

Using AI to Improve Breast Cancer Diagnosis and Its Effects on Health Care Cost

By Jeffrey Zhang

AUTHOR BIO

Jeffrey Zhang is a student at Thomas S. Wootton high school in Maryland. Since his childhood he has always been interested in technology. He is particularly interested in Artificial Intelligence and the social and economic problems it could help to solve. Recently, his family went through an urgent medical episode and was burdened with a large bill. Partly because of this incident, he attended a research seminar on healthcare and data science held by Professor Ramezani from UCLA. This has led him to this research paper on AI and healthcare costs.

ABSTRACT

AI and Machine learning have been taking over the modern world. Machine learning is a branch of artificial intelligence (AI) which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. It is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, and to uncover key insights in data mining projects. This study evaluated machine learning and its applications in healthcare, especially in the diagnosis testing process. Decision Tree model, Random Forest model and Support Vector Classifier model were tested and compared. This study demonstrated that machine learning models can be used to optimize the diagnosis testing process. The new and improved breast cancer test is shown to be as accurate if not more than before. Test efficiency was improved by 45%. The application of machine learning should be encouraged, and such practice will lead to significant cost saving.

Keywords: Machine Learning, HealthCare, Breast Cancer, Decision Tree, Random Forest, SVC Model.

INTRODUCTION

Healthcare costs keep increasing at an alarming rate in the U.S. The Congressional Budget Office calculates that federal outlays for Social Security, Medicare, and Medicaid will rise from 6-1/2 percent of GDP in 2003 to 12-1/2 percent of GDP by 2050 (Follette, G., & Sheiner,). The share of GDP accounted for by health care spending rose from 4.5% in 1940 to 12.2% in 1990. In 2005 health care spending was nearly \$2 trillion, or \$6,697 per capita, which represents 16% of GDP (Neeraj Sood). The ever so increasing costs of healthcare are becoming more and more out of hand.

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. (IBM, 2022) Machine learning has been taking over the modern world in many aspects. Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, and to uncover key insights in data mining projects. (Wikipedia)

What role can AI play in consumer healthcare? Can it make healthcare cheaper and more sustainable? This study evaluated machine learning applications in the diagnosis testing process. Decision Tree model, Random Forest model and Support Vector Classifier model were tested and compared. This study demonstrated that machine learning models can be used to optimize the diagnosis testing process. The improved breast cancer test is shown to be as accurate if not more than before. Test efficiency was improved by 45%.

Procedure

Data

The Breast Cancer Wisconsin data from UCI Machine Learning Repository (Wolberg, William) is used in this study. This data is widely used in breast cancer diagnosis research (Street, W.H, Mangasarian)

Attributes are computed from a digitized image of a fine needle aspirate of a breast mass. They describe characteristics of the cell nuclei present in the image. This data set had a total of eleven attributes including one for identification and one for classification. (UCI Machine Learning Repository)

The 9 attributes that can be used in diagnosis were:

1. Clump Thickness
2. Uniformity of Cell Size
3. Uniformity of Cell Shape
4. Marginal Adhesion
5. Single Epithelial Cell Size
6. Bare Nuclei
7. Bland Chromatin
8. Normal Nucleoli
9. Mitosis

The Bare Nuclei had some missing values, so the data was cleaned by removing the rows that had missing data values.

The data for the 9 attributes are given values 1-10 and the “Class” column is given values 2 for benign and 4 for malignant.

	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
0	5	1	1	1	2	1	3	1	1	2
1	5	4	4	5	7	10	3	2	1	2
2	3	1	1	1	2	2	3	1	1	2
3	6	8	8	1	3	4	3	7	1	2
4	4	1	1	3	2	1	3	1	1	2
5	8	10	10	8	7	10	9	7	1	4
6	1	1	1	1	2	10	3	1	1	2
7	2	1	2	1	2	1	3	1	1	2
8	2	1	1	1	2	1	1	1	5	2
9	4	2	1	1	2	1	2	1	1	2
10	1	1	1	1	1	1	3	1	1	2
11	2	1	1	1	2	1	2	1	1	2
12	5	3	3	3	2	3	4	4	1	4

Figure 1: Data Sample

AI Models

Three AI models were tested in this study.

1. **Decision Tree:** Decision tree is a type of machine learning that makes predictions based on previous data and questions. Decision tree tends to imitate human thinking so it's easy for humans to understand and interpret the results. (W3Schools Decision Tree)

```
# Classifier imports
from sklearn import tree

# Performance metrics
from sklearn.metrics import accuracy_score, classification_report

dt = tree.DecisionTreeClassifier()

dt.fit(train_most_x, train_most_y)
test_most_y_dt_model = dt.predict(test_most_x)
print("DecisionTree Accuracy :", accuracy_score(test_most_y, test_most_y_dt_model))
```

Figure2: Decision Tree Model

2. **Random Forest:** Random Forest model is essentially a classification algorithm that consists of many decision tree models which use bagging and feature randomness when building each individual tree. Bagging is a learning technique that helps improve the performance and accuracy of machine learning algorithms. (Yiu, Tony)

```
from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier()

rf.fit(train_most_x, train_most_y)
test_most_y_rf_model = rf.predict(test_most_x)
print("RandomForest Accuracy :", accuracy_score(test_most_y, test_most_y_rf_model))
```

Figure3: Random Forest Model

3. **SVC:** Support Vector Classifier, is a supervised machine learning algorithm typically used for classification tasks. SVC works by mapping data points to a high-dimensional space and then finding the optimal hyperplane that divides the data into two classes (Datatechnote)

```
from sklearn.svm import SVC

rbf = SVC(kernel='rbf')
rbf.fit(train_most_x, train_most_y)

test_most_y_rbf_model = rbf.predict(test_most_x)
print("SVM with Radial Basis Kernel Accuracy :", accuracy_score(test_most_y, test_most_y_rbf_model))
```

Figure 4: SVC Model

The python module is imported from the scikit-learn library (Scikit-learn)

Attribute Selection

Attribute selection is the process of reducing the number of input variables when developing a predictive model.

It is desirable to reduce the number of attributes to both reduce the computational cost and, in some cases, to improve the model performance (Brownlee, Jason).

Data was split into training data and testing data. Correlation coefficient of each individual attributes was calculated. The high correlated attributes would be better predictors. SVC model was used for attributes selection.

SVC Model Accuracy	SVC Model Accuracy
5 Attributes	All Attributes
0.9766	0.9708

Figure 5: Attribute Selection

SVC model with 5 attributes produced accuracy of .9766 and SVC model with all attributes produced accuracy of .9708.

It turned out that removing the attributes that added little value to classifying the tumor actually raised the accuracy. This led to these selected attributes.

Selected Attributes

- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape
- Bare Nuclei
- Normal Nucleoli

Less Impactful Attributes:

- Marginal Adhesion
- Single Epithelial Cell Size
- Bland Chromatin
- Mitosis

Model Comparison

A new dataset was created with the selected attributes, and the model was tested for 30 trials.

Decision Tree	Random Forest	SVC
92.40%	96.78%	97.08%
93.57%	96.49%	97.08%
92.69%	96.49%	97.08%
92.40%	95.91%	97.08%
92.98%	96.49%	97.66%
92.69%	95.91%	97.66%
92.40%	96.49%	95.91%
92.11%	96.49%	97.66%
93.86%	97.08%	96.78%
92.98%	96.78%	97.37%
92.40%	95.91%	96.20%
93.27%	96.19%	96.78%
93.27%	95.91%	97.37%
92.11%	95.91%	96.78%
92.40%	96.49%	96.78%
93.27%	96.49%	97.66%
92.69%	96.49%	95.03%
92.69%	96.49%	98.25%
93.27%	96.49%	97.37%
93.27%	95.61%	97.66%
92.40%	95.61%	98.25%
93.57%	95.61%	96.78%
92.40%	96.49%	96.49%
91.81%	96.78%	95.91%
92.98%	95.91%	97.08%
92.69%	95.91%	96.49%
92.40%	96.78%	96.49%
92.69%	96.49%	96.78%
92.69%	96.78%	96.78%
92.77%	96.32%	96.98%

Figure 6: Model Results

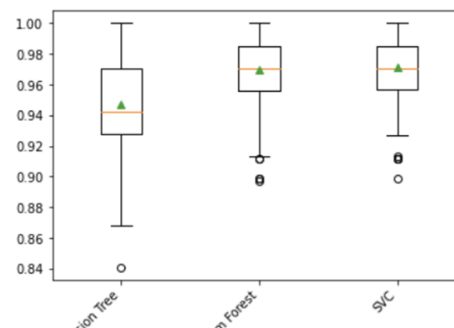


Figure7: Box Plot of Model Accuracy

Decision Tree model had an average accuracy of about 92.77%

Random Forest model had an average accuracy of 96.32%

SVC model had an average accuracy of 96.98%.

All models produced consistent results and performed well. SVC model performed best.

Discussion

This study demonstrated that machine learning models improve breast cancer diagnosis accuracy and efficiency. Reducing the number of attributes needed from 9 to 5 has a positive effect on the accuracy. Accuracy was increased from 97.1% to 97.7%.

Which leads to our other question: Is prediction power greater than or the same as clinical prediction accuracy? In a paper titled “Wisconsin Well Woman Program Procedure Code Quick Reference”, it shows that the clinics test multiple aspects of the tumor so the least number of tests while still being able to maintain the highest possible accuracy is the best outcome (Wisconsin Well Woman Program).

For the sake of simplification, let us assume a similar cost for each attribute test. If so, testing costs can be decreased by 45%. Another side benefit would be speed, the shorter time due to the fewer required tests means a quicker turnaround. Cancer, like many other medical conditions, is time sensitive. Early diagnosis improves cancer outcomes by providing care at the earliest possible time. (WHO, 2022)

This study demonstrates that machine learning models improved breast cancer diagnosis accuracy and efficiency. The number of attributes measured was reduced from 9 to 5. Allowing quicker testing turnarounds and less testing all together. And the prediction accuracy improved from 97.1% to 97.7%. The testing process is a large contributor to healthcare costs and through machine learning testing costs of breast cancer can be decreased by about 45%.

The application of machine learning should be encouraged, and such practice will lead to significant cost saving.

Conclusion

REFERENCES

- Brownlee, Jason
How to Choose a Feature Selection Method For
Machine Learning
<https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/>
- Datatechnote Classification Example with Linear
SVC in Python
<https://www.datatechnotes.com/2020/07/classification-example-with-linearsvm-in-python.html>
- Follette, G., & Sheiner, L.

The Sustainability of Health Spending Growth.
<https://www.federalreserve.gov/pubs/feds/2005/200560/200560pap.pdf>
- IBM 2022
Machine Learning
<https://www.ibm.com/cloud/learn/machine-learning>
- Mangasarian, Breast cancer diagnosis and prognosis
via linear programming.
- Operations Research, 43(4), pages 570-577,
July-August 1995.
- Neeraj, Sood
The Effect of Health Care Cost Growth on the U.S.
Economy
<https://aspe.hhs.gov/reports/effect-health-care-cost-growth-us-economy-0>
- Scikit-learn Machine Learning in Python
<https://scikit-learn.org/stable/>
- Street, W.H.
Nuclear feature extraction for breast tumor diagnosis.
IS&T/SPIE 1993 International Symposium on
Electronic Imaging: Science and Technology, volume
1905, pages 861-870, San Jose, CA, 1993.
- UCI Machine Learning Repository.
[https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+\(diagnostic\)](https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic))
- W3Schools Decision Tree
https://www.w3schools.com/python/python_ml_decision_tree.asp
- Wolberg, William, Street, W. & Mangasarian, Olvi.
(1995). Breast Cancer Wisconsin (Diagnostic). UCI
Machine Learning Repository.
WHO 2022
<https://www.who.int/activities/promoting-cancer-early-diagnosis#:~:text=Early%20diagnosis%20improves%20cancer%20outcomes,different%20strategy%20than%20early%20diagnosis.>
- Wikipedia, Machine Learning
https://en.wikipedia.org/wiki/Machine_learning
- Wisconsin Well Woman Program Procedure Code
Quick Reference. (n.d.).
<https://www.dhs.wisconsin.gov/publications/p4/p43029a.pdf>
- Yiu, Tony, Random Forest
<https://towardsdatascience.com/understanding-random-forest-58381e0602d2>