Machine Learning Methods for Breast Cancer Diagnosis

Matthew Lee[1] and Zhaonan Sun[#]

[1]Kristin School, Albany, Auckland, New Zealand
[#]Advisor

ABSTRACT

As many modern diseases begin to surface especially as of late, such as the Ebola and COVID-19 epidemics, scientists have begun developing new and innovative tactics to combat them. While new medicine and vaccines may be developed, one area that needs special attention is the diagnosis of diseases – this is because without a proper and speedy diagnosis, scientists wouldn't be able to detect diseases, rendering treatment ineffective. Scientists have begun using machine learning algorithms to help ensure an accurate and speedy diagnosis. One specific disease that has seen frequent testing around machine learning diagnosis is breast cancer. Breast cancer is one of the deadliest and common cancers around the world for women, and due to its effects, the doctrine of speed in diagnosis is essential. This study will attempt to find out, out of three machine learning algorithms (neural networks, logistic regression and K-nearest neighbours), which one is the most effective at diagnosing breast cancer using the Wisconsin Breast Cancer Dataset. Results suggest that neural networks perform the best in diagnosing breast cancer, however only by a small margin compared to other results.

Introduction

Breast cancer is the most prevalent form of cancer and the second leading cause of cancer deaths and amongst women in the world. In the year 2021, nearly 300,000 cases were diagnosed and 44,000 deaths were reported in just the United States alone, according to the American Cancer Society (2021). It develops mainly in women and a very small percentage of men when certain breast cells begin to grow abnormally, causing them to grow more and form lumps (Boughey, n.d.)

Due to its prevalence and its relatively high fatality rate, a rapid diagnosis and identification of malignant cases is essential to combating breast cancer and preventing deaths. As such, breast cancer has – especially as of late – been a widely focused topic in the field of machine learning, where machine learning techniques are being applied to ensure a rapid and accurate diagnosis.

However, due to the variety of techniques that are present within machine learning, it is important to establish which ones are the most effective at diagnosing breast cancer. While manual diagnosis within laboratories is an option, due to the structural similarities between benign and malignant tumours there runs the risk of the diagnosis being both slow and inaccurate. This is very clearly shown by the fact that a study conducted by Yale University (Hathaway, 2020) shows that a high estimate of 250,000 people die each year from accidents by medical professionals – surpassing that caused by both Alzheimer's and Diabetes combined according to the United States CDC (2020). Therefore, it is abundantly clear why there is a need for better methods of diagnosis.

While various machine learning techniques have been applied to this problem, it is unclear whether there is a specific model that works for breast cancer the best. Therefore, the objective of this study is to compare the prediction power of different models in diagnosing breast cancer. In order to do so, this study has selected neural networking, logistic regression and K-nearest neighbours. Previous research papers such as (Potdar,

2016) and (Sharma Et Al, 2017) all have done similar studies as this paper. Both, however, did not use neural networking as one of the possible methods, and the reason the three diagnostic methods are used here is to compare them towards neural networking. The objective of this essay is to clearly identify, out of the three models, which is the most effective at diagnosing breast cancer.

Experimental DATA

The data used in this study was sourced from the Wisconsin Breast Cancer Database, and all coding was done with Python version 3.10.2. The dataset contained 33 columns in total, among which 32 columns are patient specific information and the last one indicates the diagnostic result. Patient specific information includes concavity mean, smoothness mean, radius mean, and area mean etc.

The dataset was imported to python using the Pandas library. The column containing the patient IDs was removed. The diagnostic result column was transformed to a binary label with malignant being 1 and benign being 0. All other columns were treated as predictors in the model development process. The dataset was split into training and validation sets, with 75% being training and 25% being validation. The numerical predictors were then scaled so that they are all on the same order of magnitude. After that, the dataset was considered processed and ready for training.

A sequential model with six layers in total was used to apply the neural network onto the dataset. Three used the 'relu' function, two used the 'tanh' function, and one used the sigmoid activation function (Fig 1).

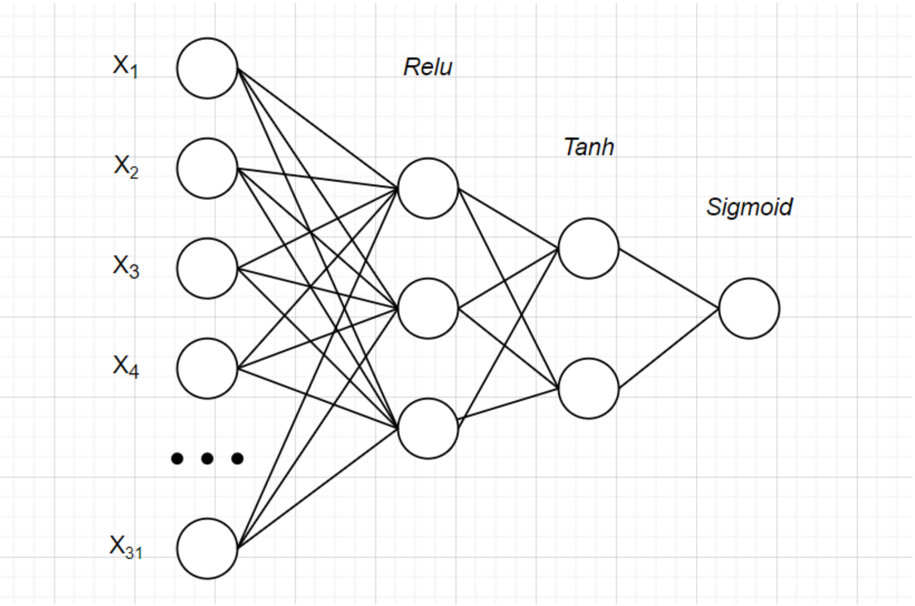


Figure 1. Neural Network Illustration

The five hidden layers were simply removed, and a sigmoid activation function kept for logistic regression. Both models were used on an epoch of 400. The K-nearest neighbours code, meanwhile, was imported via the scikit-learn package. and was examined when K = 5 and had the current dataset applied to it.

After model training, validation was performed by applying the trained model to predict the diagnostic result of the patients in the validation set. Precision, recall, and accuracy were used to evaluate the prediction power of the models (Equations 1, 2, 3).

$$Precision \ = \frac{True\ Positive}{True\ Positive + False\ Positive}(1)$$

$$Recall\ = \frac{True\ Positive}{True\ Positive + False\ Negative}\ (2)$$

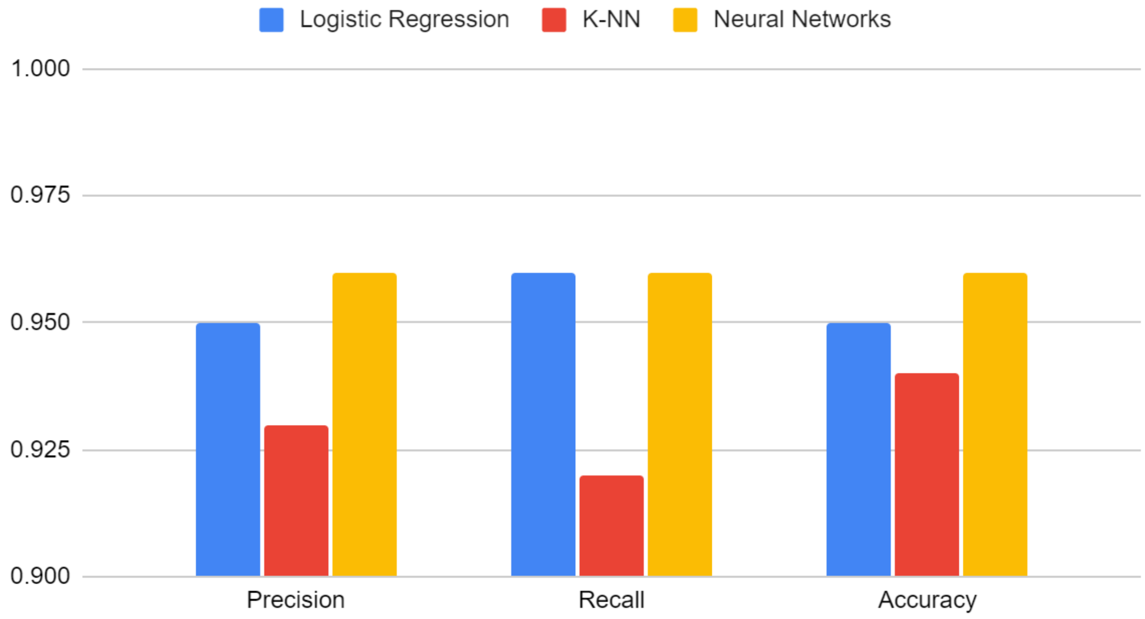
$$Accuracy\ = \frac{True\ Negative + True\ Positive}{True\ Positive + False\ Positive + True\ Negative + False\ Negative}\ (3)$$

Evaluation was done on both training and testing datasets.

Results

After running the three models, it can be seen that they all possessed similar results. In order to observe how well each model did, three metrics were considered: accuracy, precision and recall – the latter two of which, due to separate training and validation sets being used, are calculated via the weighted average. All results were taken from the validation set (Table 1, Figure 2).

Table 1. Performance of the models in cancer diagnosis



Metric.	Logistic Regression	K-Nearest Neighbours	Neural Network
Precision	0.95	0.93	0.96

Recall	0.96	0.92	0.96
Accuracy	0.95	0.94	0.96

Figure. 2 Comparison between the performance of the machine learning techniques used in this study.

Discussion

After comparing the various metrics used to determine how good a prediction is, the Neural Network model has proven itself to be the most superior out of the three models, followed by Logistic Regression and KNN. This confirms the study's hypothesis, as Neural Networks are known to be more powerful than both Logistic Regression and KNN, hence achieving better results. This is most likely due to the fact that the more layers a model has (or more complex a model is), the stronger its computing power. The results reflect this as, within the code, Logistic Regression is essentially a Neural Network model without any hidden layers, and the Neural Network performed better. KNN is also widely regarded as a simple model (Yildrim, 2020), which further matches this statement. There doesn't appear to be any correlation on how well each model did between each of the three metrics, with results simply fluctuating.

However, the predictions achieved have shown some unexpected results. It is to be noted that the differences between the results are very minimal. For example, for the accuracies of different models, the Neural Network had an accuracy of 0.96, followed by Logistic Regression with 0.95 and KNN with 0.94. The small difference between the results suggest that the dataset may be too simple to be analysed via machine learning. This is further reinforced by the fact that previous studies using the same algorithms on different datasets have typically shown larger differences of 6-8% (Potdar, 2016), (Sharma et al., 2017).

In conclusion, this study has shown that neural networks are superior to both logistic regression and KNN when it comes to processing numerical data. Using the Wisconsin Breast Cancer Dataset, neural networks have managed to achieve an accuracy of 0.96, compared to that of logistic regression (0.95) and KNN (0.94) on their respective validation sets. This study has also shown that, while it is quite easy and simple to achieve a high accuracy, most machine learning algorithms still aren't able to achieve a perfect accuracy of 1, despite the neural network in this case being immensely powerful with 6 hidden layers, suggesting that it may be impossible in any case. However, the incredibly high accuracy may also simply suggest that the dataset is too simple to be calculated by powerful algorithms such as neural networks. Therefore, it may be more suitable to only use algorithms such as logistic regression to save computing power.

As such, there are definitely some limitations that would need to be addressed in order to improve the current study. The dataset is extremely simple and could be made more suitable by further extending the dataset by replicating some parts or simply using another dataset as a whole. Another limitation would be that the data range may be too close together, so scaling the dataset may not have proven to be entirely necessary. Future studies could improve and extend this study by introducing other machine learning algorithms for more comparison, potentially finding out if there is another algorithm that is more suitable than neural networks. If this model were to be used, then future studies could try to find a model that balances complexity with accuracy, trying to strive for a middle ground between logistic regression and neural networks if possible.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

Boughey, J. C. (n.d.). *Breast Cancer: Symptoms and Causes*. Mayoclinic. Retrieved April 4, 2022, from https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470

Cancer Facts & Figures 2021. (2021). American Cancer Society. Retrieved April 4, 2022, from https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2021/cancer-facts-and-figures-2021.pdf

Deaths and Mortality. (2020). Faststats. Retrieved April 4, 2022, from https://www.cdc.gov/nchs/fastats/deaths.htm

Hathaway, B. (2020, January 28). *Estimates of preventable hospital deaths are too high, new study shows*. YaleNews. Retrieved April 4, 2022, from https://news.yale.edu/2020/01/28/estimates-preventable-hospital-deaths-are-too-high-new-study-shows

Kumar, M., & Choi, M. (Eds.). (n.d.). Breast Cancer Wisconsin (Diagnostic) Data Set. *Kaggle*. Retrieved April 4, 2022, from https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data?select=data.csv

Potdar, K. (2016, September). *A Comparative Study of Machine Learning Algorithms applied to Predictive Breast Cancer Data*. Retrieved April 4, 2022, from https://www.researchgate.net/publication/308725638_A_Comparative_Study_of_Machine_Learning_Algorithms_applied_to_Predictive_Breast_Cancer_Data

Sharma, A., Kulshrestha, S., & Daniel, S. (2017, December). *Machine learning approaches for breast cancer diagnosis and prognosis*. Retrieved April 4, 2022, from https://www.researchgate.net/publication/322944323_Machine_learning_approaches_for_breast_cancer_diagnosis_and_prognosis

U.S. Breast Cancer Statistics. (n.d.). Breastcancer.org. Retrieved April 4, 2022, from https://www.breastcancer.org/facts-statistics

Yildirim, S. (2020, March 1). *K-Nearest Neighbors (kNN) — Explained*. Towards Data Science. https://towardsdatascience.com/k-nearest-neighbors-knn-explained-cbc31849a7e3