Weight-elimination Neural Networks Applied to Coronary Surgery Mortality Prediction

Colleen M. Ennett, *Student Member, IEEE*, and Monique Frize, *Senior Member, IEEE*

Abstract— **The objective was to assess the effectiveness of the weight-elimination cost function in improving classification performance of artificial neural networks (ANNs), and to observe how changing the *a priori* distribution of the training set affects network performance. Back-propagation feed-forward ANNs with and without weight-elimination estimated mortality for coronary artery surgery patients. The ANNs were trained and tested on cases with 32 input variables describing the patient's medical history; the output variable was in-hospital mortality (mortality rates: training 3.7%, test 3.8%). Artificial training sets with mortality rates of 20, 50 and 80% were created to observe the impact of training with a higher-than-normal prevalence. When the results were averaged, weight-elimination networks achieved higher sensitivity rates than those without weight-elimination. Networks trained on higher-than-normal prevalence achieved higher sensitivity rates at the cost of lower specificity and correct classification. The weight-elimination cost function can improve the classification performance when the network is trained with a higher-than-normal prevalence. A network trained with a moderately high artificial mortality rate (artificial mortality rate of 20%) can improve the sensitivity of the model without significantly affecting other aspects of the model's performance. The ANN mortality model achieved comparable performance as additive and statistical models for coronary surgery mortality estimation in the literature.**

Index Terms— **Coronary artery bypass graft surgery, decision-making, neural networks, pattern classification.**

I. Introduction

Risk models are developed to help identify the factors that increase the likelihood of a particular outcome. The search for an effective method of mortality risk stratification for coronary artery surgery patients began in 1986 after the Health Care Financing Administration in the United States began releasing raw statistics on the mortality rate of Medicare coronary artery bypass grafting (CABG) patients in American hospitals. The stated objective was to inform patients about the quality of care at various hospitals and to help them make knowledgeable decisions to attain the best service and treatment possible [1]. This data, from the point of view of the hospitals and surgeons, did not consider the patient's severity of illness before undergoing surgery.

Categorizing the patients into different levels of risk provides a more accurate view of the quality of surgical care, and can potentially be used as a decision-aid to assess a patient's risk of mortality before surgery. The models are used to observe changes in the characteristics of the patient population over a period of years, effects of changes in surgical, pre- and postoperative procedures, and statistical variations from institution to institution. The heart surgery patient population is a particularly difficult group to classify as there are few defining characteristics that easily identify whether a patient will survive the surgery or not [2].

Cardiac surgery mortality models can easily make accurate estimations about low-risk patients, while the higher-risk patients are poorly stratified. Because the models are not yet sufficiently accurate, a patient should not be withheld treatment even if the mortality risk model suggests a high-risk outcome. These models are decision-aid tools, and the final decision should always be based on a clinician's expertise.

II. Background

There are several challenges associated with CABG patient databases. Since surgeons are generally successful at identifying patients who are unsuitable for surgery, CABG databases have low mortality rates (usually less than 5%). This means that little information exists about the patients who do not survive CABG surgery, making it difficult to develop a model that discriminates well between survivors and nonsurvivors. Current models have difficulty classifying high-risk patients, therefore this work was focused particularly on increasing the sensitivity of the model, that is, the correct classification of the nonsurvivors (those who died in-hospital following surgery).

Building on work of our research group [3]-[7], this paper

Manuscript received February 27, 2001; revised January 28, 2002. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada with Grant 202972-01.

C. M. Ennett is with the Systems and Computer Engineering Department, Carleton University, Ottawa, ON, K1S 5B6 Canada (e-mail: ennett@canada.com).

M. Frize is with the Systems and Computer Engineering Department, Carleton University, Ottawa, ON, K1S 5B6 Canada (phone: 613-520-2600 Ext. 8229; fax: 613-520-2109; e-mail: monique_frize@carleton.ca) and the School of Information Technology and Engineering, University of Ottawa, Ottawa, ON, K1N 6N5 Canada (phone: 613-562-5800 Ext. 6065; fax: 613-562-5175; e-mail: frize@site.uottawa.ca).

summarizes the effect of changing the *a priori* distribution of the training set on the performance of the artificial neural network (ANN), and presents findings on the classification performance of weight-elimination ANNs compared to their no weight-elimination counterparts. Artificially changing the *a priori* statistics of the datasets is an attempt to combat the problems that arise with the low representation of the nonsurvivors of CABG surgery when estimating in-hospital mortality using ANNs. Changing the distribution of the training dataset by entering the cases of the under-represented class multiple times into the artificial dataset can result in higher sensitivities for ANNs trained on datasets with higher *a priori* statistics. This indicates that "doping" a training set may aid in the classification of the nonsurvivors in the test set meaning that the model's sensitivity with the test set will be higher [7].

III. Methodology

Previous approaches to mortality risk stratification for CABG patients include additive models (such as Parsonnet [8] and the Cleveland Clinic [9]), statistical models (Bayesian models [10] and logistic regression [11]) and artificial neural networks (probabilistic neural networks based on Bayesian theory [2] and multilayer perceptrons [12]). Pliam et al. [13] completed a comparative analysis with the San Francisco Heart Institute database (also used for the current experiments) using the Parsonnet, Cleveland Clinic, Bayesian and logistic regression models. These will be used as a standard for comparison of our ANN's performance.

A. Network Architecture

The ANN used in the experiments carried out for this paper had the following characteristics: Back-propagation training algorithm, fully-connected feed-forward weight connections, hyperbolic tangent transfer function, three-layer architecture (input-hidden-output layers), and the weight-elimination cost function.

1) Weight-elimination cost function: One approach to solve the problem of overfitting is to add a complexity term to the cost function. Two cost functions have proven to reduce memorization: weight-decay and weight-elimination [14], [15]. Weight-decay limits the size of the connection weights, thereby penalizing large weights. The effect is a more stable network, because the output has less variance. Weight-decay is actually contained within the weight-elimination formula. Contrary to weight-decay, weight-elimination tries to reduce the small weights to zero (in other words, possibly eliminating the variables associated with these weights from consideration). This approach is well-suited for network pruning by eliminating variables that offer little or no assistance in estimating the correct outcome [3], [4], [14], [16]. The small weights only add unwanted "white noise" to the model. These cost functions work best when using a large initial network structure, relatively small initial weights, and a relatively small learning rate [14]. We followed these guidelines in our experiments.

By reducing the number of connection weights and hence the model's complexity using the weight-elimination penalty term, we expect to improve the network's classification performance. The weight-elimination cost function is shown in (1). The penalty term in weight-elimination (the second term) "counts the number of parameters, and minimizes the sum of performance error and the number of weights by backpropagation" [17].

$$E(W) = E_0(W) + \lambda \sum_{ij} \frac{\frac{w_{ij}^2}{w_0^2}}{1 + \frac{w_{ij}^2}{w_0^2}} \qquad (1)$$

E(W) is the combined cost function that includes the initial cost function, $E_0(W)$ (here, the sum of squared errors), and the weight-elimination term (the second term). Here, *W* represents the weight vector, λ is the weight-decay constant, and w_{ij} indicates the individual weight of the ANN.

The role of the weight-decay constant, λ, is to determine the relative importance of the weight-elimination term. Larger values of λ mean that a weight must be closer to zero to be considered a part of the "noise" distribution and increase the "pressure" on small weights to further reduce their size. Choosing a value of λ that is too small will not affect the network. When λ is too large, all weights are forced to zero [17]. The value of λ is generally chosen ad hoc. Several trials need to be run to observe how the network is responding. The value is increased if the weights are not diminishing or decreased if all the weights are forced to zero.

The scale parameter, w_0, defines the sizes of "large" and "small" weights. This scale parameter must be chosen by the user. When w_0 is small, the small weights will be forced to zero resulting in fewer large weights (i.e., weight-elimination). A large w_0 causes many small weights to remain, and limits the size of large weights (i.e., weight-decay) [17].

B. The Coronary Surgery Database

The San Francisco Heart Institute cardiac database has 7050 patients who underwent all types of open-heart surgery between January 1, 1985 and June 30, 1994 [13]. All of the variables available for this research are categorical except the patient's age and the date of surgery, which are continuous. The only surgical cases included in this analysis were those patients who underwent CABG surgery, CABG plus valve surgery, or CABG plus repair surgery [13]. The total number of cases in this reduced dataset was 6325. Out of those 6325 cases, there were 248 deaths giving an overall mortality rate of 3.9%.

To reduce the impact of changing patient profiles and clinical practice guidelines over time, only patients who underwent surgery between 1986 and 1991 were included in this analysis. There were only nine patient cases collected in 1985, so these cases were deleted because they were not representative of the CABG surgeries in 1985. As well, due to the technical advances introduced at the hospital in 1990-91, patients in 1992 and afterwards were also excluded (2790 cases). After removing cases with missing values, a more homogeneous database spanning the years of 1986-1991 with 3427 patient cases remained (mortality rate of 3.7%, or 127 deaths). The network had 32 input variables with information about the patient's medical history, and the output variable was in-hospital mortality [6], [7]. The input variables were chosen based on univariate analysis, and other variables that are commonly used in coronary surgery mortality risk models [13]. Table I presents the prevalence of each of the categorical or binary variables in the database. If the patient had the particular condition, the variable was scored as 1, if not it was scored as -1.

The application of the ANN algorithm with weight-elimination to a medical database has given successful results in a previous study [3]. For CABG surgery outcomes, the patients who do not survive the operation generally represent less than 5% of the database. This poses serious difficulties for the intended ANN to be investigated, as shown by preliminary results [3], [4]. A region of inconsistent network performance (i.e., sometimes the network would learn patterns and at other times it would classify everything as belonging to the group with the highest *a priori* probability) was discovered when the under-represented class made up between 8 and 15% of the cases [4]. At representations of less than 8%, the ANN failed to learn anything. Instead, it classified all patients as belonging to the larger outcome class. The technique of changing the *a priori* statistics of the training set provides a solution to some of the challenges faced when dealing with a drastically under-represented outcome in a two-class problem [4], [18].

C. *Creation of the Artificial Datasets*

There are two methods of artificially altering the patient database: randomly remove the survivor cases until the representation of the nonsurvivor cases is sufficiently high [18]-[20], or randomly copy the nonsurvivor cases until their representation in the database is sufficiently high. We chose the latter technique because there was no loss of potentially valuable information about the survivors that may enhance their correct classification, and we had not found articles detailing experiments using this approach. Given the limited number of cases in this database, in particular cases of nonsurvivors, it was preferable to increase the number of nonsurvivors rather than decrease the number of survivors in the database so that we maintained as large a database as possible.

TABLE I
VARIABLE DESCRIPTIONS

Variable Definition	Prevalence (%)
date of surgery	[a]
patient's age	[a]
female gender	24.7
mitral value disease operation	2.1
aortic value disease operation	3.1
emergent/urgent priority for surgery	22.9
failed percutaneous transluminal coronary angioplasty prior to surgery	4.7
reoperation	10.0
renal disease	6.2
ventricular aneurysm	0.3
left main disease	19.5
ejection fraction:	
normal	66.0
moderate	25.0
severe	9.0
mitral valve regurgitation	4.1
aortic value stenosis	2.5
tricuspid valve disease	0.1
hypertension	62.1
pre-operative intraaortic balloon pump	4.2
previous myocardial infarction	53.2
evolving myocardial infarction	1.3
history of congestive heart failure	10.2
unstable angina	62.9
cerebrovascular disease	11.9
peripheral vascular disease	12.1
triple vessel disease	74.0
obesity	10.7
small stature	1.8
chronic obstructive pulmonary disease	13.2
diabetes	22.8
cardiogenic shock	0.7
hypercholesterolemia	53.1
previous cerebrovascular accident	3.9
anemia	16.5

[a] These are continuous variables.
Date of surgery: range Jan 1/86 to Dec 31/91, mean Apr 22/89
Patient's age: range 22 to 91 years, mean 63.8 years

The objective was to artificially increase the number of nonsurvivors (and hence the mortality rate) in the training set. To achieve this goal, the datasets were separated according to their outcome: death or survival. This approach was necessary due to the small number of nonsurvivors, and to ensure that the desired distribution could be achieved [21]. First, the database was randomly separated into training and test sets (two-thirds and one-third, respectively), and then these sets were further subdivided into those who survived the surgery and those who did not. The artificial sets were formulated from these "categorized" datasets.

To artificially increase the percentage of nonsurvivors in the datasets, a simple program that performed random sampling with replacement was used. Since the sample size can affect the performance of a model, the total number of cases in the training and test sets was kept constant. Given the number of cases, and the desired percentage of nonsurvivors for the particular dataset, it was possible to determine how many patients who did not survive the surgery would be included in that particular set. Artificial training sets with mortality rates of 20, 50 and 80% were created. The nonsurvivor and survivor cases were chosen separately, but the patient records were chosen randomly with replacement. In addition to these artificial datasets, another 30 different artificial test sets with the true distribution were created. These additional test sets were used for the bootstrap approach to provide a number of datasets upon which to test the ANN. The results were averaged over the 31 test sets that were unknown to the network.

D. *Adjustable Network Parameters and Measures of Performance*

The ANN parameters that were optimized to the best of the ability of this network were the learning rate, momentum, weight-elimination scale factor, weight-decay constant, error ratio, number of hidden nodes, and output error weighting factor. Sensitivity was chosen as the key measure of performance for these experiments, because nonsurvivors (high-risk patients) are more difficult to classify than survivors (low-risk patients). The important criteria for identifying the best-performing ANN were: highest sensitivity, highest specificity, and highest correct classification rate. Although a constant predictor (a simple statistical tool that classifies all cases as belonging to the class with the highest *a priori* probability) would have a higher correct classification rate (accuracy), it was not considered as clinically useful as a model with a higher sensitivity. The area under the Receiver Operating Characteristic (ROC) curves was also recorded for the training and test sets.

Table II shows a confusion matrix to help visualize and interpret the results. The correct classification rate identifies the rate at which the model correctly classifies the data into their proper categories. This is calculated by summing the number of cases that were correctly classified into their respective classes (the number of true positives, TP, plus the number of true negatives, TN, and dividing by the total number of cases in the dataset (TP+TN)/total cases). The sensitivity of the model identifies the percentage of subjects who die following surgery and are correctly classified as dying. The formula for sensitivity is TP/(FN + TP) where FN represents the number of false negatives. Specificity is the percentage of patients who survive and are correctly classified as surviving, and can be calculated using the following formula: TN/(TN + FP) where FP is the number of false positives. Both the sensitivity and specificity are affected by the prevalence of the situation under investigation [22].

TABLE II
Confusion Matrix

		Correct Output	
		Not Present	Present
Model	Not Present	true negative (TN)	false negative (FN)
Output	Present	false positive (FP)	true positive (TP)

The area under the ROC curve assesses the ability of the model to discriminate between outcomes. Since this measure does not require a predefined decision threshold, it may also be used to discover the optimal cutpoint for the test [23]. It is a plot of the model's sensitivity versus one minus its specificity. The generated ROC curve is a visual description of the operating points and potential tradeoff between the true and false positive rates. This curve is obtained by varying the threshold value of the output node across its range of values [24]. Despite being a well-accepted measure of performance, the area under the ROC curve can be influenced by the prevalence of the outcome of interest [25].

ROC curve analysis only works for two output class problems since the generated curve is two-dimensional. A perfect model would be represented by an ROC curve that is a step function. This would indicate that all values of the true positive rate are equal to one (i.e., no classification error). The ROC curve of a redundant classifier is a 45-degree positive diagonal line, where the true positive rate equals the false positive rate, and offers no improvement over random guessing. Thus, the closer the ROC curve resembles a step function, the better the model is. Typical ROC values for coronary artery surgery models range from 0.72 to 0.76 [12], [26].

IV. Results

Tables III and IV summarize the parameter settings and the performance of the networks. Table III provides the parameter settings at which the best performance was achieved: number of layers, initial weights, number of hidden nodes, learning rate and its adjustable parameters, weight-decay constant, weight-elimination scale factor, momentum, error ratio, output error weighting factor, and the cut off value. Table IV provides a summary of the performance measures based on the test set performance: sensitivity, specificity, accuracy, and area under the ROC curve. The results are the average (mean) performance of the 31 different test sets for each experiment, and the standard deviation of those experiments. All test sets were the same size with the same prevalence.

A. *Different Training Set Distributions*

This analysis involves a comparison of the ANN performance on the true distribution test sets when training with the true distribution (i.e., a mortality rate similar to the test sets) or artificial training sets with 20, 50 and 80% mortality rates. The question here is: Does training with a higher-than-normal prevalence improve the ANN performance on the test sets with the true mortality rate?

The results reported in Table IV show that training with a higher-than-normal prevalence of the under-represented outcome noticeably improved the mean sensitivity of the ANN using the test data at the cost of poorer results for the

TABLE III
NETWORK SETTINGS

Experiment	Train MR = 3.7% Test MR = 3.8%	Train MR = 20% Test MR = 3.8%	Train MR = 50% Test MR = 3.8%
number of hidden layers	1	1	1

initial weights[a]	W1=rands()*1
	B1=rands()*1
	W2=rands()*0.1
	B2=rands()*0.1
		W1=rands()*1
		B1=rands()*1
		W2=rands()*0.1
		B2=rands()*0.1
			W1=rands()*1
			B1=rands()*1
			W2=rands()*0.1
			B2=rands()*0.1

hidden nodes	9	7	4
learning rate	0.001	0.001	0.0001
learning rate increment	1.003	1.003	1.003
learning rate decrement	1	1	1.001
weight-decay constant[b]	0.0003 from the beginning	0.0001 from the beginning	0.0008 initially, after 200 epochs 0.0009
weight-elimination scale factor	0.15	0.10	0.20
momentum	0.75	0.88	0.80
error ratio	1.02	1.02	1.03
output error weighting factor	1	1	1.1
Cutoff value	0	0	0

MR = mortality rate

[a] Matlab code where W1 refers to the weights connecting the input layer to the hidden layer and B1 is the bias weight of this layer. W2 refers to the weights connecting the hidden layer to the output layer and B2 is the bias weight. Rands()*1 and rands()*0.1 indicate that the initial weights are assigned random values which are multiplied by either 1 or 0.1 to scale the weights appropriately.

[b] Weigend *et al.* [14] recommended increasing the value of the weight-elimination constant after a period of time to improve the pruning of the weights. this approach was not always useful, hence the reason that it was not used during experiments with MR = 3.7% and MR = 20%.

other performance measures; we also observe that there is only a moderate reduction in mean specificity and accuracy when the training set has an artificial mortality rate of 20%[7]. However, when training with a 50% mortality rate, the specificity and accuracy drop off dramatically. The results from training with an 80% mortality rate are not shown in Table IV because in every model developed, there was 100% sensitivity and 0% specificity – not a useful classifier because there is no distinction between survivors and nonsurvivors.

B. Weight-elimination Technique

The weight-elimination networks with the true mortality rate and the 20% mortality rate were compared with their no weight-elimination counterparts that used only the sum of squared errors cost function, and the results of this analysis are presented in Table V. The goal of the weight-elimination cost function is to reduce the size of the connection weights to eliminate less useful input variables. Theoretically, this course of action of reducing the number of weights in the network, and hence the network's complexity, is expected to improve the network's classification performance [3], [5], [14]. Of course, the question here is: Does weight-elimination improve ANN classification performance with this particular database?

The results of Table V indicate that the use of the weight-elimination cost function did improve the ANN's classification performance of the nonsurvivors (the sensitivity) without a dramatic effect (if any) on the other performance parameters. Although the error bounds overlap when training and testing with the true mortality rates, the mean sensitivity for the weight-elimination ANNs was higher than that of the networks not using the weight-elimination cost function (with weight-elimination, 10.85 ± 4.38%, without weight-elimination 8.94 ± 4.54%). The improved sensitivity using the weight-elimination cost function is more pronounced when comparing the networks trained using the artificial test set with a 20% mortality rate and tested on the true mortality rate (with weight-elimination 43.55 ± 7.59%, without weight-elimination 24.05 ± 7.89%). Here, the error bounds do not overlap.

C. Comparison with Other Models in the Literature

In order to assess the ANN model's performance with respect to other models in the literature, we compared our results to those of Pliam et al. [13] whose database we used. One difference is that our model used a subset of the database to account for changes in clinical practice, whereas Pliam and his colleagues used the entire database of patients who underwent CABG surgery, CABG plus valve surgery and CABG plus repair surgery. There are no models for estimating in-hospital mortality following CABG surgery in the literature that were developed using neural networks. Table VI shows the results of the ANN models presented here and the results for the models based on the Parsonnet model, the Cleveland Clinic model, Bayesian theory and logistic regression as found by Pliam et al. [13].

The standard deviation of the results for the ANN and the statistical models overlap in most cases. This indicates that the differences between the models is not significant. Although the mean ROC values are lower for the neural networks, as mentioned previously, ROC values are still affected by the prevalence of the outcomes [25]. It was possible to attain slightly higher ROC values for the ANN

TABLE IV
EXPERIMENTAL TEST RESULTS WITH WEIGHT-ELIMINATION

Experiment	Train MR=3.7% Test MR=3.8%	Train MR=20% Test MR=3.8%	Train MR = 50% Test MR=3.8%
Sensitivity (%)	10.9 ± 4.4	43.6 ± 7.6	71.7 ± 7.3
Specificity (%)	98.3 ± 0.4	93.8 ± 0.8	69.1 ± 1.2
Accuracy (%)	95.0 ± 0.4	91.9 ± 0.8	69.2 ± 1.0
Area under ROC (test)	0.57 ± 0.03	0.72 ± 0.04	0.75 ± 0.04
Area under ROC (training)	0.63	0.90	0.87

MR = mortality rate
ROC = receiver operating characteristic curve

TABLE V
EVALUATION OF THE WEIGHT-ELIMINATION COST FUNCTION

Experiment	Train MR = 3.7% Test MR = 3.8%		Train MR = 20% Test MR = 3.8%	
	WE	No WE	WE	No WE
Sensitivity (%)	10.9 ± 4.4	8.9 ± 4.5	43.6 ± 7.6	24.1 ± 7.9
Specificity (%)	98.3 ± 0.4	99.3 ± 0.2	93.8 ± 0.8	93.2 ± 0.7
Accuracy (%)	95.0 ± 0.4	95.9 ± 0.3	91.9 ± 0.8	90.6 ± 0.7
Area under ROC (test)	0.57 ± 0.03	0.55 ± 0.03	0.72 ± 0.04	0.69 ± 0.04
Area under ROC (training)	0.63	0.66	0.90	0.93

MR = mortality rate
ROC = receiving operating characteristic curve
WE = weight-elimination

TABLE VI
COMPARISON OF ANN PERFORMANCE WITH OTHER MODELS FROM THE LITERATURE

Researchers	Risk models	Area under ROC curve
Models presented in this article	Train MR = 3.7% Test MR = 3.8%	0.57 ± 0.03
	Train MR = 20% Test MR = 3.8%	0.72 ± 0.04
	Train MR = 50% Test MR = 3.8%	0.75 ± 0.04
Pliam *et al.* [13]	Parsonnet	0.80 ± 0.02
	Cleveland Clinic	0.80 ± 0.02
	SFHI Bayesian	0.83 ± 0.02
	SFHI logistic regression	0.80 ± 0.02

MR = mortality rate
ROC = receiver operating characteristic curve

(near 0.80), however, the classification performance suffered and the network learning was erratic. Based on the learning pattern, classification performance showed progressive learning, high sensitivity and high specificity for the test set, but reported a slightly lower ROC value.

V. INTERPRETATION OF THE RESULTS

An important point to consider is the cost of misclassification. For example, predicting that a patient will not survive surgery (but the patient actually lives) has a different associated cost than foretelling survival, when in actual fact the patient dies. When there are different misclassification costs, accuracy is not necessarily the best measure of model performance [27]. With the CABG data, accuracy was reported knowing the limitations of its interpretation. A satisfactory balance between the sensitivity, specificity and accuracy would be ideal.

One of the limitations of this study is the decreased specificity that occurred when the mortality rate was increased, and sometimes when weight-elimination was used. Certainly, the best model would have high sensitivity and high specificity, whereas the models presented here permitted a lower specificity at the cost of a higher sensitivity. To deal with this dilemma, we need to develop a new criterion for evaluating the classification performance of models. This new criterion should attempt to maximize sensitivity while maintaining high specificity. The resulting models would have high correct classification rates, and presumably larger areas under the ROC curves indicating overall good performance.

The ANN trained with a moderately higher prevalence (mortality rate of 20%) is of more clinical value than the model trained with a 50% mortality rate because more of the nonsurvivors are correctly classified without a significant impact on the classification of survivors.

VI. CONCLUSIONS

Training a feed-forward back-propagation ANN with weight-elimination with a moderately higher-than-normal prevalence of the under-represented outcome (artificial mortality rate of 20%) can improve the sensitivity of an ANN without dramatically affecting other aspects of the network's performance. When the artificial mortality rate of the training set is increased to 50%, the increase in sensitivity of the model does not outweigh the increased misclassification cost of survivors (i.e. decreased specificity). As well, a training set mortality rate which is too high (high mortality rate and low survival rate) skews the data towards estimating death and makes accurate estimations difficult. This is similar to the problem when the data are highly skewed towards survival (low mortality rate and high survival rate).

When the results are averaged, weight-elimination ANNs achieved higher sensitivity rates than ANNs using only the sum of squared errors cost function. These results show that the weight-elimination cost function can improve the correct classification of nonsurvivors in this CABG patient database, since more nonsurvivors were correctly classified when using the weight-elimination cost function.

Compared to additive and statistical models in the literature, our neural network achieved similar classification performance results. In the case of all prediction models, the prevalence of the outcome influences the classification ability of the model. When a database is highly skewed toward one outcome, the predictive ability of the model may be weakened

[4]-[7], [28].

A. *Future Work*

Our research group is working toward developing a new criterion for assessing network performance. This new criterion will attempt to optimize both the sensitivity and the specificity of a model to ensure that the best results are achieved.

B. *Recommendations*

Hospitals should begin to include in their CABG database patients who are refused surgery [6]. If this data were available, it would be possible to compare the patients who are refused CABG surgery with those who undergo the surgery and die. It may be possible to use this information to improve identifying characteristics that put a CABG surgery patient at a higher risk of death.

Acknowledgment

We would like to thank Dr. Richard E. Shaw for use of the San Francisco Heart Institute database.

References

[1] P. A. Ebert, "Keynote address from American College of Surgeons, Chicago, IL," *Circulation*, vol. 79(Suppl 1), p. I2, 1989.

[2] R. K. Orr, "Use of a probabilistic neural network to estimate the risk of mortality after cardiac surgery," *Med. Decis. Making,* vol. 17(2), pp. 178-185, Apr.-Jun. 1997.

[3] M. Frize, H. C. E. Trigg, F. G. Solven, M. Stevenson, B. G. Nickerson, "Decision-support systems designed for critical care," in *Proc. AMIA. Symp.,* 1997, p. 855.

[4] C. M. Ennett and M. Frize, "An investigation into the strengths and limitations of artificial neural networks: an application to an adult ICU patient database," in *Proc. AMIA Symp*, 1998, p. 998.

[5] M. Frize, L. Wang, C. M. Ennett, B. Nickerson, F. G. Solven, M. Stevenson, "New advances and validation of knowledge management tools for critical care using classifier techniques," in *Proc. AMIA. Symp.,* 1998, pp. 553-558.

[6] C. M. Ennett, M. Frize and R. E. Shaw, "Methodologies for predicting coronary surgery outcomes," in *Proc. IEEE EMBS-BMES,* 1999.

[7] C. M. Ennett and M. Frize, "Selective sampling to overcome skewed *a priori* probabilities with neural networks," in *Proc. AMIA. Symp.,* 2000, pp. 225-229.

[8] V. Parsonnet, D. Dean and A. D. Bernstein, "A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease," *Circulation*, vol. 79(Suppl 1), pp. I3-I12, Jun. 1989.

[9] T. L. Higgins, F. G. Estafanous, F. D. Loop, G. J. Beck, J. M. Blum and L. Paranandi, "Stratification of morbidity and mortality outcome by preoperative risk factors in coronary artery bypass patients. A clinical severity score," *JAMA*, vol. 267, pp. 2344-2348, May 1992.

[10] F. H. Edwards and G. M. Graeber, "The theorem of Bayes as a clinical research tool," *Surg. Gyn. Obstet.,* vol. 165, pp. 127-129, 1987.

[11] D. F. Kleinbaum, L. L. Kupper, K. E. Muller and A. Nizam, *Applied Regression Analysis and Other Multivariable Methods,* Pacific Grove: Duxbury Press, 1998, p. 656.

[12] R. P. Lippmann and D. M. Shahian, "Coronary artery bypass risk prediction using neural networks," *Ann. Thorac. Surg.,* vol. 63, pp. 1635-1643, Jun. 1997.

[13] M. B. Pliam, R. E. Shaw and Z. Zapolanski, "Comparative analysis of coronary surgery risk stratification models," *J. Invas. Cardiol.,* vol. 9, pp. 203-222, 1997.

[14] A. S. Weigend, D. E. Rumelhart and B. A. Huberman, "Back-propagation, weight-elimination and time series prediction," in *Proc. 1990 Connectionist Models Summer School*, San Mateo: Morgan Kaufmann, pp. 105-116, 1990.

[15] A. Krogh and J.A. Hertz, "A simple weight decay can improve generalization," in Lippmann RP, Moody J, Touretzky, eds. *Advances in Neural Information Processing Systems 4 (NIPS'91),* San Matteo: Morgan Kaufmann, pp. 950-957, 1992.

[16] M. Frize, C. M. Ennett, M. Stevenson and H. C. E. Trigg, "Clinical decision-support systems for intensive care units using artificial neural networks," *Med. Eng. Physics,* vol. 23(3), pp. 217-225, Apr. 2001.

[17] A. S. Weigend, D. E. Rumelhart and B. A. Huberman, "Generalization by weight-elimination with application to forecasting," in R. P. Lippmann, J. Moody and D. S. Touretzky, eds, *Advances in Neural Information Processing Systems (NIPS*90)*, San Mateo: Morgan Kaufmann, vol. 3, pp. 875-882, 1991.

[18] L. Ohno-Machado, H. S. Fraser and A. Ohrn, "Improving machine learning performance by removing redundant cases in medical data sets," in *Proc. AMIA. Symp*., 1998, pp. 523-527.

[19] W. G. Baxt and J. Skora, "Prospective validation of artificial neural networks trained to identify acute myocardial infarction," *Lancet*, vol. 347, pp. 12-15, 1996.

[20] D. L. Hudson and M. E. Cohen, *Neural Networks and Artificial Intelligence for Biomedical Engineering*, San Francisco: IEEE Press and John Wiley & Sons, 2000, p. 336.

[21] S. Katz, A. S. Katz, N. Lowe and R. C. Quijano, "Neural net-bootstrap hybrid methods for prediction of complications in patients implanted with artificial heart valves," *J. Heart Valve Dis.,* vol. 3(1), pp. 49-52, Jan. 1994.

[22] W. Penny and D. Frost, "Neural networks in clinical medicine," *Med. Decis. Making*, vol. 16, pp. 386-398, 1996.

[23] J. J. Forsstrom and K. J. Dalton, "Artificial neural networks for decision support in clinical medicine," *Annals of Medicine*, vol. 27, pp. 509-517, 1995.

[24] K. Woods and K. W. Bowyer, "Generating ROC curves for artificial neural networks," *IEEE Trans. on Medical Imaging*, vol. 16(3), pp. 329-337, 1997.

[25] T. G. Buchman, K. L. Kubos, A. J. Seidler and M. J. Siegforth, "A comparison of statistical and connectionist models for the prediction of chronicity in a surgical intensive care unit," *Crit. Care Med.*, vol. 22(5), pp. 750-762, 1994.

[26] R. E. Clark, "Calculating risk and outcome: The Society of Thoracic Surgeons database," *Ann. Thorac. Surg*., vol. 62, pp. S2-5, 1996.

[27] M. W. Kattan and J. R. Beck, "Artificial neural networks for medical classification decisions," *Arch. Pathol. Lab. Med.*, vol. 119, pp. 672-677, 1995.

[28] J. S. Turner, C. J. Morgan, B. Thakrar and J. R. Pepper, "Difficulties in predicting outcome in cardiac surgery patients," *Crit. Care Med.*, vol. 23, pp. 1843-1850, 1995.

Colleen M. Ennett (S'96) was born in Hanover, ON, Canada, in 1974. She received the B.Sc.(Eng.) degree in biological engineering from the University of Guelph, Guelph, ON, Canada, in 1997, and the M.A.Sc. degree in electrical engineering from the University of Ottawa, Ottawa, ON, Canada, in 1999. She is currently working toward the Ph.D degree in electrical engineering at Carleton University, Ottawa, ON, Canada. Her Ph.D. dissertation relates to her novel approach for replacing missing values in a medical database with a hybrid pattern recognition system by combining artificial neural networks with case-based reasoning.

She has published 21 papers in refereed journals and conference proceedings about her work on artificial neural networks and medical databases. Her current research interests include pattern recognition and analysis techniques using digital signal processing and artificial intelligence tools with decision-making and modelling capabilities in the medical engineering domain.

Monique Frize (M'75-SM'93) received the B.A.Sc. degree in electrical engineering from the University of Ottawa, Ottawa, ON, Canada, in 1966, the M.Phil. degree from Imperial College of Science and Technology, London, U.K., in 1970, the M.B.A. degree from the Université de Moncton, Moncton, NB, Canada, in 1986, and the Doctorate degree from Erasmus Universiteit, Rotterdam, The Netherland, in 1989.

She has published more than 100 papers in refereed journals and conference proceedings in the field of biomedical engineering and medical informatics.

Ennett CM, Frize M. Weight-elimination neural networks applied to coronary surgery mortality prediction. *IEEE Trans Info Technol Biomed* (Accepted Oct 9/02).

Dr. Frize has received four honorary degrees since 1992, was inducted as a Fellow of the Canadian Academy of Engineering in 1992 and as an Officer of the Order of Canada in 1993.