Write a report comparing 5 classification algorithms – SVM, Decision Trees, Random Forrest, Boosted Trees and Neural Network

AI Tech Systems
https://ai-techsystems.com/
Parth Pathak
Student, Bachelor of Technology, Computer Science, School of Engineering and Technology
Ansal University, Gurugram, India
parthpathak1998@gmail.com

Abstract. About 1.7M new breast cancer cases were diagnosed in 2012. As of 2018, nearly 12.4% women in US are expected to develop invasive breast cancer over their lifetime. Mammography has always been the most effective technique for the screening of breast cancer. But, the low positive predictive value of breast biopsy which results from the interpretation of mammogram leads to nearly 70% unnecessary biopsies with benign outcomes. To solve this problem, supervised machine learning classification algorithms can be applied to develop a machine learning models which can predict the rigorousness of a mammographic mass with the help of BI-RADS attributes and the patient's age.

830 records with a total of 6 attributes were recorded in the dataset to check the nature of mammographic masses. The study investigates 5 different classification models Each model is evaluated on the basis of confusion matrix, standard metrics of Accuracy, Precision, Recall and F-measure.

The research work was aimed to assess performance of various classification algorithms introduced in recent years to design a predictive model for breast cancer identification on data obtained from full field digital mammograms.

Keywords. Benign, Malignant, Mammographic Mass, Classification, Prediction, Confusion Matrix, Principal Component Analysis, Breast Cancer, Mammography, Mammograms.

1 Introduction

The present era tells us that breast cancer is actually the most commonly occurring cancer in women. It is also the second-most common cancer prevalent in today's time. Over 2 million new cases of this cancer have been recorded in the year 2018. For the year of 2018, Belgium had the highest rate of breast cancer with age-standardized rate of 100,000 equal to 113.2, followed by Luxembourg at 109.3. A mammographic mass may potentially be an abnormality on a mammogram which may or may not always be cancer. Mammography or mammogram refers to the X-ray of the breasts to examine them for diagnosis and screening. The aim of mammography is to identify breast cancer at its early stages, typically through characteristic mass detection (or micro-calcifications).

However, interpretation of mammogram generally leads to about 70% unnecessary breast biopsy which results in benign outcomes. To deal with the situation, technology has come into role with the development of computer-aided diagnosis (CAD) systems in the last years. These systems have aided the physicians to decide on their decision of whether to perform breast biopsy or not.

Our dataset predicts rigorousness (benign or malignant) of mammographic mass lesion on the basis of BI-RADS attribute and patient's age. Each record in the dataset has an associated BI-RADS assessment where it's values range from 1 to 5, 1 being "definitely benign"; 5 being "highly malignant". This research paper highlights the applications and effectiveness of seven different classification algorithms (including both machine learning and deep learning techniques) for the prediction of breast cancer severity to help experts in the healthcare domain.

2 Methods

The mammography provided us with a total of 6 attributes, information of which is given in Table 1. The total number of instances recorded is 830. The attribute “Severity” was identified as the goal field or predictable attribute with 0 referring to patient having benign and/or 1 referring to patient having malignant. The “BI-RADS” attribute is an ordinal attribute with values ranging from 1 to 5, with 1 being “definitely benign”; and 5 being “highly malignant”. The “Shape” attribute refers to the shape of the mass where 1 represents “round” shape, 2 represents “oval” shape, 3 represents “lobular” shape and 4 represents “irregular” shape. The “Margin” attribute is another nominal attribute depicting the margin of the masses where 1 is for “circumscribed”, 2 is for “lobulated”, 3 is for “obscured”, 4 is for “ill-defined” and 5 is for “spiculate”. Density of the mass is depicted by “Density” attribute where 1 is for “high density”, 2 is for “iso”, 3 is for “low” and 4 is for “fat-containing”.

S. No.	Attributes	Description	Types
1	BI-RADS	Breast Imaging-Reporting & Data System (1-5)	Ordinal
2	Age	Age of patient in years	Integer
3	Shape	Shape of mass (round, oval, lobular, irregular)	Nominal
4	Margin	Margin of mass (circumscribed, microlobulated, obscured, ill-defined, spiculated)	Nominal
5	Density	Density of mass (high, iso, low, fat-containing)	Ordinal

Table 1 – List of attributes with their description

On the basis of above five attributes, I predict the “Severity” attribute which gives results in the form of two classes:

S. No.	Class	Type	Value
1	Benign	Binominal	0

2	Malignant	Binominal	1

Table 2 – Predictable attribute "Severity"

I compare and evaluate the behaviour of different classification models based on Confusion Matrix. It is a contingency table that contains information about actual and predicted classifications done by a classification model and also aids us in calculating standard metrics of accuracy, precision, recall and F-measure.

Predicted Class	**Positive**	**Negative**
Positive	True Positive	False Negative
Negative	False Positive	True Negative

Table 3 – Confusion Matrix

Table 3 shows a confusion matrix where the result contains two classes. The total number of correct and incorrect predictions are abridged with count values and broken down by each class.

- True Negative(TN) – Number of correct predictions that a record is negative
- True Positive (TP) – Number of correct predictions that a record is positive
- False Negative(FN) – Number of incorrect predictions that a record is negative
- False Positive(FP) – Number of incorrect predictions that a record is positive

3 Formulas for Evaluating Model's Performance

For a classification problem consisting of two classes:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1-Score = (2 * Precision * Recall) / (Precision + Recall)

4 Results

Performance analysis on a well-preprocessed data containing only the strong attributes (estimated using Principal Component Analysis) has shown that SVMs has outperformed all other classification algorithms including Artificial Neural Network, with overall accuracy of 80.72%. When it comes to predicting benign, Neural Network has proved its performance as the best classifier by predicting with accuracy of 82.3%. The research was designed to evaluate and compare the performance of Support Vector Machine, Decision Tree Induction, Random Forest Classification, Neural Networks and boosted trees.

5 Discussion

We make use of various classification models to predict the nature of cancer. Models used are – Support Vector Machine, Decision Tree, Random Forrest, Boosted Trees and Neural Networks.

To have a clearer understanding to the results derived from all these models we have used confusion matrix as the one shown in Table 3. Further the performance analysis is carried on by measuring some other factors like – Precision, Recall, F1-Score and Support.

Now as we look at confusion matrix for each model,

Support vector machine predicted benign with an accuracy of 73.07% and predicted malignant with an accuracy of 79.5%. Overall the accuracy of this model is slightly lower than previous one with 76.5%.

Decision Tree was able to predict benign with an accuracy of 78.35% and malignant with an accuracy of 68.18% with an overall accuracy of 71.08%.

Random Forrest was able to predict benign with an accuracy of 73.07% and was able to predict malignant with an accuracy of 72.7% with an overall accuracy 72.89% slightly higher than previous model.

Neural Network was trained consisting of two hidden layers. The activation function used at hidden layer is Rectified Linear Unit (ReLU), whereas Sigmoid function has been applied at the output layer. The neural network was trained on a batch size of 5 and 100 epochs. The resulting network predicted benign with an accuracy of 82.3% and malignant with an accuracy of 75.3% giving an overall accuracy of 78.9%.

Table 4 summarizes the confusion matrix for all models.

Algorithm	**Actual Class**	**Predicted Class**	
		0	1
1.Boosted Trees	0	65	14
	1	22	65
2.Support Vector Machine	0	65	14
	1	20	67
3.Decision Tree	0	63	16
	1	26	61
4.Random Forrest	0	66	13

	1	21	66
5.Artificial Neural Networks	0	75	15
	1	13	63

Table 4 – Confusion Matrix for various models

Table 5 shows the calculated evaluation metrics we have chosen to compare performance of all the seven models.

Model	Precision	Recall	F1-Score	Support
1.Boosted Trees	0.79	0.78	0.78	166
2.Support Vector Machine	0.80	0.80	0.80	166
3.Decision Tree	0.75	0.75	0.75	166
4.Random Forrest	0.80	0.80	0.80	166
5.Artificial Neural Network	0.55	0.55	0.55	166

Table 5 – Performance measures for comparison of models

From all these observations we can infer that Neural Networks outperformed all other models by predicting benign with the highest accuracy of 82.3% and malignant was predicted best by SVM with an accuracy of 86.3%.

6 Conclusion

The analysis shows that the Neural Network performed better in predicting benign while SVM gave the best results for predicting malignant cancer. Talking about overall performance, SVM provides the best accuracy of 80.72%, followed by Neural Networks which provides overall accuracy of 78.9%. Now since benign is more common than malignant and number of benign cases are more than malignant ones, it is advised to adopt Neural Network for the prediction.

7 Conflicts of Interest

There exists no conflict of interest for the present study.

8 References

1. Breast Cancer Statistics, American Institute for Cancer Research |
   https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics
2. Mammograms, Cleveland Clinics |
   https://my.clevelandclinic.org/health/diagnostics/4751-what-if-they-think-they-see-something-on-my-mammogram
3. Mammography, Healthline |
   https://www.healthline.com/health/mammography
4. Mammographic Mass Data Set, UCI |
   http://archive.ics.uci.edu/ml/datasets/mammographic+mass
5. Mammography, Wikipedia
   https://en.wikipedia.org/wiki/Mammography