Comparing Statistical and Machine Learning Classifiers: Alternatives for Predictive Modeling in Human Factors Research

Brian Carnahan, Auburn University, Auburn, Alabama, **Gérard Meyer,** Carnegie Mellon Driver Training and Safety Institute, Lemont Furnace, Pennsylvania, and **Lois-Ann Kuntz,** University of Maine at Machias, Maine

Multivariate classification models play an increasingly important role in human factors research. In the past, these models have been based primarily on discriminant analysis and logistic regression. Models developed from machine learning research offer the human factors professional a viable alternative to these traditional statistical classification methods. To illustrate this point, two machine learning approaches – genetic programming and decision tree induction – were used to construct classification models designed to predict whether or not a student truck driver would pass his or her commercial driver license (CDL) examination. The models were developed and validated using the curriculum scores and CDL exam performances of 37 student truck drivers who had completed a 320-hr driver training course. Results indicated that the machine learning classification models were superior to discriminant analysis and logistic regression in terms of predictive accuracy. Actual or potential applications of this research include the creation of models that more accurately predict human performance outcomes.

INTRODUCTION

Background

As defined by Clancy (1997), classification analysis seeks to identify mathematical and/or statistical relationships between independent variables (discrete or continuous) that can effectively distinguish or characterize various levels within a nominal dependent variable (categorical variable). Once identified, these relationships can be used for descriptive assessment or predictive modeling. Researchers in the areas of human factors, ergonomics, safety, and psychology have for some time used multivariate classification analysis to expand the body of knowledge in their respective fields. Traditionally, discriminant analysis and logistic regression have been the most commonly used statistical methods for carrying out classification in these fields.

In industrial ergonomics, for example, these methods have been used to predict or identify carpal tunnel syndrome status (Babski-Reeves & Crumpton-Young, 2002; Matias, Salvendy, & Kuczez, 1998), jobs that pose high risk for work-related low-back disorders (Marras et al., 1993), lumbar discomfort while sitting (Vergara & Page, 2002), and the risk of slipping and falling while walking on a sloping surface (Hanson, Redfern, & Mazumdar, 1999). The field of transportation safety research has used these methods to identify factors that contribute to airline pilot error incidents (McFadden, 1997), red-light-running behavior in urban settings (Porter & England, 2000), driver drowsiness in virtual driving tasks (Verwey & Zaidel, 2000), and injury severity in motor vehicle accidents (Al-Ghamdi, 2002). In psychology and human factors research, these same methods have been used to predict success or failure in activities such as lifting and carrying criterion tasks (Rayson, Holliman, & Belyavin, 2000), air force basic training (Lubin, Fielder, & Van Whitlock,

Address correspondence to Brian Carnahan, Industrial and Systems Engineering, Auburn University, 207 Dunstan Hall, Auburn, AL 36849; carnahan@eng.auburn.edu. **HUMAN FACTORS,** Vol. 45, No. 3, Fall 2003, pp. 408–423. Copyright © 2003, Human Factors and Ergonomics Society. All rights reserved.

1999), and nuclear reactor operation (Mackieh & Cilingir, 1998).

Although their use is widespread in the human performance field, discriminant analysis and logistic regression do have their limitations. Principal among these is their dependence on a fixed, underlying model or functional form. Discriminant analysis uses linear summation of independent variables to differentiate one category from another (Huberty & Lowman, 1997). Logistic regression also makes use of linear summations of independent variables, incorporated into a logistic function (Myers, 1990). Koza (1992) made the observation that both techniques use regression merely to discover numerical coefficients for predetermined models. In human factors research, however, these parametric models may not be the most appropriate ones for classification tasks involving some human performance activities, especially those for which outcomes cannot be differentiated or predicted based on a simple linear summation of independent variables or for which functional form cannot be established a priori. Such circumstances would necessitate the use of alternative classification modeling approaches.

Machine Learning Alternatives for Classification Model Development

Beyond the plethora of traditional statistical classification techniques available, of which discriminant analysis and logistic regression are but two, machine learning research offers the human factors professional a viable alternative for classification model development. Decision tree induction and genetic programming are two of these machine learning approaches. Unlike the traditional statistical methods previously discussed, these methods do not rely on predetermined models using linear summations of independent variables.

In decision tree induction (Quinlan, 1986), the decision tree is constructed from a data set composed of a series of cases. Each case is composed of a set of values (discrete or continuous) associated with a fixed set of descriptive independent variables and a corresponding outcome-dependent variable (i.e., category or class) associated with those values. Based on the specific values of these descriptive variables, the tree partitions the cases into subsets that are homogeneous in terms of the outcome variable. Each of these subsets represents a single path or branch along the decision tree. Each path represents a specific pattern of descriptive variable values common to a number of cases that possess the same outcome. Once constructed, the tree can classify future cases in which the outcome is not known, based on the values of its descriptive variables. Decision tree induction has been used in the past to identify common patterns associated with specific automotive accident outcomes (Clarke, Forsyth, & Wright, 1998b; Sohn & Shin, 2001).

Genetic programming (Koza, 1992, 1994; Koza, Bennett, Andre, & Keane, 1999) is a search method based on the Darwinian principles of natural selection and survival of the fittest. This approach uses fitness-based selection and solution recombination to produce a population of increasingly effective computer programs designed to solve a particular problem. This technique is one form of heuristic search algorithms found in a field of computer science research known as evolutionary computation (Sebald & Fogel, 1994). By relying on an evolving population of programs driven by natural selection, genetic programming uses multiple points that can climb in different directions in parallel as well as "jump" to different locations within the solution space. For classification modeling, genetic programming can be used to solve problems in symbolic regression. As defined by Koza (1992), symbolic regression entails discovering a mathematical expression that provides the best fit between independent and corresponding dependent variables within a finite data sample.

Kishore, Patnaik, Mani, and Agrawal (2000) identified two distinct advantages that genetic programming would have over traditional statistical methods for classification. First, genetic programming does not require advanced knowledge or assumptions concerning the statistical distribution of the data. Second, genetic programming does not use any specific predetermined model but can detect an underlying relationship between independent and dependent variables and express that relationship through a mathematical model constructed to fit the data.

Evolutionary computation techniques, such as genetic programming, have been used to discover solutions to various ergonomics design problems involving control panels (Pham & Onder, 1992), lifting tasks (Carnahan & Redfern, 1998a), job rotation schedules (Carnahan, Redfern, & Norman, 2000), and assembly line balances (Carnahan, Norman, & Redfern, 2001). In terms of classification, transportation safety researchers have made use of these techniques to discover rules that could classify and predict the outcome of automotive accidents (Clarke, Forsythe, & Wright, 1998a). In addition, Carnahan and Redfern (1998b) have used genetic programming to develop models that accurately classify lifting tasks as posing high or low risk of occupational back injury.

Project Purpose

The purpose of the current project was to conduct a methodological pilot study that compared traditional statistical classification methods (discriminant analysis and logistic regression) with methods based on machine learning (decision tree induction and genetic programming) in an area relevant to human factors/human performance research. The basis for comparison will be data associated with commercial-driver training and subsequent examination. The hypothesis being tested was that the machine learning classification models could yield performances superior to those of discriminant analysis or logistic regression when called upon to accurately predict human performance outcome.

METHOD

Participants

Data were collected from 37 trainees (36 men, 1 woman) who enrolled in a novice truck driver training course offered by the Carnegie Mellon Driver Training and Safety Institute. The trainees' ages ranged from 18 to 58 years (average = 38.0; standard deviation = 10.1). All trainees had previously received a high school or general equivalency diploma. Ethnic composition was not diverse (all participants were Caucasian). All trainees had passed drug toxicology and medical screenings as required by the U.S. Department of Transportation, Federal Highway Administration. All trainees received full tuition assistance from state funding sources.

Data Collection Procedure and Classification Task

Performance data from each of the 37 participants were based on their completion on an 8-week, 320-hr course designed to teach novice drivers basic truck driving skills with emphasis in the areas of safety and accident prevention. The curriculum was composed of six major components: classroom instruction (Byrnes & Fox, 1999), range driving, road driving, simulation training, controlling a truck through a skid or slide, and physical health/wellness training. By completing the curriculum, each participant accumulated 10 scores that reflected his or her performance in the truck driver training course. A description of these scores is shown in Table 1.

Referring to Table 1, Curriculum Areas 1 through 5 (basic operations, driving techniques, vehicle systems, operation skills, and professional driver) covered textbook knowledge commonly found in commercial driver training courses. Scores for these five areas were based on multiple-choice tests of textbook chapter materials (Byrnes & Fox, 1999). Curriculum Area 6, driver development, covered information concerning a personal health program designed to teach ways of improving and maintaining health and physical fitness. Scores for this area were based on a multiple-choice exam covering topics related to physical wellness and lifetime management. Curriculum Areas 7 through 9 (skill development, range skills, pretrip inspections, and road skills) covered those driving skills that are relevant to the commercial drivers' license (CDL) examination. Scores in these curriculum areas were based on examinations that used the same range tests, road-driving tests, and guidelines as those used by actual CDL examiners. Each score was divided by its corresponding maximum value (the maximum score possible in the curriculum area), thereby normalizing the scores for all 10 variables. These 10 normalized scores were selected to represent the set of dependent variables for the classification task.

Once training was completed, each trainee took the State of Pennsylvania's CDL examination at one of six nearby testing locations. This examination consisted of a trained evaluator

TABLE 1: Description of Curriculum Areas, Content, and Evaluation Score Used in Driving Training Course

Curriculum Area	Curriculum Content	Basis for Evaluation Score
1. Basic operation	Orientation, control systems, vehicle inspection, basic controls, backing, coupling and uncoupling, proficiency development	20-item multiple-choice test
2. Driving techniques	Visual search and communication, speed and space management, night operations, extreme driving, hazard perception, emergency maneuvers, skid control recovery, proficiency development	20-item multiple-choice test
3. Vehicle systems	Vehicle systems functions, diagnosing and reporting malfunctions, preventative maintenance, shifting gears, fire prevention and safety, accident procedures	20-item multiple-choice test
4. Operations skills	Cargo documentation, handling cargo, hours of service requirements, trip planning, hazardous materials, computers in trucking	20-item multiple-choice test
5. Professional driver	Job placement and succeeding as a truck driver, public and employer relations, personal health and safety, U.S. Department of Transportation regulations	100-item multiple-choice test covering all previous curriculum areas, including 5
6. Driver development	Physical wellness, lifetime management	50-item multiple-choice test
7. Skill development	Truck backing, parallel parking, alley docking, and right/left hand turns	23-item skill assessment in simulator and field tests
8. Range skills and pretrip inspection	Inspection/testing of all truck systems and safety/emergency equipment	33-item skill assessment in field test similar to CDL examination
9. Road skills	Truck maneuvers covered in CDL examination	76-item skill assessment in road test
10. Attendance	Class attendance covering all topic Areas	Percentage of class hours attended

using a standard form to assess and score the trainee in three skill areas: pretrip inspection, range maneuvers, and road skills. Of the 37 trainees, 25 passed the CDL on the first attempt and the remaining 12 passed on their second or third attempt. This binary categorical variable (those who passed the CDL on their first attempt and those who did not) was selected as the dependent variable for the classification task. Thus the classification task was to accurately predict whether or not a trainee would pass his or her first CDL examination based on the 10 scores the individual received after completing the truck driver training course. A number of research studies have searched for factors that are predictive of driving behavior. Some of these factors have been used to predict accident propensity (Reason, Manstead, Stradling, Baxter, & Campbell, 1990; Simon & Corbett, 1996), risky behaviors (Deery & Fildes, 1999; Meadows, Stradling, & Lawson, 1998), or perceptual and cognitive abilities (Avolio, Kroeck, & Panek, 1985; French, West, Elander, & Wilding, 1993; Myers, Ball, Kalina, Roth, & Goode, 2000). However, to date, there has been no empirically validated test that may be utilized to predict CDL examination performance outcome.

Once collected, the performance data were partitioned into two subsets. Five cases in which the trainee passed and 5 cases in which the trainee failed his or her initial CDL examination were randomly selected from the original data set. This collection of 10 cases constituted the test or validation data set. The remaining 27 cases constituted the training data set. Each approach (discriminant analysis, logistic regression, decision tree induction, and genetic programming) used the training data set to develop its corresponding classification model. Each model was then validated using the 10 cases of the test set. Clancy (1997) has recommended this procedure of data partitioning, model development, and validation for assessing the true predictive accuracy of classification models used in ergonomics research. What follows is a description of the four approaches used to create the classification models.

Discriminant Analysis

Discriminant analysis (Huberty & Lowman, 1997) accomplishes multivariate classification through the use of linear functions, as expressed in the equation

$$Si = W_{1i}X1i + W_{2i}X2i + \ldots + W_{10i}X10i + c_i, \quad (1)$$

in which Si = resultant classification score for the ith category, $X1$ through $X10$ = the 10 scores acquired by completing the truck driver training curriculum, W_{1i} through W_{10i} = weighting values corresponding to the independent variables $X1$ through $X10$ for the ith category, and c_i = constant for the ith category.

The classification functions, the number of which corresponds to the number of categories, can be used to directly compute classification scores for new observations. Specifically, newly observed cases are classified to that group whose classification function yields the highest classification score, Si. The discriminant analysis established two classification functions of the form shown in Equation 1. These functions allowed the model to provide a binary response that predicted the passing or failing of the CDL examination based on the training curriculum scores.

Logistic Regression

The logistic regression model for predicting outcome of the CDL examination is expressed in the equation

$$Y = \frac{1}{1 + e^{-(\beta0 + \beta1\mathrm{X}1 + \beta2\mathrm{X}2 + \ldots + \beta10\mathrm{X}10)}}, \quad (2)$$

in which Y = the probability that the input vector belongs to the "pass" category, $\beta0$ = regression model constant, and $\beta1$ through $\beta10$ = coefficients corresponding to the independent variables $X1$ through $X10$.

Using Equation 2, the dependent variable Y output of .5 or greater would result in the model classifying the case as "passing the CDL examination on the first attempt." An output of below .5 results in the model classifying the case as "failing the CDL examination on the first attempt." This same rule was applied when validating the model using the 10 cases of the test data set. Two logistic regression models were created: one that included a constant and one that did not. The reasoning behind this decision was that if the logistic models differed in form, they might also differ in terms of subsequent predictive performance on the test data set.

Decision Tree Induction: The C4.5 Algorithm

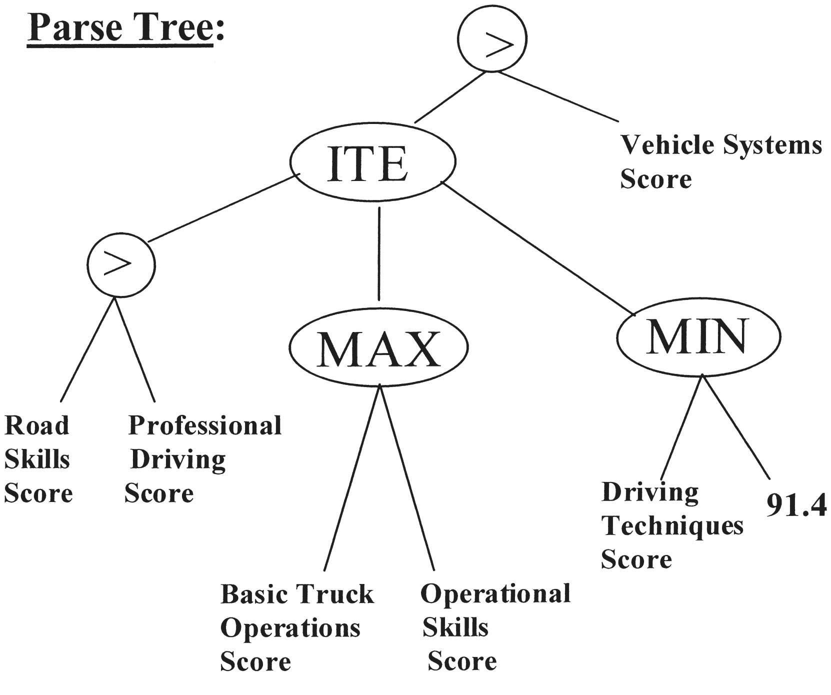
The decision tree for predicting trainee success or failure on the CDL examination was constructed using the C4.5 algorithm (Quinlan, 1993). The algorithm was a recursive greedy heuristic that selected independent variables for membership within the tree. Whether or not an independent variable was included within the tree was based on the value of its information gain. As a statistical property, information gain measured how well the variable (curriculum area score) separated training cases into subsets in which the outcome or dependent variable value (passing or failing the CDL examination) was homogeneous. Given that curriculum area scores were all continuous variables, a threshold value had to be established within each score variable so that it could partition the training cases into subsets. These threshold values for each variable were established by rank ordering the values within each variable from lowest to highest and repeatedly calculating the information gain using the arithmetical midpoint between all successive values within the rank order. The midpoint value with the highest information gain was selected as the threshold value for its score variable. That variable with the highest information gain (information being the most useful for classification) was then selected for inclusion in the decision tree.

The algorithm continued to build the tree in this manner until it accounted for all training cases. Ties between variables that were equal in terms of information gain were broken using a random number generator embedded in the algorithm. The algorithm was run 10 times, using a different random number seed each time. By altering the random number seed, the algorithm could produce decision trees that differed in structure and classification performance.

Genetic Programming

The genetic programming algorithm searched for accurate classification models by using five distinct stages: initialization, fitness evaluation, selection, reproduction, and replacement.

Initialization. In this stage, the algorithm generated a population of 500 computer programs. A parse tree represented each computer program in the population. Each parse tree was constructed from a random selection of both function nodes and terminal nodes. The function nodes consisted of arithmetical operators (addition, +; subtraction, –; multiplication, ×; division, ÷; return maximum value, MAX; and return minimum value, MIN) and Boolean operators (greater than, >; less than, <; logical and, AND; logical or, OR; if-then-else, ITE). The terminal nodes consisted of the trainee scores from the truck driver training course (Curriculum Areas 1–10) and random numbers ranging from 1 to 100. Figure 1 shows an example of a randomly generated parse tree and its translation into pseudo computer code.



Pseudo Computer Code:

```
BEGIN PROGRAM
IF          Condition 1- Road Skill Score is greater than Professional Driving Score
THEN        Value = Maximum (Basic Truck Operations Score, Operational Skills Score)
ELSE        Value = Minimum (Driving Techniques Score, 91.4)

IF          Condition 2 - Value is greater than Vehicle Systems Score
THEN        Predict trainee will pass CDL examination
ELSE        Predict trainee will fail CDL examination
END PROGRAM
```

Figure 1. Example of a single parse tree accompanied by its translation into pseudo computer code.

As shown in Figure 1, the parse tree represents a computer program that predicts whether or not a trainee will pass the CDL examination on the first attempt. The program bases its decision on the values assigned to the following terminal nodes: road skill score, professional driving score, basic truck operations score, operational skills score, driving techniques score, and vehicle systems score. In order to predict performance, each parse tree must be translated into a corresponding computer program (shown in the lower half of Figure 1). The root node of each parse tree created at initialization was always a Boolean operator (>, <, AND, OR, ITE). This allowed each program to act as a binary classifier that predicted a trainee would either pass the CDL examination on the first attempt (i.e., the program returns a Boolean output of TRUE) or fail it on the first attempt (i.e., the program returns a Boolean output of FALSE).

The program's decision would be based on the variables, constants, arithmetical operators, and Boolean operators that composed the rest of the tree. Each parse tree is a random combination of no more than 30 input variables, numerical constants, and arithmetical operators. Constraints used in strongly typed genetic programming (Montana, 1995) were incorporated in the algorithm to maintain the feasibility of all parse trees in the population during all algorithmic stages. These constraints allowed the genetic programming algorithm to discover computer programs that seamlessly integrated arithmetical operators, Boolean operators, independent variables, and numerical constants into an overall classification model.

Fitness evaluation. In this stage, each of the 500 computer programs in the population was assigned a fitness value. This fitness value was simply the percentage of 27 cases in the training data set that the program was able to correctly classify. After a fitness value had been assigned to a program, the program was then called upon to classify the 10 cases in the test data set. The specifications of those programs in the population that possessed 80% accuracy across both data sets were saved to a file. This procedure enabled the genetic programming algorithm to retain those classification models that performed well in terms of training and validation.

Selection. In this stage of the algorithm, candidates from the population of computer programs were selected for survival. Using stochastic selection with replacement (Goldberg, 1989), a new population of surviving programs was created from the current population. Those programs with the highest fitness values (i.e., most accurate classification of the training data set) had the greatest probability of being repeatedly chosen for survival. Thus the survivor population was biased in the sense that there were more copies of higher-fit programs than of lower-fit programs. These surviving programs participated in the two remaining stages of the genetic programming algorithm.

Reproduction. In this stage, surviving computer programs produce offspring, using the mechanisms of crossover and mutation. These offspring (new programs) would go on to represent the next generation of programs in the population. In the crossover mechanism, 60% of the surviving programs (300 parse trees) were randomly matched to form 150 pairs. Within each pair of programs, randomly chosen branches (i.e., subtrees) from each program's parse tree were exchanged. The result was the creation of two new programs, known as offspring, which possessed characteristics of both parents. In the mutation mechanism, 1% of the surviving programs (5 parse trees) had a randomly chosen branch erased and then replaced with a randomly generated branch. This procedure created a single offspring program that was similar to, but different from, its parent. For both the crossover and mutation mechanisms, the offspring programs replaced their parents in the surviving population.

Replacement. In this stage of the algorithm, the surviving computer programs, some of which have been altered because of crossover and mutation, become the new programs of the next generation.

The completion of a single iteration of the fitness evaluation, selection, reproduction, and replacement stages constituted a single generation within the genetic programming algorithm. Within a single run, the algorithm completed 1000 of these generations. Given the heuristic nature of this algorithm, 10 runs were completed, with each run starting with a different random number seed.

Criteria for Classification Model Comparison

Comparisons among the discriminant analysis model, the two logistic regression models, the most fit program found using genetic programming, and the most accurate decision tree discovered using the C4.5 algorithm were based on their classification of the 10 cases constituting the test data set. The classification performance results on the test data set were described in terms of the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), as shown in Table 2. The performance of each model was evaluated using the following epidemiological equations (Hennekens & Buring, 1987), in which *accuracy* is the percentage of all cases accurately classified by the model; *sensitivity* is the ability of the model to identify those trainees who passed the CDL test on the first attempt; *specificity* is the ability of the model to identify those trainees who failed the CDL test on the first attempt; and *validity* is an overall measure of the model's sensitivity and specificity performance.

$$Accuracy = [(TP+TN)/(TP+TN+FP+FN)] \times 100 \quad (3)$$
$$Sensitivity = [TP/(TP + FN)] \times 100 \quad (4)$$
$$Specificity = [TN/(TN + FP)] \times 100 \quad (5)$$
$$Validity = (Sensitivity + Specificity - 1) \times 100 \quad (6)$$

In addition to these epidemiological criteria, the classification accuracy of each of the models on the training set data was also noted and recorded.

RESULTS

Discriminant Analysis Model Description

The descriptive characteristics of the classification model developed using discriminant analysis are summarized in Table 3. As shown in Table 3, three individual regressors – operation skills, professional driver, and road skills – were statistically significant at the $\alpha = .05$ level in terms of discriminating between those trainees who passed the CDL examination on the first attempt and those who did not.

Logistic Regression Analysis Model Description

The descriptive characteristics of the classification models developed using logistic regression are summarized in Tables 4 and 5. Table 4 summarizes the descriptive characteristics of the logistic regression model that included a constant (Model 1). The overall model was statistically significant, $\chi^2(10) = 22.19$, $p = .014$. None of the models' individual regressors, however, were statistically significant at the = .05 level. As shown in Table 4, increases in the likelihood of membership in the "pass CDL examination" class were most influenced by increases in professional driver, road skills, and skill development score values. Table 5 summarizes a similar logistic regression model, one that did not include a constant (Model 2). As with logistic regression Model 1, the overall model was also statistically significant, $\chi^2(10) = 23.67$, $p = .009$; however, the effects of individual regressors were not. As shown in Table 5, increases in the road and operation skills had the greatest impact in terms of increasing the likelihood of membership in the "pass CDL examination" class.

Decision Tree Model Description

Across the 10 runs, the most accurate decision tree found by the C4.5 algorithm made use of only 6 of the 10 curriculum area scores in predicting passing or failing the CDL examination. An illustration of this decision tree and its corresponding translation into pseudo code is

TABLE 2: Results of Binary Classification Performance

Model Predicted Outcome	Actual CDL Exam Outcome	
	Trainee Passed	Trainee Failed
True (passed)	True positive (TP)	False positive (FP)
False (failed)	False negative (FN)	True negative (TN)
Total cases	TP + FN	FP + TN

TABLE 3: Descriptive Characteristics of the Discriminant Analysis Classification Model

				Classification Function Coefficients	
Variable	Wilks's Lambda	Univariate *F* Ratio	Significance (*p*)	Pass CDL	Fail CDL
Basic operation	0.98	0.40	.53	21.88	22.058
Driving techniques	1.00	0.11	.74	6.22	6.329
Vehicle systems	1.00	0.01	.94	10.69	10.82
Operations skills	0.85	4.57	.04	42.52	42.22
Professional driver	0.85	4.27	.05	–82.59	–83.11
Driver development	0.97	0.72	.40	2178.44	2173.08
Skill development	0.95	1.40	.25	–23.23	–23.31
Range skills and pretrip inspection	1.00	0.10	.75	37.85	37.97
Road skills	0.86	4.11	.05	–18.59	–18.84
Attendance	1.00	0.08	.78	42.68	42.80
Constant	N/A	N/A	N/A	–110 275.05	–110 760.34

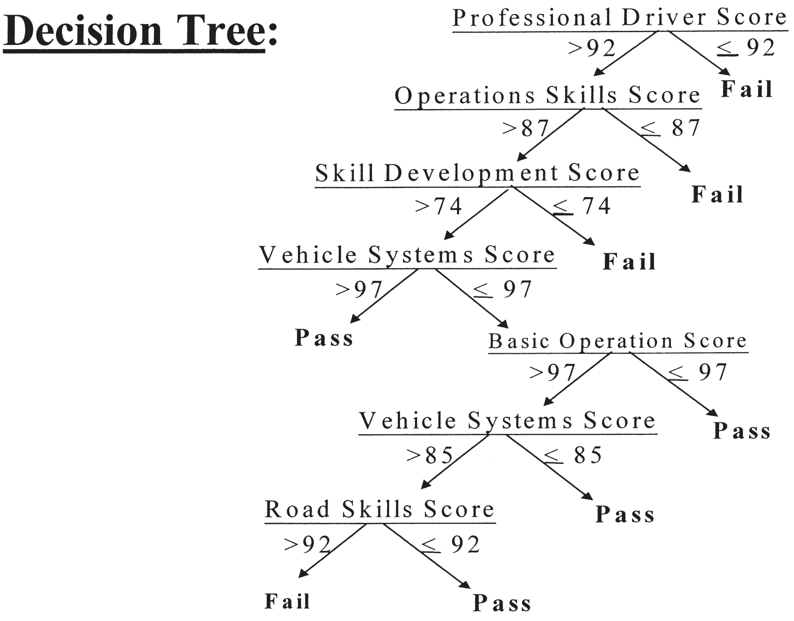
TABLE 4: Descriptive Characteristics of Logistic Regression Model 1

Variable	β	Standard Error	Wald χ^2	Significance (*p*)	Exp(β)
Basic operation	–1.64	1.65	1.00	.32	0.19
Driving techniques	–0.60	0.79	0.58	.45	0.55
Vehicle systems	0.21	0.30	0.50	.48	1.24
Operations skills	0.24	0.41	0.36	.55	1.28
Professional driver	2.77	2.88	0.92	.34	15.94
Driver development	–28.64	107.18	0.07	.79	0.00
Skill development	0.92	1.07	0.73	.39	2.50
Range skills/pretrip	–1.41	1.26	1.25	.26	0.24
Road skills	1.14	0.76	2.21	.13	3.11
Attendance	0.07	0.29	0.06	.80	1.08
Constant	2701.63	10692.74	0.06	.80	N/A

TABLE 5: Descriptive Characteristics of Logistic Regression Model 2

Variable	β	Standard Error	Wald χ^2	Significance (*p*)	Exp(β)
Basic operation	–0.30	0.24	1.61	.20	0.74
Driving techniques	–0.54	0.47	1.33	.25	0.58
Vehicle systems	0.40	0.39	1.02	.31	1.49
Operations skills	0.71	0.43	2.66	.10	2.02
Professional driver	0.25	0.33	0.63	.43	1.29
Driver development	–1.36	0.80	2.90	.09	0.26
Skill development	–0.19	0.39	0.24	.63	0.83
Range skills/pretrip	–0.54	0.50	1.18	.28	0.58
Road skills	1.42	1.01	2.00	.16	4.14
Attendance	0.26	0.29	0.75	.39	1.29

shown in Figure 2. As shown in Figure 2, the decision tree made its prediction of passing or failing the CDL examination based on the trainees' scores in the following curriculum areas: professional driver, operations skills, skill development, vehicle systems, basic operation, and road skills. These variables constituted the decision nodes of the tree. The decision tree depicted in Figure 2 identified four scenarios associated with trainees who failed the CDL examination on their first attempt: (a) their professional driving scores were at or below 92%; (b) their operational skill scores were at or below 87%; (c) their skill development scores were at or below 74%; or (d) their vehicle systems scores were between 97% and 86% and their basic truck operations and road skills examination scores were greater than 97% and 92%, respectively. If a trainee's performance did not fit any of these four scenarios, the decision



Decision Tree:

Pseudo Computer Code:

```
BEGIN PROGRAM
IF          Condition 1 – Professional Driver Score is less than or equal to 92
OR          Condition 2 – Operational Skill Score is less than or equal to 87
OR          Condition 3 – Skill Development Score is less than or equal to 74
THEN        Predict trainee will fail CDL examination
ELSE
IF          Condition 4 – Vehicle Systems Score is greater than 97
OR          Condition 5 – Basic Operation Score is less than or equal to 97
OR          Condition 6 – Vehicle Systems Score is less than or equal to 85
OR          Condition 7 – Road Skill Score is less than or equal to 92
THEN        Predict trainee will pass CDL examination
ELSE        Predict trainee will fail CDL examination
END PROGRAM
```

Figure 2. Decision tree for predicting passing or failing the CDL examination accompanied by its translation into pseudo computer code.

tree would predict that he or she would pass the CDL examination.

Genetic Programming

Across generations within each run, the genetic programming algorithm was successful in discovering computer programs that were increasingly accurate in classifying cases from the training data set, as Figure 3 illustrates. As shown in Figure 3, when the stages of selection, crossover, mutation, and replacement are iteratively applied from generation to generation, the percentage of cases correctly classified by the most accurate classifier in the population increases. Across the 10 runs, the computer program that was most accurate in classifying both the training and test data set cases made use of only 6 of the 10 input variables available: basic operation, driver technique, vehicle systems, operations skills, professional driver, and road skills. A depiction of the program's parse tree, and its corresponding translation into pseudo code, is shown in Figure 4.

Review of the parse tree and corresponding pseudo code in Figure 4 revealed that the computer program made use of two separate classification rules for predicting whether or not a trainee would pass the CDL examination on the first attempt. From the parse tree's root node (ITE), Branch 1 represents the conditions that the program used to decide which classification rule to employ. These two classification rules are represented by Branches 2 and 3 in the parse tree. Using Branches 1 and 2, the program predicted that trainees whose road skill score was greater than their professional driving score, *or* whose vehicle systems score was greater than their operations skills score, would pass the CDL examination if they achieved a road skill score greater than 93.5. It should be noted that the constant (93.5) utilized by the computer program and displayed in Figure 4 was not established a priori but was discovered as a consequence of the evolutionary search steps employed by the genetic programming algorithm.

For trainees whose road skill score was less than or equal to their professional driver score and whose vehicle systems score was less than or equal to their operations skill score, the program would use the classification rule described by Branch 3 in the parse tree to predict their performance on the CDL examination. Using this

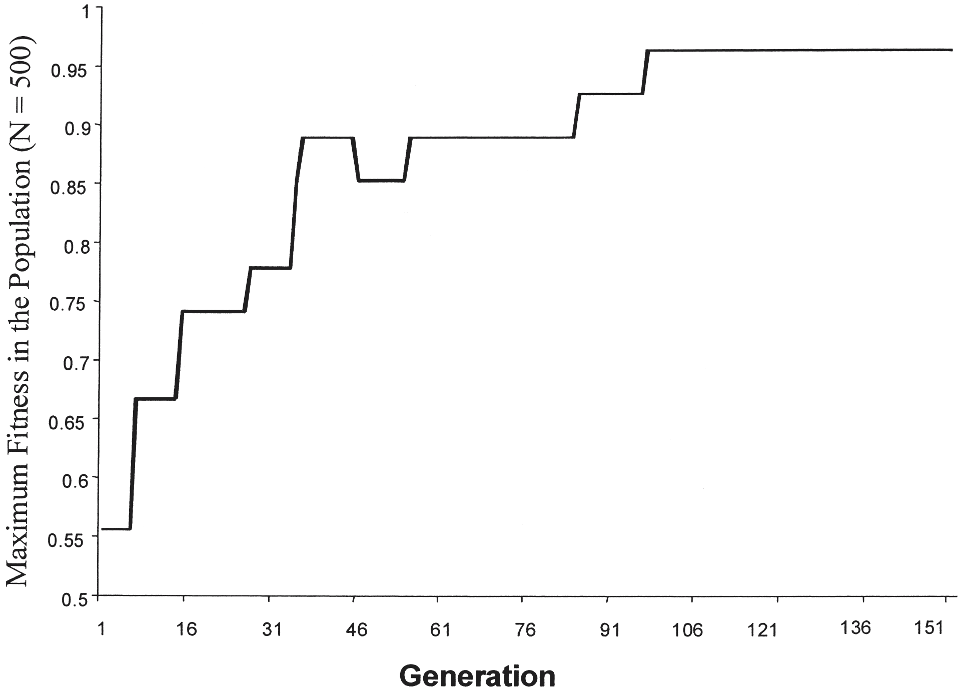
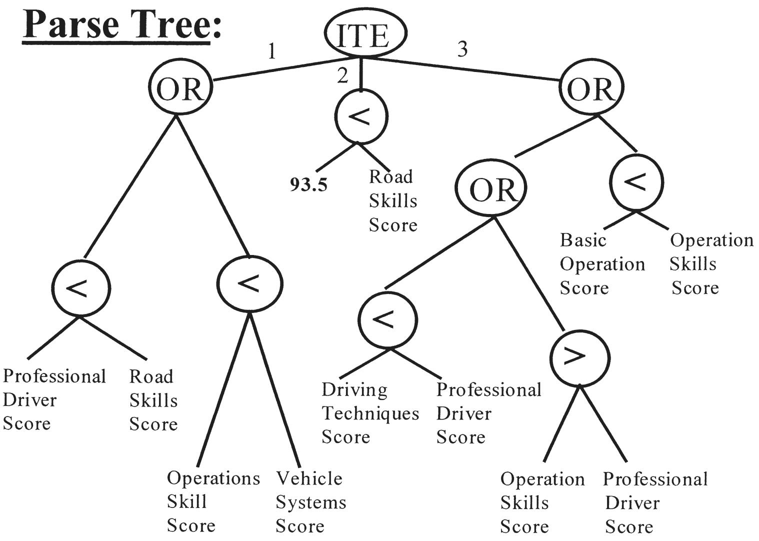


Figure 3. The maximum fitness among computer programs as a function of generation.



Pseudo Computer Code:

```
BEGIN PROGRAM
// Branch 1 in the parse tree
IF          Condition 1 - Road Skill Score is greater than Professional Driver Score
OR          Condition 2 - Vehicle Systems Score is greater than Operational Skill Score
THEN        // Use Branch 2 in the parse tree

IF          Condition 3 - Road Skill Score is greater than 93.5
THEN        Predict trainee will pass CDL examination
ELSE        Predict trainee will fail CDL examination

ELSE        // Use Branch 3 in the parse tree

IF          Condition 4 – Professional Driver Score is greater than Driving Techniques Score
OR          Condition 5 – Operations Skills Score is greater than Professional Driver Score
OR          Condition 6 – Operations Skill Score is greater than Basic Operation Score
THEN        Predict trainee will pass CDL examination
ELSE        Predict trainee will fail CDL examination
END PROGRAM
```

Figure 4. A parse tree that predicts success or failure on the CDL examination accompanied by its translation into pseudo computer code.

branch/rule, the program would predict that trainees would pass the CDL examination on the first attempt if their performance in the training curriculum met any of the following criteria: (a) their professional driver score was greater than their driving techniques score; (b) their operation skills score was greater than their professional driver score; (c) their operations skill score was greater than their basic operations score. If a trainee's curriculum performance met none of these three criteria, the branch/rule would predict that he or she would fail the CDL examination.

Comparisons between Statistical and Machine Learning Classification Models

All four models were compared based on their classification performance on the 10 cases within the test data set. The results of this comparison are summarized in Table 6. As shown in Table 6, all five models performed reasonably well in terms of classifying the cases of the

TABLE 6: Performance Comparison of Classification Models

Classification Performance Criteria	Discriminant Analysis	Logistic Regression Model		Genetic Program	C4.5 Algorithm
		1	2		
Training set accuracy	88.9	92.6	92.6	96.3	100
Test set accuracy	50	40	50	90	80
Test set sensitivity	60	60	80	100	80
Test set specificity	40	20	20	80	80
Test set validity	0	–20	0	80	60

Note. Criterion values given in percentages.

training data set. However, when called upon to generalize (i.e., predict) the outcome of cases that were not used in model development, the models developed using machine learning clearly outperformed those based on discriminant analysis and logistic regression. This advantage is especially evident in terms of specificity (i.e., the model's ability to identify trainees who failed the CDL examination on their first attempt). Comparisons between machine learning techniques revealed that the genetic programming and decision tree induction approaches were comparable, with only a single case separating their performances on both the training and test data sets.

The predictive capabilities of the classification models differed in terms of true positives, true negatives, false positives, and false negatives. These differences are illustrated in Figure 5. As shown in Figure 5, in terms of sensitivity (i.e., true positives), performance was weakest for discriminant analysis and for logistic regression Model 1. These models had the highest percentages of false negatives of all the classification models and were able to identify only three of the five test cases (60%) in which the student driver passed the CDL exam. The genetic program model possessed the strongest sensitivity: It correctly identified all five test cases in which the student passed. In terms of specificity

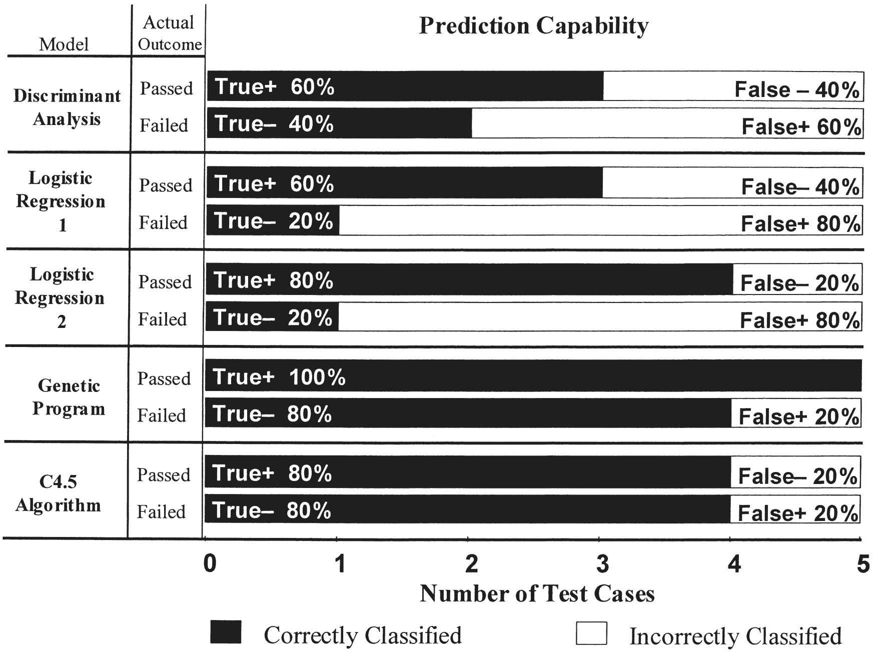


Figure 5. Predictive capability of classification models in terms of true positives (true+), true negatives (true–), false positives (false+), and false negatives (false–).

(i.e., true negatives), performance was weakest for both logistic regression models. These models had the highest percentages of false positives and were able to identify only one of the five test cases (20%) in which the student driver failed the CDL exam. Specificity performance was highest for the machine learning models (genetic program and C4.5 algorithm). These models were able to correctly classify four of the five test cases (80%) in which the student failed the CDL exam.

DISCUSSION

Results from the current study suggest that the machine learning approaches of the genetic programming algorithm and the C4.5 algorithm were more accurate at classifying driver performance on the CDL examination than were discriminant analysis and logistic regression. Overall, the traditional statistical models tended to make false positive identifications, predicting that those drivers who failed the CDL exam would pass. The classification models based on machine learning, however, did not demonstrate this error tendency. In this example, the machine learning methods employed were able to identify relationships between the independent and dependent variables that were consistent across different data sets, whereas discriminant analysis and logistic regression were not. Unlike discriminant analysis and logistic regression, the machine learning approaches were not constrained by predetermined models based on a linear summation of independent variables.

The decision tree, discovered by the C4.5 algorithm, relied on a linked series of fixed threshold values for the curriculum area scores when making its prediction of CDL examination outcome. The parse tree, discovered using genetic programming, made use of a threshold value and a series of relative comparisons between curriculum area scores in order to make its prediction. The decision and parse trees' corresponding computer programs made their decisions by assessing the trainees' curriculum area scores and, based on this assessment, determined which classification rules to use in the prediction of CDL examination outcome. This ability to switch between rules afforded the decision tree and genetic program a level of flexibility not possessed by the traditional statistical approaches used in this study. These findings, albeit limited in nature, support the notion that machine learning techniques, such as genetic programming and decision tree induction, could be used as a viable alternative to traditional statistical approaches in human-factors-related research.

Although successful in its scope, the current study has limitations that should be noted. First, the size of the data set used to develop and compare the classification models was limited to the performances of 36 men and only 1 woman. Given this limitation, it is unknown whether or not these models would be applicable to the driving performances of female truck drivers. Second, the outputs of all models developed and tested were binary (pass or fail CDL examination) in nature. These models did not provide insight into what aspects of the CDL exam a driver might fail, given his or her performance in the training curriculum. This specific information could prove to be valuable to instructors by allowing them to more effectively address the trainee's educational needs. Third, in their current form, these classification models can only identify trainees who may have considerable difficulty on their CDL examination. They do not suggest or prescribe specific interventions that an instructor can use to improve the trainees' performance on the examination.

Fourth, the performance comparison of these approaches was based on a single data source. One would expect classification performance differences between these approaches to vary depending on the data used for model development and validation. For example, Carnahan and Redfern (1998b) found that their classification model, discovered using genetic programming, outperformed logistic regression when predicting low-back injury risk for a set of occupational lifting tasks. Sohn and Shin (2001), however, found no difference in the accuracy of logistic regression and the C4.5 algorithm in classifying automotive accidents in terms of their severity outcomes. Further research is necessary to fully assess the potential advantages machine learning classifiers would have over their traditional statistical counterparts in the human factors area.

In order to address these limitations, future research will make use of larger training data sets and test data sets from more diverse populations. This will allow for statistical repeatability in training and testing various models. Future research will also focus on the expansion of the classification models beyond simple binary outcomes into multicategory classification. With this type of approach, one could attempt to predict what specific elements of the CDL examination would prove most difficult for certain trainee drivers. Finally, future research will focus on applying these (and other) machine learning techniques to other classification problems in the fields of safety, ergonomics, human factors, and human performance.

REFERENCES

Al-Ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis and Prevention, 34,* 729–741.

Avolio, B. J., Kroeck, K. G., & Panek, P. E. (1985). Individual differences in information-processing ability as a predictor of motor vehicle accidents. *Human Factors, 27,* 577–587.

Babski-Reeves, K. L., & Crumpton-Young, L. L. (2002). Comparisons of measures for quantifying repetition in predicting carpal tunnel syndrome. *International Journal of Industrial Ergonomics, 30,* 1–6.

Byrnes, M., & Fox, D. (Eds). (1999). *Bumper-to-bumper: The complete guide to tractor-trailer operations* (3rd ed.). Corpus Christi, TX: Byrnes & Associates.

Carnahan, B. J., Norman, B., & Redfern, M. S. (2001). Incorporating physical demand criteria into assembly line balancing. *IIE Transactions, 33,* 875–887.

Carnahan, B. J., & Redfern, M. S. (1998a). Application of genetic algorithms to the design of lifting tasks. *International Journal of Industrial Ergonomics, 21,* 145–158.

Carnahan, B. J., & Redfern, M. S. (1998b). Building a low back injury risk classifier using evolutionary computation. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 881–885). Santa Monica, CA: Human Factors and Ergonomics Society.

Carnahan, B. J., Redfern, M. S., & Norman, B. (2000). Designing safe job rotation schedules using optimization and heuristic search. *Ergonomics, 43,* 543–560.

Clancy, E. A. (1997). Factors influencing the resubstitution accuracy in multivariate classification analysis. *Ergonomics, 40,* 417–427.

Clarke, D. D., Forsyth, R., & Wright, R. (1998a). Behavioural factors in accidents at road junctions: The use of a genetic algorithm to extract descriptive rules from police case files. *Accident Analysis and Prevention, 30,* 223–234.

Clarke, D. D., Forsyth, R., & Wright, R. (1998b). Machine learning in road accident research: Decision trees describing road accidents during cross-flow turns. *Ergonomics, 41,* 1060–1079.

Deery, H. A., & Fildes, B. N. (1999). Young novice driver subtypes: Relationship to high-risk behavior, traffic accident record, and simulator driving performance. *Human Factors, 41,* 628–643.

French, D. J., West, R. J., Elander, J., & Wilding, J. M. (1993). Decision-making style, driving style, and self-reported involvement in road traffic accidents. *Ergonomics, 36,* 627–644.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning.* New York: Addison Wesley.

Hanson, J. P., Redfern, M. S., & Mazumdar, H. (1999). Predicting slips and falls considering required and available friction, *Ergonomics, 42,* 1619–1633.

Hennekens, C. H., & Buring, J. E. (1987). *Epidemiology in medicine.* Boston: Little, Brown.

Huberty, C. J., & Lowman, L. L. (1997). Discriminant analysis via statistical packages. *Educational and Psychological Measurement, 57,* 759–784.

Kishore, J. K., Patnaik, L. M., Mani, V., & Agrawal, V. K. (2000). Application of genetic programming for multicategory pattern classification. *IEEE Transactions on Evolutionary Computation, 4,* 242–258.

Koza, J. R. (1992). *Genetic programming: On the programming of computers by means of natural selection.* Cambridge, MA: MIT Press.

Koza, J. R. (1994). *Genetic programming II: Automatic discovery of reusable programs.* Cambridge, MA: MIT Press.

Koza, J. R., Bennett, F. H., Andre, D., & Keane, M. A. (1999). *Genetic programming III: Darwinian invention and problem solving.* San Francisco: Morgan Kaufmann.

Lubin, B., Fielder, E. R., & Van Whitlock, R. (1999). Predicting discharge from airforce basic training by battery of affect. *Journal of Clinical Psychology, 55,* 71–78.

Mackieh, A., & Cilingir, C. (1998). Effects of performance shaping factors on human error. *International Journal of Industrial Ergonomics, 22,* 285–292.

Marras, W. S., Lavender, S. A., Leurgans, S., Sudhakar, L. R., Allread, W. G., Fathallah, F., & Ferguson, S. (1993). The role of dynamic three dimensional trunk motion in occupationally related low back disorders. *Spine, 18,* 617–628.

Matias, A. C., Salvendy, G., & Kuczez, T. (1998). Predictive models of carpal tunnel syndrome causation among VDT operators. *Ergonomics, 41,* 213–226.

McFadden, M. (1997). Predicting pilot-error incidents of U.S. airline pilots using logistic regression. *Applied Ergonomics, 28,* 209–212.

Meadows, M. L., Stradling, S. G., & Lawson, S. (1998). The role of social deviance and violations in predicting road traffic accidents in a sample of young offenders. *British Journal of Psychology, 89,* 417–431.

Montana, D. J. (1995). Strongly typed genetic programming. *Evolutionary Computation, 3,* 199–230.

Myers, R. H. (1990). *Traditional and modern regression with applications* (2nd ed.). Belmont, CA: Duxbury.

Myers, R. S., Ball, K. K., Kalina, T. D., Roth, D. L., & Goode, K. T. (2000). Relation of useful field of view and other screening tests to on-road driving performance. *Perceptual and Motor Skills, 91,* 279–290.

Pham, D. T., & Onder, H. H. (1992). A knowledge-based system for optimizing workplace layouts using a genetic algorithm. *Ergonomics, 35,* 1479–1487.

Porter, B. E., & England, K. J. (2000). Predicting red-light running behavior: A traffic safety study in three urban settings. *Journal of Safety Research, 31,* 1–8.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1,* 81–106.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning.* San Mateo, CA: Morgan Kaufmann.

Rayson, M., Holliman, D., & Belyavin, A. (2000). Development of physical selection procedures for the British Army Phase 2: Relationship between physical performance test and criterion tasks. *Ergonomics, 43,* 73–105.

Reason, J., Manstead, A., Stradling, S., Baxter, J., & Campbell, K. (1990). Errors and violations on the roads: A real distinction? *Ergonomics, 33,* 1315–1332.

Sebald, A. V., & Fogel, L. J. (Eds.). (1994). *Proceedings of the 3rd Conference on Evolutionary Computation.* Princeton, NJ: World Scientific.

Simon, F., & Corbett, C. (1996). Road traffic offending, stress, age, and accident history among male and female drivers. *Ergonomics, 39,* 757–780.

Sohn, S., & Shin, S. (2001). Pattern recognition for road traffic accident severity in Korea. *Ergonomics, 44,* 107–117.

Vergara, M., & Page, A. (2002). Relationship between comfort and back posture and mobility in sitting posture. *Applied Ergonomics, 33,* 1–8.

Verwey, W. B., & Zaidel, D. M. (2000). Predicting drowsiness accidents from personal attributes, eye blinks, and ongoing behavior. *Personality and Individual Differences, 28,* 123–142.

Brian J. Carnahan is an assistant professor in the Department of Industrial and Systems Engineering at Auburn University. He received his Ph.D. in industrial engineering in 1999 at the University of Pittsburgh.

Gérard Meyer is president of Chez Gérard Consulting, Pittsburgh, Pennsylvania. He received his doctoral degree in ergonomic physiology in 1973 at the Conservatoire National des Arts et Métiers (Paris), France.

Lois-Ann Kuntz is an assistant professor in the Division of Education and Behavioral Sciences at the University of Maine at Machias. She received her Ph.D. in sensory processes and cognitive psychology in 1996 at the University of Florida.

Date received: October 31, 2001
Date accepted: February 24, 2003