LR	CART	MARS	GAM	NNET	BOOST	MDA
		1.6	5.0	3.3	4.0	1.5	1.0	0.5	2.0	9.8	**0.3**
		0.2	42.3	40.5	39.8	**17.8**	26.8	23.3	24.8	40.3	23.8
		0.5	35.0	35.5	34.0	**15.5**	26.8	22.0	23.8	33.5	16.5
	50:50	0.8	26.8	27.3	26.0	**6.0**	18.0	15.0	16.3	25.5	7.5
		1.6	12.3	12.3	11.0	1.3	2.5	1.0	5.0	11.0	**0.0**
		0.2	23.0	23.5	17.5	**10.5**	17.0	13.5	12.8	20.0	19.0
		0.5	20.0	19.8	16.0	**9.3**	17.0	13.3	12.5	19.0	14.3
Complex	75:25	0.8	17.3	15.3	15.0	**5.0**	13.5	10.8	10.3	17.8	7.0
		1.6	8.0	7.5	7.0	1.5	1.8	0.3	3.3	12.8	**0.0**
		0.2	10.0	10.3	8.8	**4.8**	7.0	5.8	6.0	10.0	15.0
		0.5	10.0	10.0	8.0	**4.5**	7.0	5.8	5.5	10.0	13.0
	90:10	0.8	8.5	7.8	7.0	**3.3**	6.8	5.0	4.8	10.0	4.8
		1.6	5.0	3.3	4.0	1.0	1.0	**0.3**	2.0	10.0	**0.3**

Note. LDA = linear discriminant analysis; QDA = quadratic discriminant analysis; LR = logistic regression; CART = classification and regression trees; MARS = multivariate adaptive regression splines; GAM = generalized additive models; NNET = neural networks; BOOST = boosting; MDA = mixture discriminant analysis.

increased, the overall percentage of misclassified cases decreased (Breckenridge, 2000; deCraen et al., 2006; Finch & Schneider, 2006; Holden & Kelley, 2010; Lei & Koehly, 2003). The effect of model complexity differed depending on the classification method used. In general, for more complex models the classification methods made fewer misclassification errors. This result was particularly notable for CART, MARS, GAM, NNET, and MDA. Across the board, the fewest misclassification errors were made by CART, followed by GAM and NNET, whereas the most errors were made by LDA, QDA, LR, and BOOST. MDA and MARS yielded the largest gains in accuracy as model complexity increases. Also of note is that MDA consistently demonstrated very high classification accuracy at very high effect sizes ($d = 1.6$).

Table 3 contains the misclassification percentages by classification technique, group size ratio, sample size, and effect size. It is again evident that the percentage of misclassified cases decreases as effect size and group size ratio increased. In addition, across the simulated conditions, CART had the lowest levels of misclassification, followed by NNET and GAM, whereas LDA, QDA, and LR continued to show the highest misclassification rates. It should be noted that CART had markedly lower misclassification rates at $N = 100$ than any of the other approaches, particularly in the equal group size ratio condition. In contrast, for the largest sample size condition, the methods all produced much more similar rates of misclassification across group size ratios, with the exception of MDA, which tended to yield higher rates, particularly for the unequal–N conditions.

Looking at the effect of sample size, however, reveals interesting differences between the classification techniques. For LDA and QDA, as sample size increased, the percentage of misclassified cases decreased, a result already documented by many studies (Holden & Kelley, 2010; Lei & Koehly, 2003). However, for CART, GAM, NNET, and MDA, larger sample sizes were associated with higher percentages of misclassified cases. The only exception to this result is that for MDA, when sample sizes are very large ($N = 1{,}000$) the misclassification percentage decreases instead of increases.

In summary, when looking at overall misclassification rates, larger effect sizes, unequal group size ratios, and greater model complexity tended to decrease the number of misclassification errors. In addition, larger sample sizes increased the accuracy of some classification techniques (LDA, QDA) but decreased the accuracy of others (CART, GAM, NNET, MDA). Finally, across the vast majority of study conditions, CART demonstrated the highest classification accuracy, whereas LDA, QDA, LR, and BOOST demonstrated the lowest.

Smaller and Larger Group Misclassification

In addition to looking at the overall misclassification rate, the results were also examined further by the percentage misclassified for each group. As for the overall misclassification rates, a factorial ANOVA was run testing all four-way, three-way, two-way interactions, and main effects for both small-group misclassification and

Table 3. Misclassification Percentages Broken Down by Method, Group Size Ratio, Sample Size, and Effect Size

Group Size Ratio	Sample Size	Effect Size	Method								
			LDA	QDA	LR	CART	MARS	GAM	NNET	BOOST	MDA
50:50	100	0.2	51.0	51.3	43.0	**16.7**	41.3	28.0	31.7	43.3	35.3
		0.5	39.7	39.7	35.7	**14.0**	34.0	25.7	27.3	35.7	26.0
		0.8	30.0	29.0	27.3	**11.3**	23.0	19.3	19.0	27.3	16.0
		1.6	14.3	15.3	12.7	5.0	6.3	**4.0**	8.3	13.0	5.3
	500	0.2	44.7	44.0	43.7	**35.0**	38.3	35.7	35.7	43.3	35.3
		0.5	36.3	36.0	36.0	**25.0**	33.0	32.0	28.7	35.7	25.7
		0.8	28.3	29.0	25.7	**13.3**	26.0	22.3	20.7	28.0	15.7
		1.6	13.3	13.3	13.0	**5.3**	7.3	6.0	8.7	13.0	**5.3**
	1,000	0.2	44.0	42.0	43.3	38.0	38.0	37.3	36.3	43.3	**35.3**
		0.5	33.0	32.0	33.3	23.3	27.7	27.0	25.7	32.7	**20.7**
		0.8	28.3	26.7	28.3	**15.3**	22.7	22.3	20.7	28.0	15.7
		1.6	13.3	14.0	13.0	6.0	7.3	6.3	9.0	13.0	**5.3**
75:25	100	0.2	24.0	26.0	21.3	**12.7**	18.0	15.0	16.7	22.3	26.0
		0.5	24.7	25.3	22.3	**15.7**	23.0	19.7	19.0	23.3	21.3
		0.8	20.3	19.3	18.0	**9.0**	16.7	13.0	13.0	19.0	10.0
		1.6	10.7	11.0	9.3	4.3	5.3	**3.0**	5.7	13.3	3.3
	500	0.2	24.3	24.3	22.7	**19.7**	22.3	21.0	20.3	23.3	34.7
		0.5	22.3	21.7	21.0	**15.7**	21.0	19.7	18.7	22.0	24.7
		0.8	20.3	17.3	19.3	**11.3**	17.0	16.7	15.3	20.7	15.7
		1.6	9.0	8.7	8.3	**4.3**	5.0	4.7	6.0	12.3	5.3
	1,000	0.2	24.3	24.0	22.7	**20.7**	22.3	21.0	21.0	23.3	34.3
		0.5	22.0	21.0	21.0	**18.0**	21.3	19.7	18.7	22.0	24.3
		0.8	19.0	18.0	18.3	**11.7**	16.0	16.0	15.0	19.7	15.7
		1.6	10.0	8.7	10.0	**4.7**	5.3	5.0	6.3	14.0	5.3

(continued)

Table 3. (continued)

Group Size Ratio	Sample Size	Effect Size	Method								
			LDA	QDA	LR	CART	MARS	GAM	NNET	BOOST	MDA
90:10	100	0.2	10.0	11.0	9.3	7.7	9.0	8.0	**7.0**	10.0	32.0
		0.5	10.0	11.0	9.3	7.3	9.0	7.3	**6.7**	10.0	22.0
		0.8	10.0	9.3	8.7	5.7	8.3	**5.3**	5.7	10.0	12.7
		1.6	5.7	5.0	5.0	3.7	3.0	**2.3**	3.7	10.0	5.3
	500	0.2	10.0	10.0	9.7	**7.7**	9.0	8.7	8.3	10.0	34.0
		0.5	10.0	10.0	9.3	**6.3**	9.0	8.3	8.0	10.0	24.7
		0.8	9.0	8.0	8.7	**5.0**	8.7	7.7	7.0	10.0	15.7
		1.6	5.7	4.3	5.0	**2.3**	3.0	**2.3**	3.3	10.0	5.0
	1,000	0.2	10.0	10.0	10.0	**8.3**	9.0	8.7	9.0	10.0	34.0
		0.5	10.0	9.7	9.3	**7.7**	9.0	8.7	8.3	10.0	24.0
		0.8	9.0	8.0	8.7	**5.3**	8.7	8.0	7.0	10.0	15.7
		1.6	5.7	4.0	5.0	**2.3**	3.0	3.0	3.7	9.7	5.3

Note. LDA = linear discriminant analysis; QDA = quadratic discriminant analysis; LR = logistic regression; CART = classification and regression trees; MARS = multivariate adaptive regression splines; GAM = generalized additive models; NNET = neural networks; BOOST = boosting; MDA = mixture discriminant analysis.

large-group misclassification, separately, to identify significant effects. The same significant four-way interaction was identified for both groups: Classification method × Model complexity × Effect size × Group size ratio ($p < .01$, $\eta^2 = .570$ for small-group, $\eta^2 = .604$ for large-group misclassification). A significant two-way interaction was also found between sample size and classification method ($p < .01$, $\eta^2 = .503$ for small-group; $p < .05$, $\eta^2 = .145$ for large-group misclassification) was also found. The remainder of this article will focus on these two results.

Tables 4 and 5 contain the smaller and larger group misclassification rates by classification method, model complexity, effect size, and group size ratio. It is immediately evident by looking at Tables 4 and 5 that for the majority of methods, when group sizes were unequal, the classification techniques do not misclassify equal numbers of cases from the smaller and larger groups. In particular, all the studied methods, with the exception of MDA, demonstrated a classification bias in favor of the larger group. In other words, they misclassified a larger percentage of cases from the smaller group as belonging to the larger group than the larger group as belonging to the smaller group. This pattern became more pronounced as the effect size and discrepancy between group sizes increased. Worthy of note is the result that for unequal groups with low to moderate effect sizes (0.2 and 0.5), the majority of models in the linear condition misclassified more than 70% of the cases in the smaller group. A more striking pattern is seen for the simple interaction condition with very discrepant group sizes (90:10), in which most methods had misclassification rates more than 90%, except for CART, NNET, and MDA. As the models increased in complexity, both the smaller and larger group misclassification rates dropped, which is consistent with the previous finding that the overall misclassification rate decreases as model complexity increases.

It is evident that small- and large-group misclassification rates were very much dependent on the group size ratio and model complexity of the study. Regarding large-group misclassification, when group sizes were equal and the model was linear or simple, CART demonstrated the most accuracy in classifying individuals into the larger group. When group sizes were equal and the model was complex, CART, MARS, GAM, and NNET were the most accurate. When group sizes were unequal, however, large-group misclassification reduced sharply for all methods, resulting in perfect or near perfect classification of the larger group for all methods except MDA.

With regard to small-group misclassification, across the board MDA, CART, and GAM most accurately classified the cases in the smaller group. It should be noted, however, that when group sizes were unequal and effect sizes were small, the majority of methods (with the exception of CART and MDA) misclassified nearly 100% of the cases from the smaller group. As effect size and model complexity increased, however, misclassification of the smaller group declined. Indeed, for an effect size value of 1.6, misclassification rates for all of the methods except BOOST were less than 40%, with several less than 10%, even in the most unequal-groups case.

MDA displayed a different pattern from the other models tested in this study. MDA was the only method for which the smaller and larger group misclassification rates

Table 4. Small-Group Misclassification Percentages Broken Down by Method, Model, Complexity, Effect Size, and Group Size Ratio

Model Complexity	Group Size Ratio	Effect Size	LDA	QDA	LR	CART	MARS	GAM	NNET	BOOST	MDA
Linear	50:50	0.2	49.5	53.0	45.0	32.3	27.3	**31.5**	40.5	45.0	44.8
		0.5	39.8	43.3	38.0	**27.8**	34.5	34.0	35.8	38.0	38.0
		0.8	32.0	33.5	32.3	**23.8**	28.0	28.3	26.0	31.8	31.3
		1.6	16.5	17.0	16.0	**12.0**	14.5	14.3	14.5	19.0	16.0
	75:25	0.2	100.0	100.0	100.0	66.8	97.5	98.8	90.0	100.0	**40.0**
		0.5	91.5	94.0	90.3	71.8	89.5	71.3	79.3	91.8	**38.3**
		0.8	71.5	76.5	69.5	46.3	69.8	67.0	59.3	74.8	**28.3**
		1.6	33.3	36.5	31.8	24.0	34.0	29.0	30.5	52.8	**14.3**
	90:10	0.2	100.0	100.0	100.0	75.5	100.0	99.0	93.0	100.0	**43.8**
		0.5	100.0	100.0	100.0	71.3	99.0	96.5	92.0	100.0	**36.8**
		0.8	97.3	99.3	95.0	63.5	93.5	87.3	85.0	100.0	**29.8**
		1.6	54.5	58.3	52.3	39.5	55.8	49.5	48.3	99.3	**14.0**
Simple	50:50	0.2	48.8	48.0	45.0	46.3	65.0	43.0	45.5	45.0	**37.8**
		0.5	37.3	33.3	35.8	**15.8**	30.3	31.3	31.8	36.5	18.0
		0.8	31.8	31.0	30.0	**8.5**	18.8	18.0	22.3	29.3	9.0
		1.6	16.5	18.0	13.0	1.8	3.0	1.3	8.5	18.0	**0.0**
	75:25	0.2	81.8	77.0	80.8	56.8	74.5	73.0	71.3	82.3	**38.5**
		0.5	82.5	72.0	81.5	37.0	82.5	76.3	57.0	84.5	**18.5**
		0.8	59.0	49.8	58.0	16.5	38.5	37.0	36.0	69.3	**8.0**
		1.6	26.8	24.3	22.8	3.5	4.5	2.8	13.0	51.3	**0.0**
	90:10	0.2	100.0	100.0	100.0	71.5	98.8	97.8	92.8	100.0	**38.0**
		0.5	98.5	91.3	97.8	49.5	95.5	90.5	77.0	100.0	**22.3**
		0.8	78.3	73.0	79.3	27.3	90.3	61.5	44.3	99.3	**8.5**
		1.6	36.3	31.3	33.8	7.8	6.8	4.5	17.8	99.5	**0.8**

(continued)

Table 4. (continued)

Model Complexity	Group Size Ratio	Effect Size	LDA	QDA	LR	CART	MARS	GAM	NNET	BOOST	MDA
Complex	50:50	0.2	33.5	**24.8**	33.3	26.5	48.3	39.5	48.5	34.5	27.3
		0.5	30.8	28.5	31.0	24.0	54.5	33.5	44.3	32.5	**17.0**
		0.8	26.3	27.3	26.0	**6.5**	37.5	15.3	25.5	28.0	7.3
		1.6	13.8	19.0	11.0	1.5	3.0	1.0	6.0	15.8	**0.0**
	75:25	0.2	90.8	82.3	69.8	33.8	49.0	43.5	49.8	80.0	**25.3**
		0.5	75.8	58.5	62.5	29.8	50.8	41.5	48.0	77.3	**15.3**
		0.8	57.3	49.5	49.8	8.8	44.8	34.8	38.3	70.3	**6.3**
		1.6	29.8	29.5	23.3	5.3	5.8	1.0	13.8	64.8	**0.0**
	90:10	0.2	100.0	98.8	84.3	39.3	55.8	42.3	58.3	100.0	**23.3**
		0.5	97.0	85.5	78.3	35.5	58.8	39.0	51.5	100.0	**12.3**
		0.8	78.3	66.5	66.8	19.8	56.5	37.5	41.0	100.0	**5.8**
		1.6	37.5	28.8	31.8	5.5	8.0	1.8	18.0	99.3	**0.8**

Note. LDA = linear discriminant analysis; QDA = quadratic discriminant analysis; LR = logistic regression; CART = classification and regression trees; MARS = multivariate adaptive regression splines; GAM = generalized additive models; NNET = neural networks; BOOST = boosting; MDA = mixture discriminant analysis.

Table 5. Large-Group Misclassification Percentages Broken Down by Method, Model, Complexity, Effect Size, and Group Size Ratio

Model Complexity	Group Size Ratio	Effect Size	LDA	QDA	LR	CART	MARS	GAM	NNET	BOOST	MDA
Linear	50:50	0.2	49.5	53.0	45.0	**34.0**	59.0	55.3	39.5	45.0	44.8
		0.5	39.5	43.8	37.5	**28.0**	40.3	34.0	32.5	37.0	37.3
		0.8	31.3	33.5	30.8	**23.5**	30.8	29.3	29.3	29.5	30.8
		1.6	16.8	17.3	16.0	**12.5**	16.0	14.3	14.5	13.0	16.0
	75:25	0.2	**0.0**	0.5	**0.0**	4.5	**0.0**	**0.0**	1.3	**0.0**	39.5
		0.5	3.0	3.8	2.8	4.5	**2.5**	16.0	5.0	2.8	37.3
		0.8	7.3	8.0	6.8	7.5	5.5	5.5	6.5	**4.8**	27.3
		1.6	6.3	6.8	6.0	5.3	5.0	5.3	4.5	**2.0**	14.5
	90:10	0.2	**0.0**	**0.0**	**0.0**	2.0	**0.0**	**0.0**	**0.0**	**0.0**	45.8
		0.5	**0.0**	**0.0**	**0.0**	2.0	**0.0**	**0.0**	0.3	**0.0**	35.5
		0.8	0.3	0.5	0.0	2.3	**0.0**	**0.0**	0.5	**0.0**	32.0
		1.6	2.0	2.0	2.0	2.0	1.0	1.8	1.5	**0.0**	14.0
Simple	50:50	0.2	48.8	45.5	44.0	**16.3**	23.0	32.5	31.0	45.3	38.0
		0.5	33.8	24.8	31.8	**14.0**	29.8	30.0	18.0	30.5	19.0
		0.8	25.8	15.3	25.0	**8.5**	22.8	19.0	14.5	23.0	9.0
		1.6	8.5	5.5	10.0	1.8	1.8	1.3	4.8	6.0	**0.0**
	75:25	0.2	4.8	4.3	5.0	6.3	4.3	**3.3**	4.8	4.0	38.8
		0.5	3.3	4.5	**2.0**	5.8	**2.0**	2.8	4.5	**2.0**	19.8
		0.8	5.3	2.3	4.0	4.0	5.0	4.8	3.8	**2.0**	8.3
		1.6	3.0	0.3	3.0	1.3	0.8	0.8	1.0	0.5	**0.0**
	90:10	0.2	**0.0**	0.3	**0.0**	2.0	**0.0**	**0.0**	**0.0**	**0.0**	39.8
		0.5	**0.0**	1.3	**0.0**	2.3	**0.0**	**0.0**	0.8	**0.0**	21.0
		0.8	1.3	1.0	1.0	1.8	0.3	0.5	1.5	**0.0**	8.0
		1.6	1.3	**0.0**	1.0	0.5	0.3	**0.0**	**0.0**	**0.0**	**0.0**

(continued)

Table 5. (continued)

Model Complexity	Group Size Ratio	Effect Size	LDA	QDA	LR	CART	MARS	GAM	NNET	BOOST	MDA
Complex	50:50	0.2	52.5	56.5	46.5	8.8	4.8	8.0	**1.3**	46.0	20.0
		0.5	39.3	40.0	37.3	7.5	3.0	11.0	**3.0**	34.8	15.8
		0.8	28.0	27.3	26.5	**6.0**	13.0	15.0	6.8	23.0	6.3
		1.6	10.3	7.5	11.0	1.5	1.8	1.3	3.8	6.8	**0.0**
	75:25	0.2	**0.0**	3.8	**0.0**	2.5	**0.0**	**0.0**	**0.0**	**0.0**	16.8
		0.5	1.3	6.8	1.0	2.8	**0.0**	0.3	0.5	**0.0**	14.0
		0.8	3.8	3.5	3.3	5.8	0.5	1.5	1.5	**0.0**	7.0
		1.6	2.8	**0.0**	3.3	0.8	0.3	0.3	0.8	**0.0**	0.3
	90:10	0.2	**0.0**	0.3	**0.0**	0.8	**0.0**	**0.0**	**0.0**	**0.0**	14.0
		0.5	**0.0**	1.0	**0.0**	0.8	**0.0**	**0.0**	**0.0**	**0.0**	8.8
		0.8	0.8	1.0	**0.0**	1.3	**0.0**	**0.0**	0.5	**0.0**	4.8
		1.6	1.3	**0.0**	1.0	0.3	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**

Note. LDA = linear discriminant analysis; QDA = quadratic discriminant analysis; LR = logistic regression; CART = classification and regression trees; MARS = multivariate adaptive regression splines; GAM = generalized additive models; NNET = neural networks; BOOST = boosting; MDA = mixture discriminant analysis.

remained relatively equal regardless of the ratio of group sizes. MDA misclassified a slightly higher percentage of cases from the smaller group; however, the values for smaller and larger group misclassification were generally within a percentage of one another. In comparison with the other methods, MDA displayed higher rates of misclassification for the larger group, except for an effect size of 1.6. On the other hand, it consistently outperformed the other techniques with respect to the smaller group.

Table 6 contains the small- and large-group misclassification percentages by sample size and method. The bias toward small-group misclassification for all methods except MDA can be seen very clearly. The tendency for MDA to equally misclassify cases to the smaller and larger group is also clearly apparent. The effect of sample size differed depending on the method of classification. For LDA and QDA, larger sample size slightly decreased the percentage of misclassified cases. On the other hand, for LR, CART, MARS, GAM, NNET, BOOST, and MDA, increasing sample size increased the percentage of misclassified cases. It should be noted that this pattern was most evident for the smaller group, whereas misclassification percentages for the larger group varied relatively little across sample sizes and were generally low.

In an effort to summarize the results of this study, refer to Figures 1 through 3, which display the misclassification rates for the traditional LDA approach to prediction with that of CART, which had the lowest such rates for many conditions studied here, and of MDA, which generally had the lowest misclassification rates for the smaller group. Results in Figure 1 indicate that the overall misclassification percentage is lower for CART than LDA across nearly all the simulated conditions. However, the difference in performance is much more noticeable when the underlying model has interactions present (simple or complex structure) and when the groups were of equal size. In addition, MDA had the highest misclassification percentages of the three methods in the linear case but comparable percentages to those of CART in the complex condition, particularly when the groups were of equal size. In addition, MDA was the most affected by the effect size. Figures 2 and 3 include misclassification results for the large and small groups respectively. For the large group, misclassification percentages were nearly identical for CARD and LDA, and were generally less than 10%, whereas for MDA, they were much larger, particularly for the smaller sample size condition. On the other hand, in the smaller group case the percent misclassified for CART was lower than that of LDA, which was most notable for the models containing interactions and when the group size ratio was 90:10. In addition, neither CART nor LDA was as accurate for the smaller group as was MDA, particularly for the 90:10 group size ratio.

Real-Data Analysis

To demonstrate their prediction accuracy in an applied social science context, the methods examined in this study were applied to a data set described in Tabachnick and Fidell (2007, p. 277). These data include a categorical variable indicating the level of masculinity (high and low) based on scores on the Bem Sex Role Inventory, along with several continuous scale scores including attitudes toward the role of women,

Table 6. Small- and Large-Group Misclassification Percentages by Sample Size

	Sample Size	Method								
		LDA	QDA	LR	CART	MARS	GAM	NNET	BOOST	MDA
Smaller Group	100	59.3	58.8	53.9	25.6	48.4	38.9	40.9	66.7	**18.3**
	200	60.6	58.9	56.6	22.8	49.9	43.5	47.0	68.8	**20.4**
	500	59.9	55.7	56.9	35.2	51.5	46.3	47.2	70.0	**20.5**
	1000	59.4	55.4	56.6	42.3	52.8	46.1	48.6	68.9	**20.2**
Larger Group	100	13.4	13.4	11.6	6.8	7.1	8.9	**6.5**	10.3	17.4
	200	12.0	11.6	11.1	6.4	8.6	8.8	**6.3**	9.9	19.4
	500	11.4	11.5	11.1	**6.8**	7.6	7.8	**6.8**	10.0	19.9
	1000	11.0	10.3	10.9	**5.8**	7.1	7.1	6.8	9.7	19.2

Note. LDA = linear discriminant analysis; QDA = quadratic discriminant analysis; LR = logistic regression; CART = classification and regression trees; MARS = multivariate adaptive regression splines; GAM = generalized additive models; NNET = neural networks; BOOST = boosting; MDA = mixture discriminant analysis.

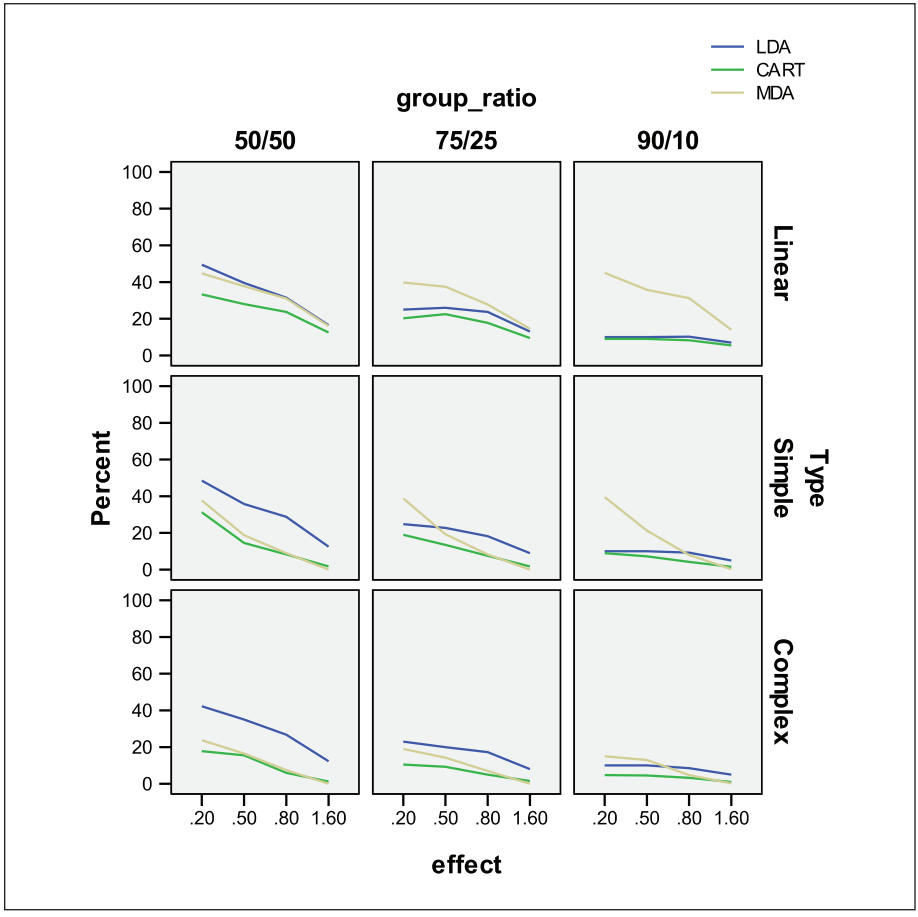

Figure 1. Overall misclassification percentages for LDA, CART, and MDA by effect size, sample size ratio, and type of underlying model
Note. LDA = linear discriminant analysis; CART = classification and regression trees; MDA = mixture discriminant analysis.

self-esteem, locus of control, neuroticism–stability index, introversion–extroversion, and socioeconomic level. The sample consisted of 369 females who participated in the study in 1975. This data set was selected for this example both because it is available to anyone interested in using it to gain experience with these methods by replicating these analyses and because it has a social science context.

For this application, continuous scale scores were used to predict category on the masculinity variable (high or low) for each of the prediction models studied here. A

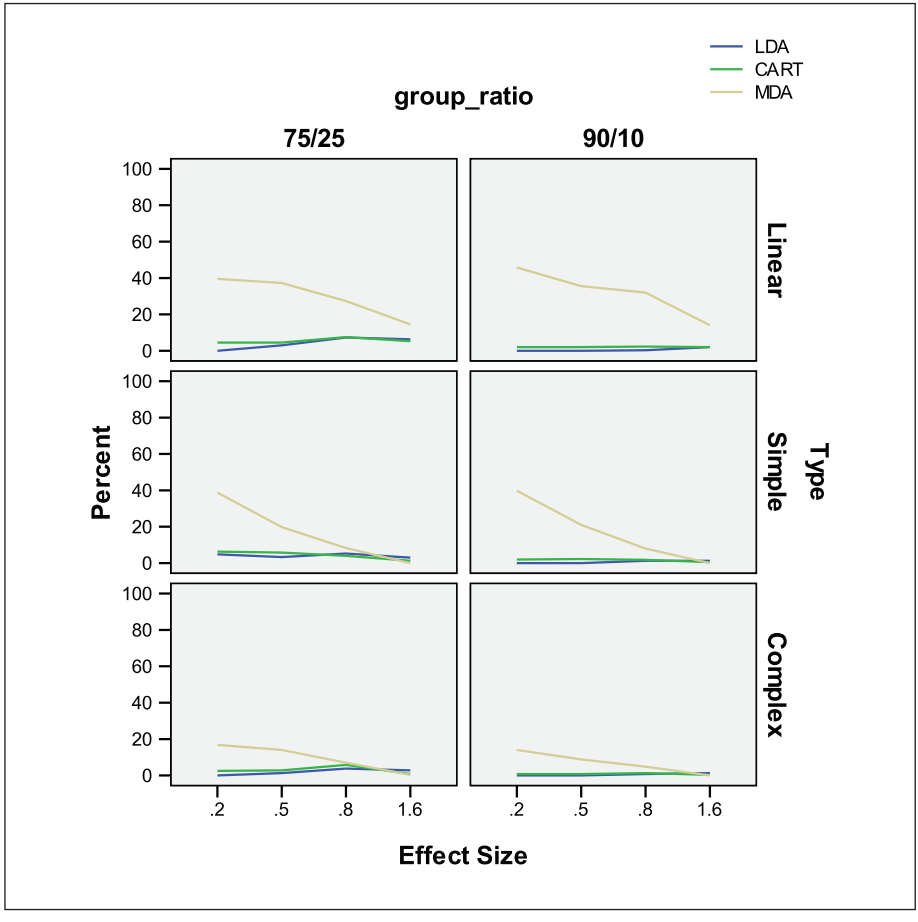

Figure 2. Large-group misclassification percentages for LDA, CART, and MDA by effect size, sample size ratio, and type of underlying model
Note. LDA = linear discriminant analysis; CART = classification and regression trees; MDA = mixture discriminant analysis.

randomly selected training sample of 269 from the full data set was used to develop the prediction models for each method, which were then applied to a cross-validation sample made up of the 100 individuals not included in the training sample. Rates of overall and group misclassification for each technique appear in Table 7. In general, these results are similar to those reported for the simulation study. CART and GAM provided the most accurate predictions overall, whereas MDA and CART were most accurate for the smaller group. LDA and QDA performed very similarly, with LR

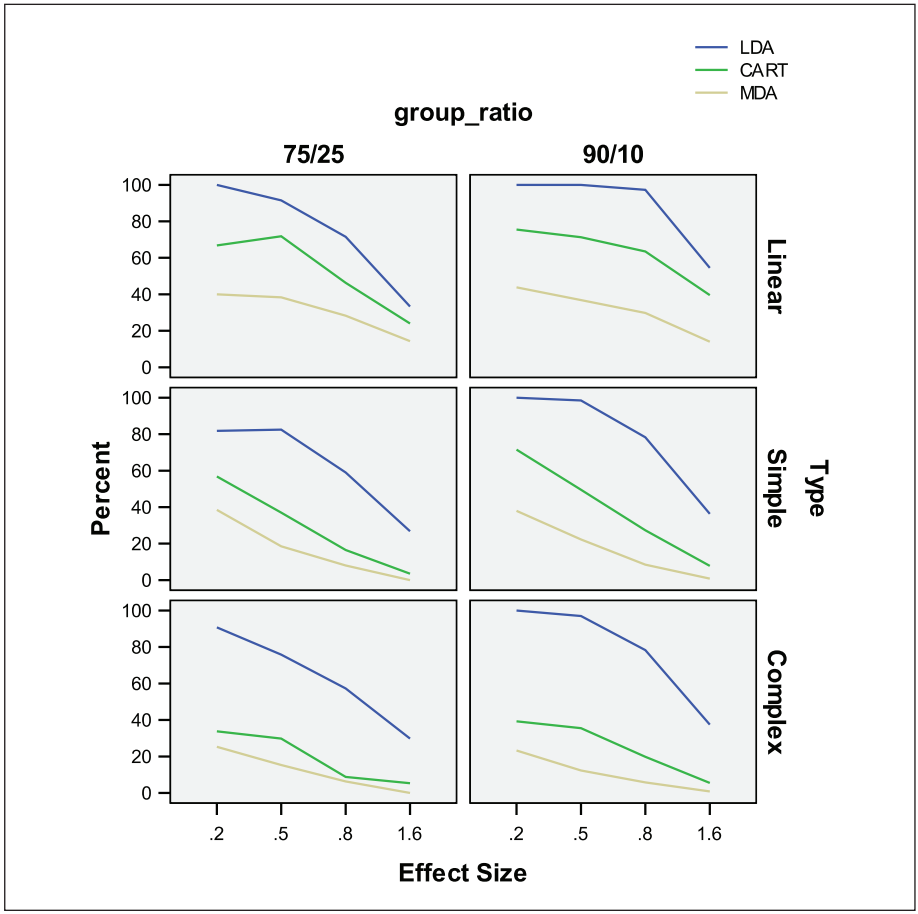

Figure 3. Small-group misclassification percentages for LDA, CART, and MDA by effect size, sample size ratio, and type of underlying model
Note. LDA = linear discriminant analysis; CART = classification and regression trees; MDA = mixture discriminant analysis.

predicting with slightly greater accuracy. BOOST was the least accurate method, which was also reported for the simulation study.

Discussion

The results of this study have several implications for researchers interested in two-group classification. Although literature in the social and behavioral sciences on classification remains dominated by traditional methods such as LDA and LR, it appears that several of the alternative methods investigated here might be more

Table 7. Misclassification Rate for Masculinity by Prediction Method

Method	Total (*N* = 100)	Large-Group (*N* = 66)	Small-Group (*N* = 34)
LDA	.26	.13	.49
QDA	.27	.17	.47
LR	.23	.12	.46
CART	.20	.13	.34
GAM	.20	.09	.41
MARS	.24	.12	.48
NNET	.29	.16	.54
BOOST	.32	.01	.93
MDA	.28	.26	.32

Note. LDA = linear discriminant analysis; QDA = quadratic discriminant analysis; LR = logistic regression; CART = classification and regression trees; GAM = generalized additive models; MARS = multivariate adaptive regression splines; NNET = neural networks; BOOST = boosting; MDA = mixture discriminant analysis.

appropriate in many situations. At the same time, there are occasions when none of these models will be particularly successful in terms of correctly classifying individuals.

The existing literature failed to reach consensus regarding optimal model choice for classification accuracy. Results of the present study found that CART consistently produced the greatest accuracy across the widest array of conditions. CART produced the greatest accuracy when interactions were present in the data, and it was also among the most accurate methods in the case of strictly linear population models. Although some of the other approaches (e.g., GAM, NNET, MARS, MDA) provided comparable or superior accuracy to LDA, QDA, and LR in certain specific cases, in other instances they were not more accurate than these traditional tools. On the other hand, CART was typically the best, or among the best, performer regardless of sample size, group size ratio, effect size, and type of model and virtually always provided more accurate predictions than LDA, QDA, or LR.

Another clear result was that the linear methods, BOOST, LDA, QDA, and LR consistently had difficulty in classifying individuals when the simulated models contained interactions. The only exception was for the 90:10 group ratio condition in which case all methods were fairly comparable except for MDA. Thus, a researcher interested in group prediction may want to consider carefully whether to use these more traditional methods when it may be that alternative techniques, particularly CART, GAM, MARS, and NNET, may prove superior. This superiority was particularly notable for the most complex models. Also considering that boosting has rarely been compared with other classification methods, its poor performance is of considerable note.

As has been shown in previous research (e.g., Finch & Schneider, 2007; Holden & Kelley, 2010), when groups are of unequal size, prediction was more accurate for the larger group for all the methods studied here, except MDA. MDA provided more accurate predictions for the smaller group than any of the other models and indeed was

generally as accurate for this group as for the larger. Although it did not perform as well in this regard as MDA, CART did have lower misclassification rates for the smaller group in the 75:25 and 90:10 group size ratio conditions than did the other techniques. Considering that MDA is based on a linear model like LDA and LR, it is noteworthy that it demonstrated such strong performance when classifying the smaller group even in the nonlinear condition, especially noting that it is the only linear model to demonstrate a classification bias in favor of the smaller group.

Implications for Practice

Given the results described above, several implications for categorical prediction in practice seem to emerge. First of all, researchers should consider using an alternative to the traditional discriminant function analysis or LR, even when they know or strongly suspect that the relationship between group membership and the predictors is linear. Second, when this relationship is more complex, methods such as CART, GAM, or NNET would seem most appropriate. Third, in those cases where the groups are of unequal size in the population and there is particular interest in correctly identifying members of the smaller group, researchers should consider using MDA. Finally, when the degree of group separation is quite large (effect size values in excess of 0.8), the prediction methodology used may not matter a great deal, as they will all provide fairly accurate results.

Although the results of this study generally support the use of one or more of the alternative prediction methods, they are not without some drawbacks. Perhaps foremost of these is that they are typically not automatic to use and require some experience to obtain optimum results. Methods such as NNET, GAM, and MARS all require the user to decide on the relative degree of model complexity that is appropriate. CART analysis typically involves the creation of an initial tree, followed by a winnowing of its complexity in a process known as pruning. In short, all these methods require an investment of time beyond that required for the more traditional approaches. A second drawback to these alternative prediction methods is that, with the exception of CART, they do not typically provide the user with useful information regarding the relative importance of predictors in group separation or with an easily digested equation for this purpose. Finally, several of these methods are available only in specialized software such as R and not the more commonly used SPSS or SAS. Although this may not represent a tremendous problem for many researchers, it does require users to become familiar with a new computing environment.

Directions for Future Research

This study represents a first investigation into the use of these alternative methods of prediction in the social sciences, and as such further work extending it needs to be done. For example, the nonlinear models associated with group membership used here are just two of a great many that could be used. Thus, future studies need to expand the

types of such population models that are examined. In this regard, future studies should also examine the impact of model misspecification on the performance of these methods (e.g., having a population model that takes the form $Y = X_1 + X_2$, but where the sample model is $Y = X_1 + X_2 + X_3 + X_4$). In addition, the settings for the alternatives used here were those that are recommended in the general literature. However, in practice researchers typically make adjustments to the various tuning parameters and prune CART trees before arriving at a final model. Therefore, a next step in this line of research would be to investigate how results change when such adjustments are made.

This study also limited classification to the two-group case. However, not surprisingly, classification becomes increasingly difficult as the number of groups increases. Therefore further research into the behavior of these models when multiple groups are present is warranted. On a related note, classification also becomes increasingly complex when the accuracy of the observed data is in question. Each of the methods described in this study would be considered to be supervised classification methods—that is, methods that rely on knowledge of true-group classification. However, when the accuracy of the training data is in question, the accuracy of supervised classification methods is debatable. Authors have previously discussed the accuracy of classification with misclassified training data for discriminant function methods and LR (Chhikara & McKeon, 1984; Grayson, 1987; Holden & Kelley, 2010; Lachenbruch, 1966, 1974, 1979; Lei & Koehly, 2003; McLachlan, 1972); however, the topic of training data misclassification has yet to be studied for these newer classification methods. The results of the current study indicate that alternative methods of classification, particularly CART, may provide better classification accuracy, especially for complex models. Thus, it would also be of interest to discover if these methods also provide higher accuracy under circumstances of misclassified training data.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Arabie, P., Hubert, L. J., & De Sote, G. (1996). *Clustering and classification*. River Edge, NJ: World Scientific.

Bailly, J. S., Arnaud, M., & Puech, C. (2007). Boosting: A classification method for remote sensing. *International Journal of Remote Sensing, 7*, 1687-1710.

Baker, F. B. (1974). Stability of two hierarchical grouping techniques. Case 1: Sensitivity to data errors. *Journal of the American Statistical Association, 69,* 440-445.

Berk, R. A. (2008). *Statistical learning from a regression perspective.* New York: Springer.

Blashfield, R. K. (1976). Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *Psychological Bulletin, 83*, 377-388.

Breckenridge, J. M. (2000). Validating cluster analysis: Consistent replication and symmetry. *Multivariate Behavioral Research, 35*, 261-285.

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

Breiman, L., Friedman, J., Olshen, R. A.., & Stone, C. J. (1984). *Classification and regression trees.* Monterey, CA: Wadsworth & Brooks/Cole.

Buhlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science, 22*, 477-505.

Buhlmann, P., & Yu, B. (2003). Boosting with the L2 loss: Regression and classification. *American Statistical Association, 98*, 324-339.

Chhikara, R. S., & McKeon, J. (1984). Linear discriminant analysis with misallocation in training samples. *Journal of the American Statistical Association, 79*, 899-906.

Clayton, K., Blumberg, F., & Auld, D. P. (2010). The relationship between motivation, learning strategies and choice of environment whether traditional or including an online component. *British Journal of Educational Technology, 41*, 349-364.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Hillsdale, NJ: Lawrence Erlbaum.

Curram, S. P., & Mingers, J. (1994). Neural networks, decision tree induction and discriminant analysis: An empirical comparison. *Journal of the Operational Research Society, 45*, 440-450.

deCraen, S., Commandeur, J. F., Frank, L. E., & Heiser, W. J. (2006). Effects of group size and lack of sphericity on the recovery of clusters in k-means cluster analysis. *Multivariate Behavioral Research, 41*, 127-145.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*, 1-38.

Do, A, Q, & Grudnitski, G (1992). A neural network analysis of the effect of age on housing values. *The Journal of Real Estate Research, 8*, 253-264.

Dudoit, S., Fridlyand, D., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association, 97*, 77-87.

Finch, H., & Schneider, M. K. (2006). Misclassification rates for four methods of group classification: Impact of predictor distribution, covariance inequality, effect size, sample size and group. *Educational and Psychological Measurement, 66*, 240-257.

Finch, H., & Schneider, M. K. (2007). Classification accuracy of neural networks vs. discriminant analysis, logistic regression and classification and regression trees. Three- and five-group cases. *Methodology, 3*, 47-57.

Freund, Y., & Schapire. R. E. (1997). Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the Thirteenth International conference, 1996.*

Garson, G. D. (1998). *Neural networks: An introductory guide for social scientists.* London: Sage.

Grassi, M., Villani, S., & Marinoni, A. (2001). Classification methods for the identification of "case" in epidemiological diagnosis of asthma. *European Journal of Epidemiology, 17*, 19-29.

Grayson, D. A. (1987). Statistical diagnosis and the influence of diagnostic error. *Biometrics, 43*, 975-984.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction.* New York: Springer.

Harrell, F. E., & Lee, K. I. (1985). A comparison of the discrimination of discriminant analysis and logistic regression under multivariate normality. In P. K. Sen (Ed.), *Biostatistics: Statistics in biomedical, public health and environmental sciences* (pp. 333-343). New York, NY: Elsevier Science.

Hart, A. (1992). Using neural networks for classification tasks: Some experiments on datasets and practical advice. *Journal of the Operational Research Society, 43*, 215-226.

Hastie, T., & Tibshirani, R. (1996). Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society: Series B (Methodological), 58*, 155-176.

Holden, J. E., & Kelley, K. (2010). The effects of initially misclassified data on the effectiveness of discriminant function analysis and finite mixture modeling. *Educational and Psychological Measurement, 70, 36-55.*

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics, 15,* 651-674.

Keogh, B. K. (2005). Revisiting classification and identification. *Learning Disability Quarterly, 28*, 100-102.

Krzanowski, W. J. (1976). The performance of Fisher's linear discriminant function under non-optimal conditions. *Technometrics, 19*, 191-200.

Kuhnert, P. M., Mengersen, K., & Tesar, P. (2003). Bridging the gap between different statistical approaches: An integrated framework for modeling. *International Statistical Review, 71*, 335-368.

Lachenbruch, P. A. (1966). Discriminant analysis when the initial samples are misclassified. *Technometrics, 8*, 657-662.

Lachenbruch, P. A. (1974). Discriminant analysis when the initial samples are misclassified II: Non-random misclassification models. *Technometrics, 16*, 419-424.

Lachenbruch, P. A. (1979). Note on initial misclassification effects on the quadratic discriminant function. *Technometrics, 21*, 129-132.

LeBlanc, M., & Tibshirani, R. (1996). Combining estimates on regression and classification. *Journal of the American Statistical Association, 91,* 1641-1650.

Lee, T., Chiu, C., Chou, Y, & Lu, C. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics and Data Analysis, 50,* 1113-1130.

Lei, P, & 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*, 25-49.

Lillvist, A. (2010). Observations of social competence of children in need of special support based on traditional disability categories versus a functional approach. *Early Child Development and Care, 18*, 1129-1142.

Liu, D., & Chun, Y, (2009). The effects of different classification models on error propagation in land cover change detection. *International Journal of Remote Sensing, 20, 5345-5364.*

Mammarella, I. C., Lucangeli, D., & Cornoldi, C. (2010). Spatial working memory and arithmetic deficits in children with nonverbal learning difficulties. *Journal of Learning Disabilities, 43*, 455-468.

Marshall, D. B., & English, D. J. (2000). Neural network modeling of risk assessment in child protective services. *Psychological Methods, 5,* 102-124.

McLachlan, G. J. (1972). Asymptotic results for discriminant analysis when initial samples are misclassified. *Technometrics, 14*, 415-422.

Moisen, G. G., & Frescino, T. S. (2002). Comparing five modeling techniques for predicting forest characteristics. *Ecological Modeling, 157*, 209-225.

Nguyen, N., & Cripps, A. (2001). Predicting housing value: A comparison of multiple regression analysis and artificial neural networks. *Journal of Real Estate Research, 22*, 313-336.

Preatoni, D. G., Nodari, M., Chirchella, R., Tosi, G., Wauters, L. A., & Martinoli, A. (2005). Identifying bats from time-expanded recordings of search calls: Comparing classification methods. *Journal of Wildlife Management, 69*, 1601-1614.

Rausch, J. R., & Kelley, K. (2009). A comparison of linear and mixture models for discriminant analysis under nonnormality. *Behavior Research Methods, 41*, 85-98.

Ripley, B. D. (1994). Neural networks and related methods for classification. *Journal of the Royal Statistical Society: Series B (Methodological), 3*, 409-456.

Reibnegger, G., Weiss, G., Werner-Felmayer, G., Judmaier, G., & Wachter, H. (1991). Neural networks as a tool for utilizing laboratory information: Comparison with linear discriminant analysis and with classification and regression trees. *Proceedings of the National Academy of Science, 88*, 11426-11430.

Russell, J. A. (2008). A discriminant analysis of the factors associated with the career plans of string music educators. *Journal of Research in Music Education, 56*, 204-219.

Schumacher, M., Robner, R., & Vach, W. (1996). Neural networks and logistic regression: Part I. *Computational Statistics and Data Analysis, 21,* 661-682.

Simonoff, J. S. (1996). *Smoothing methods in statistics*. New York, NY: Springer.

Smith, A. E., & Mason, A. K. (1997). Cost estimation predictive modeling: Regression versus neural network. *The Engineering Economist, 42*, 137-161.

Tabachnick, B. G., & Fidell, L. S. (2007). Using multivariate statistics (5th ed). Boston, MA: Pearson and Allyn and Bacon.

Tibshirani, R., & LeBlanc, M. (1992). A strategy for binary description and classification. *Journal of Computational and Graphical Statistics, 1*, 3-20.

Ture, M., Kurt, I., Kurum, A. T., & Ozdamar, K. (2005). Comparing classification techniques for predicting essential hypertension. *Expert Systems With Applications, 29*, 583-588.

Waller, N. G., & Jones, J. A. (2009). Correlation weights in multiple regression. *Psychometrika, 75*, 58-69. Retrieved from http://www.springerlink.com/content/e3572021626270x6/

West, P. M., Brockett, P. L., & Golden, L. L. (1997). A comparative analysis of neural networks and statistical methods for predicting consumer choice. *Marketing Science, 16*, 370-391.

Williams, C. J., Lee, S. S., Fisher, R. A., & Dickerman, L. H. (1999). A comparison of statistical methods for prenatal screening for down syndrome. *Applied Stochastic Models in Business and Industry, 15,* 89-101.

Wood, S. N. (2006). *Generalized additive models.* New York: Chapman & Hall.

Yoon, Y., Swales, G., Jr., & Margavio, T. M. (1993). A comparison of discriminant analysis versus artificial neural networks. *Journal of the Operational Research Society, 44*, 51-60.

Zigler, E., & Phillips, L. (1961). Psychiatric diagnosis: A critique. *Journal of Abnormal and Social Psychology, 63*, 607-618.