Journal of Biomedical Informatics **34,** 28–36 (2001)
doi:10.1006/jbin.2001.1004, available online at http://www.idealibrary.com on IDEAL®

# A Comparison of Machine Learning Methods for the Diagnosis of Pigmented Skin Lesions

Stephan Dreiseitl,*,† Lucila Ohno-Machado,* Harald Kittler,‡ Staal Vinterbo,* Holger Billhardt,* and Michael Binder*,‡

*Decision Systems Group, Brigham and Women's Hospital, Division of Health Sciences and Technology, Harvard Medical School, Massachusetts Institute of Technology, Boston, Massachusetts; †Department of Software Engineering for Medicine, Polytechnic University of Upper Austria, A-4232 Hagenberg, Austria; and ‡Department of Dermatology, University of Vienna Medical School, Vienna, Austria*

*Received January 2, 2001; published online March 13, 2001*

We analyze the discriminatory power of *k*-nearest neighbors, logistic regression, artificial neural networks (ANNs), decision tress, and support vector machines (SVMs) on the task of classifying pigmented skin lesions as common nevi, dysplastic nevi, or melanoma. Three different classification tasks were used as benchmarks: the dichotomous problem of distinguishing common nevi from dysplastic nevi and melanoma, the dichotomous problem of distinguishing melanoma from common and dysplastic nevi, and the trichotomous problem of correctly distinguishing all three classes. Using ROC analysis to measure the discriminatory power of the methods shows that excellent results for specific classification problems in the domain of pigmented skin lesions can be achieved with machine-learning methods. On both dichotomous and trichotomous tasks, logistic regression, ANNs, and SVMs performed on about the same level, with *k*-nearest neighbors and decision trees performing worse. © 2001 Academic Press

*Key Words:* machine learning; decision support; image classification; neural networks; support vector machines.

## 1. INTRODUCTION

The increasing number of electronic data bases containing medical data has led to an increasing interest in their utilization for building classification models that can "learn" from examples. A variety of statistical and machine learning approaches to classification tasks are currently available, but few comparisons among different models have been done on the same data sets. Although the potential advantages and disadvantages of utilizing each of these methods have been defined theoretically, given certain assumptions about data distribution, characteristics of the classification task, signal-to-noise ratio, etc., it is often the case that these assumptions cannot be verified in practice. Under these circumstances, empirical comparison of classification performance using standard metrics to describe discrimination and calibration is necessary. Utilities associated with misclassification in any direction (e.g., false positives or false negatives) can be built into the models, or treated separately. A final selection of the "best model" for a given classification task can only be concluded after considering the tradeoffs between classification performance, costs, and model interpretability.

Since the first step toward the selection of a class of models for a particular data set is based on classification performance, this area was the focus of our investigations. We compare the discriminatory performance of five methods (*k*-nearest neighbors, logistic regression, artificial neural networks, decision trees, and support vector machines) on the

1532-0464/01 $35.00

task of classifying pigmented skin lesions (PSLs) as being common nevi, dysplastic nevi, or melanoma. The same data set was used for all models; it was split differently 100 times into training and test sets to eliminate the effect of having particularly good or bad combinations of cases in training and test sets. All algorithms were run on all 100 splits. The task of classifying PSLs is complex and involves automated feature measurements obtained from digital images, as well as clinical and demographical data collected by dermatologists.

The motivation for using PSL data as testbed for the classification algorithms is the fact that incidence of melanoma has risen dramatically in recent years. Therefore, it is increasingly important to accurately diagnose PSLs. This classification task is difficult, as can be seen from the fact that the diagnostic performance of even expert dermatologists is far from optimal, with accuracy of diagnosing early melanoma reported to be only slightly higher than 60% [1].

Epiluminescence microscopy was developed as a tool to aid in the diagnostic process, and expert performance increases when using this method [2]. The availability of digital PSL images raises the question as to whether machine-learning methods can perform better than human experts. An answer to this question would help to determine whether an increase in performance could be gained from automated decision-support tools.

In this paper, we study the performance of five machine-learning methods on a data set of 1619 PSLs. Section 2 gives details about the data set and briefly outlines the characteristics of the different learning algorithms. The results of the experiments are presented in Section 3. A discussion follows in Section 4; concluding remarks are given in Section 5.

## 2. MATERIAL AND METHODS

The data set used for the experiments in this paper was collected at the pigmented lesion unit of the Department of Dermatology, Division of General Dermatology, University of Vienna Medical School. A total of 1619 PSL images in three classes (common nevi, dysplastic nevi, melanoma) were selected. The distribution of cases in the data set was 1290 common nevi, 224 dysplastic nevi, and 105 melanoma. The diagnosis of the lesions was established by histopathology (melanoma and dysplastic nevi) and 1-year follow-up examinations (common nevi), respectively. In addition to the image, six clinical data items were recorded for each lesion.

The images of the lesions were obtained by using a digital epiluminescence microscopy system (MoleMax II, Derma-Instruments, Austria). Images of lesions were taken by a hand-held video microscopy unit with a CCD color sensor at a resolution of 752 × 582 pixels and stored in 24-bit resolution. From these images, 107 morphometric features were extracted using an adaptation of gray level tresholding to three-dimensional color space (hue, saturation, value). The algorithm for feature extraction was developed at the Department of Computer Graphics and Vision, Technical University Graz. The morphometric features were categorized in global and local features. Global features contained basic features, shape features, color features, normalized color features, quantized color features, and border features. Local features contained segment features, quantized color features, and ratios of features.

The data set was split into training and test sets for the machine learning algorithms. The training sets contained 600 common nevi, 144 dysplastic nevi, and 65 melanoma, and the test sets contained 690 common nevi, 80 dysplastic nevi, and 40 melanoma. To determine the influence of different data set splits on the methods, a total of 100 different splits were used in each method. The data set splits were the same for each of the methods used. To ensure that no method was given an advantage due to the scaling of the data set, each variable was transformed to be zero-mean and unit variance over the whole data set.

### *k-Nearest Neighbors*

The $k$-nearest-neighbors algorithm [3] is a popular density estimation algorithm for numerical data. In contrast to the other methods, this algorithm does not implement a decision boundary, but uses the elements of the training set to estimate the density distribution of the data. They implicitly combine this information with class prevalences in Bayes' rule to obtain the posterior (class membership) probability estimates of a data point. The density estimation uses a distance measure (usually Euclidean distance). For a given distance measure, the only parameter of the algorithm is $k$, the number of neighbors. The parameter $k$ determines the smoothness of the density estimation: larger values consider more neighbors, and therefore smooth over local characteristics. Smaller values consider only limited neighborhoods. Generally, the choice of $k$ can only be determined empirically. In our experiments, we used values of $k = 10, 20, \ldots, 100$.

In medicine, most applications use nearest-neighbor algorithms as benchmarks for other machine-learning techniques [4–6].

### *Logistic Regression*

Logistic regression is an algorithm that constructs a separating hyperplane between two data sets, using the logistic function to express distance from the hyperplane as a probability of class membership.

Logistic regression is widely used in medical applications for the ease with which the parameters in the model can be interpreted as changes in log odds, for the variable selection methods that are often available in commercial implementations, and for allowing the interpretation of results as probabilities. Although the model is linear-in-parameters and can thus only calculate linear decision boundaries, it is a widely used predictive model in medical applications [7–9].

In our experiments, we used the SAS system (SAS Institute, Cary, NC) to derive logistic regression models. The significance level for entry and removal of a variable in the model was set to 0.05, and only the eight most significant variables were included in the model.

### *Artificial Neural Networks*

Artificial neural networks (ANNs) represent a means to calculate posterior class membership probabilities by minimizing a cross-entropy error function [10].

The ANN consists of several small processing units (the artificial neurons) that are highly interconnected. Information flow in an ANN is modeled after the human brain, in that information is propagated between neurons, with the information stored as connection strengths (called weights) between neurons. The minimization process is implemented as an update rule for the weights in the network. Since this iterative process requires many presentations of the training set, the system is said to learn from examples.

For medical applications, a major drawback of ANNs is the fact that the parameters in the model are not directly interpretable, so that no additional understanding of a data set can be derived from a neural network model. Nevertheless, the ability to calculate nonlinear decision boundaries makes them attractive in several medical problem domains [11, 12].

For our experiments, we used the NETLAB (Neural Computing Research Group, Aston University, UK) implementation of conjugate gradient optimization, with all parameters set to default values.

### *Decision Trees*

The decision tree paradigm constructs classifiers by dividing the data set into smaller and more uniform groups, based on a measure of disparity (usually entropy). It does this by identifying a variable and a threshold in the domain of this variable that can be used to divide the data set into two groups. The best choice of variable and threshold is the one that minimizes the disparity measures in the resulting groups. The advantage of decision trees over many of the other methods used here is that small decision trees can be interpreted by humans as decision rules. They therefore offer a way to extract decision rules from a database. This makes them especially well suited for medical applications, and advantages and disadvantages of decision trees in medicine have been widely investigated [13–15].

We used the See5 decision tree software by Rulequest Research (St. Ives, New South Wales, Australia) for our simulations, with the parameters set to default values.

### *Support Vector Machines*

Support vector machines (SVMs) are a machine learning paradigm based on statistical learning theory [16, 17]. Although the theory of support vector machines was developed more than 20 years ago, this paradigm has only recently been widely applied by the machine learning community. Few applications of this method in the medical domain have been reported so far [18]. The most attractive feature of this paradigm is that it is possible to give bounds on the generalization error of the model, and to select the best model from a class using the principle of *structural risk minimization* [16].

Support vector machines calculate separating hyperplanes that maximize the margin between two sets of data points. By using Lagrange multipliers, the problem can be formulated in such a way that the only operations on the data points are the calculation of scalar products. While the basic training algorithm can only construct linear separators, *kernel functions* can be used to calculate scalar products in higher-dimensional spaces. If the kernel functions are nonlinear, the separating boundary in the original space will be nonlinear. Because there are many different kernel functions, there is a wide variety of possible SVM models.

In this work, we used the SVM-Light implementation (Department of Computer Science, University of Dortmund, Germany) with polynomial kernels of degrees 1–3 and Gaussian radial basis function kernels with $\gamma$ (inverse variance) parameters between $10^{-2}$ and $10^{-6}$.

## 3. EXPERIMENTS

Each of the five algorithms presented above was run on each of the 100 different splits of the data set into training and test data. Since the classification problem is trichotomous, there are several possibilities to report the results of the different algorithms. Three of the five methods (*k*-nearest neighbors, neural networks, logistic regression) give results that can easily be interpreted as probabilities; we therefore analyzed the discriminatory power of these methods over all three classes by means of three-way ROC analysis [19].

Three-way ROC analysis is an extension of ROC analysis to trichotomous tests. It summarizes the discriminatory power of a trichotomous test in a single value, called the *volume under surface* (VUS) in analogy to the area under curve (AUC) value for dichotomous tests. Just as the AUC value for dichotomous tests is equivalent to the probability of correctly ranking a given pair of normal and abnormal cases, the VUS value for trichotomous tests is equivalent to the probability of correctly distinguishing three cases, where each case is from a different class. A trichotomous test that discriminates perfectly has a VUS value of 1, whereas an uninformative test has a value of only 1/6. This means that VUS values which correspond to poor AUC values are often quite good. For example, a VUS value of 0.6–0.7 corresponds to an AUC value of 0.83–0.88 [20]. Since it is possible to visualize the general shape of a three-way ROC surface, but not possible to easily obtain meaningful information from its two-dimensional image on paper, we do not show three-way ROC surfaces here. The important aspect of these surfaces is that they measure the discriminatory power of a classification method. The discriminatory power is summarized by numerical VUS value, which is given in Table 1 for the methods considered here.

Of the other two methods, decision trees can be used for multiclass discrimination; however, the See5 software does not support the calculation of probabilities for all classes in the model. We therefore reduced the problem to two dichotomous classification tasks: First, to discriminate common nevi from the other two lesion types (dysplastic nevi and melanoma), and second, to discriminate melanoma from common and dysplastic nevi. Standard ROC analysis [21–23] was used to summarize the results of both these classification tasks. For support vector machines, there are extensions for multiclass discrimination [24], but the basic algorithm is strictly dichotomous. Although the outputs of support vector machines cannot be interpreted as probabilities, it is still possible to calculate AUC values and thus compare results by using the equivalence of the AUC measure and the *c*-index [25, 21]. The classification tasks for this method were the same as for the decision trees.

TABLE 1

Performance Comparison of *k*-Nearest Neighbors, Logistic Regression, and Artificial Neural Networks on the Trichotomous Problem of Classifying PSLs as Common Nevi, Dysplastic Nevi, or Melanoma

| | *k*-nearest neighbors | | | | Log | |
|---|---|---|---|---|---|---|
| | $k = 10$ | $k = 40$ | $k = 70$ | $k = 100$ | regression | ANN |
| Avg VUS | 0.5858 | 0.6042 | 0.5788 | 0.5804 | 0.6708 | 0.6821 |
| Std dev | 0.0357 | 0.0434 | 0.0416 | 0.0432 | 0.0373 | 0.0347 |
| Min VUS | 0.4773 | 0.4972 | 0.4833 | 0.4779 | 0.5544 | 0.5968 |
| Max VUS | 0.6668 | 0.7156 | 0.6929 | 0.6968 | 0.7414 | 0.7698 |

*Note*. VUS denotes the volume under the ROC surface.

As dichotomous tasks to compare all five methods, we used the problem of distinguishing common nevi from dysplastic nevi and melanoma, and the problem of distinguishing melanoma from common and dysplastic nevi. For these tasks, we already had the results from decision trees and support vector machines. To obtain results from *k*-nearest neighbors, logistic regression, and artificial neural networks, it was sufficient to suitably combine the probabilities (outputs) of the trichotomous tasks to arrive at dichotomous probabilities.

A summary of the results of the three methods for which three-way ROC analysis can be used is given in Table 1. The results for all five methods on the task of discriminating common nevi from dysplastic nevi and melanoma are shown in Table 2; the results on the task of discriminating melanoma from common and dyplastic nevi are shown in Table 3. The entries in Table 1 show the discriminatory power of the methods, as measured by the average of the volume under the ROC surface (VUS) over the 100 test sets. Furthermore, Table 1 gives the standard deviations of VUS values, and minimum and maximum VUS values.

The results of the two dichotomous tasks, which were used by all five methods, are given in Tables 2 and 3. The entries are the following, for each method and task: Average AUC over 100 tests sets, standard deviation of AUC, maximum and minimum AUC value, as well as average maximum sensitivity and specificity, as measured at the optimal threshold on the ROC curve (closest to upper-left corner).

ROC curves for the best and worst methods (support vector machines and decision trees, respectively) on both dichotomous tasks are shown in Figs. 1 and 2. The curves were obtained by averaging the sensitivities and specificities of all 100 data set splits at 200 thresholds for each split. We show only these curves, since the results of the other methods lie between those of decision trees and support vector machines. In particular, since the results of logistic regression and neural networks are almost the same as those of support

TABLE 2

Performance Comparison of *k*-Nearest Neighbors, Logistic Regression, Artificial Neural Networks, Decision Trees, and Support Vector Machines on the Task of Distinguishing Common Nevi from Dysplastic Nevi and Melanoma

| | | | | | SVM | |
|---|---|---|---|---|---|---|
| | *k*-NN | Log regression | ANN | Decision trees | Polynomial | Gaussian |
| Avg AUC | 0.7943 | 0.8288 | 0.8263 | 0.7751 | 0.8131 | 0.8305 |
| Std dev | 0.0219 | 0.0168 | 0.0177 | 0.0250 | 0.0185 | 0.0149 |
| Min AUC | 0.7482 | 0.7899 | 0.7866 | 0.7035 | 0.7574 | 0.7939 |
| Max AUC | 0.8404 | 0.8726 | 0.8730 | 0.8311 | 0.8508 | 0.8623 |
| Avg sens | 0.7292 | 0.7686 | 0.7714 | 0.7312 | 0.7571 | 0.7820 |
| Avg spec | 0.7287 | 0.7395 | 0.7417 | 0.7020 | 0.7183 | 0.7360 |

*Note*. For nearest neighbors, $k = 50$. For SVM, the optimal polynomial kernel was linear, with $C = 100$, and the optimal Gaussian RBF kernel had inverse variance $\gamma = 10^{-4}$ and $C = 100$.

vector machines, their ROC curves are virtually indistinguishable.

We now briefly discuss the results of the different methods on the data sets.

### *k*-Nearest Neighbors

The distance metric used for this method was the standard Euclidian distance on real vectors. Since the data had been normalized to zero mean and unit variance, every variable contributed equally to the distance measure.

It is interesting to note that the *k*-nearest-neighbors algorithm is very robust on this problem; i.e., the classification results vary only little with the choice of parameter *k*. Not shown in Table 1 are the results for $k = 20$, $k = 30$, $k = 50$, $k = 60$, $k = 80$, and $k = 90$ because these vary only slightly from those shown in the table. The VUS results over all values of *k* ranged from 0.5788 ($k = 70$) to 0.6042 ($k = 40$). While these results are not as good as those of logistic regression or neural networks, they are a good starting point against which the other methods can be measured.

For the dichotomous tasks summarized in Tables 2 and 3, it can be seen that the results of the *k*-nearest-neighbors algorithm are only slightly inferior (3 to 4 percentage points) to those of the better methods. On the other hand, there is a larger difference of 7 to 8 percentage points in VUS values (see Table 1). This discrepancy is due to the different scalings of the measurements: In the range of the VUS values for trichotomous tests in Table 1 (0.60 for *k*-nearest neighbors to 0.68 for neural nets), 8 percentage points correspond to only 4 percentage points in AUC values for a dichotomous test. Thus, the difference in results for the trichotomous classification task is comparable to the difference in results for the two dichotomous classification tasks.

TABLE 3

Performance Comparison of *k*-Nearest Neighbors, Logistic Regression, Artificial Neural Networks, Decision Trees, and Support Vector Machines on the Task of Distinguishing Melanoma from Common and Dysplastic Nevi

| | | | | | SVM | |
|---|---|---|---|---|---|---|
| | *k*-NN | Log regression | ANN | Decision trees | Polynomial | Gaussian |
| Avg AUC | 0.9332 | 0.9677 | 0.9680 | 0.8857 | 0.9184 | 0.9700 |
| Std dev | 0.0234 | 0.0175 | 0.0122 | 0.0421 | 0.0268 | 0.0132 |
| Min AUC | 0.8469 | 0.8948 | 0.9371 | 0.7615 | 0.8276 | 0.9282 |
| Max AUC | 0.9775 | 0.9949 | 0.9929 | 0.9714 | 0.9709 | 0.9936 |
| Avg sens | 0.8493 | 0.9240 | 0.9143 | 0.7985 | 0.8448 | 0.9205 |
| Avg spec | 0.9044 | 0.9405 | 0.9397 | 0.9000 | 0.8845 | 0.9497 |

*Note*. For nearest neighbors, $k = 50$. For SVM, the optimal polynomial kernel was linear, with $C = 100$, and the optimal Gaussian RBF kernel had inverse variance $\gamma = 10^{-4}$ and $C = 100$.

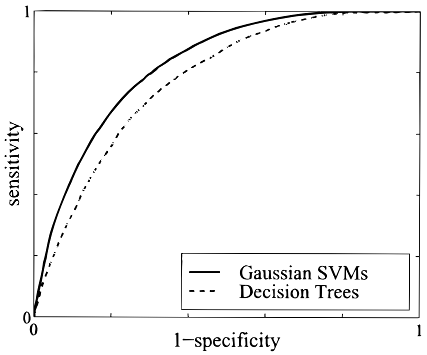

**FIG. 1.** Averaged ROC curves for support vector machines with Gaussian kernels and decision trees on the task of distinguishing common nevi from dysplastic nevi and melanoma. The AUC value is 0.8305 for the SVMs and 0.7751 for the decision trees.

### *Logistic Regression*

Although logistic regression is a linear-in-parameters method that can only implement linear separating hyperplanes between data points, it is nevertheless widely used in medicine. The two main advantages this method has over other algorithms is its ease of use (it is implemented in numerous software packages), and its variable-selection capability. The latter has only limited importance for the PSL classification tasks, since the input variables are obtained from an image segmentation algorithm and are not directly interpretable as humanly visible features of an image. Nevertheless, it is desirable to eliminate input variables that contribute only random correlations to the overall result.

In all three classification tasks (summarized in Tables 1 to 3), logistic regression performs on about the same level as artificial neural networks and support vector machines, which are both capable of implementing nonlinear separating surfaces.

### *Artificial Neural Networks*

This machine learning method has received considerable interest over the past decade for its promise to automatically "learn" structure from data. However, many of the early implementations required a significant amount of parameter-tuning to achieve satisfactory results, a process that needed too much time and expertise for a nonexpert. Over the past few years, statistically motivated Bayesian methods [26] and implementations of faster learning algorithms [10] have allowed nonexperts the use of sophisticated methods that require little to no parameter-tuning. For the experiments in this paper, we used a conjugate gradient algorithm that required no additional parameters to be set. We used 20 nodes in the hidden layer; sample runs with 10 nodes showed similar results.

The results obtained by neural networks were in the same range as those of logistic regression and support vector machines. The training times were comparable to most of the other methods as well, with only a few seconds for each of the 100 data set splits on a standard workstation. For the training process, 200 common nevi, 64 dysplastic nevi, and 20 melanoma were randomly selected from the training set to form a holdout set. To avoid overtraining, the networks were then trained on the remaining PSL images in the training set until the error on the holdout set started to increase.

### *Decision Trees*

Decision trees are not ideally suited for the task of classifying PSL images, as can be seen from the results in Tables 1 to 3. The reason for this is that almost all the variables in the data set represent continuous data. This makes it hard to find the optimal thresholds needed to construct the decision tree.

Given this fundamental disadvantage, it is not surprising to see that decision trees perform poorest of all the methods investigated for this paper. The main advantage that this paradigm has over the other methods — the human interpretability of the results, the trees themselves — is not applicable in this domain, since the input variables are machine-generated (from the vision segmentation system) and do not correspond directly to visible features of the lesion. Separate experiments with a different image segmentation system that

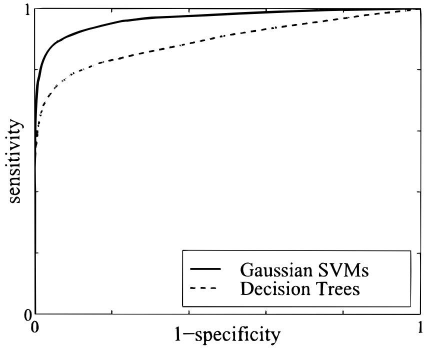

**FIG. 2.** Averaged ROC curves for support vector machines with Gaussian kernels and decision trees on the task of distinguishing melanoma from common and dysplastic nevi. The AUC value is 0.9700 for the SVMs and 0.8857 for the decision trees.

focused on the interpretability of the variables (such as color distribution on the lesion, asymmetry) gave even poorer results, both for decision trees and for the other algorithms. From the point of view of optimal classification results, it seems that the images can be represented better by variables that are not easily interpreted by the human eye.

### *Support Vector Machines*

While SVMs only implement separating hyperplanes, they can effectively construct nonlinear decision boundaries by mapping the data into a higher-dimensional space in a nonlinear manner by using kernel functions. Since it is not possible to determine a priori which kernel function works best for which data set, considerable time is spent on trying different kernel functions. The most popular kernel functions are polynomials and Gaussian radial basis functions (RBFs).

For polynomial kernels, the adjustable parameter is the degree of the polynomial; for Gaussian RBF kernels, it is the inverse variance. For any kernel function, it is also necessary to specify a cost factor $C$ that determines the importance of misclassifications on the training set.

In our experiments, we used polynomial kernels of degrees 1 to 3, Gaussian RBF kernels with parameter $\gamma = 10^{-6}$, $10^{-5}, \ldots, 10^{-2}$, and cost factor parameter values of $C = 100$ and $C = 1000$. The results for both these cost parameter settings were similar, with the $C = 100$ models performing slightly better than the others. Therefore, we report only results for $C = 100$. Training times were about an order of magnitude slower than for the neural network models, but still in the range of only a few minutes. For the polynomial kernels, convergence times depended heavily on the degree of the kernel polynomial, with times for degree four kernels too slow to be included here. Gaussian RBF kernels were generally fast to converge, and did not depend as heavily on the choice of precision parameter $\gamma$.

Since few results comparing the performance of SVM models with other machine learning paradigms on medical data sets are available, we report the results of the SVM models on the dichotomous classification tasks in more detail in Tables 4 and 5. For the polynomial kernels, it is interesting to note that the linear kernel function performs better than the polynomial kernels of degrees 2 and 3. In light of the good performance of the logistic regression model, it is not surprising that a linear model should do well. It is surprising, however, that the higher-degree polynomial kernels did not perform at the same level. For the Gaussian RBF kernels, the best results were obtained for $\gamma = 10^{-4}$. The classification performance decreases for smaller and larger values of $\gamma$. The results for $\gamma = 10^{-5}$ and $\gamma = 10^{-3}$ are not listed in the tables because they are less than the best results, but better than those for $\gamma = 10^{-6}$ and $\gamma = 10^{-4}$, respectively.

## 4. DISCUSSION

As mentioned previously, choosing a "best model" for a given classification task depends not only on discriminatory power, but also on other factors such as cost of model construction and model interpretability. In this paper, we focused solely on determining the classification performance and disregarded the latter two points. This is because we used the same data set for all methods, so that the cost of collecting data is the same for each method. Furthermore, the variables in the data set were automatically derived from an image segmentation system and are not human interpretable, so that interpretability of the model itself is not an issue.

Of the five methods investigated in this paper, the top three (logistic regression, artificial neural networks, and support vector machines) give almost identical results, whereas the other two ($k$-nearest neighbors and decision trees) drop off considerably on some of the classification tasks. Even the worst of the five (decision trees) achieves sensitivity and specificity values that are comparable to human experts [1]. The top three (logistic regression, artificial neural networks, and support vector machines) obtain results that are well above this level.

With the experimental setup of this paper of using 100 different data set splits for training and test sets, it is not possible to check for statistically significant differences in classification performance. This is due to the fact that the 100 different splits are highly correlated, and thus the results obtained from these runs are not independent. Statistical tests would have to consider the dependencies introduced by using overlapping training sets, making formal statistical tests extremely difficult.

What *can* be said about the results of the runs is that the data set was large enough (or well-behaved enough) so that for almost all methods on all the tasks, there were no outliers in the results. By this, we mean that the results of the 100 splits are nicely distributed almost within two standard deviations around a mean value.

Furthermore, it is surprising to note that the nonlinear methods are not able to give better results than logistic regression, which is "only" a linear method. The good performance of logistic regression cannot even be attributed to the high-dimensional input space, since we used only the eight

TABLE 4
Performance Comparison of Different SVM Models on the Task of Distinguishing Melanoma from Common and Dysplastic Nevi

| | Polynomial kernel | | | Gaussian RBF kernel | | |
|---|---|---|---|---|---|---|
| | $d = 1$ | $d = 2$ | $d = 3$ | $\gamma = 10^{-6}$ | $\gamma = 10^{-4}$ | $\gamma = 10^{-2}$ |
| Avg AUC | 0.9184 | 0.8544 | 0.9051 | 0.9644 | 0.9700 | 0.9471 |
| Std dev | 0.0268 | 0.0390 | 0.0346 | 0.0132 | 0.0132 | 0.0178 |
| Min AUC | 0.8276 | 0.7593 | 0.7863 | 0.9337 | 0.9282 | 0.9100 |
| Max AUC | 0.9709 | 0.9551 | 0.9849 | 0.9892 | 0.9936 | 0.9852 |
| Avg sens | 0.8448 | 0.7820 | 0.8370 | 0.9114 | 0.9205 | 0.8903 |
| Avg spec | 0.8845 | 0.8855 | 0.8950 | 0.9141 | 0.9497 | 0.9124 |

*Note*. All models used cost factor settings of $C = 100$.

most significant variables for this model. It seems that in this domain, as with other real-world problems, there is not much to be gained by including nonlinearities in the models.

The good results of support vector machines indicate that this paradigm is going to be investigated and used more frequently in medical domains. It seems to be a viable alternative to logistic regression and neural networks, especially since there are theoretical bounds on the generalization error in SVM models [16].

## 5. CONCLUSION

We investigated the use of five machine-learning paradigms on the problem of automatically classifying pigmented skin lesions as common nevi, dysplastic nevi, or melanoma. While the decision tree paradigm is not well suited for this problem domain (most of the input variables are continuous), the other methods performed well ($k$-nearest neighbors) to very well (logistic regression, artificial neural networks, and support vector machines) on the data sets.

Although it is not desirable to replace dermatologists in the diagnostic procedure, the results of this paper indicate that decision support tools could be used to increase the performance of human experts. One possible application area is in intelligent training tools. Such tools could be designed as tutoring systems for dermatologists, with large repositories of lesion images and gold standard diagnoses for these images. Trained models could then provide reference probability assessments and, for a given lesion from the repository, present lesions with similar degrees of malignancy. Similarity matching on the lesion features could also be used to present features that are not only similar in diagnosis, but also similar in appearance. Further work will be needed to investigate these ideas in detail.

## ACKNOWLEDGMENTS

Financial support for this study was provided in part by Grants J-1661-INF (S.D.) and P-11735-MED (M.B.,H.K.) from the Austrian

TABLE 5
Performance Comparison of Different SVM Models on the Task of Distinguishing Common Nevi from Dysplastic Nevi and Melanoma

| | Polynomial kernel | | | Gaussian RBF kernel | | |
|---|---|---|---|---|---|---|
| | $d = 1$ | $d = 2$ | $d = 3$ | $\gamma = 10^{-6}$ | $\gamma = 10^{-4}$ | $\gamma = 10^{-2}$ |
| Avg AUC | 0.8131 | 0.7379 | 0.7377 | 0.8189 | 0.8305 | 0.7863 |
| Std dev | 0.0185 | 0.0223 | 0.0334 | 0.0184 | 0.0149 | 0.0183 |
| Min AUC | 0.7574 | 0.6702 | 0.5000 | 0.7699 | 0.7939 | 0.7488 |
| Max AUC | 0.8508 | 0.7909 | 0.7814 | 0.8659 | 0.8623 | 0.8257 |
| Avg sens | 0.7571 | 0.7043 | 0.7209 | 0.7362 | 0.7820 | 0.7459 |
| Avg spec | 0.7183 | 0.6598 | 0.6402 | 0.7443 | 0.7360 | 0.7164 |

*Note*. All models used cost factor settings of $C = 100$.

Science Foundation FWF, by the Max Kade Foundation (M.B.), by NLM/NHLBI Contract 467-MZ-802289 (L.O.M.), by Project Grant 107409/320 from the Norwegian Research Council (S.V.), by NLM Grant R29 LM06538-01 (L.O.M., S.V.), and by Grants FIS 97/0267 and CICYT TEL97-1073-C02-01 (HB). The authors thank H. Ganster, M.Sc., Department of Computer Graphics and Vision, Technical University Graz, for his work on the image segmentation algorithm. This work was also supported in part by Grant P-11735-MED of the FWF.

## REFERENCES

1. Grin C, Kopf A, Welkovich B, Bart R, Levenstein M. Accuracy in the clinical diagnosis of malignant melanoma. Arch Dermatol 1990; 126:763–766.
2. Binder M, Schwarz M, Winkler A, Steiner A, Kaider A, Wolff K, Pehamberger H. Epiluminescence microscopy. A useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists. Arch Dermatol 1995; 131:286–291.
3. Dasarathy B, editor. Nearest neighbor (NN) norms: NN pattern classification techniques. IEEE Computer Society Press, Los Alamitos, CA, 1991.
4. Decaestecker C, Salmon I, Dewitte O, Camby I, Ham PV, Paseels J, Brotchi J, Kiss R. Nearest-neighbor classification for identification of aggressive versus nonaggressive low-grade astrocytic tumors by means of image cytometry-generated variables. J Neurosurg 1997; 86:532–537.
5. el Kwae E, Fishman J, Bianchi M, Pattany P, Kabuka M. Detection of suspected malignant patterns in three-dimensional magnetic resonance breast images. J Digit Imaging 1998; 11:83–93.
6. Handels H, Ross T, Kreusch J, Wolff H, Poppl S. Feature selection algorithm for optimized skin tumor recognition using genetic algorithms. Artif Intell Med 1999; 16:283–297.
7. Harrel F, Lee K. Regression modelling strategies for improved prognostic prediction. Stat Med 1984; 3:143–152.
8. Spiegelhalter D. Probabilistic prediction in patient management and clinical trials. Stat Med 1986; 5:421–433.
9. Altman D. Practical statistics for medical research, London/New York: Chapman & Hall, 1991.
10. Bishop C. Neural networks for pattern recognition. London: Oxford Univ. Press, 1995.
11. Burke H. Artificial neural networks for cancer research: outcome prediction. Semin Surg Oncol 1994; 10:73–76.
12. Dybowski R, Weller P, Change R, Gant V. Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. Lancet 1996; 347:1146–1150.
13. Zorman M, Stiglic M, Kokol P, Malcic I. The limitations of decision trees and automatic learning in real world medical decision making. J Med Syst 1997; 21:403–415.
14. Clark D. Computational methods for probabilistic decision trees. Comput Biomed Res 1997; 30:19–33.
15. McKenzie D, McGorry P, Wallace C, Low L, Copolov D, Singh B. Constructing a minimal diagnostic decision tree. Methods Inf Med 1993; 32:161–166.
16. Vapnik V. Statistical learning theory. New York Wiley, 1998.
17. Vapnik V. The nature of statistical learning theory. 2nd ed. Berlin: Springer Verlag, 1995.
18. Morik K, Imhoff M, Brockhausen P, Joachims T, Gather U. Knowledge discovery and knowledge validation in intensive care. Artif Intell Med 2000; 19:225–249.
19. Mossman D. Three-way ROCs. Med Decis Making 1999; 19:78–89.
20. Dreiseitl S, Ohno-Machado L, Binder M. Comparing three-class diagnostic tests by three-way ROC analysis. Med Decis Making 2000; 20:323–331.
21. Hanley J, McNeil B. The meaning and use of the area under the receiver operating characteristic (ROC) curve. Radiology 1982; 143:29–36.
22. Hanley J, McNeil B. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983; 148:839–843.
23. DeLong E, DeLong D, Clarke-Pearson D. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 1988; 44:837–845.
24. Bredensteiner E, Bennett K. Multicategory classification by support vector machines. Comput Opt Appl 1999; 12:53–79.
25. Harrell F, Califf R, Pryor D, Lee K, Rosati R. Evaluating the yield of medical tests. J Am Med Assoc 1982; 247:2543–2546.
26. Neal R. Bayesian learning for neural networks. New York: Springer, 1996.