Training Artificial Neural Network for Breast Cancer Detection: a High Accuracy Model to Compare Different Features of Breast Cell Nuclei in Breast Cancer Diagnosis

Lily Zheng

Received March 10, 2024
Accepted September 28, 2024
Electronic access October 31, 2024

Breast cancer is a prevalent cancer in the globe, killing hundreds of thousands of people annually. The incorporation of Artificial Intelligence (AI) in early cancer detection shows great potential in recent years. In the research, an Artificial Neural Network (ANN) model is trained on Breast Cancer Wisconsin (Diagnostic) Data Set that contains information about various characteristics of the cell nuclei present in a digitized image of a fine needle aspirate (FNA) of a breast mass[1]. By adjusting the number of hidden layers and calculating the corresponding loss functions, the study aims to achieve high-accuracy predictions based on the 30 features provided by the dataset. A model with 3 hidden layers is developed, with 22 nodes in each layer. With 100 epochs, the model can achieve a 97.48% accuracy, 97.6% precision and 97.6% recall on the validation dataset. The impact of individual characteristics of the cell nuclei on the accuracy of diagnosis is then analyzed by using the same model but only including data that are relevant to the studied characteristics as features. The prediction accuracy is the highest when concave points are the only feature (94.4%) and lowest when only the fractal dimension or texture is considered (72.0%). Therefore, the research suggests that concave points have the strongest correlation with breast cancer diagnosis whereas fractal dimension and texture play the least significant role.

Introduction

Cancer is a pressing disease in the 21st century. Breast cancer, specifically, is the most prevalent cancer in women in the US except for skin cancer[2]. Around 240,000 women and 2100 men are diagnosed with breast cancer every year in the United States, out of which about 42,000 women and 500 men died because of the disease[2]. A key to mitigating and curing breast cancer is early detection. While there are many methods that assist the diagnosing process such as breast exam, mammogram, and breast ultrasound, the only definitive way to diagnose breast cancer is biopsy. During a biopsy, a core tissue from the suspicious area in patients' breast is extracted by a specialized needle device guided by X-Ray or other imaging tests. Experts then analyze the imaged biopsy sample to determine whether the cells are cancerous[3].

With the rise of Artificial Intelligence (AI), the incorporation of AI into predicting the susceptibility of diagnosis for breast cancer early detection shows a promising future. AI models are capable of imitating human intelligence that uses existing data to find meaningful patterns in new conditions[4]. They are able to analyze the biopsy samples quantitatively that are not subjected to human bias, helping human experts to make data-driven calls when a sample can be interpreted differently. They can also detect subtle abnormalities and interpret ambiguous features that humans might not be able to notice immediately. Therefore, it has the potential of assisting the human experts' job of analyzing the data collected from biopsy samples and making a highly accurate prediction. Even though the use of AI in cancer detection is still a relatively new field, multiple machine learning models, specifically deep learning, have been tested and deployed. Deep learning refers to a type of machine learning that automatically determines feature representations from input data by using learning representation. Unlike traditional machine learning models such as K-nearest neighbors (KNN) and random forest (RF), deep learning models can perform optimally without human-engineered features. Some popular deep learning models used in breast cancer detection include Convolutional Neural Network (CNN), Recursive Neural Network (RNN) and Artificial Neural Network (ANN)[5]. Like all neutral networks, CNN has an input layer, one or more hidden layers and an output layer. CNN performs exceptionally when working with images, speech and audio signals by taking the spatial structure of data points into account[6]. There are three main types of hidden layers in CNN: Convolutional layer, Pooling layer and Fully-connected (FC) layer. The complexity of the model increases with the addition of layers, allowing it to identify greater portions of the image[7]. RNN specializes in dealing with sequential data, voice, and natural language. RNNs can keep a nuanced record of previous computations when processing sequential inputs.

The integration of the temporal layer allows RNN models to learn complex changes with their recurrent cells[5]. ANN on the other hand, is dominant when image input is not necessary and the dataset is limited[8]. While CNNs and ANNs both have the ability to address numerical data, CNNs are commonly used in image recognition and computer vision since they are best suited for visual data whereas ANNs are more suited for quantitative datasets[9] Additionally, ANNs are dominant when datasets are limited, while a lot more data inputs are needed for CNNs to achieve their high accuracy. In this paper, the dataset used only contains numerical data on the biopsy screening results of 569 patients. Due to the limitations in the data points available and the unique advantage of ANNs in dealing with this type of data, the study uses the ANNs model.

Results: The Optimal Model and Comparisons Between Different Features in Their Predictive Power

As mentioned above, models with different numbers of hidden layers are experimented to optimize the accuracy, precision, and recall. Following the basic structure of ANN, the model consists of the input layer, the hidden layer(s) and the output layer. The number of hidden layers are determined through comparing the accuracy, precision and recall of models with different numbers of hidden layers while holding the number of nodes in each hidden layer constant. While the number of hidden holds depends on domain-specific factors and there are no mathematically founded theoretical rules for predicting it, some heuristics prove to produce models with decent accuracy. Three of the most-commonly accepted heuristics used for determining the number of hidden nodes are as follows[10]. (1) 75 percent of the quantity of input nodes (2) 50 percent of the quantity of input and output nodes (3) 2n + 1 hidden layer nodes where n is the number of nodes in the input layer Models with each number of hidden layers will be tested three times using each of the three numbers of hidden nodes computed using these methods. Refer to Table 1 for data on each model. The quantity of input nodes is 30 since the input data consists of 30 breast cell nuclei features. Since the output layer is activated by the sigmoid function, the quantity of output nodes is 1. Thus, following the three heuristics listed above, the three optimal number of nodes in the hidden layers are as follows.

(1) $75\% \times 30 = 22$ (rounded)
(2) $50\% \times (30+1) = 16$ (rounded)
(3) $2 \times 30 + 1 = 61$

Every model is tested 5 times for accuracy, precision and recall, and their averages are used for the purpose of the research. Refer to Table 1 for data on each model. As Table 1 below shows, the models' accuracy, precision, and recall, on average, gradually increase as the number of hidden layers increases until it reaches three hidden layers. It starts to gradually decrease afterwards. This can be a result of overfitting (introduced in the Method section) where the complexity of the model exceeds that of the problem, resulting in the network matching the training data points so closely that it loses its generalization ability over data it has never been exposed to. The model with three hidden layers and 22 nodes in each hidden layer is shown to be the most effective. Therefore, this model is used in the following study of individual characteristics of breast mass cell nuclei.

To study the characteristic of the breast mass cell nuclei that performs the best in predicting whether the breast cancer is benign and malignant, any of the 30 features that are irrelevant to the characteristic studied are dropped. This process is repeated for all 10 characteristics. A feature is considered irrelevant if it does not provide direct data for the mean, the standard error and the "worst" for the characteristics studied. For example, when studying the characteristic "radius", only radius mean, radius standard error and radius worst are considered relevant. The other 27 features in the dataset such as texture mean, area standard error and concavity worst would be dropped. Table 2 shows the accuracy, precision and recall scores of the model that predicts based on each of the 10 characteristics. Based on these results, concave points of the breast cell nuclei are the most effective in predicting whether the breast cancer is benign and malignant. This is because it has the highest accuracy (94.4%), precision (95%) and recall (94%) when used as the only feature of a model. Models using only fractal dimension and texture as features tied for the least accuracy (72.0%), precision (72%) and recall (72%).

To provide an example of the model developed in the research predicting the diagnosis of breast cancer, three sets of input data are randomly selected, representing three digitized image breast cell nuclei. The model is able to accurately predict the diagnosis for all three samples. The 30 features of each breast cell nuclei are as follows.

Discussion

In conclusion, the study trained an Artificial Neural Network on the "Wisconsin (Diagnostic) Data Set" that documents 30 features of digitized images of the fine needle aspirate of 569 patients' breast mass[11]. Using accuracy, precision and recall as metrics, the research shows that a model with an input layer, 3 hidden layers with 22 nodes each and an output layer is the most effective. Concave points of patients' breast mass cell nuclei are best able to predict the diagnosis of breast cancers when used as the model's sole feature, whereas fractal dimension and texture show the least accuracy. The significance of concave points in breast cancer diagnosis is supported by biological evidence. Studies have shown that size-related features such as area, perimeter and concave points are statistically important

Table 1: Accuracy, Precision and Recall for Models with Different Number of Hidden Layers and Different Number of Nodes in Each Hidden Layer

Number of Hidden Layers	Number of Nodes in Each Hidden Layer	Accuracy	Precision	Recall
1	22	96.22%	96.6%	96.6%
1	16	96.64%	97.0%	97.0%
1	61	96.92%	97.0%	97.2%
2	22	97.06%	97.2%	97.2%
2	16	97.34%	97.4%	97.4%
2	61	96.22%	96.4%	96.2%
3	22	97.48%	97.6%	97.6%
3	16	97.34%	97.4%	97.4%
3	61	96.36%	96.4%	96.6%
4	22	96.78%	96.8%	96.4%
4	16	95.80%	95.8%	95.8%
4	61	96.50%	96.6%	96.8%
5	22	95.33%	95.6%	95.4%
5	16	95.10%	95.4%	95.4%
5	61	95.80%	96.0%	96.0%

Table 2: Accuracy, Precision and Recall of Models Using Individual Characteristics to Predict the Diagnosis of Breast Cancers

Characteristics Used to Predict whether the Breast Cancer is benign or malignant	Accuracy	Precision	Recall
Radius	90.2%	91%	90%
Texture	72.0%	72%	72%
Perimeter	87.4%	88%	87%
Area	88.1%	89%	88%
Smoothness	74.8%	74%	75%
Compactness	83.9%	85%	84%
Concavity	93.7%	94%	94%
Concave Points	94.4%	95%	94%
Symmetry	72.7%	72%	73%
Fractal Dimension	72.0%	72%	72%

and demonstrate significant differences between benign breast lesions and carcinoma. Thereforefore, they are the appropriate parameters to differentiate between benign breast lesions and infiltrative ductal carcinoma (most common type of invasive breast cancer). Concave points, specifically, are proven to be highly significant in differentiating hyperplasia (a condition where the number of cells in an organ or tissue increases but is not a cancer) from carcinoma [12]. The result of this study can be potentially used to adjust the weights of the 30 features to obtain a more effective model.

While the study gives us a better understanding of machine learning's ability to assist diagnosing breast cancer, its limitations must be acknowledged. The limitations include but are not limited to the instability of prediction accuracy, the existence of biases, and the inability to handle minority cases. Future research can be done to further investigate the application of machine learning in breast cancer diagnosis. In addition to the numerical data used in the study, other types of data can also serve as the model's features. For example, convolutional neural networks, which are extremely effective in processing images, can be trained on mammograms or digitalized images of biopsy to predict the diagnosis. A model trained on both numerical

Table 3: Features of Breast Cell Nuclei Samples Randomly Selected to Demonstrate the Model

Feature	Breast Cell Nuclei Sample 1	Breast Cell Nuclei Sample 2	Breast Cell Nuclei Sample 3
Radius Mean	13.71	13.54	19.79
Texture Mean	20.83	14.36	25.12
Perimeter Mean	90.2	87.46	130.4
Area Mean	577.9	566.3	1192
Smoothness Mean	0.1189	0.09779	0.1015
Compactness Mean	0.1645	0.08129	0.1589
Concavity Mean	0.09366	0.06664	0.2545
Concave Points Mean	0.05985	0.04781	0.1149
Symmetry Mean	0.2196	0.1885	0.2202
Fractal Dimension Mean	0.07451	0.05766	0.06113
Radius Standard Error	0.5835	0.2699	0.4953
Texture Standard Error	1.377	0.7886	1.199
Perimeter Standard Error	3.856	2.058	2.765
Area Standard Error	50.96	23.56	63.33
Smoothness Standard Error	0.008805	0.008462	0.005033
Compactness Standard Error	0.03029	0.0146	0.03179
Concavity Standard Error	0.02488	0.02387	0.04755
Concave Points Standard Error	0.01448	0.01315	0.01043
Symmetry Standard Error	0.01486	0.0198	0.01578
Fractal Dimension Standard Error	0.005412	0.0023	0.003224
Radius Worst	17.06	15.11	22.63
Texture Worst	28.14	19.26	33.58
Perimeter Worst	110.6	99.7	148.7
Area Worst	897	711.2	1589
Smoothness Worst	0.1654	0.144	0.1275
Compactness Worst	0.3682	0.1773	0.3861
Concavity Worst	0.2678	0.239	0.5673
Concave Points Worst	0.1556	0.1288	0.1732
Symmetry Worst	0.3196	0.2977	0.3305
Fractal Dimension Worst	0.1151	0.07259	0.08465
Diagnosis	Malignant	Benign	Malignant
Predicted Diagnosis	Malignant	Benign	Malignant

data and images has the potential of showing greater accuracy in predicting whether the breast cancer is benign or malignant.

Methods

The "Wisconsin (Diagnostic) Data Set" is used in the study. It documents the information computed from the digitized image of a fine needle aspirate (FNA) of a breast mass of 569 patients [1]. The dataset includes 30 features which describe the characteristics of the cell nuclei present in the images for each patient. The 30 features belong to 10 major categories, including radius, perimeter and concave points. In each category, the mean, standard error and "worst" (the mean of the three largest values) are computed. Some examples of the features are *radius_mean*, *radius_worst* and *perimeter_se*. The datasets also provide information about whether the tumors are benign or malignant, which is the target of my model.

The extraction and data processing method for the 30 features are as follows. The picture of the drop of fluid from breast tumor is taken by a JVC TK-1070U color video camera and projected into the camera with a 63x objective and a 2.5x ocular. It is captured by a ComputerEyes/RT color frame grabber board as a 512x480, 8-bit-per-pixel Targa file. A graphical user interface was developed for inputting approximate initial boundaries for cell nuclei. Using the approximate boundaries as a reference, the actual boundaries of the cell nucleus is located by an active contour model known as a "snake" [13]. As a deformable spline, a snake seeks to minimize an energy function defined over the arclength of a closed curve. It is thus able to deform to the exact shape of the nucleus, allowing for precise analysis and quantification of nuclear shape, size and texture [14]. Figure 1 shows the snake after converging to the cell nucleus boundaries. The methods to determine each of the 10 characteristics are as follows [14].

1. Radius: the average length of the radial line segments defined by the centroid of the snake and the individual snake points.

2. Perimeter: total distance between snake points

3. Area: number of pixels on the interior of the snake (adding $\frac{1}{2}$ of the pixels on the perimeter)

4. Compactness: $\frac{\text{perimeter}^2}{area}$
5. Smoothness: distance between the length of the radial line and the mean length of the surrounding lines
6. Concavity: the extent to which the boundary of the nucleus lies on the inside of the chord drawn between non-adjacent snake points
7. Concave points: concavity measured in numbers instead of magnitude
8. Symmetry: the length difference between lines perpendicular to the longest chord through the center to the cell boundaries in both directions
9. Fractal Dimension: "coastline approximation"
10. Texture: the variance of the grey scale intensities in the component pixels

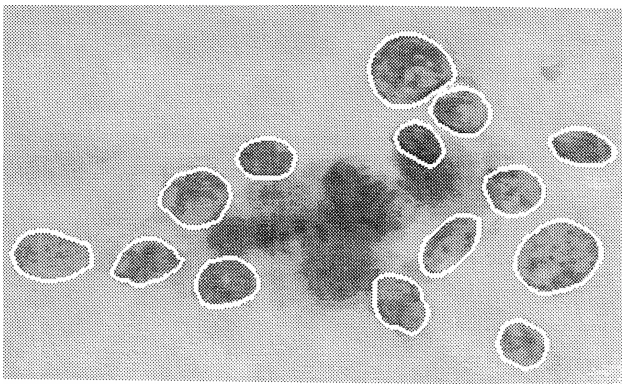

Figure 1: Cell Nucleus bounded by the converged "snakes" [14]

The 30 features of the breast mass cell nuclei are used to train the machine learning model used to predict whether the diagnosis is benign and malignant. Since the diagnosis in the raw data set is qualitative - either "benign" or "malignant" - one-hot encoding is applied to the categorical data to transform it into a format that can be fed into the machine learning algorithm. I use 1 to replace malignant diagnosis and 0 for "benign". The dataset is split into a training set and validation set to avoid overfitting. Overfitting refers to an undesirable situation where the model is over-trained that it is highly accurate on the training data but unable to give accurate predictions on new data. Since the model is only trained on the training set, its performance on the validation dataset accurately describes its ability to make predictions on new data. Thus, the validation set not only serves as an accurate performance metric but also allows for adjusting hyperparameters to prevent overfitting.

After the features and the targets are scaled, a sequential model is constructed with one input layer and one output layer.

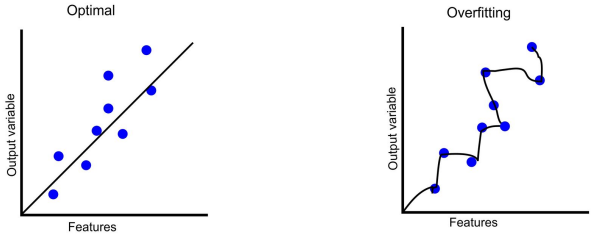

Figure 2: Diagrams for optimal model and overfitting [15].

The output layer is activated by the "sigmoid" function, for which the mathematical expression is:

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

The sigmoid function maps any input values into a value between 0 and 1, and it is especially useful for binary classification. Hidden layers are then added to improve the accuracy and precision of the model. The hidden layers are activated using the "relu" function, which can be mathematically expressed as

$$relu(x) = max(0,x)$$

Different numbers of hidden layers and nodes in each layer are experimented and contrasted to find the most accurate model.

Since the model is unbalanced, its effectiveness is measured by accuracy, precision, and recall. Unbalanced models are models with classifications datasets that have skewed class proportions. In the dataset used, there are 357 benign results and 212 malignant results for the cancer diagnosis - benign results constitute 62.7% of all target data. When the model is unbalanced, solely using accuracy as a metric for the effectiveness of the model might lead to biases towards the majority class since it has more samples to train on. Thus, in addition to accuracy, it is also necessary to calculate its precision and recall using the equations shown in Table 4.

Finally, a loss function is deployed to calculate the errors in the predicted values. The loss function uses cross entropy error since it is best suited for binary classification [16]. The loss is minimized to find the optimal model. The loss function can be expressed as

$$J = \frac{-1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]$$

y_i is the label of the i^{th} sample in a dataset with n samples; $\hat{y}_i$ is the predicted label of the i^{th} sample in the dataset

Table 4: Equations for calculating the three metrics used to measure the model's effectiveness

Metric	Equation
Accuracy	$\frac{\#\ of\ correct\ prediction}{\#\ of\ predictions}$
Precision	$\frac{\#\ of\ true\ positives}{\#\ of\ positives}$
Recall	$\frac{\#\ of\ true\ positives}{\#\ of\ positive\ predictions}$

The loss function is applied to the dataset used in this research, which contains 569 samples. The optimal model has 100 epochs. The number of epochs is the number of times that the entire dataset goes through the model. As Figure 3 below shows, as epochs approach 100, the cross entropy error approaches zero, which means the accuracy of the model approaches 100%.

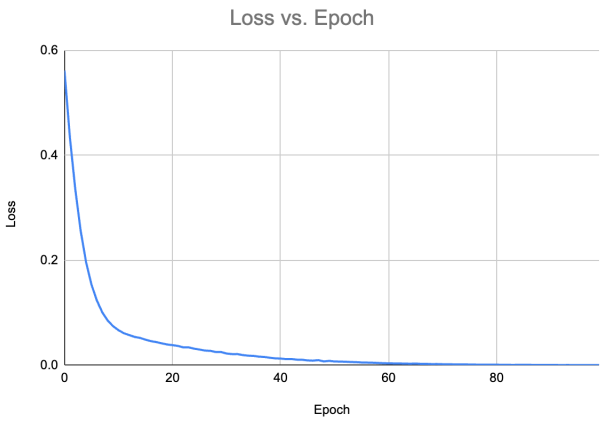

Figure 3: The Loss Function of the ANN Model in the Study as It Approaches 100 Epochs

References

1 W. Wolberg, O. Mangasarian, N. Street and W. Street, *Breast Cancer Wisconsin (Diagnostic)*, UCI Machine Learning Repository, 1995.

2 Centers for Disease Control and Prevention, *Basic Information About Breast Cancer*, https://www.cdc.gov/cancer/breast/basic_info/index.htm, Accessed: 2023-12-02.

3 Mayo Clinic, *Breast cancer - Diagnosis and treatment*, https://www.mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475, Accessed: 2023-12-02.

4 S. Greengard, *Communications of the ACM*, 2016, **59**, 29–31.

5 M. Nasser and U. Yusof, *Diagnostics*, 2023, **13**, 161.

6 P. Chintarungruangchai and I. Jiang, *Publications of the Astronomical Society of the Pacific*, 2019, **131**, 1–15.

7 IBM, *What are Convolutional Neural Networks?*, https://www.ibm.com/topics/convolutional-neural-networks, Accessed: 2023-12-02.

8 V. Meel, *ANN and CNN: Analyzing Differences and Similarities - viso.ai*, Viso Suite, https://viso.ai/deep-learning/ann-and-cnn-analyzing-differences-and-similarities/, Accessed: 2023-12-02.

9 M. Vawda, R. Lottering, O. Mutanga, K. Peerbhay and M. Sibanda, *Sustainability*, 2024, **16**, 1051.

10 S. Walczak and N. Cerpa, *Encyclopedia of Physical Science and Technology (Third Edition)*, Academic Press, 2003.

11 W. Wolberg, O. Mangasarian, N. Street and W. Street, *Breast Cancer Wisconsin (Diagnostic)*, UCI Machine Learning Repository, 1995.

12 A. Narasimha, B. Vasavi and H. Kumar, *International Journal of Applied and Basic Medical Research*, 2013, **3**, 22–26.

13 M. Kass, A. Witkin and D. Terzopoulos, *International Journal of Computer Vision*, 1988, **1**, 321–331.

14 W. Street, W. Wolberg and O. Mangasarian, *Electronic Imaging*, 1993.

15 freeCodeCamp, *What is Overfitting in Machine Learning?*, https://www.freecodecamp.org/news/what-is-overfitting-machine-learning/, 2023, Accessed: 2023-12-02.

16 J. Brownlee, *How to Choose Loss Functions When Training Deep Learning Neural Networks*, https://machinelearningmastery.com/how-to-choose-loss-functions-when-training-deep-learning-neural-networks/, 2020, Accessed: 2024-01-16.