Empirical Evaluation of Supervised Machine Learning Algorithms on Healthcare Classification Tasks

Laura Fleig
University of California, San Diego
Department of Cognitive Science
COGS 118A Final Project

Abstract

This study presents a systematic evaluation of three prominent supervised machine learning algorithms (Random Forests, Support Vector Machines, and Neural Networks) across three healthcare-related classification tasks. Following the methodological framework established by Caruana and Niculescu-Mizil (2006), we assessed these algorithms' performances on heart disease prediction, breast cancer diagnosis, and Parkinson's disease detection. Our experimental design incorporated multiple train-test splits (20-80, 50-50, 80-20) with three independent trials per configuration, evaluating performance through accuracy and ROC-AUC metrics. The results consistently showed Random Forests achieving superior or competitive performance across datasets and splits, particularly in scenarios with limited training data. SVMs demonstrated comparable performance, while Neural Networks generally performed slightly below the other two methods. These findings reinforce Caruana and Niculescu-Mizil's observations about the relative strengths of different learning algorithms.

1 Introduction

The application of machine learning to healthcare problems represents one of the most promising intersections of artificial intelligence and real-world impact. As healthcare systems digitize and collect more patient data, the ability to accurately diagnose conditions becomes increasingly important. However, the selection of appropriate machine learning algorithms remains a critical challenge, as different methods may vary in effectiveness.

This study builds upon the work of Caruana and Niculescu-Mizil (2006) [1], who conducted a comprehensive empirical evaluation of multiple learning algorithms across many datasets. In this paper, we focus specifically on healthcare applications: heart disease, breast cancer, and Parkinson's disease. We aim to answer the questions of how RF, SVM, and NN compare in their ability to handle medical data, how training dataset size affects relative performance, and to what extent Caruana and Niculescu-Mizil's findings generalize to healthcare classification tasks.

2 Methods

We conducted a comprehensive empirical evaluation of three widely used supervised learning algorithms: Random Forests (RF), Support Vector Machines (SVM), and Neural Networks (NN). Following the methodology outlined in Caruana and Niculescu-Mizil (2006) [1], we implemented a rigorous experimental framework to assess classifier performance across different training set sizes and multiple trials. This framework was carried out across three datasets from the UCI Machine Learning Repository.

2.1 Classifier Overview

Random Forests (RF) construct multiple decision trees during training and output the majority vote for classification tasks. Each tree is built using a random subset of the training data and a random subset of features, which helps prevent overfitting and improves generalization. Support Vector Machines (SVM) work by finding an optimal hyperplane that maximally separates different classes in a high-dimensional space. SVMs can find non-linear relationships in the data by using kernel functions. Neural Networks (NN) are inspired by biological neural systems and consist of interconnected layers of nodes. Each node combines input values using learned weights, applies an activation function, and then passes the result to the next layer. Through backpropagation and gradient descent, the network learns to map inputs to desired outputs.

2.2 Implementation Framework

Our experimental framework consisted of three main components: data partitioning and cross-validation, hyperparameter optimization, and performance evaluation.

For each dataset, we created three different train-test splits (20-80, 50-50, 80-20) to evaluate how the algorithms perform with varying amounts of training data. For each split ratio, we conducted three independent trials to ensure robust results. Within each trial, we employed cross-validation on the training data to optimize hyperparameters before evaluating on the held-out test set.

2.3 Algorithm Configuration

We implemented each classifier using scikit-learn with the following hyperparameter spaces:

Random Forest

- Number of estimators: 100
- Maximum depth: [3, 5, 10]
- Minimum samples for split: [5, 10]
- Maximum features: sqrt

Neural Network

- Hidden layer sizes: [(10,), (20,)]
- Alpha (regularization): [0.01, 0.1]
- Initial learning rate: 0.01
- Early stopping: enabled
- Validation fraction: 0.2

Support Vector Machine

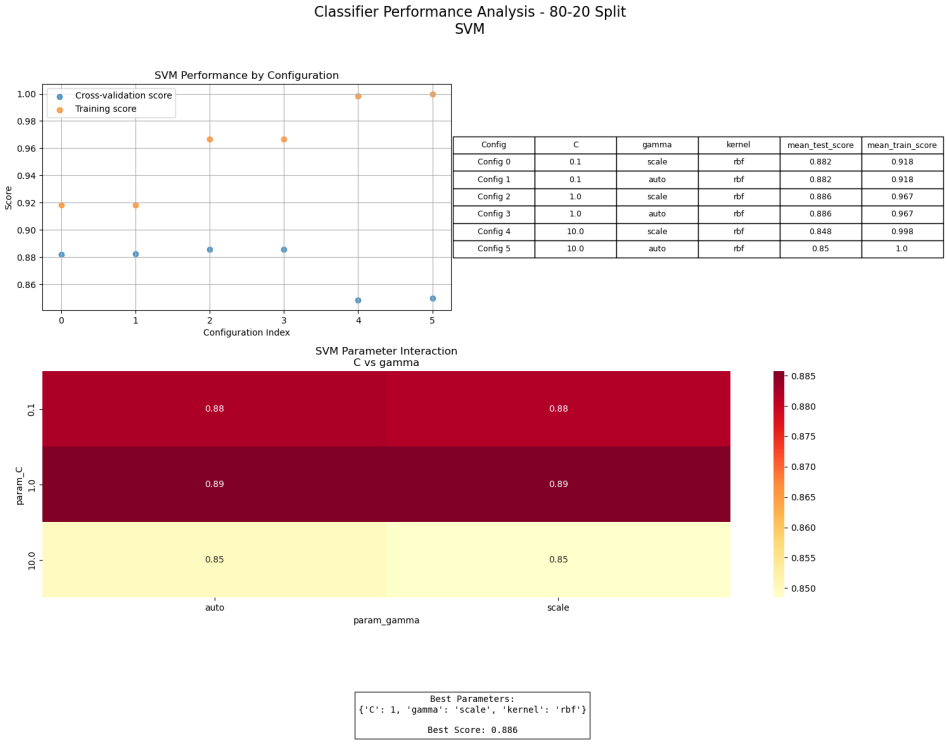
- Kernel: RBF
- C (regularization): [0.1, 1, 10]
- Gamma: [scale, auto]

Fig.1 shows one example iteration of this process. For the third trial of the 80-20 split on Dataset 1, SVM cycles through all different parameter configurations, conveniently placed in a lookup table. The scores for each configuration are in a graph. Two selected hyperparameters, in this case C and gamma, are also visualized in a heatmap. Please refer to the attached GitHub repository (see Supplementary Material below) to see all visualizations, as they have been omitted from this paper for conciseness.

2.4 Performance Metrics

We employed two complementary metrics to evaluate classifier performance:

Figure 1:



Config	C	gamma	kernel	mean_test_score	mean_train_score
Config 0	0.1	scale	rbf	0.882	0.918
Config 1	0.1	auto	rbf	0.882	0.918
Config 2	1.0	scale	rbf	0.886	0.967
Config 3	1.0	auto	rbf	0.886	0.967
Config 4	10.0	scale	rbf	0.848	0.998
Config 5	10.0	auto	rbf	0.85	1.0

1. Accuracy: to measure overall classification performance
2. ROC-AUC: to assess discrimination ability independent of decision threshold

Accuracy measures the proportion of correct predictions among all predictions made, providing an intuitive measure of overall performance. The Receiver Operating Characteristic Area Under the Curve (ROC-AUC) measures the classifier's ability to discriminate between classes across all possible classification thresholds. ROC-AUC scores range from 0 to 1, with 1 indicating perfect prediction and 0.5 representing random chance (so this is a "higher is better" metric). ROC-AUC is particularly useful because it is independent of classification threshold and is robust to class imbalance.

3 Experiments

3.1 Datasets

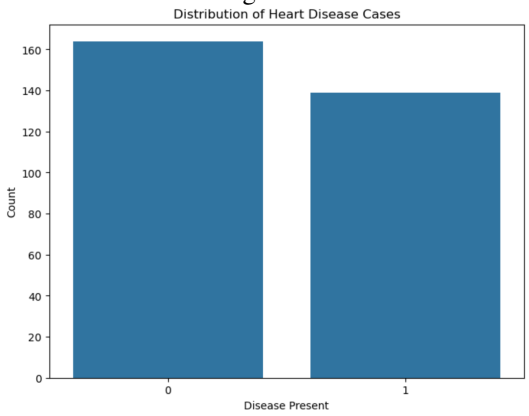
For each dataset, we implemented a standardized preprocessing pipeline: loading and initial cleaning, binary target variable conversion (if necessary), feature normalization, missing value handling, and train-test splitting according to specified ratios.

3.1.1 Dataset 1: Heart Disease

The UCI[1] Cleveland Heart Disease dataset [2] is a widely used machine learning dataset that contains medical and demographic information for 303 patients, along with whether they have heart disease. The dataset comprises 13 features, including demographic attributes (e.g., age, sex), clinical measurements (e.g., blood pressure, cholesterol), and medical test results (e.g., ECG readings, thalassemia type). The target variable was binarized to represent the presence (1) or absence (0) of heart disease (see Fig.2).

[1] https://archive.ics.uci.edu/dataset/45/heart+disease

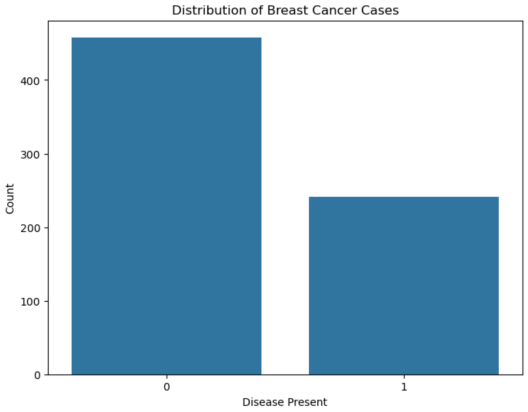
Figure 2:



3.1.2 Dataset 2: Breast Cancer

The UCI[2] Breast Cancer Wisconsin (Original) dataset [3] is a widely-used medical dataset for binary classification. It contains 699 instances of breast tumors, characterized by 9 features scored on a scale of 1-10. These features include clump thickness, cell size and shape uniformity, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses. Each instance is classified as either benign or malignant, making it a binary classification problem (see Fig.3).

Figure 3:



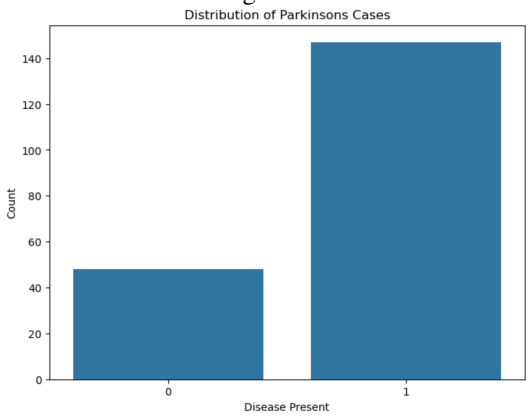
3.1.3 Dataset 3: Parkinson's Disease

The UCI[3] Parkinson's Disease Detection dataset [4] is a biomedical voice analysis dataset designed to support early detection of Parkinson's disease through voice recordings. The dataset comprises 195 voice recordings from 31 individuals (23 with Parkinson's disease and 8 healthy controls), with approximately six recordings per person. Each recording is characterized by 22 voice measures including vocal fundamental frequency, measures of variation in amplitude and frequency, nonlinear dynamical complexity measures, signal fractal scaling exponents, and nonlinear measures of fundamental frequency variation. These features capture various aspects of vocal impairment, a common early indicator of Parkinson's disease. The classification task is binary, with the target variable indicating presence (1) or absence (0) of Parkinson's disease (see Fig.4).

[2]https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original

[3]https://archive.ics.uci.edu/dataset/174/parkinsons

Figure 4:



3.2 Results

3.2.1 Results for Dataset 1: Heart Disease

The experimental results for the Cleveland Heart Disease dataset revealed several interesting patterns across different training set sizes and classifiers. The average performance metrics across the three trials for each classifier and split ratio are shown in Fig.6 and Fig.7.

The best-performing hyperparameters for each classifier were:

- Random Forest: max depth=3, max features=sqrt, min samples split=5
- Neural Network: hidden layer sizes=(20,), alpha=0.01, early stopping=True
- SVM: C=10, gamma=scale, kernel=rbf

These results suggest that while all three classifiers can achieve strong performance on the heart disease prediction task (see Fig.5), Random Forests offer the best combination of consistency and accuracy across different training set sizes, consistent with the results of Caruana and Niculescu-Mizil (2006) [1]. The SVM classifier showed particularly strong performance when given access to larger training sets, outperforming Random Forests in terms of accuracy for the 80-20 split (although Random Forests had higher ROC-AUC for the 80-20 split). Neural Networks, while competitive, generally performed slightly below the other methods in both metrics.

Figure 5:

Average Performance:

split	classifier	accuracy	roc_auc
20-80	Neural Network	0.699588	0.837345
	Random Forest	0.781893	0.896913
	SVM	0.720165	0.814701
50-50	Neural Network	0.815789	0.900868
	Random Forest	0.822368	0.903819
	SVM	0.809211	0.923090
80-20	Neural Network	0.819672	0.917026
	Random Forest	0.852459	0.949353
	SVM	0.901639	0.941810

3.2.2 Results for Dataset 2: Breast Cancer

The experimental results for the second dataset, Breast Cancer, yielded similar results. Random Forests outperformed both SVM and Neural Networks, although RF and SVM were very close. For all three train-test splits, all three classifiers achieved >95% accuracy. The average performance for this dataset are shown in Fig.8, while the evaluation metrics are in Fig.9 and Fig.10.

The best hyperparameters were:

- Random Forest: max depth=5, max features=sqrt, min samples split=5

Figure 6:

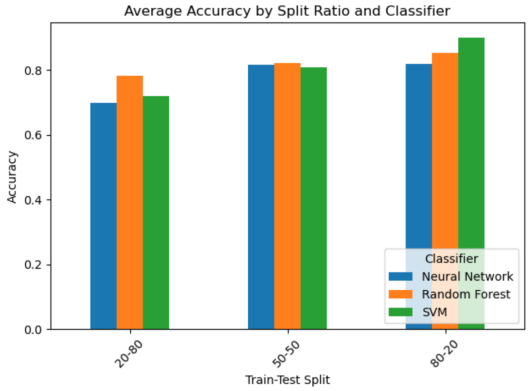
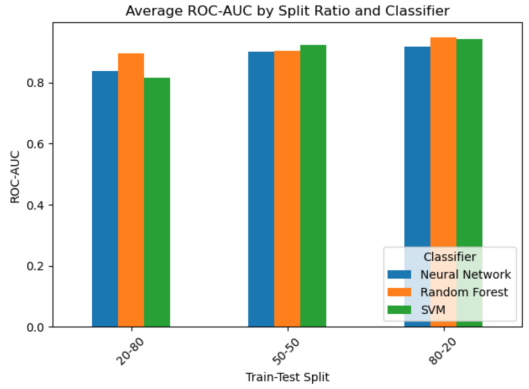


Figure 7:



- Neural Network: hidden layer sizes=(10,), alpha=0.01, early stopping=True, validation fraction = 0.2
- SVM: C=0.1, gamma=scale, kernel=rbf

Figure 8:

Average Performance:

split	classifier	accuracy	roc_auc
20-80	Neural Network	0.964286	0.992202
	Random Forest	0.966071	0.992393
	SVM	0.960714	0.992145
50-50	Neural Network	0.957143	0.988117
	Random Forest	0.962857	0.990258
	SVM	0.960000	0.989366
80-20	Neural Network	0.964286	0.991579
	Random Forest	0.971429	0.994854
	SVM	0.964286	0.995556

3.2.3 Results for Dataset 3: Parkinson's Disease

Finally, the third dataset also gave similar results to the previous datasets and to Caruana and Niculescu-Mizil (2006) [1]: Random Forests showed the strongest performance, with SVM close behind (see Fig.11, and Fig.12 and Fig.13 for metrics). Neural Networks performed noticeably worse on this dataset, particularly with a low train split.

The best hyperparameters for this dataset were:

- Random Forest: max depth=3, max features=sqrt, min samples split=5

Figure 9:

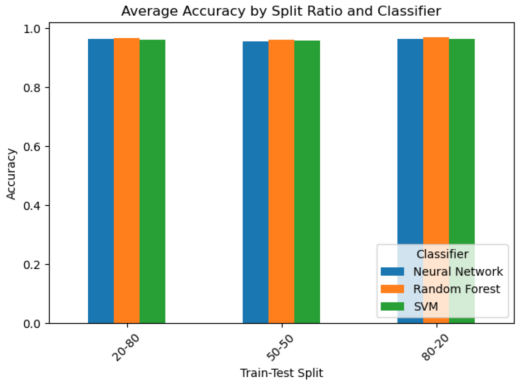
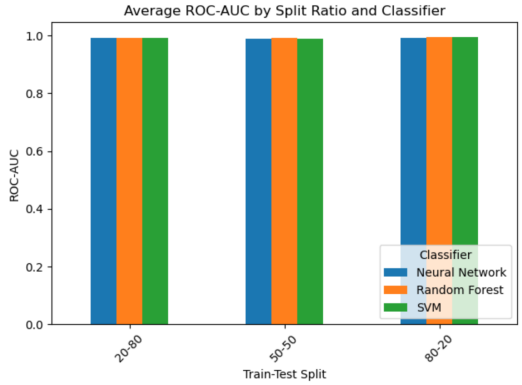


Figure 10:



- Neural Network: hidden layer sizes=(20,), alpha=0.01, early stopping=True, validation fraction = 0.2
- SVM: C=10, gamma=scale, kernel=rbf

Figure 11:

Average Performance:

split	classifier	accuracy	roc_auc
20-80	Neural Network	0.717949	0.561473
	Random Forest	0.878205	0.931405
	SVM	0.826923	0.880342
50-50	Neural Network	0.836735	0.864928
	Random Forest	0.948980	0.962319
	SVM	0.928571	0.974493
80-20	Neural Network	0.871795	0.879464
	Random Forest	0.948718	0.933036
	SVM	0.948718	0.959821

4 Conclusion

Our experimental results largely align with the findings of Caruana and Niculescu-Mizil (2006) [1]. Random Forests consistently demonstrated the highest performance, particularly excelling in scenarios with limited training data. This reinforces Caruana and Niculescu-Mizil's observation about the effectiveness of ensemble methods.

RFs showed notable consistency in terms of results. SVMs were competitive, particularly in the heart disease dataset where SVMs achieved the highest accuracy with the 80-20 split. NNs generally performed slightly worse, which contrasts somewhat with recent deep learning successes. This might be attributable to training size though. All three algorithms showed improved performance with

Figure 12:

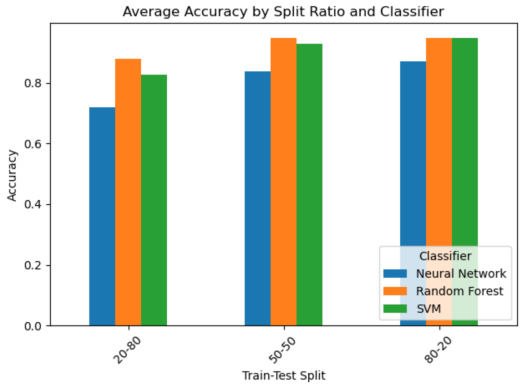
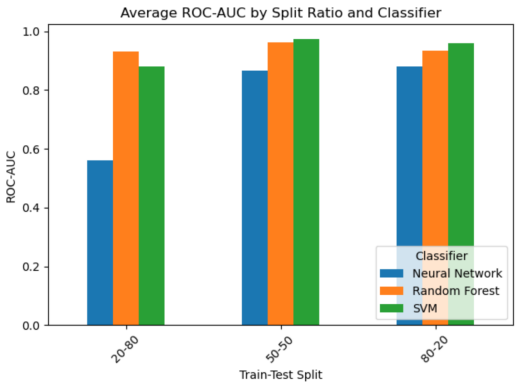


Figure 13:



increased training data, as expected, which highlights the importance of considering data availability when selecting algorithms.

Extensions of this work could include more recent deep learning models and a broader range of datasets, particularly including unstructured data.

5 Supplementary Material

Please find our code at https://github.com/laurafleig/cogs118a-final.

References

[1] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 161–168.

[2] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *The American journal of cardiology*, vol. 64, no. 5, pp. 304–310, 1989.

[3] O. L. Mangasarian and W. H. Wolberg, "Cancer diagnosis via linear programming," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 1990.

[4] M. Little, P. McSharry, E. Hunter, J. Spielman, and L. Ramig, "Suitability of dysphonia measurements for telemonitoring of parkinson's disease," *Nature Precedings*, pp. 1–1, 2008.