# Performance Analysis of Machine Learning Algorithms on Cancer Cell Detection

*UIUC ECE 380 Final Project Report*

Usha Tripuramallu, Tony Xu, Conan Wen

## Abstract

In this research project, several machine learning algorithms were analyzed to see their effects on cancer cell detection. In doing so, we researched common ways of using machine learning for analyzing different images from other imaging modalities as well as the steps needed to take before an image can be processed. Finally, we used our own algorithms to decide which method of machine learning was most effective in cancer cell detection through the analysis of ROC and Precision-Recall output. We found that the random forest model has the best performance and the support vector machine performed the worst.

## Introduction

There are many biomedical imaging modalities that are used to detect cancer cells. Current practices in oncology may lack accuracy in determining the prognosis of cancer cells. Machine learning algorithms can help increase the chances of diagnosing cancer cells determine the  growth of the disease.

## Background

*Background Research*

Many research papers were reviewed in seeing how institutions have taken steps to use machine learning in the field of biomedical imaging. A particular paper that we used was by Javaria Amin and Muhammad Sharif from COMSATS University Islamabad. In this paper, the researchers highlighted the many steps that were required before computer analysis of cancer cells, which included image enhancement through filtering as well as feature extraction.

The researchers believed that doctors could diagnose cancer through the physical examination of images, but that it would take a significant amount of time to do so especially if the image has uneven lighting and other forms of noise. It is also very time consuming to segment the affected area. The solution to their challenge, which was brain tumor detection, was to implement machine learning algorithms as a faster and more reliable method to segment and prognose cancer.

Before the different algorithms could be implemented, the image needed to be filtered first. Traditional filters, such as the high-pass or low-pass filters including the hamming filter often cause blurring or speckle artifacts. To counteract this, a

Wiener filter is often used. The Wiener filter helps preserve the image while filtering it by minimizing the error between a random signal and a predicted signal. Afterwards, potential field clustering was used to segment the tumor and provide a subset of pixels that could be analyzed. From these segments, features were extracted depending on the intensity of these pixels. An important feature that was extracted is called the Gabor Wavelet, which is a basis used in the fourier transform that is more accurate than the traditional sine and cosine basis. The Gabor wavelet minimizes the standard deviation and decreases uncertainty associated with the fourier transform. Finally, the researchers implemented common machine learning algorithms on their dataset such as the random forest and support vector machine. Our research will go into more regarding how well each of these algorithms could classify the dataset and the performance analysis of each.

*Procedure*

The first step was to conduct research on the various types of machine learning algorithms present that are used for analyzing tumor cells. We found that the most optimal approach was to use supervised machine learning techniques.

Following this, a dataset with was acquired from the University of Wisconsin; breast cancer cells were imaged using a method knowns as fine needle aspirate. This is a minimally invasive tissue sampling method that traditionally uses ultrasound as a targeting instrument. however, other targeting instruments can be used including CT, MRI or fluoroscopy. Once the cells are sampled, they are analyzed using microscopy techniques in order to generate various features.

Once the data set was gathered basic pre-processing was conducted to analyze the characteristics of the dataset. The set consisted of a total of 569 cells in which there were 357 benign cells and 212 malignant cells. For each cell, the following 10 basic features were described: radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths) , compactness (perimeter$^2$ / area - 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, and fractal dimension ("coastline approximation" - 1). Along with these 10 baseline features, for each feature for each cell the mean, standard error, and the largest mean of the three largest values was taken totalling in 30 features apart from the diagnosis describing each cell. These 30 features were used to predict the diagnosis of the cell.

The final step taken was splitting the data into training and test sets and running various algorithms. The models were trained on 80% of the data and then tested on the remaining 20%. This split ratio was used in order to minimize the possible variance with both parameter estimates and performance statistics. Performance metrics for each of the supervised machine learning algorithms was conducted after the model was constructed. An overall comparison was conducted to determine the best performing model.

## Technical Descriptions

Various supervised machine learning algorithms were used to classify the cells. A supervised machine learning algorithm works based on knowing prior output values. The goal is to construct a function that best approximates the relationship between the input characteristics and the output in the observable data. Four of these such algorithms were used: a linear classifier, a decision tree, a random forest, and a support vector machine. The performance of these models was determined by analyzing true positive, false positive, true negative, and false negative rates on a receiver operating characteristic curve and a precision-recall curve.

*Linear Classifier*

A linear classifier, also known as a logistic regression, is an optimization problem that works to minimize the cost function. Various different cost functions are present including $l_1$ penalized, $l_2$ penalized and elastic-net. Each one is based on a different cost function, and the elastic-net method combines the $l_1$ and $l_2$ penalization schemas. The $l_2$ regularization method was used in this research; the equation for this cost function is given by:

$$\min_{w,c} \frac{1}{2}w^T w + C\sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1).$$

With a linear model, each feature is given a weight that indicates how much it contributes to the overall classification of a datapoint. The output of the linear model is a series of weights for each feature and an intercept (a bias term that is added to decision function).

*Decision Tree*

Decision tree is a type of supervised machine learning where the data set is broken down into smaller and smaller subsets while a decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. The leaves are the decisions or the final outcomes and the decision nodes are where the data is split. Compared to other algorithms, decision trees requires less effort for data preparation during pre-processing and does not require normalization of data. However, a small change in the data can cause a large change in the structure of the decision tree which results in instability.

*Random Forest*

Random Forest is an ensemble of decision trees. Each individual decision tree in the random forest gives a class prediction and the class with the most votes becomes the model's prediction. Building multiple decision trees and merging them together results in a more accurate and stable prediction. Random Forest works well with both categorical and continuous variables and can handle nonlinear parameters efficiently. It also has greater stability and is less impacted by noise. When a new data point is introduced to the dataset, the overall algorithm is not affected much since only one tree is impacted. However, this algorithm requires much more computational power and resources due to its complexity. Since the Random Forest algorithm generates a lot of trees, much more time is required to train compared to a decision tree.

*Support Vector Machine (SVM)*

In a support vector machine, data is classified and separated as far as possible from each other using a kernel function. This function can take two sets of data and distributes them in a way that can be easily analyzed and classified. In our case, we used a Radial Basis Function kernel (RBF) and classified between cancerous and non-cancerous. The advantages of using the SVM is that it is able to classify data in higher dimensions as well as stay effective when there are more dimensions than samples. Furthermore, different kernel functions can provide flexibility depending on the particular dataset. This flexibility, however, comes at the price of complexity and longer processing time.

*Receiver Operating Characteristic (ROC) Curve*

This curve is a way to measure the performance of classification problems at different threshold settings. This curve shows how well a model is able to distinguish between the various classes. The curve is plotted with the false positive rate on the x-axis and the true positive rate on the y axis. The higher the area under the curve (AUC) of the ROC the better the model. An area that is close to 1 means that the model has a good measure of how separate the different classes are.

*Precision-Recall Curve*

Similar to the ROC curve, the precision recall curve is a way to measure the performance of a classification model. Like the ROC, having a large area under the curve corresponds to high performance of a given model. A large area is defined by both high precision and high recall. High precision relates to low false positive rates and high recall relates to low false negative rates.

## Results

| Model | Accuracy | TP | FP | FN | TN | ROC AUC |
|---|---|---|---|---|---|---|
| Linear | 0.9385 | 72 | 2 | 5 | 35 | 0.9859 |
| Decision Tree | 0.9649 | 75 | 2 | 2 | 35 | 0.9599 |
| Random Forest | 0.9649 | 75 | 2 | 2 | 35 | 0.9791 |
| SVM | 0.6754 | 77 | 37 | 0 | 0 | 0.9070 |

*Table 1: Performance output of each model*

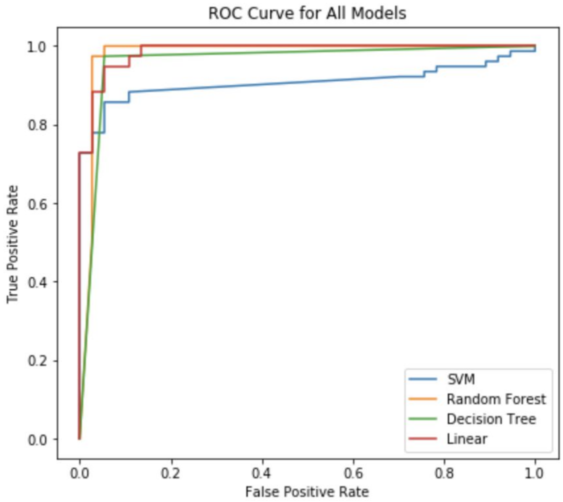

*Figure 1: Combined graph of each ROC curve*

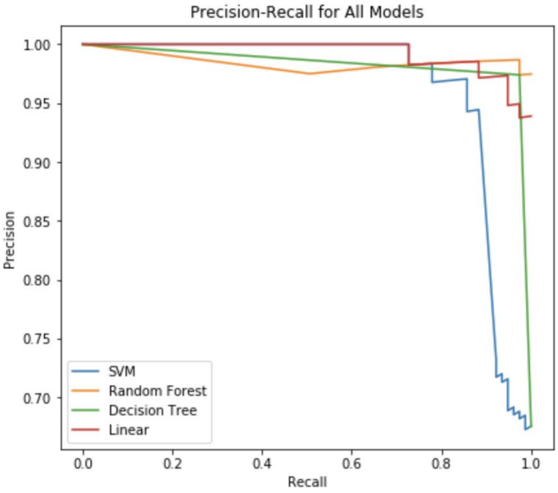

*Figure 2: Combined Graph of each Precision-Recall curve*

The different metrics that were measured were accuracy, the number of true positives, the number of false positives, the number of false negatives, the number of true negatives, and finally the area under the ROC curve. The accuracy of the model is given by the following equation:

$$\texttt{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

Where $\hat{y}$ is the predicted value and $y$ is the corresponding true value. The TP, FP, FN, and TN values are the number of predictions in each category for the training set.

In addition to these values, the ROC curve and the he precision-recall curves were generated. Both require the output sample and the probabilities of each class. Different thresholds are used to compute each point on the curve. While ROC compares the true positive rate and false positive rate, the precision and recall and computed based on the following equations give how each of these values is computed:

$$P = \frac{T_p}{T_p + F_p} \quad R = \frac{T_p}{T_p + F_n}$$

The results show that in terms of accuracy, both the decision tree and the random forest models performed equally well. However, when further is analysis is performed, it can be seen that the random forest model has a higher value for the AUC for the ROC. By this factor, it can be concluded that the random forest is the most ideal model when for training on breast cancer images. The reason that the random forest could have performed better than the decision tree could be due to the fact that random forests account for overfitting on the dataset with the voting from various trees. Since each tree constructed has its own decision boundary and is built on a random sample and a random set of features are considered for splitting, there is diversity among the different trees that are built. This diversity allows for the problem of overfitting to be solved.

The model that performed the worst was the SVM model. This was both in terms of accuracy as well as AUC. Possible reasons for this is an inappropriate kernel function or the dimensionality of the data was not high enough; there were only 30 features that the model was predicted on and a total of 569 cells.

It can also be observed that the linear classifier also performed well in terms of both the accuracy and the ROC. Although there was the highest ROC, the value for the accuracy was not as high.

From Figure 1, the disparity between the ROC curves can be seen. The SVM line is significantly lower than  the other models. From Figure 2, it can be seen that the precision-recall values for each model are closer together. However, even in this case, the SVM model curve is farthest to the left, meaning there would be a lower area for this, just like the ROC.

Overall, the linear model, decision tree, and random forest models had a high accuracy level. However, given the application of these models being diagnosis of cancer cells, the accuracy may not be high enough. In order to achieve a higher accuracy, many factors can change. For example, the penalty function and amount can be changed for the linear model. For the random forest, the number of trees used to predict can be increased and for the SVM a different kernel function can be used.

## Conclusion

From the results, it can be concluded that machine learning algorithms would be useful in determining the diagnosis of breast cancer cells. Although the models didn’t perform well enough to be used directly when classifying cells, the models can be used in order to analyze what areas and which cells could potentially develop cancer. Each model outputs and associated “weight” for each feature. Using this, doctors would be able to determine what features are most influence a cell being malignant or benign.

## Further Research

There are many aspects that allow this research to be continued in various ways. Other tumor cells can be analyzed using a similar approach. Once cells are imaged, the only processing that is required is feature extraction. Any biomedical imaging modality would be appropriate given the images can be analyzed in the appropriate way.

Another way this research can be extended is by using other metrics when analyzing the performance of each model. For example, more analysis can be performed for the precision and recall of each model. Another metric that could be used is computing the average precision of each model.

This research can also be extended by using other model to construct the predictions. For example, neural networks would add a layer of complexity that could outbeat the performance of the random forest model. Another way to extend the research is potentially reorganize the dataset and construct models using an unsupervised learning approach.

## Contributions

Each member contributed an equal amount of time and effort into the research, experiment, and final report. Specifically, Usha was responsible for researching and experimenting with the machine learning algorithms and performance analysis, Tony was responsible for background research and research on the different machine learning algorithms, and Conan was responsible for researching specific classification strategies in machine learning.

## Acknowledgements

We would like to thank Professor Gao and Jorge for leading a great semester and introducing the field of biomedical imaging for us; Thank you for teaching us what we needed to know for this project and provide motivation for further research in this field.

## References

- Amin, J, Sharif, M, Raza, M, Saba, T, Anjum, M 2019, ‘Brain Tumor Detection Using Statistical and Machine Learning Method’, *Computer Methods and Programs in Biomedicine*, vol. 177, pp. 69-79

- Abel-Maksoud, E, Elmogy, M, El-Awadi, R 2015, ‘Brain Tumor Segmentation Based on a Hybrid Clustering Technique’, *Egyptian Informatics Journal,* vol. 16, pp. 71-81

- API design for machine learning software: experiences from the scikit-learn project, Buitinck *et al.*, 2013.

- Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011

- Wolberg, W, Street, N, Mangasarian, O. *Breast Cancer Wisconsin (Diagnostic) Dataset, 2016*