Comparing and Contrasting Machine Learning Tools

Edward Chen, Michael Cirrito, Jainam D Shah, and Efthimios Vlahos

ABSTRACT

Does machine learning have a one size fits all technique? Does data coming in all different modalities make any difference? In this experimental study, several machine learning tools, more specifically classification tools, are investigated. Two gender classification datasets with different data types, i.e., image and numerical, are used to compare the performance of those tools. The efficiency of the features and the misclassification cases are analyzed.

1. INTRODUCTION

Machine learning is to feed the training data to a machine learning tool such that the trained tool can fulfill some intended tasks upon receiving novel data. Machine learning can accomplish quite a few tasks, such as classification, regression, detection, ranking, etc. Over the past decades, a lot of machine learning tools have emerged. Before deep learning, many of them had shown quite impressive performance.

Can we rely on just one method to solve all the problems in machine learning? Or does the choice of tool depend on the task and the data modalities used? Several interesting comprehensive survey papers on machine learning tools had shed some light. [1] presented a comprehensive view on these machine learning algorithms that can be applied to enhance the intelligence and the capabilities of an application. The paper explained the principles of different machine learning techniques and their applicability in various real-world application domains. Based on the study, the challenges and potential research directions were also highlighted. A number of machine learning tools were discussed in [2], as well as how they are applied to different tasks. The paper gave a brief overview of those machine learning tools and their key features. Those tools that can be used to solve real-world problems were also presented. Different parameters and highlights features were examined. Specific frameworks that can be used with the processing platforms were provided as well.

Published sixteen years ago, [3] conducted a large-scale empirical comparison between supervised learning methods. A variety of performance criteria to evaluate the learning methods were used. The authors noted that 1) it is possible that learning methods perform well on one metric but poorly on another, 2) performance by metric has found that calibration has significant impact on tool performance, 3) performance by model has found that there is no universally best learning algorithm. Even the best models perform poorly on some problems, and models that have poor average performance perform well on a few problems or metrics.

Motivated by the aforementioned papers, in this study, we aim to find out which tools out of SVM (Support Vector Machines), Random Forests, and ANN (Artificial Neural Networks) work best for classification on different data modalities. More specifically, the goal is 1).to find the best performance of each classification tool on the datasets, 2).to compare the performance of each machine learning tool to classify the same modality and the different modalities, 3).to analyze the efficiency of those features, and 4).to investigate the misclassification samples.

The rest of this report is organized as follows. In Section 2, the datasets, the machine learning tools, and the performance evaluation metrics will be briefly introduced. The experiments, results, and analysis will be presented in Section 3. Conclusion will be drawn in Section 4. Contribution of each team member is described in Section 5.

2. DATASET, CLASSIFICATION TOOLS, AND PERFORMANCE METRICS

2.1 Dataset

The gender classification dataset from Kaggle.com will be used in this experimental study. This dataset has two independent sub-datasets, one is an image dataset [4], another is a numerical dataset [5].

2.1.1 Image dataset

The image dataset is of cropped images of about 28,500 male and 28,500 female faces. Some sample images are shown in Figure 1.

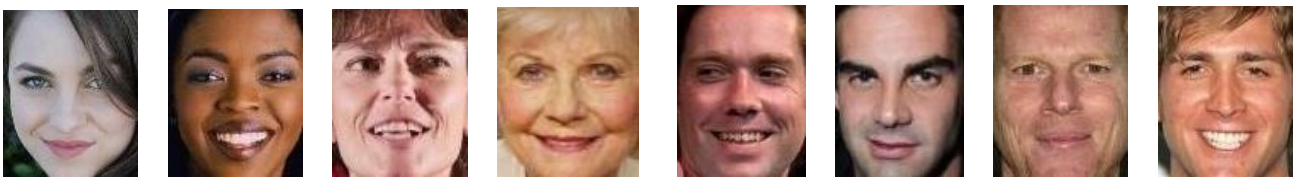

Figure 1. Some sample images used in this study (female first four and male last four).

2.1.2 Numerical dataset

The numerical dataset contains a label and seven descriptive features for each of the 5,001 subjects. The label is the gender of a subject, which is either ”Male” or ”Female”. The features are described as follows:

longhair - If the subject has ”long hair”, this feature is 1. Otherwise, this feature is 0.

foreheadwidthcm - This is the width of the subject’s forehead in centimeters. foreheadheightcm -

This is the height of the subject’s forehead in centimeters.

nosewide - If the subject has a ”wide nose”, this feature is 1. Otherwise, this feature is 0. noselong - If

the subject has a ”long nose”, this feature is 1. Otherwise, this feature is 0. lipsthin - If the subject has

”thin lips”, this feature is 1. Otherwise, this feature is 0. distancenosetoliplong - 1 if the subject has

”long distance between nose and lips” and 0 otherwise.

2.2 Classification Tools

SVMs, ANNs, and Random Forests, all classic classification tools [6], are used in this experimental study.

2.3 Performance Metrics

In this study, three performance metrics Accuracy, F1-score [7], and AUC-ROC curve [8] are used to gauge the comparison of the above classification tools.

3. EXPERIMENTS AND RESULTS

3.1 Data Preparation

3.1.1 Image Data

The images were preprocessed into a dataframe so they can be inputted into the machine learning algorithms. Each image was opened and converted to greyscale. We then resized the images so that each image was the same size 50×50 pixels. We normalized the pixel data by dividing each pixel value by 255. We then reshaped the data of each image to a single array consisting of 2,500 factors containing the normalized information for each pixel in the image.

3.1.2 Numerical Data

Among the seven descriptive features in the Numerical dataset, no preprocessing was applied to the five binary features. For each of the two real number features, i.e., foreheadwidthcm and foreheadheightcm, the value of each feature V was normalized by using the maximum value V_{max} and the minimum value V_{min} of that feature as $(V - V_{min})/(V_{max} - V_{min})$ such that all the normalized values are in the range of [0.0, 1.0].

3.2 Experiment Processes

In this study, we have used grid search to find the hyperparameters for each of those classification tools.

For SVMs, we have tested a few combinations of C and gamma on the polynomial, RBF, and sigmoid kernels.

For ANNs, we have tested a few combinations of number of hidden layers, number of neurons in each layer, and activation functions.

For Random Forests, we have tested a few combinations of n estimators and max depth.

Once the hyperparameter set has been found for a classification tool, that hyperparameter set is used to train the classification tool and the trained classification tool is used to test the unseen data.

3.3 Experiment Results

The Accuracy, F1-score, and AUC/ROC of each of the classification tools (including different kernels) on the test dataset are given in Table 1. The elapsed time of each case is also given in Table 1.

Table 1. Performance metrics of all the test cases (**GPU time, others are CPU time).

Dataset	Perf Metric	SVM-Poly	SVM-Rbf	SVM-Sigmoid	ANN	Random Forest
Numerical	Accuracy	0.9770	0.9780	0.9790	0.9800	0.9740
	F1-Score	0.9766	0.9777	0.9790	0.9799	0.9736
	AUC/ROC	0.9981	0.9981	0.9977	0.9976	0.9943
	Time (Milisec)	62.13	95.40	73.48	1,770.24	749.40
Image	Accuracy	0.9196	0.9270	0.8143	0.9574	0.8646
	F1-Score	0.9213	0.9279	0.8187	0.9576	0.8684
	AUC/ROC	0.9696	0.9779	0.8788	0.9903	0.9362
	Time (Sec)	3,739.26	6,936.98	1,028.12	1,695.29**	10,224.12

To compare the performances of different classification tools on different modalities, the ROC curves of each of the tools are grouped into two plots, one on the numerical dataset and the other on the image dataset, and shown in Figure 2. To compare the performances of a classification tool on different modalities, the ROC curve on numerical dataset and that on image dataset of each of the tools are grouped together. Figure 3 shows five plots, each of which is for a specific tool (or kernel).

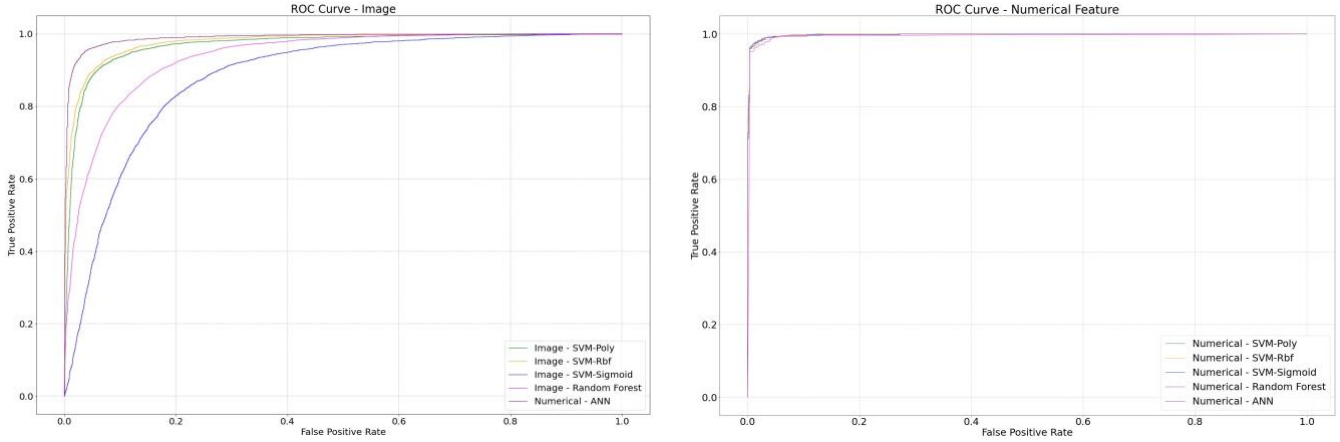

Figure 2. ROC curves across all the classification tools.

For numerical dataset, the performances of all the tools are very comparable. For image dataset, ANN has better performance than all other classification tools. Apparently, ANN takes much longer time than all the other tools (please notice, on image dataset, ANN uses GPU while others just use regular CPU). It is also clear that time elapsed on image dataset is way longer than that on numerical dataset.

3.4 Feature Efficiency and Misclassification Analysis of Numerical Dataset

The efficiency of each feature is shown in Table 2.

Table 2. Efficiency of numerical feature.

Feature	long hair	forehead width	forehead height	nosewide	noselong	lipsthin	distnosetoliplong
Efficiency	0.003627	0.063485	0.046414	0.292320	0.189607	0.148448	0.256099

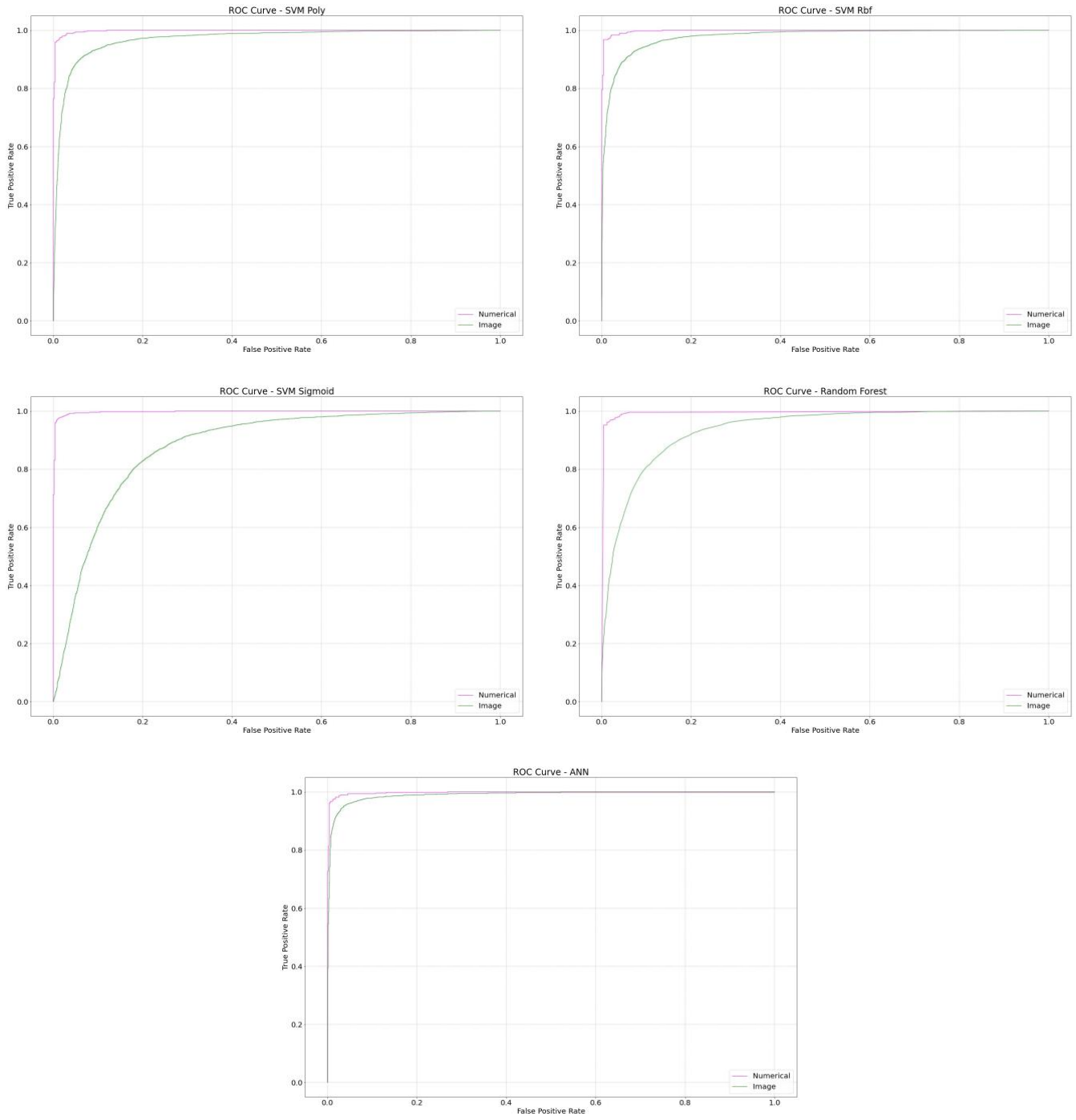

Figure 3. ROC curve of each of the classification tools.

It has been found that long hair is very insignificant to determine if a subject is female or not and the four most significant features are wide nose, long nose, thin lips, and long distance nose to lip. When a female is at least three out of those four significant features, the female is misclassified as a male. If a male either has at most one of those four significant features, or at most two of those features plus a forehead width between 12.8cm and 12.9cm, the male is misclassified as a female.

Features of some misclassified subjects are shown in Table 3. The first two females are misclassified as male while the last five males are misclassified as female by all the training classification tools.

Table 3. Features of some misclassified subjects.

Gender	long hair	forehead width	forehead height	nosewide	noselong	lipsthin	distnosetoliplong
Female	1	12.4	6.3	0	1	1	1
Female	1	14.3	5.9	1	0	1	1
Male	1	14.1	6.2	0	0	0	1
Male	1	12.9	5.4	0	1	0	1
Male	1	12.8	6.0	0	0	1	1
Male	1	11.7	6.3	1	0	0	0
Male	1	12.9	5.6	1	1	0	0

3.5 Misclassification Analysis of Image Dataset

There are quite many mislabelled images in the image dataset. Some examples are shown in Figure 4. Those mislabelled images are very easy to be misclassified. The samples shown in Figure 4 are all misclassified.

Figure 4. Some mislabelled images (male mislabelled as female first four and female mislabelled as male last four).

We analyzed the misclassified images, excluding those mislabelled (and thus misclassified). It has been observed that a lot of images belong to people displaying an emotion. For example, many males misclassified as female seem to be smiling. It is also easy to be misclassified if certain features of a face are obstructed. For instance, many females misclassified as male seem to have glasses on. Males with thinner eyebrows seem to be misclassified as female. Females with wider noses seem to be misclassified as male (interestingly this observation coincides the "nosewide" feature in the numerical dataset). It is worthy noticing that there are also a few faces on which we as human cannot determine their gender. Figure 5 shows some misclassified sample images.

Figure 5. Some misclassified images (female misclassified as male first four and male misclassified as female last four).

4. CONCLUSIONS AND FUTURE WORK

Machine learning has been speeding up its advancement in recent years, due to great advancement of computational power, such as GPU. ANNs have become the backbone of machine learning tools nowadays. We have seen in this experimental study that ANNs have outperformed other tools in classification, though with limited data. However, the elapsed time is still a concern of ANNs.

From this experimental study, it has been found that the tools performs much better on the numerical dataset than on the image dataset, in terms of the performance metrics and also elapsed time. This is because of human interaction with the machine, where the human tells the machine which features may be important and should be looked at. Therefore, when we combine human intelligence and artificial intelligence we get the best results. This could become a future research such that more human intelligence is involved in machine learning in an organic way.

Moreover, there is still room for the numerical features to improve. Although there are four significant numerical features, other numerical features are still in the play. Some other relevant features can be added in the future, though they might be less significant. One such example is the measure of the eyebrows.

5. CONTRIBUTION OF TEAM MEMEBRS

5.1 Edward

Edward's contribution: Reviewed literature, composed part of poster slides, investigated and provided codes for AUC/ROC.

5.2 Michael

Michael's contribution: Performance metric investigation, image preprocessing and postprocessing, numerical SVM grid search and model fitting, numerical data misclassification analysis, image data misclassification analysis, and assisted in the composition of the final report.

5.3 Jainam

Jainam's contribution: Set-up pre-trained deep learning models (AlexNet and ResNet50) for image dataset, hyperparameter search to find learning rate and number of epochs, trained the ANNs and got predictions on test data. Prototyped code for the models (SVMs, ANNs, Random Forests) used for numerical data.

5.4 Efthimios

Efthimios' contribution: Coordinated the collaboration, worked with the team to pick up the topic, to review the literature, to compose and submit abstract, proposal, report, and poster, preprocessed the numerical dataset, worked on the Random Forest coding, training, and test, participated coding, training, and test on other classification tools and tally such as result collection, feature efficiency analysis, and misclassification analysis

6. REFERENCES

1. I. H. Sarker, Machine Learning: Algorithms, Real-World Applications and Research Directions, SN Computer Science 2.3 (2021): 1-21.
2. S. Sarumathi, M. Vaishnavi, S. Geetha, P. Ranjetha, Comparative Analysis of Machine Learning Tools: A Review, International Journal of Computer and Information Engineering 15.6 (2021): 354-363.
3. R. Caruana, A. Niculescu-Mizil, An Empirical Comparison of Supervised Learning Algorithms, Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006.
4. https://www.kaggle.com/datasets/cashutosh/gender-classification-dataset.
5. https://www.kaggle.com/datasets/elakiricoder/gender-classification-dataset.
6. https://www.wikipedia.org.
7. J Korstanje, The F1 score. https://towardsdatascience.com/the-f1-score-bec2bbc38aa6.
8. R Draelos, Measuring Performance: AUC (AUROC). https://glassboxmedicine.com/2019/02/23/measuringperformance-auc-auroc/.