Proceedings of ICBioMed 2025 Symposium: AI for Healthcare: Advanced Medical Data Analytics and Smart Rehabilitation
DOI: 10.54254/2753-8818/2025.AU28004

Diagnosing Breast Cancer with Machine Learning

Haoran Wang

School of Mathematics and Statistics, Huazhong University of Science and Technology, Wuhan, China
u202210201@hust.edu.cn

Abstract. To develop a model that establishes if a patient's breast cancer is benign or malignant, it uses them in the patient data set. The issue discussed in this article is part of a larger problem that is classified as supervised learning. Python is employed for coding. This study makes use of Keras, which is linked to TensorFlow and serves as one of the most widely used free modules. It's a categorization problem that is binary, and neurons are employed to create the model. The sigmoid function is employed in the output layer as a means of activation. The article attempts to create different models with different numbers of hidden layers. A set for verification (25%) and the training data (75%), respectively, are created from the collected data. The validation set is used to assess each model's success after it has been first trained on the set that was used for training. The article's conclusion states that the model with no additional layer is the most effective.

Keywords: Machine learning, Supervised learning, Binary classification problem, Logistic regression, Neural network, Breast cancer

1. Introduction

Machine learning, a subset of AI, has seen rapid growth due to big data and increased computing power.It automates the creation of analytical models by taking data from training sessions and not by programming explicitly.The system identifies patterns in the input data and generates outputs based on this learning. In this case, the system becomes more intelligent, smarter, and wiser with time without human involvement. Machine learning relieves humans of the burden of explicating and formalising their knowledge into a machine-accessible form. It makes it possible to increase the precision and effectiveness of forecasts or choices. Machine learning has significantly impacted fields such as facial recognition, iris scans, handwritten recognition, speech recognition, medical imaging analysis, natural language processing, stock pricing predictions, scientific research, marketing campaigns, and autonomous driving [1-4].

The most prevalent type of cancer among women is breast cancer, and it is primarily responsible for the majority of cancer-related deaths. With estimates for 2023 indicating 300,590 new invasive cases and 43,700 fatalities, it ranks second in the United States for female cancer diagnoses and accounts for almost 30% of all female cancer cases [5]. This article aims to create a machine learning model to help diagnose breast cancer accurately and efficiently.

DOI: 10.54254/2753-8818/2025.AU28004

The dataset comprises 569 patient entries, each with 32 variables. One variable, "diagnosis," indicates whether the tumor is malignant ("M") or benign ("B"). The remaining 31 variables are features derived from FNA images, characterizing cell nuclei morphology. A fundamental aspect of machine learning is supervised learning, which uses algorithms to identify data patterns from both independent and dependent variables to forecast the values of the latter [6]. Data in supervised learning comprises patient examples with features (radius, texture, perimeter measurements) and labels (malignant/benign breast cancer). This article analyzes a 569-patient breast cancer dataset using neural networks implemented with Keras, demonstrating machine learning's predictive power for real-world cancer diagnosis applications.

2. Methodology

2.1. Data preprocessing

It is considered that the status of the patient's breast cancer has no significant correlation with their "id". Therefore, the column "id" is removed from the dataset. Besides, for the convenience of the computer reading the data, the "M" (malignant) and "B" (benign) in the column "diagnosis" are replaced with 0 and 1.

In the dataset, the number of cancer diagnoses is 212, and the number of non-cancer diagnoses is 357.They can be regarded as balanced, which means it is appropriate to use this dataset to create an accurate model [7,8].

2.2. Training set and validation

A training set and a validation set are created by randomly dividing the sample data. Seventy-five percent of the examples are in the one used for training set, while the other 25 percent are in the one used for validation set. The training set is used only for the development of computational models, while the set for validation is only used to evaluate the models' performance. The next sections will have more specific information.

2.3. Binary classification problem

One of the broad areas of machine learning and statistics is data classification. Classification problems are often divided into two categories: binary and multiclass [4]. The instances in binary classification are divided into two groups and given labels of either 0 or 1. In the dataset, benign breast cancer is represented by a value of 1 and malignant breast cancer by a value of 0. These examples are designated as category zero and category 1, accordingly.

A model that takes the properties of an instance as input and outputs an amount that ranges from 0 to 1 is said to be dealing with a binary classification issue. This is customary. The features are denoted by x_1 , x_2 , ..., x_{30} . Therefore, x_1 is the radius mean, x_2 is the texture mean, etc. The output is denoted by $\hat{y} = \hat{y}(x_1, x_2, \ldots, x_{30})$. An example will be predicted to belong to the category 0 if $\hat{y} \leq 0.5$ and the category 1 if $\hat{y} > 0.5$.

2.4. Error in entropy of binary crosses

Suppose that an illustration's actual title is represented by the letter y, and y is either 0 or 1. Let $\hat{y}$ represents the model's predicted probability that the example belongs to category 1, such that

DOI: 10.54254/2753-8818/2025.AU28004

$\hat{y} \in [0,1]$. The exact number of lost neurons is described as:

$$BCE(\hat{y}, y) = -y\log(\hat{y}) - (1-y)\log(1-\hat{y}) \tag{1}$$

In the creation of the model, the error is not calculated from isolated instances, but instead from a collection of examples. The average of the binary cross entropy errors on a set of examples is the mean of the binary cross entropy errors on each instance in the set. The model's inaccuracy is derived not from individual cases, but from a compilation of examples. The average of the cross-entropy binary errors for an array of examples is the standard deviation of the binary cross-entropy errors for each occurrence within the set.

2.5. Logistic regression

Logic regression is a quantitative analytical tool that creates a statistical model that describes the relationship between a binary or dichotomous (yes/no type) outcome and a set of independent variables or parameters [9]. While this is not the technique used to create the models in this article, it is instructive to start by explaining it.

The function known as the sigmoid is defined as

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{2}$$

In the logistic regression model, the prediction function $\hat{y}$ is believed to have the following structure

$$\hat{y} = \hat{y}(x_1, x_2, \ldots, x_{30}) = \sigma(w_1x_1 + w_2x_2 + \ldots + w_{30}x_{30} + b)$$

2.6. Neural network

Logistic regression, while simple, has limitations. It has a poor performance when the label isn't a function of the combination's linear attributes. Neural networks, an extension of logistic regression, can overcome these limitations, potentially providing more accurate predictions [10].

As in logistic regression, neural network assumes that the prediction $\hat{y}$ has a certain functional form of the features, i.e. $\hat{y} = \hat{y}(x_1, x_2, \ldots, x_{30})$, but this functional form is not easy to describe. Network is used to describe this functional form. That is why this method is called a neural network.

Before going into the details of neural networks, it is necessary to introduce some frequently used activation functions. One of them is the sigmoid function, as mentioned above. Other functions include the tanh(Hyperbolic tangent) function and the ReLU(rectified linear function).

The hyperbolic tangent function is

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{3}$$

The definition of the ReLU function is the process

$$\mathrm{ReLU}(x) = \begin{cases} x, & if\ x \geq 0 \\ 0, & if\ x < 0 \end{cases} \tag{4}$$

DOI: 10.54254/2753-8818/2025.AU28004

The structure of a neural network can be articulated as follows: The structure consists of multiple layers; each layer comprises nodes; every node within a layer is interconnected with every other node in that layer via edges; the initial layer is designated as the input layer; the terminal layer is referred to as the output layer; layers that are neither the input nor the output layer are termed layers that are hidden.

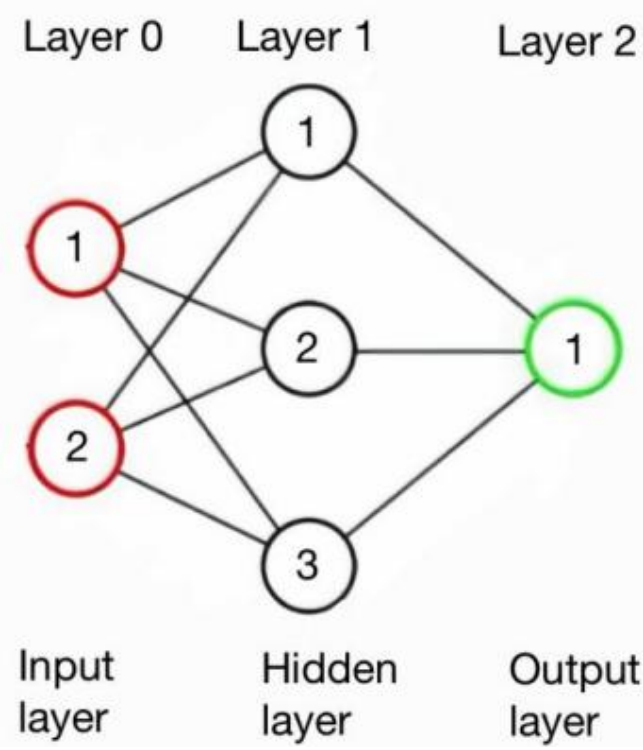

Figure 1. The design of a neural network's architechture

Every point in the network is connected with a weight (refer to Figure 1). Furthermore, every layer in the non-input areas is assigned a bias value. The input layer's dimension corresponds to the number of features per sample, with each node signifying a distinct attribute. The task's categorical character necessitates a singular node in the output layer.

Every layer in the concealed set corresponds to a particular task. The layer that outputs data utilises the sigmoid coefficient due to its dual nature.

The procedure for assessing predictions is uncomplicated. The feature values from the examples are entered into the respective nodes of the input layer. These values are subsequently disseminated downward and modified in accordance with rules not addressed in this article. The numerical number linked to the output layer's node represents the prediction. Analogous to logistic regression, the neural network methodology determines the parameters that minimise the average binary cross-entropy error of the training dataset. No detailed information regarding the methodologies employed to ascertain these metrics will be disclosed. These values are generally obtained from widely utilised free software libraries, such as the Keras library within TensorFlow.

3. Measuring the quality of the model

The set of validation results is utilised to evaluate how well the model works. A reduced average linear cross-entropy error on this dataset signifies superior model performance. A prevalent metric

DOI: 10.54254/2753-8818/2025.AU28004

for evaluating machine learning models is accuracy [11], which refers to the ratio of accurate predictions provided by the algorithm. In binary classification, accuracy is calculated by dividing the total number of correct predictions, both positive and negative, by the total number of predictions made. This produces a number ranging from 0 to 1. A greater accuracy value (approaching 1) signifies superior prediction performance, whereas a lower value (approaching 0) denotes inferior performance. In a balanced dataset, as presented in this paper, accuracy can be a dependable metric for assessing model efficacy.

4. Overfitting

Overfitting refers to a modelling approach that fails to generalise from seen data to unobserved data. The model exhibits overfitting, resulting in excellent performance on the learning dataset but inadequate performance on the set used for validation. Overestimating transpires as a statistical predictive system assimilates both the systematic and stochastic (noise) elements of the training data to such an extent that it adversely impacts the model's performance on novel data. The statistic predictive approach has considerable flexibility concerning both the signal and the noise within the training data [12,13].

5. Applications to the diagnosis of breast cancer

This part utilises the neural network methods outlined earlier to examine data from people potentially affected by breast cancer, subsequently training the model to evaluate the malignancy of the breast cancer as either cancerous or non-cancerous. Python is utilised for programming. This study utilises Keras, a widely-used open-source library connected with TensorFlow.

Table 1 displays the results obtained from models with different numbers of hidden layers. More details of the model, i.e. the numbers of the nodes in hidden layers and the activation functions associated with the hidden layers, can also be seen from Table 1.

Table 1. The results of the models

The number of hidden layers	0	1	1	1	1	2
The number of nodes in the hidden layer	\	1	2	1	2	1(first hidden layer) 1(second hidden layer)
Activation function of the hidden layer	\	relu	relu	tanh	tanh	relu(first hidden layer) tanh(second hidden layer)
Training error	0.058262262	0.039697997	0.011838581	0.025759002	0.03474618	0.040282782
Validation error	0.08050053	0.14174409	0.144471	0.15344304	0.09476268	0.12748282
Training accuracy	0.99	0.99	1.00	1.00	0.99	0.99
Validation accuracy	0.96	0.94	0.98	0.96	0.96	0.97

The principal aim of developing and applying methods for machine learning is to forecast unobserved data not utilised in the learning phase of the method; therefore, the validation error,

DOI: 10.54254/2753-8818/2025.AU28004

representing the overall error in predicting future samples, should be minimised rather than the training error, which pertains specifically to the data employed for model training [13].

Therefore, considering both the mean binary cross entropy error and the accuracy, the model with no hidden layer can be seen as performing best. Its mistakes are small, and the accuracies are close to 1. It works quite well on both the training set and the validation set. It does not show signs of overfitting. It should be effective at predicting the degree to which a breast cancer is aggressive or indolent. It should be effective and precise when employed on breast cancer predictions.

6. Conclusion

Machine learning models and libraries have been developed to evaluate the likelihood of a patient having breast cancer. The instruments and these "Sequential" model in the machine learning libraries TensorFlow and Keras are utilised to construct the models. The models are constructed and assessed using a patient dataset. The dataset is divided into a training set comprising 75% and a validation set comprising 25%. The former is utilised to train the models through a neural network.

The trained models are then employed on the validation set. Models with different numbers of layers of invisibility are created. Due to their performance on the data set, the one with no hidden layer is considered optimal. There are no signs of an overfitting problem. This study demonstrates machine learning's predictive capabilities, specifically using the Keras library for neural network model creation. Further optimization is possible through advanced techniques and libraries.

References

[1] Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. Electronic Markets, 31(3), 685-695. https: //doi.org/10.1007/s12525-021-00459-5

[2] Sharifani, K., & Amini, M. (2023). Machine learning and deep learning: A review of methods and applications. World Information Technology and Engineering Journal, 10(07), 3897-3904. https: //doi.org/10.18063/witej.v10i07.3897

[3] Sharma, N., Sharma, R., & Jindal, N. (2021). Machine learning and deep learning applications-a vision. Global Transitions Proceedings, 2(1), 24-28. https: //doi.org/10.1016/j.glt.2021.09.004

[4] Goar, V., & Yadav, N. S. (2024). Foundations of machine learning. In Intelligent Optimization Techniques for Business Analytics (pp. 25-48). IGI Global. https: //doi.org/10.4018/978-1-7998-8105-9.ch002

[5] Wang, J., & Wu, S.-G. (2023). Breast cancer: An overview of current therapeutic strategies, challenges, and perspectives. Breast Cancer: Targets and Therapy, 721-730. https: //doi.org/10.2147/BCTT.S411789

[6] Tiwari, A. (2022). Supervised learning: From theory to applications. In Artificial Intelligence and Machine Learning for EDGE Computing (pp. 23-32). Academic Press. https: //doi.org/10.1016/B978-0-12-819187-2.00003-7

[7] Ghavidel, A., & Pazos, P. (2025). Machine learning (ML) techniques to predict breast cancer in imbalanced datasets: A systematic review. Journal of Cancer Survivorship, 19(1), 270-294. https: //doi.org/10.1007/s11764-025-01145-7

[8] Meliboev, A., Alikhanov, J., & Kim, W. (2022). Performance evaluation of deep learning based network intrusion detection system across multiple balanced and imbalanced datasets. Electronics, 11(4), 515. https: //doi.org/10.3390/electronics11040515

[9] Das, A. (2024). Logistic regression. In Encyclopedia of Quality of Life and Well-Being Research (pp. 3985-3986). Springer International Publishing. https: //doi.org/10.1007/978-3-319-69909-7_1504

[10] Hassanipour, S., et al. (2019). Comparison of artificial neural network and logistic regression models for prediction of outcomes in trauma patients: A systematic review and meta-analysis. Injury, 50(2), 244-250. https: //doi.org/10.1016/j.injury.2018.09.019

[11] Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media, Inc.

[12] Ying, X. (2019). An overview of overfitting and its solutions. Journal of Physics: Conference Series, 1168, 012013. https: //doi.org/10.1088/1742-6596/1168/1/012013

DOI: 10.54254/2753-8818/2025.AU28004

[13] Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Overfitting, model tuning, and evaluation of prediction performance. In Multivariate Statistical Machine Learning Methods for Genomic Prediction (pp. 109-139). Springer International Publishing. https: //doi.org/10.1007/978-3-030-75622-8_5