Breast Cancer 6th Stage Prediction Based on Machine Learning Models

Junpeng Yang

College of Science, Purdue University, 610 Purdue Mall, West Lafayette, IN 47907, USA

yang1828@purdue.edu

Abstract. The differences between each 6th stage of the breast cancer are subtle, and doctors' judgement alone is not sufficient to determine the 6th stage accurately. 6th stage is the different levels of breast cancer development and it represents the current status of the cancer. Therefore, it is crucial to determine it correctly in order to conduct corresponding treatments. The incorrect categorization of the 6th stage and misuse of treatments can be catastrophic, and there are currently no such models to help doctors predicting the 6th stage. The dataset Seer Breast Cancer Data is used which include features like race, t-stage, n-stage, etc. This paper proposed to use random forest and K Nearest Neighbor (KNN) methods to build models and use features related to the patients and their cancer as training data. The random forest model achieved a predictive result of 99% for precision, recall, and f1 score after data normalization. The only mistake this model made is when differentiating stage IIIA and IIIB. The KNN model achieved an accuracy of 95% after normalization. The result shows that Random Forest model is best suited for predicting the 6th stage. The random forest model with 99% accuracy can effectively help doctors determine the 6th stage when they are having difficulties.

Keywords: Breast cancer diagnosis using AI, improved breast cancer diagnosis accuracy.

1. Introduction

Breast cancer is a type of cancer that develops from the breast tissue [1]. It is the second most common cause of cancer death in women, and over 250,000 women are diagnosed with breast cancer every year in the United States alone [2]. There are two major types of breast cancer namely Invasive Ductal Carcinoma (IDC) and Invasive Lobular Carcinoma (ILC). Although male also have a slight chance of getting breast cancer, female are the most common victims of it. The risk factors of breast cancer are plenty, such as high body mass index, first birth at age greater than 30, and family history [2]. When compared to other types of breast cancer, breast cancer has a relatively low fatality rate of about 2.6%, and this percentage is continuing to fall as screening technology and improved awareness [3]. Despite the low death rate, the most common question that breast cancer patients will consider is whether they would fall into the 2.6% death category. Therefore, an accurate prognosis of the breast cancer is very important regarding the patient's survival chance. Different stages require different treatments; hence, a correct perdition of the breast cancer stage based on advanced models is crucial for surviving the disease.

The stage group is a combination of tumor, node, and metastasis, and it is able to reflect the development of breast cancer so that doctors are able to apply proper treatment to it. The preferred methods for doctors to determine the stage are imaging methods, such as mammography, ultrasonography, magnetic resonance imaging [4]. These methods will generate an image of the breast cancer which will be used by doctors to determine the stage of the cancer. The differences between some stages may be minor, yet the required treatment can have variations. For example, in stage IIB and IIIA: the tumor is bigger than 50mm, IIB has not spread to the axillary node (i.e., T3, N0, M0), but IIIA has spread to the axillary node (i.e., T3, N1, M0) [5]. The differences are so small that could lead to an incorrect diagnosis since image alone is not sufficient for doctors to determine the precise stage. Machine learning has been broadly used in breast cancer field in the recent years. Different machine learning algorithms and models are created to help improving the accuracy of diagnosis and prognosis of breast cancer. Like [6], discusses the application on breast cancer and the authors explore how different machine learning skills can be applied on different aspects of breast cancer treatments and research. Similarly, Wang and his colleges focus on whether the patients have breast cancer or not using machine learning techniques while this research will focus on those who have breast cancer, but uncertain of which stage they are on [7].

Unlike some studies that focus on a broader scope of machine learning applications on breast cancer, this research aims specifically to design models that are capable of determine the stage group of breast cancer based on information known about the patients and the condition of their cancer [6, 7]. This research uses the dataset "SEER Breast Cancer Data". Based on parameters such as tumor size, T stage, and number of positive regional node, machine learning models are built to predict the breast cancer group stage a certain patient is in. This research focuses on machine learning algorithms e.g., Support Vector Machines (SVM) [8], KNN [9] and random forest [10]. These methods are conducted separately and used to build models to train the dataset and obtain accuracies. The parameter "6th stage" is used as target, and other parameters are used as features of the training model. An accurate prediction of the prognosis can provide useful assistance for doctors to proceed treatment.

2. Method

2.1. Dataset information and preprocessing

In this research, Seer Breast Cancer Data from Kaggle is used to train and test the prediction models [11]. This dataset was obtained from SEER program of NCI which provides population-based information on cancer statistics [11]. It includes female patients diagnosed with IDC and ILC from 2006 to 2010 [11]. There is a total of 4,024 patients included in this dataset, and a total of 16 features are obtained, such as age, race, and information regarding the patient's breast cancer.

Most of the features are represented in string format and strings cannot be used to train machine learning models. Therefore, label encoder is used to transform strings to integers where every integer represents a status of a feature. For the example of marital status, 0 represents divorced, 1 represents married, 2 represents separated, 3 represents single, 4 represents widowed. The detailed feature label encoder is shown in the Table 1. Then, the dataset is split into a train and test subset according to the ratio of 7:3.

Table 1. LabelEncoder references

	0	1	2	3	4
Race	Black	Other	White		
Marital Status	Divorced	Married	Separated	Single	Widowed
T Stage	T1	T2	T3	T4	
N Stage	N1	N2	N3		
6th Stage	IIA	IIB	IIIA	IIIB	IIIC
Differentiate	Moderately	Poorly	Undifferentiated	Well	
A Stage	Distant	Regional			
Estrogen Status	Negative	Positive			
Progesterone Status	Negative	Positive			
Status	Alive	Dead			

2.2. Proposed approach – random forest

The goal of this research is to classify the patient's 6th stage so that corresponding treatments can be used. In order to find out which 6th stage the patient belongs to, classification algorithms like random forest are needed. A random forest is consisted of many individual decision trees which each one maps a process of deciding an output. Each decision tree will output a predicted value and the value that appears the most will become the result of a random forest.

In order to generate the most reliable results, the dataset was normalized first. Standard scaler was used to normalize the data. Random forest parameters are also adjusted to produce the most accurate result. The most important parameter is n_estimators which decides the number of decision trees in a random forest. After adjusting the n_estimators value, the change in accuracy is very subtle. Therefore, it is unnecessary to increase the value of n_estimators. Then, the normalization scaler is changed to min max scaler and the steps above are repeated. However, changing the scaler to min max scaler decreased the prediction accuracy.

2.3. Proposed approach – K Nearest Neighbor

K Nearest Neighbor (KNN) algorithm is also a very useful algorithm when the task is to predicting results. KNN predicting result by calculating the distance between the given data point and k nearest neighbors from different classes that has been classified. The class that appears the most in the data point's k neighbor will be the class this data point belongs to.

Standard scaler was used first in KNN prediction. Since there are five different 6th stages included in the dataset, n_neighbors was also set to the same number. After n_neighbors was increased, the accuracy stays unchanged. However, the highest accuracy is achieved after decrease the original n_neighbors value. Then, the normalization scaler is changed to min max scaler and the steps above are repeated. In contrast to random forest algorithm, using min max scaler increased the accuracy. N_neighbors were increased, but the accuracy was lowered as n_neighbors increased.

3. Results and Discussion

3.1. Random forest model

Table 2. Random Forest Classification Report (StandardScaler)

	precision	recall	f1 - score	support
0	1.00	1.00	1.00	389
1	0.99	0.97	0.98	344
2	0.96	1.00	0.98	310
3	1.00	0.67	0.80	18
4	1.00	1.00	1.00	147
accuracy			0.99	1208
macro avg	0.99	0.93	0.95	1208
weighted avg	0.99	0.99	0.99	1208

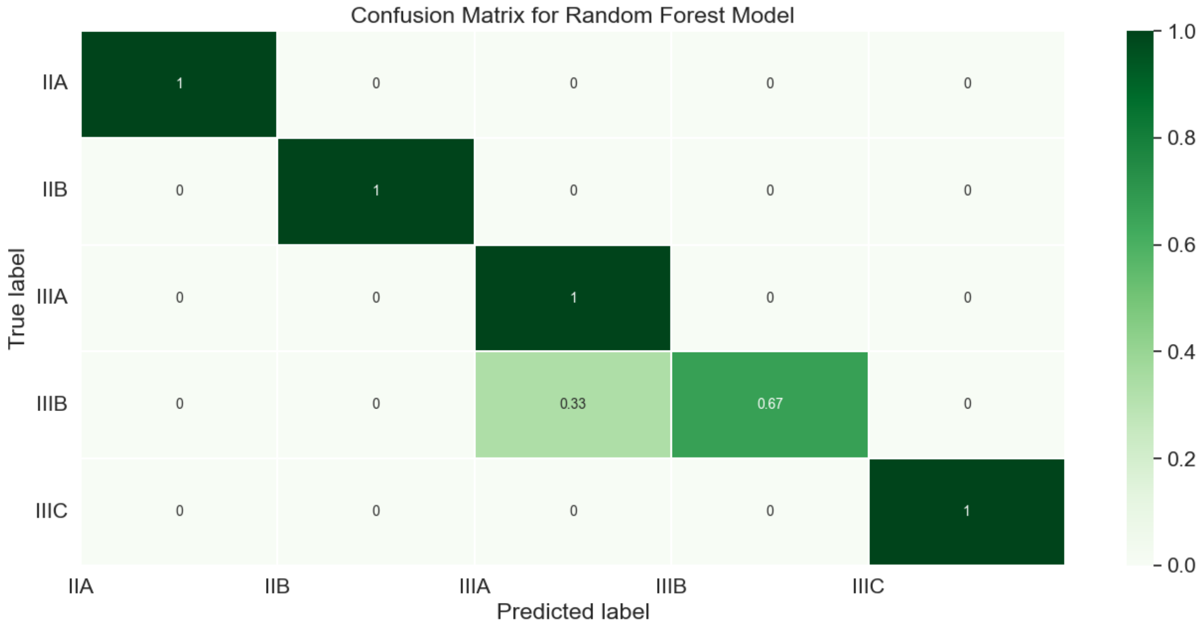

Figure 1. Random forest model confusion matrix

After building the random forest model, the model prediction accuracy and confusion matrix were obtained as shown in Table 2 and Figure 1. These results were obtained by using standard scaler to normalize the dataset. The random forest model consists of 8 independent decision trees had a prediction accuracy of 99%. And in this model, a small portion of stage IIIB are wrongly recognized as stage IIIA.

3.2. K Nearest Neighbor model

Table 3. KNN Classification Report (MinMaxScaler)

	precision	recall	f1 - score	support
0	0.95	0.99	0.97	389
1	0.96	0.94	0.95	344
2	0.93	0.94	0.93	310
3	1.00	0.28	0.43	18
4	1.00	0.97	0.98	147
accuracy			0.95	1208
macro avg	0.97	0.82	0.85	1208
weighted avg	0.95	0.95	0.95	1208

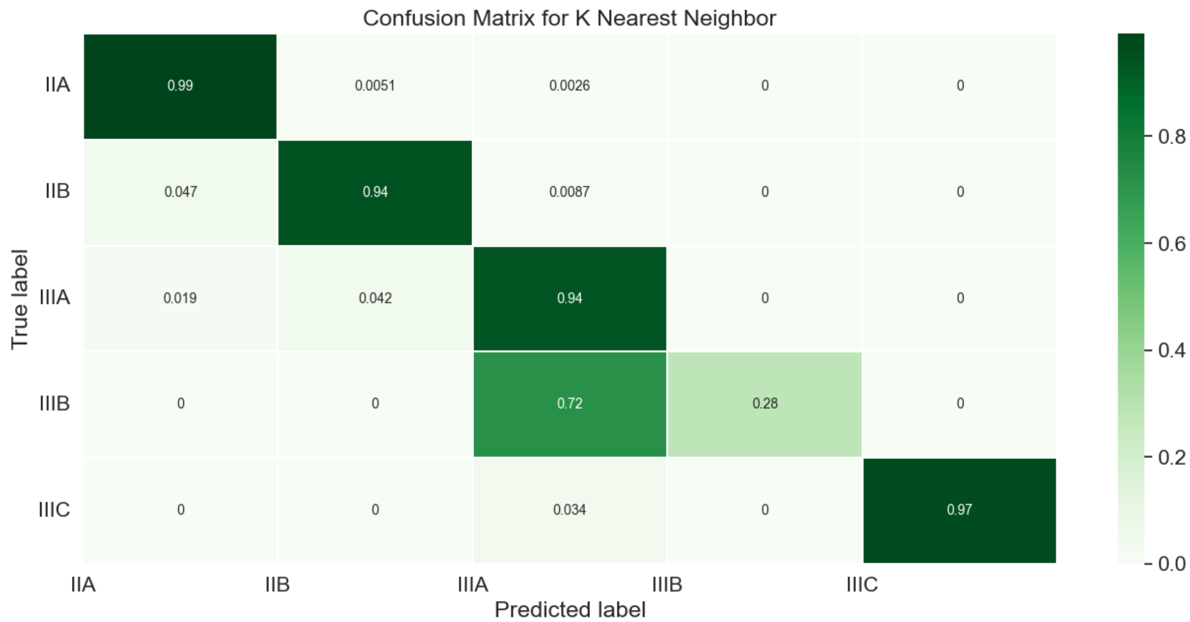

Figure 2. KNN model confusion matrix

After building the random forest model, the model prediction accuracy and confusion matrix were obtained as shown in Table 3 and Figure 2. These results were obtained by using min max scaler. The KNN model produced an accuracy of 95%, and the majority of prediction errors happened how recognizing stage IIIB as stage IIIA.

Table 4. KNN Classification Report (StandardScaler)

	precision	recall	f1 - score	support
0	0.94	0.99	0.97	389
1	0.94	0.95	0.95	344
2	0.89	0.90	0.89	310
3	1.00	0.33	0.50	18
4	0.99	0.89	0.94	147
accuracy			0.93	1208
macro avg	0.95	0.81	0.85	1208
weighted avg	0.94	0.93	0.93	1208

3.3. Discussion

After obtained accuracies of KNN and Random Forest models, Random Forest model clearly has a better accuracy comparing to KNN. For better comparison, Table 4 is the classification report of KNN using standard scaler. As a result, Random Forest still has a better accuracy under the same normalization scaler. From Figure 2, the two models made a precise prediction on stage IIA, IIB, and IIIC. However, both models made mistakes on stage IIIA and IIIB since the difference between them are subtle. The only difference is that whether the tumor has spread to other parts of the body [5]. This is an understandable mistake since there is no such feature in the dataset indicating if tumor has spread to other parts of the body.

4. Conclusion

This paper mentioned that it is crucial for doctors to correctly determine the 6th stage of breast cancer in order to provide corresponding treatments, and doctors alone cannot be 100% accurate due to some limitations. Machine learning models can be useful in terms of predicting the 6th stage based on other information about breast cancer. This paper developed two predicting models using random forest and KNN used to predict 6th stage of the breast cancer. Further experiments were conducted to evaluate and compare different modes. The predicted results correctly predicted stage IIA, IIB, and IIIC; only made minor mistakes in distinguishing stage IIIA and IIIB due to the subtle feature differences. The results also shows that random forest method had an overall higher predicting accuracy comparing to the KNN method. However, these methods may be limited due to the features fed to the training model. Other features that can affect the 6th stage may not be included in the training and testing data. More relating features can be studied in the future.

References

[1] Wikipedia 2022 Breast cancer classification https://en.wikipedia.org/wiki/Breast_cancer#Classification.

[2] Watkins E J 2019 Overview of breast cancer. Journal of the American Academy of Physician Assistants 2019 Volume 32 Issue 10 p 13-17.

[3] Cancer 2022 How common is breast cancer deaths. https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html.

[4] He Z et al. 2020 A review on methods for diagnosis of breast cancer cells and tissues Cell Proliferation 53(7).

[5] Cancer Net 2021 Breast Cancer: Stages https://www.cancer.net/cancer-types/breast-cancer/stages September 2021.

[6] Yue W 2018 Machine learning with applications in breast cancer diagnosis and prognosis. Designs, 2(2), 13.

[7] Wang H 2018 A support vector machine-based ensemble algorithm for breast cancer diagnosis European Journal of Operational Research 267.2 687-699.

[8] Byvatov E et al. 2003 Support vector machine applications in bioinformatics Applied bioinformatics 2.2 67-77.

[9] Yu Q et al. 2020 Clustering Analysis for Silent Telecom Customers Based on K-means++ 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). Vol. 1. IEEE.

[10] Qi Y 2012 Random forest for bioinformatics Ensemble machine learning. Springer, Boston, MA, 307-323.

[11] Namdari, Rihanna. “SEER Breast Cancer Data.” Kaggle, 8 Aug. 2022, https://www.kaggle.com/datasets/reihanenamdari/breast-cancer?resource=download.