**An Exploration of Machine Learning Algorithms for Breast Cancer Detection**

Harsh Patel

Faculty of Applied Science and Technology, Sheridan College

Personal Email: [harshpatel2312002@gmail.com](mailto:harshpatel2312002@gmail.com)

**Abstract**

Breast cancer remains a critical global health concern, necessitating effective early detection strategies to improve patient outcomes. This research explores the application of machine learning algorithms, specifically Decision Trees (DT) and Neural Networks (NN), in enhancing breast cancer detection and diagnosis. Leveraging data from medical imaging sources, the study aims to compare the performance of DT and NN models in accurately identifying breast cancer indicators. The methodology involves training and testing both algorithms on a diverse dataset, with performance metrics including accuracy, precision, and recall. Results indicate a significant difference in performance between DT and NN, with DT exhibiting higher mean accuracy (0.7113) compared to NN (0.6756). Hypothesis testing confirms this disparity, highlighting the potential of DT as a viable tool for breast cancer detection due to its interpretability and classification capabilities. However, NN's ability to capture intricate patterns suggests its relevance in specific contexts. Overall, this research underscores the importance of leveraging machine learning techniques for early breast cancer detection, offering insights into optimizing diagnostic processes for improved patient care.

Table of Contents

[**Introduction** 4](#_Toc165417583)

[**Literature Review** 6](#_Toc165417584)

[**Decision Tree** 6](#_Toc165417585)

[**Neural Network** 6](#_Toc165417586)

[**Methods** 7](#_Toc165417587)

[**Dataset Clarification:** 7](#_Toc165417588)

[**Experimental Design:** 7](#_Toc165417589)

[**Repeated Experiment:** 7](#_Toc165417590)

[**Performance Evaluation:** 7](#_Toc165417591)

[**Accuracy, Precision, and Recall:** 7](#_Toc165417592)

[**Paired t-test:** 8](#_Toc165417593)

[**Results** 9](#_Toc165417594)

[**Hypothesis Testing:** 10](#_Toc165417595)

[**References** 12](#_Toc165417596)

# **Introduction**

Cancer is the leading cause of death worldwide, where breast cancer is a significant health concern and early detection of it makes it easier to treat patients. Various screening tests helps to identify abnormalities in the breast tissue promptly. Dr. Maryory Gomez explains that its early detection helps women to have a good prognosis, which eventually helps in curing it at a very early stage (LLC). Early detection allows for timely interventions such as surgery, radiation therapy, chemotherapy, or targeted therapy, which can effectively eradicate or control the cancerous cells.

According to Carrie Knox women should encourage themselves for self breast examination once a month. While self examining women should look for warning signs such as lumps, inverted nipples, irregular discharge, differences in skin texture on both breasts and dimpling (Renckens, Ali). As Knox said most important thing is to be aware and inform your doctor of any change that you may see.

In recent years, machine learning algorithms have emerged as powerful tools in the field of medical diagnostics, offering novel approaches to aid in the detection and diagnosis of breast cancer. By integrating machine learning into breast cancer detection and diagnosis processes, there is potential for enhancing accuracy, efficiency, and personalized treatment strategies, ultimately contributing to improved patient outcomes and quality of care. Decision tree and neural network algorithm are excelling choices as they are excellent in recognizing complex patterns that might exist in medical imaging data.

Decision Trees (DT) are a powerful machine learning algorithms that are particularly well suited for classification tasks, making them a valuable tool for breast cancer diagnosis. This operates by recursively portioning the data based on the class of interest. It also offers interpretable results making it easier for healthcare professionals to understand the reasoning behind the classification decisions. Pattern recognition (PR) algorithms can examine MRI images and identify certain patterns and features linked to the disease to detect and diagnose breast cancer (Shaw et al. 71).

Neural Networks (NN) are particularly well suited for identifying intricate patterns and relationships within complex datasets. These networks could analyze the medical imaging data such as mammograms and ultrasound scans to identify subtle patterns indicating malignant tumors. Their ability to extract features automatically from raw data allows them to detect nuanced abnormalities that may not be easily discernible to the human eye. According to Lancashire, L.J, the molecular classification of malignant breast tumours using high throughput technologies including expression arrays and immunohistochemistry screening on tissue microarrays (TMAs), has successfully identified several biologically relevant subgroups (2).

Since the study aims to compare the efficacy of both DT and neural networks NN, we can posit the null hypothesis that there is no significant difference in the performance of these algorithms and the alternative hypothesis posits that there exists a significant difference in the performance of DT and NN in the context of the study.

The reason behind the performance difference between the two algorithms could attribute to several factors like model complexity, data availability and quality, data complexity, and interpretability. For instance, DT are relatively simple models that create decision rules based on individual features, which may result in oversimplified or overly complex models depending on the dataset. NN, especially deep architectures, have the capacity to capture highly complex patterns by combining multiple layers of neurons, but this complexity also introduces challenges such as overfitting and computational resource requirements.

In summary, the performance difference between both the models in detecting breast-cancer can stem from variations in their ability to handle the above features.

# **Literature Review**

## **Decision Tree**

A Decision Tree (DT) is a supervised machine learning algorithm used for classification and regression tasks. A decision tree comprises of nodes and branches, forming a hierarchical structure in which the root node represents the entire dataset. Branches stem from the root node and represent decision points based on specific features. Leaf nodes are terminal points of the tree, representing the final outcomes or classifications.

The decision tree is induced from a set of labeled training instances represented by various attributes and class label. Decision tree training is a greedy, top-down and recursive process starting with the entire training data and an empty tree (Su and Zhang) [1]. The best attribute is chosen which best partitions the training data is chosen for the root, and then the training data is split into disjoint subsets. The algorithm keeps splitting the data into smaller groups until all the items in each group are identical.

## **Neural Network**

Neural Networks (NN) are a type of machine learning algorithms which consists of interconnected nodes, called artificial neurons, mimicking the network of biological neurons in our brains. These artificial neurons process information and transmit signals to each other, just like their biological counterparts. NN are layered structures which contains an input layer, hidden layer, and an output layer.

The inputs are multiplied by weights (strength of the respective signals) and then computed by a mathematical function which determines the activation of the neurons (Gershenson). Another function compares the output of the artificial neuron (Gershenson). The input is directly proportional to the weight of an artificial neuron, meaning as the higher the weight, the stronger the input which is multiplied by it.

Both decision tree and neural network are powerful tools to detect breast cancer. Both have their individual applications. When it comes to easier interpretation of the data which can be understood by healthcare professionals to drive decisions, decision tree would be a good choice as it is best for classification. On the other side the neural networks provide high accuracy rates in breast cancer detection, as they can learn complex patterns which might be dissemble by decision trees. Although, it doesn’t necessarily mean neural network is better than decision tree as it also depends upon the dataset and the type of data that is being processed by both algorithmic models.

# **Methods**

## **Dataset Clarification:**

A publicly available repository within the Weka platform will provide medical imaging data, including images from patients diagnosed with cancer and healthy individuals. Among these attributes are factors such as age, menopausal status, tumor size, involvement of lymph nodes, presence of node caps, degree of malignancy, breast characteristics, quadrant of the breast affected, presence of irradiation, and the classification of the case (i.e., malignant, or benign). It's worth noting that two attributes, age, and involvement of lymph nodes, have a unique constraint implemented, contributing to the integrity and specificity of the dataset.

## **Experimental Design:**

The dataset will be randomly divided into training and testing sets. The training set (66% of the data) will be utilized to train both the decision tree and neural network models, while the testing set (34% of the data) will be used for evaluating their performance.

## **Repeated Experiment:**

To ensure robustness and reliability of results, each machine learning model will be built 30 times based on different random samples of the dataset. Each sample will undergo the same division into training and testing subsets as described above.

## **Performance Evaluation:**

The accuracy of each model in all 30 repeats will be calculated. Statistical analysis will be employed to compare the performance of decision tree and neural network algorithms. Specifically, the mean, standard deviation, and p-value for each algorithm's accuracy will be determined.

## **Accuracy, Precision, and Recall:**

Accuracy, precision, and recall are the all the metrics used to evaluate performance of classification models in machine learning.

Accuracy measures the overall performance of the model. In other words, the proportion of the model being correct. It's calculated as the total number of correct predictions divided by the total number of predictions. Accuracy works well when you have a balanced dataset.

Precision focuses on the positives your model predicts. It tells you what proportion of the things your model labeled as positive are actually correct. High precision means your model is good at avoiding false positives.

Recall focuses on the completeness of your model for a particular class. It tells you what proportion of the actual positives your model was able to identify. High recall means your model is good at avoiding false negatives.

## **Paired t-test:**

Paired t-test is a statistical method of comparing the means between groups. It emphasizes the fact that is the difference between the response of the two groups statistically significant or not. It is a form of a hypothesis test in which the test statistic follows a Student’s t-distribution. It is typically applied when the population follows a normal distribution.

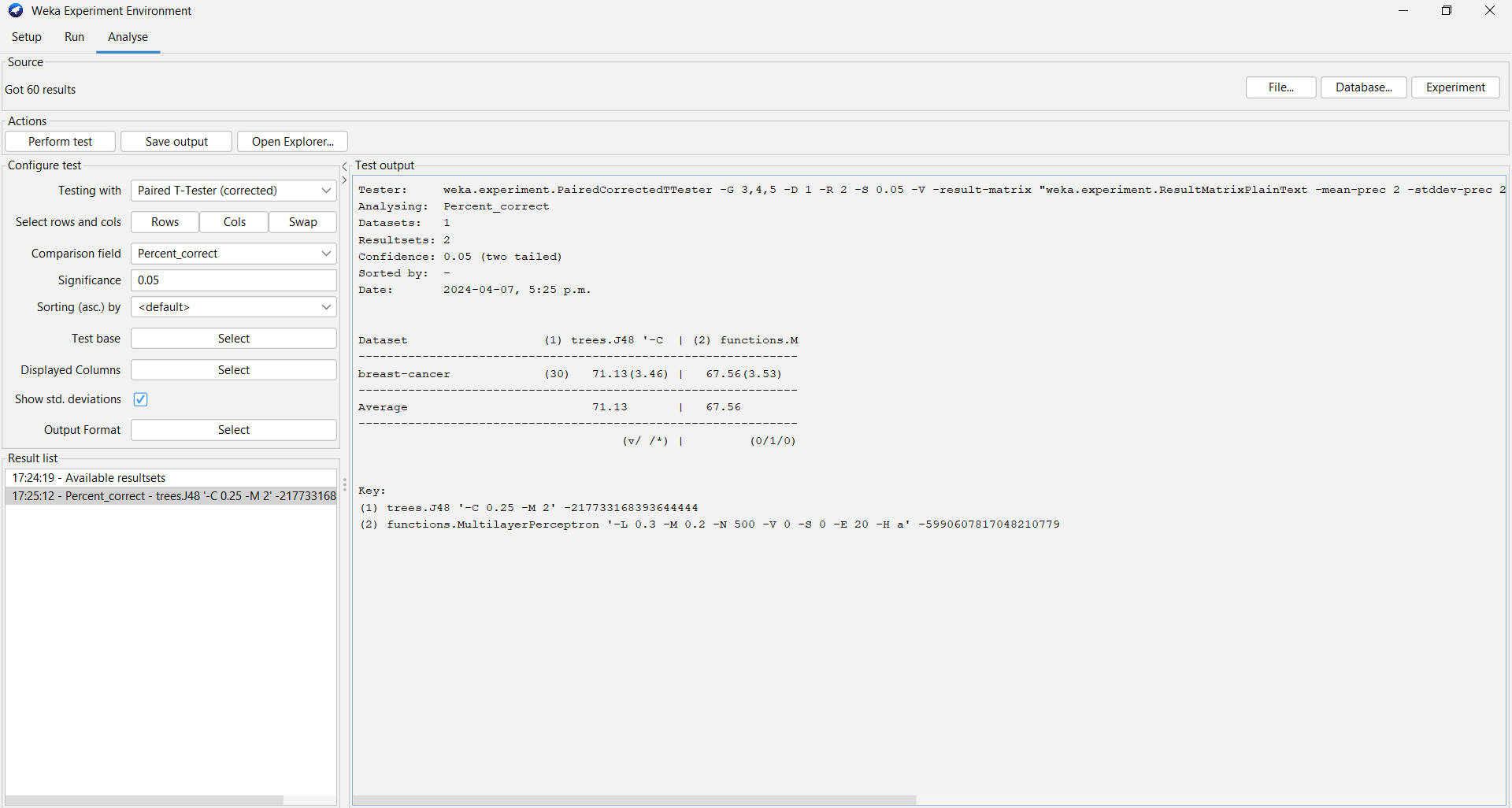
In hypothesis testing with the t-test, we're investigating if there's a meaningful difference between two groups. The null hypothesis assumes no difference, while the alternative hypothesis suggests otherwise. Typically, we use a two-tailed test for comparing two groups. We calculate a t-value based on our data and chosen confidence level, considering degrees of freedom. This t-value helps us decide if our findings are significant. However, we also rely on the p-value, indicating the likelihood of observing our results if the null hypothesis were true. By examining both, we can confidently determine if there's a genuine distinction between the two groups

# **Results**

The results of the experiment comparing the performance of decision tree and neural network algorithms in detecting breast cancer are presented below.

The table below displays the mean accuracy, standard deviation, and p-value for each machine learning algorithm test. The mean accuracy represents the average performance of the algorithm across the 30 repeats of the experiment. Standard deviation indicates the spread of data around the mean.

|  |  |  |
| --- | --- | --- |
| Performance Metrics | Decision Tree (J48) | Neural Network (Multilayer Perception) |
| Mean Accuracy | 0.7113 | 0.6756 |
| Standard Deviation | 0.0346 | 0.0353 |



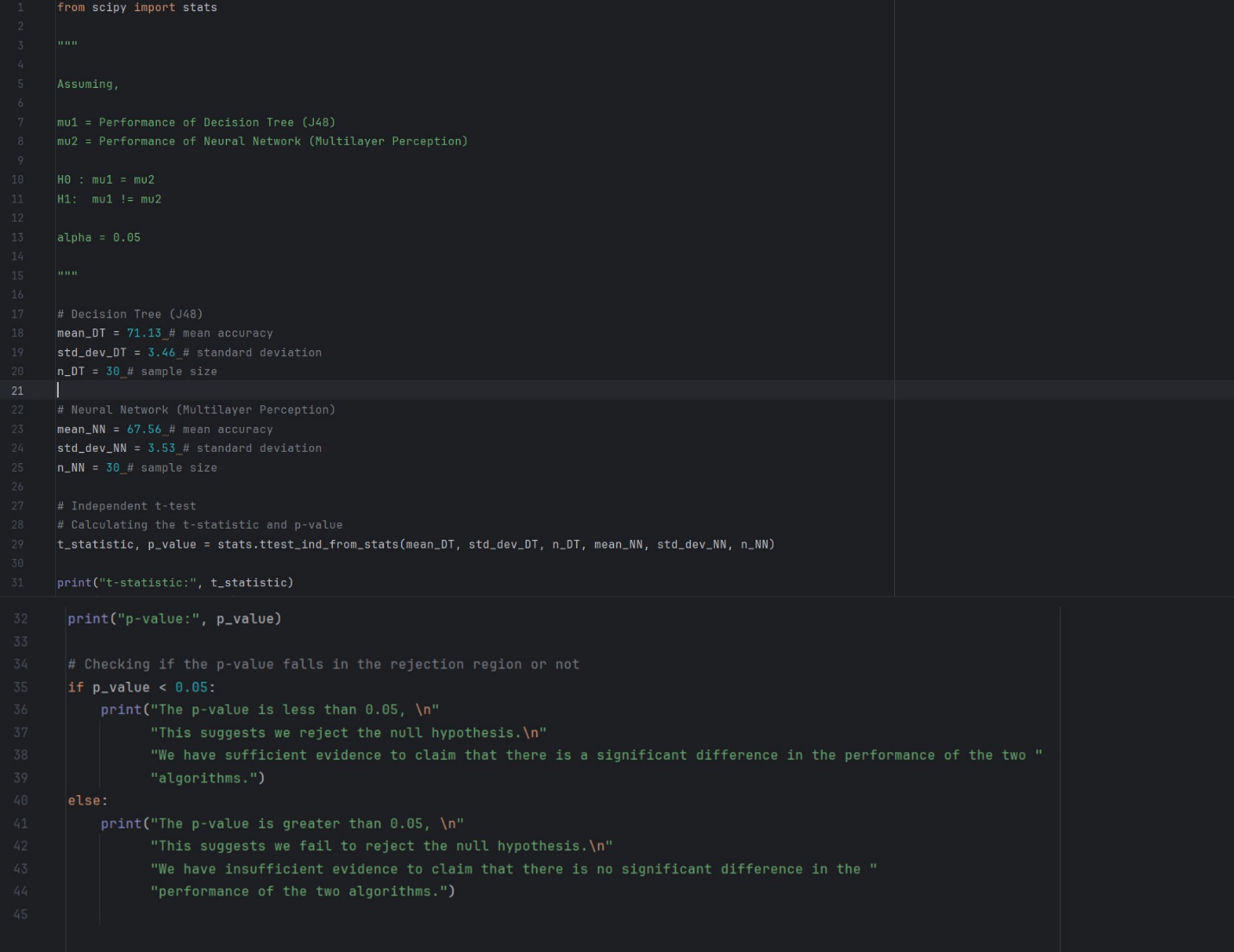
Accuracy Matrix of Both Models 1

The experiment revealed the mean accuracy of 0.7113 (SD = 0.0346) for the decision tree algorithm and 0.6756 (SD = 0.0353) for the neural network algorithm.

## **Hypothesis Testing:**

To evaluate the hypothesis regarding the difference in accuracy between the decision tree and neural network algorithms for breast cancer detection, a two-sample independent t-test was conducted using Python's ‘scipy.stats’ module.

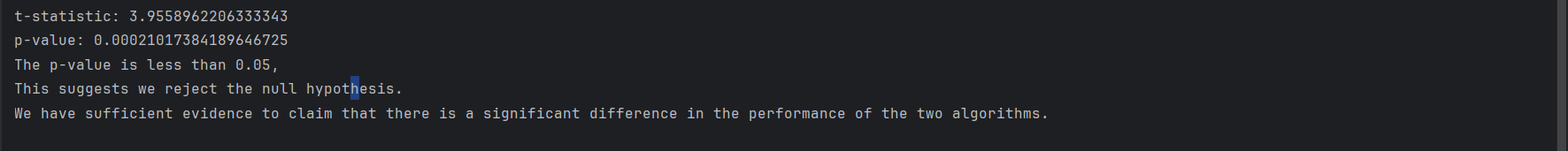
The null hypothesis states that there is no significant difference between the mean accuracy of the two algorithms (J48 and Multilayer Perception). The mean, standard deviation, and sample size was calculated through the experiment.



Code 1

The designed Python program calculated the following statistics:

1. t-statistic - 3.9558962206333343
2. p-value - 0.00021017384189646725



Result 1

According to the analysis the p-value is less than the significance level (0.05), it indicates that there is sufficient evidence to reject the null hypothesis and conclude that there is a significant difference in the performance of both the models.

# **References**

[Importance of detecting breast cancer early](https://www.proquest.com/docview/2448807019?parentSessionId=NaN%2FVzhSsvU6A%2BP63Aa%2FzxjaCTdIZVBPXPYtTW0D%2Ftw%3D&pq-origsite=primo&accountid=3455&sourcetype=Wire%20Feeds)

Translated by ContentEngine LLC.  CE Noticias Financieras, English ed.; Miami. 05 Oct 2020.

[Detecting breast cancer early](https://www.proquest.com/docview/1823245063/A4922F01416D45FEPQ/1?accountid=3455&sourcetype=Newspapers)

Renckens, Ali.  The Jackson Sun; Jackson, Tenn. 19 Sep 2016: H.6.

[Breast Cancer Detection by Using Decision Tree](https://doi.org/10.1007/978-3-031-45121-8_7)

ICACIS 2023:[Advanced Communication and Intelligent System](https://link.springer.com/book/10.1007/978-3-031-45121-8) pp 71-79

[A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks](https://www.proquest.com/docview/212469412?parentSessionId=675QUbDIhIY7xXlhsrO5PY%2FveKh5Q4y6v8HbxxHLzVg%3D&pq-origsite=primo&accountid=3455&sourcetype=Scholarly%20Journals)

Lancashire, L J; Powe, D G; Reis-filho, J S; Rakha, E; Lemetre, C; et al.  Breast Cancer Research and Treatment; Dordrecht Vol. 120, Iss. 1, (Feb 2010): 83-93. DOI:10.1007/s10549-009-0378-1

[A Fast Decision Tree Learning Algorithm](https://cdn.aaai.org/AAAI/2006/AAAI06-080.pdf)

Jiang Su and Harry Zhang

Faculty of Computer Science

University of New Brunswick, NB, Canada, E3B 5A3

{jiang.su, [hzhang}@unb.ca](mailto:hzhang%7d@unb.ca)

[Artificial Neural Networks for Beginners](https://arxiv.org/abs/cs/0308031)

Carlos Gershenson

[v1] Wed, 20 Aug 2003 09:40:25 UTC

f the two algorithms.