big data and cognitive computing

MDPI

Article

Fuzzy Neural Network Expert System with an Improved Gini Index Random Forest-Based Feature Importance Measure Algorithm for Early Diagnosis of Breast Cancer in Saudi Arabia

Ebrahem A. Algehyne [1,*], Muhammad Lawan Jibril [2,*], Naseh A. Algehainy [3], Osama Abdulaziz Alamri [4] and Abdullah K. Alzahrani [5]

[1] Department of Mathematics, Faculty of Science, University of Tabuk, Tabuk 71491, Saudi Arabia
[2] Department of Computer Science, Federal University of Kashere, Gombe State P.M.B. 0182, Nigeria
[3] Department of Medical Laboratory Technology, Faculty of Applied Medical Sciences, University of Tabuk, Tabuk 71491, Saudi Arabia; nalgehainy@ut.edu.sa
[4] Department of Statistic, Faculty of Science, University of Tabuk, Tabuk 71491, Saudi Arabia; oalmughamisi@ut.edu.sa
[5] Department of Mathematics, Faculty of Science, King Abdulaziz University, Jeddah 21589, Saudi Arabia; akalzahrani@kau.edu.sa
* Correspondence: e.algehyne@ut.edu.sa (E.A.A.); lawan.jibril@fukashere.edu.ng (M.L.J.)

Abstract: Breast cancer is one of the common malignancies among females in Saudi Arabia and has also been ranked as the one most prevalent and the number two killer disease in the country. However, the clinical diagnosis process of any disease such as breast cancer, coronary artery diseases, diabetes, COVID-19, among others, is often associated with uncertainty due to the complexity and fuzziness of the process. In this work, a fuzzy neural network expert system with an improved gini index random forest-based feature importance measure algorithm for early diagnosis of breast cancer in Saudi Arabia was proposed to address the uncertainty and ambiguity associated with the diagnosis of breast cancer and also the heavier burden on the overlay of the network nodes of the fuzzy neural network system that often happens due to insignificant features that are used to predict or diagnose the disease. An Improved Gini Index Random Forest-Based Feature Importance Measure Algorithm was used to select the five fittest features of the diagnostic wisconsin breast cancer database out of the 32 features of the dataset. The logistic regression, support vector machine, k-nearest neighbor, random forest, and gaussian naïve bayes learning algorithms were used to develop two sets of classification models. Hence, the classification models with full features (32) and models with the 5 fittest features. The two sets of classification models were evaluated, and the results of the evaluation were compared. The result of the comparison shows that the models with the selected fittest features outperformed their counterparts with full features in terms of accuracy, sensitivity, and sensitivity. Therefore, a fuzzy neural network based expert system was developed with the five selected fittest features and the system achieved 99.33% accuracy, 99.41% sensitivity, and 99.24% specificity. Moreover, based on the comparison of the system developed in this work against the previous works that used fuzzy neural network or other applied artificial intelligence techniques on the same dataset for diagnosis of breast cancer using the same dataset, the system stands to be the best in terms of accuracy, sensitivity, and specificity, respectively. The z test was also conducted, and the test result shows that there is significant accuracy achieved by the system for early diagnosis of breast cancer.

Keywords: neural network; fuzzy logic; breast cancer; random forest; dataset

check for updates

Citation: Algehyne, E.A.; Jibril, M.L.; Algehainy, N.A.; Alamri, O.A.; Alzahrani, A.K. Fuzzy Neural Network Expert System with an Improved Gini Index Random Forest-Based Feature Importance Measure Algorithm for Early Diagnosis of Breast Cancer in Saudi Arabia. *Big Data Cogn. Comput.* **2022**, *6*, 13. https://doi.org/10.3390/bdcc6010013

Academic Editor: Min Chen

Received: 5 December 2021
Accepted: 13 January 2022
Published: 27 January 2022

1. Introduction

Cancer is a group of diseases that are characterized by the uncontrollable spread and growth of abnormal cells [1,2]. Therefore, if the spread and growth of abnormal cells is not controlled, it may lead to death. Breast cancer is one of the multifactorial genetic disorders or diseases that is likely associated with the effect of multiple genes in the combination of environmental factors and lifestyle [3,4]. Breast cancer is one of the most common types of cancer and one of the deadliest cancers after lung cancer [5–7]. It is also the most common cancer among women, even though it happens in men and also where the growth of the cell starts in the breast of women or men [8]. Breast cancer occurs when certain cells in the breast become abnormal and multiply uncontrollably to form a tumor [9,10]. Breast cancer is one of the common malignancies among females in Saudi Arabia [11–14]. According to the GLOBOCAN report of the year 2018, there were 3629 new cases of breast cancer, which amounts to 14.8% of the total new disease cases in Saudi Arabia, with 899 deaths, which amounts to 8.5% of the total death cases [10,15]. Breast cancer is ranked as the one most prevalent and the number two killer disease in Saudi Arabia [9,16–18].

Clinical diagnosis or decision making related to the procedure tailored to define the reason for the disease of the patients is largely complex, and its precision depends upon the interpretation of the medical personnel's elevated perception and experience [19–21]. However, the conventional clinical diagnosis process of diseases is often associated with uncertainty and ambiguity due to complexity and fuzziness in the course of diagnosis of most of the deadly diseases, such as breast cancer, coronary artery diseases, diabetes, waterborne diseases, among others [20,22,23]. Therefore, since the conventional procedure for medical diagnosis involves uncertainty and ambiguity, recently, two or more artificial intelligence techniques are combined to carry out research to overcome the limitations of underperformance and the inefficiency of using an intelligence (AI) technique for the development of intelligent systems for the diagnosis of diseases, such as multidimensionality, nonlinearity, uncertainty, and vagueness [24–27]. Hence, fuzzy logic and neural networks are being used to develop expert systems to deal with such uncertainty and ambiguity [28–30].

Expert system is an AI based system that imitates the decision-making ability of human experts to complement humans while making various decisions [19,29–32]. Expert system has three components which include knowledge acquisition, knowledge reasoning, and knowledge presentation [5,30]. The knowledge acquisition, which involves knowledge transfer from a human expert to the knowledge base of the expert, is the most challenging task during the development of the expert system [19,25,29]. Therefore, historical datasets are currently being used to generate or transfer human expertise using artificial learning techniques from it for subsequent transfer to the knowledge base of the expert system [20,33,34]. However, most of the datasets contain many features, which many of them might be insignificant if not irrelevant for diagnosis of diseases [34–36]. Hence, if these insignificant features are not removed, they might cause a heavier burden on the overlay of the network nodes of the fuzzy neural network system, which would reduce the diagnosis accuracy, increase the time needed for the training, and make the interpretation of diagnostic results of the system very difficult to be understood [37–40]. Therefore, feature selection techniques are used to determine the significant score or how useful and important the features are for predicting a target feature [41–44]. Machine learning is also one of the sub-branches of AI that deals with the ways in which machines learn from experience [45,46]. Machine learning algorithms are being used to develop the diagnostic models of many diseases and they helped the systems to learn the diagnosis data, identify useful patterns during the learning process, and minimize human interference in making decisions. The research questions that this work answered include the following:

i. How to develop an expert system that would address the uncertainty often associated with diagnosis of breast cancer?

ii. How to address the heavier burden on the overlay of the network nodes of fuzzy neural network system with feature selection technique?

iii. How to find out the five fittest features of the diagnostic wisconsin breast cancer database among 32 features of the dataset?

iv. How to justify why the five fittest features of the diagnostic wisconsin breast cancer database are more important than using all of the feature dataset with the help of machine learning algorithms?

The motivation of this work is to develop a fuzzy neural network-based expert for early diagnosis of breast cancer in Saudi Arabia that would address the uncertainty and ambiguity associated with the diagnosis process of the disease and also the heavier burden on the overlay of the network nodes of the fuzzy neural network system that often happens due to insignificant features that are used to predict or diagnose the disease.

2. Related Work

Many AI techniques have been used for the diagnosis and prognosis of breast cancer. A neuro-fuzzy based expert system for diagnosis of breast cancer has been developed in the work of [47] and the system achieved 76% sensitivity and 97% specificity. In Reference [48], a machine learning based system for cancer screening was developed; the system was designed to diagnose cancer to determine whether it is benign or malignant using the artificial neural network. The network was trained using a back-propagation algorithm and then a neuro-fuzzy system was developed based on the model. In the work of [49], a hybrid expert system that combines an incremental fuzzy neural network and fuzzy linguistics for the diagnosis of cancer to deal with the uncertainty associated with clinical decision making was developed. The system was developed on the wisconsin breast cancer screening dataset which contains 683 records, each with 9 characteristics such as clump strength, parallelism in cell size, cell shape similarity, etc., and the proposed approach had achieved 99.08% accuracy and 98.74% sensitivity. The study of [50] introduced and developed a neural network fuzzy system for skin cancer identification and classification. The system involved preliminary processing and optimization of the tumor location using the color detection method, followed by extracting the color and the appropriate region. Classification system for high blood pressure based on a neuro-fuzzy approach was proposed in the study of [30]; the system was tested using sample data of 10 patients, each with inclusion parameters, i.e., age, body mass index (BMI), blood pressure (BP), heart rate (HR), and one risk limit as an output variable for the model. In the work of [24], a fuzzy relational model with generic algorithm for early diagnosis and detection of breast cancer in Saudi Arabia was proposed. In the work of [36], a particle swarm optimization was embedded into k-nearest neighbor, naïve Bayes, and fast decision tree algorithms to improve the breast cancer prediction accuracy by the models developed with the algorithms. Fuzzy ID3 algorithm was proposed in the work of [5] for early prediction of breast cancer to increase the classification accuracy of the decision tree, and the study identified the proposed algorithm as robust and reliable. A semantic rule-based diagnostic decision-making support system for the diagnosis of breast cancer was proposed in the work of [3], and the system was found to be able to diagnose breast cancer accurately. A fuzzy approach for pre-diagnosis of breast cancer from fine needle aspirate analysis was proposed in the work of [51], and the ability of the approach to diagnose breast cancer accurately varies from 65% to 98%.

Although many works have been carried out on AI based diagnostic systems for breast cancer, especially the fuzzy neural network-based expert system, there is a need to propose more AI based methods that would address the uncertainty of diagnostic decision making and the heavier burden on the overlay of the network nodes of the system to improve the accuracy, specificity, and sensitivity of the system. Hence, in this study, a fuzzy neural network expert system with an improved gini index random forest-based feature importance measure algorithm has been proposed for early diagnosis of breast cancer in Saudi Arabia to address the below identified challenges:

i. The heavier burden on the overlay of the network nodes of fuzzy neural network-based expert system is due to many insignificant features that are used to predict or diagnose the disease

ii. The uncertainty and ambiguity are often associated with diagnostic decision making of breast cancer.

3. Methods and Materials

In this section, the methodology used for development and implementation of fuzzy neural network expert system with an improved gini index random forest-based feature importance measure algorithm for early diagnosis of breast cancer in Saudi Arabia is discussed. Figure 1 shows the materials and methods involved in the methodology.

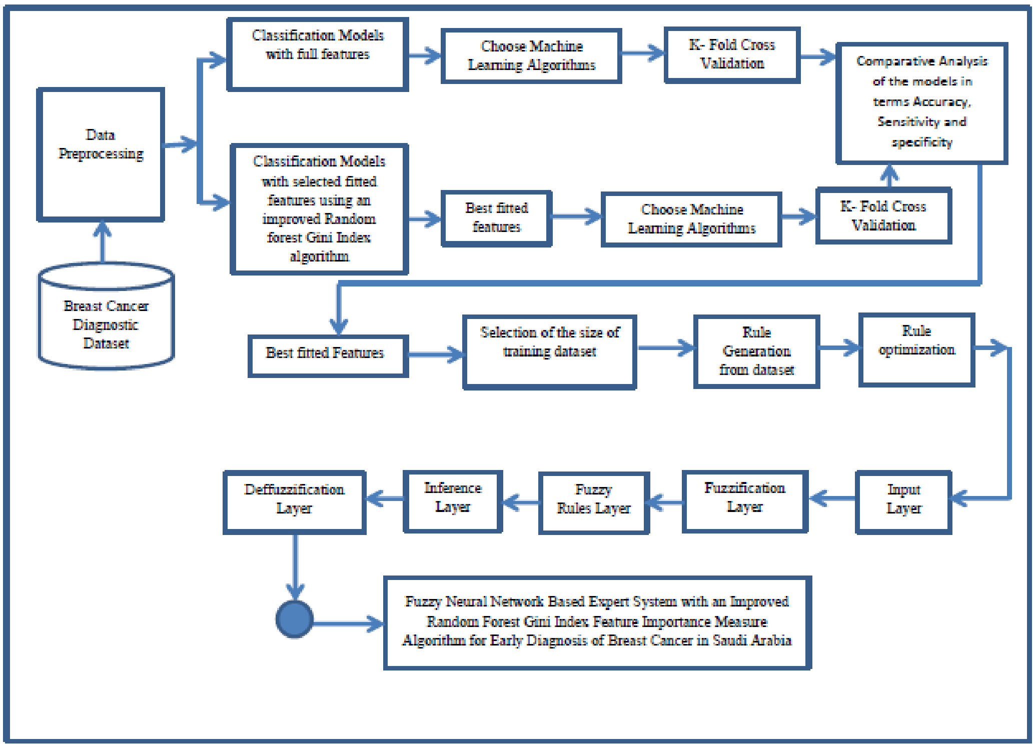

Classification Models with full features
Choose Machine Learning Algorithms
K- Fold Cross Validation
Comparative Analysis of the models in terms Accuracy, Sensitivity and specificity
Data Preprocessing
Classification Models with selected fitted features using an improved Random forest Gini Index algorithm
Best fitted features
Choose Machine Learning Algorithms
K- Fold Cross Validation
Breast Cancer Diagnostic Dataset
Best fitted Features
Selection of the size of training dataset
Rule Generation from dataset
Rule optimization
Deffuzzification Layer
Inference Layer
Fuzzy Rules Layer
Fuzzification Layer
Input Layer
Fuzzy Neural Network Based Expert System with an Improved Random Forest Gini Index Feature Importance Measure Algorithm for Early Diagnosis of Breast Cancer in Saudi Arabia

Figure 1. Research methodology.

3.1. Dataset Collection, Preprocessing, and Feature Selection

In this phase, the relevant dataset of the problem domain is collected, preprocessed, and cleaned. Therefore, only relevant data would be collected with the help and guidance of human experts of the problem domain to be modeled. Missing data are also dealt with in this phase and the redundant and derived features of the dataset are also discarded. When building a system or model in real life, it is almost rare that not all the features in the dataset are useful to build a model [52]. Adding redundant variables reduces the generalization ability of the model and may also reduce the overall accuracy of a model or system [27,53–55]. Furthermore, adding increasing variables to the system or models increases the overall complexity of the model [7,27].

We used an improved random forest gini index algorithm proposed by [56] as a feature selection technique that assigns a score to input features based on how useful and important they are for predicting a target feature of the dataset. Random forest algorithm is one of the popular machine learning approaches for analyzing high dimensional data and building models. An important reason for its popularity is the availability of feature selection or variable importance measures [57,58]. Random forest algorithm is a feature selection technique that assigns a score to input features based on how useful and important they are for predicting a target feature of the dataset [59]. The algorithm is a nonparametric technique that builds trees from a random subset of features used for classification and regression, respectively. The most widely used feature selections of random forest algorithm are impurity importance and permutation importance. The impurity importance is known as mean decrease in impurity and called the Gini index algorithm. The gini index algorithm provides multivariate feature importance scores which are relatively cheap to obtain, and the techniques have been successfully applied to high dimensional datasets arising from microarrays [39,42,52].

The Gini index algorithm was based on gini theory. The basic idea of gini index theory is that, let us assume S is a set of samples s and these samples have k classes (C_i, i =1, . . . , k). Therefore, we can divide S into k subsets according to the differences between the classes. Let us assume S_i is a sample set that belongs to class C_i and that s_i is the sample number of set S_i, then the Gini index of Set S can be obtained using equation:

$$Gini(S) = 1 - \sum_{i=1}^{m} P_i^2 \tag{1}$$

where P_i stands for the probability which is estimated with s_i/s, for any samples that belong to C_i. The minimum of Gini (S) is 0, and all of the members in the set belong to same class, indicating that maximum useful information can be obtained. Thus, with the maximum of Gini (S), indicate that the minimum useful information can be obtained.

A novel gini index algorithm based on gini index theory for text feature selection was proposed by Reference [59] and it was expressed in below equation:

$$Gini(W) = P(W)\left(1 - \sum_{i=1}^{m} P(C_i\,|W\,)^2\right) + P(\overline{W})\left(1 - \sum_{i=1}^{m} P(C_i\,|\overline{W})^2\right) \tag{2}$$

where W is a feature and C_i is i-th class among the classes. The expression eliminates the affection factor that expresses the words that do not appear and adopts a measure of purity instead of impurity to be emphasized *P(W)* factor, called *Gini–A*, as expressed in the below equation:

$$Gini(W) = P(W)\sum_{i=1}^{m} P(C_i\,|W\,)^2 \tag{3}$$

Therefore, considering the unbalanced distribution of classes, the posterity probability where the feature W appears $\sum_i P(W|C_I)^2$, to replace *P(W)*, called *Gini–B* as shown in the below equation:

$$Gini(W) = \sum_{i}^{m} P(W|C_i\ )^2\, P(C_i\,|W\,)^2 \tag{4}$$

Thus, if feature W appears in every class C_i, the maximum value of Gini Value = 1, can be obtained.

Random forest gini index algorithm has been receiving increased attention as a feature selection due to some of its advantages, which include flexibility, fastness, and robust approach toward handling high dimensional data [58,60]. However, despite its numerous advantages, random forest algorithm as a feature selection technique has some limitations, which include bias in favor toward features with many categories or many possible split points and high minor allele frequency [40,58,61,62]. However, many works have been conducted to address the issue of the bias of random forest gini index algorithm [63]. In this study, we used the improved random forest gini index algorithm proposed by [56]. The algorithm contains three reformulated gini index expressions for feature selection to address the issue of the bias of the random forest gini index algorithm. The improved algorithm was reformulated based on Shang's gini index, expressed by Reference [59] and discussed earlier. In order to address the bias of Gini index algorithm for feature selection, Equation (3), namely *IGini–A*, as shown below:

$$IGini_A(W) = \sum_{i=1}^{m} P(C_i\,|W\,)^2 \tag{5}$$

Thus, *P(W)* is eliminated from Equation (3), because most of the features have low frequencies in the dataset, the *P(W)* are very small, and the gini values are more influenced by *P(W)* than $P(C_i\,|W\,)^2$, likewise for features with high frequencies, *P(W)* is relatively higher, and thus gini values are more influenced by *P(W)*. Therefore, calculating gini values with $P(C_i\,|W\,)^2$ is more efficient and it addresses the bias of Gini index.

The Gini values obtained by the equation can be normalized by logarithm base 2 of the probability *P(W)* and its absolute value to reduce the range of *P(W)* and keep positive. This can be obtained with the below equations, namely *IGini–B*:

$$IGini_B(W) = \frac{1}{log_2 P(W)} \sum_{i=1}^{m} P(C_i \,|W\,)^2 \tag{6}$$

Equation (6) above is more efficient at estimating specific features and general features to further address the bias of the Gini index. In order to normalize the probability of *P(W* | C_i*)* and produce unbiased Gini values, below equations, namely *IGini–C* below can be used:

$$IGini_C(W) = |\frac{1}{log_2 P(W|C_i\,)^2}| \sum_{i=1}^{m} P(C_i \,|W\,)^2 \tag{7}$$

3.1.1. Classification Models with Full Features

In this phase, five renowned machine learning algorithms, which include logistic regression, support vector machine, k-nearest neighbor, random forest, and gaussian naïve bayes algorithms, would be used to develop classification models with full features of the dataset and models would be tested with the option of 10-folds cross-validation techniques. The training dataset would be used for evaluation of the models. Below is a short description of the five renowned machine learning algorithms:

i. Logistic regression algorithm is used to model the probability of an event or certain classes and it uses logistic function to model binary dependent features or variables against the independent one [34]. Logistic regression does not really have any critical hyper parameters to tune. Sometimes, you can see useful differences in performance or convergence with different solvers [64].

ii. Support vector machine (SVM) is being used to model data for classification and regression analysis, respectively [34]. SVM builds a set of hyper plane in infinite dimensional space which might be used for regression or classification or even other tasks such as the detection of outliers. SVM provides a large number of hyper parameters to tune. Perhaps the first important parameter is the choice of kernel that will control the manner in which the input variables will be projected [65].

iii. K-nearest neighbor is also a learning algorithm that stores all the existing cases and classifies new cases based on the measure of similarity and it classifies a case by a vote of the majority of its neighbors with the case that is being given the most common category among its closest neighbors, measured by a distance function. If K = 1, then that case is assigned to its nearest neighbor [16,66]. The most important hyper parameter for k-nearest neighbor is the number of neighbors [65].

iv. Random forest is a learning algorithm for regression and classification by building a multitude of decision trees at training time and producing the class that is the mode of the classes of the individual decision trees [61]. The algorithm consists of a number of individual decision trees that work as an ensemble where each decision tree in the random forest gives out a class of prediction and the class with majority votes becomes the predictive model. [66]. The most important parameter is the number of random features to sample at each split point of the maximum of features [64,67].

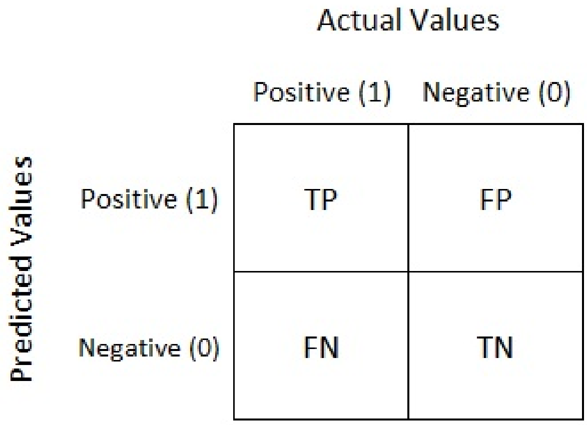
v. Gaussian naïve bayes is a learning algorithm for probabilistic classification by applying bayes theorem with strong independence assumption between dataset features [35]. The learning algorithm is highly scalable, which requires a number of parameters linear in the number of features in the learning problem [34]. Moreover, naïve Bayes has almost no hyper parameters to tune, so it usually generalizes well. One thing to note is that, due to the feature independence assumption, the class probabilities output by naïve bayes can be pretty inaccurate [65].

3.1.2. Classification Models with Selected Fittest Features

In this phase, the five fittest features selected with an improved random forest gini index algorithm proposed by [56] would be used to develop classification models with five renowned machine algorithms mentioned earlier and models would be tested with the option of 10-folds cross-validation technique.

3.1.3. Comparative Analysis of the Performance Evaluation of Classification Models

In this phase, five classification models developed with full features and five classification models also developed with the five fittest features selected with an improved random forest gini index algorithm would be compared. The classification models would be evaluated with the aid of confusion metric. The metric consists of information about the actual and predicted cases of the models. Figure 2 shows the confusion matric classification measure representation. Therefore, the classification models developed with the five fittest features selected with an improved random forest gini index algorithm would be evaluated against the classification models developed with full features based on accuracy, sensitivity, and specificity.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 2. Confusion matric.

i. Accuracy: It is used to determine the efficiency of the models in terms of the number of the corrected predicted breast cancer cases against the total number of available instances of the dataset as defined by the formula:

$$\frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

ii. Sensitivity: It is used to determine the efficiency of the models in terms of the number of predicted positive cases of the breast cancer against the total number of actual positive instances of the dataset as defined by the formula:

$$\frac{TP}{TP + FP} \quad (9)$$

iii. Specificity: It is used to determine the efficiency of the system in terms of using the predicted negative cases of the breast cancer against the total number of actual negative instances of the dataset as defined by the formula:

$$\frac{TP}{TP + FN} \quad (10)$$

3.2. Selection of the Size of Training Dataset

Selection of the size of the training dataset is also very critical for the efficiency and accuracy of diagnosis of the fuzzy neural network-based expert system [68]. Large size of the training dataset may increase the accuracy of the inference techniques, but it would however cause a heavier burden on the overlay of the network nodes. The methodology uses the limited training dataset with features of the dataset with the highest feature ranking score to build a simpler and more reliable fuzzy neural network-based expert system.

3.3. Rule Generation from Dataset

In this phase, diagnosis rules are being generated from the selected training dataset. The rules are the fundamental part of the knowledge base of the fuzzy neural network-based expert system. The training set of n inputs and a single output represent a dynamic expert system which can be defined by *n* features y_1, y_2, … y_n.

3.4. Rule Optimization

In this phase, the fuzzy rules being generated from the selected training dataset are redefined using the neural network technique. Therefore, five network neural layers are to be used for redefining the fuzzy rules and each layer defines the inference mechanism steps in the fuzzy neural network-based expert system. Figure 3 shows the five network neural layers of the fuzzy neural network-based expert system inference mechanism.

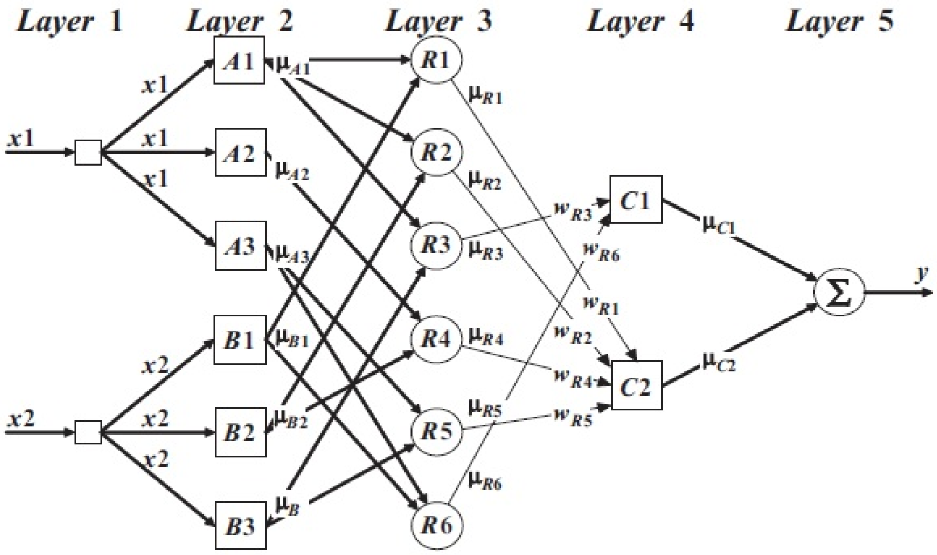


Figure 3. Five layers of fuzzy neural network-based expert inference system.

3.4.1. Input Layer

The neurons of the input layer have a linear function and it represents the system input features which are the current state of the system. Thus, the output of the neurons lying on the input layer are the corresponding input values and those neurons in this layer transmit the input features directly to their respective membership modules in the second layer.

3.4.2. Fuzzification Layer

In this layer, the input features of the system are being fuzzified into corresponding fuzzy values. Thus, the neurons in this layer convert the scalar values of the input features received from the input layers to their corresponding fuzzy values. Therefore, the membership values of each input feature are determined.

3.4.3. Fuzzy Rules Layers

This layer is also called the knowledge base of the predicator because it contains the possible combination of all input features of the system and performs logical AND operations. Therefore, the neurons in this layer represent the fuzzy rules of the system and the numbers of neurons are equal to the number of fuzzy rules. The main function of the neurons in this layer is to perform logical AND operations by the product of their respective input features of the system and pass the output to the next layer.

3.4.4. Inference Layer

In this layer, the neurons identify the firing rules corresponding to the set of input features of the system. Therefore, the neurons represent the output membership sets and each neuron combines all its input features AND operation. Therefore, the output of each neuron shows the stretcher of its input rules.

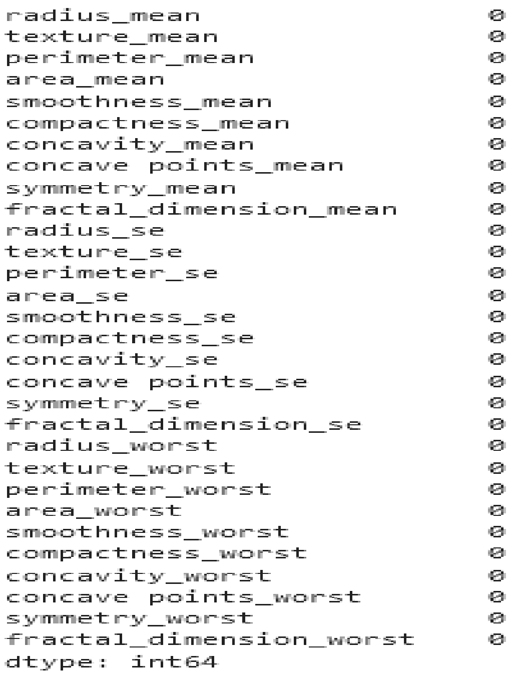
3.4.5. Defuzzification Layer

This layer has only one neuron that represents the final out of the system and it takes the input from the output fuzzy rules and combines them into single fuzzy rules. In this layer, defuzzification for converting the fuzzified output into corresponding crisp values of the input features of the system are being performed. Back propagation learning is used to rule optimization to adjust the parameters of the membership functions optimally. Therefore, the training data of input and output are compared with the system output and the errors are propagated backwards through the network from the output layer to the input layer with the aim of modifying the membership functions of the neurons in layer 2.

4. Fuzzy Neural Network Expert System with an Improved Gini Index Random Forest-Based Feature Importance Measure Algorithm

4.1. Dataset Collection, Preprocessing, and Feature Selection

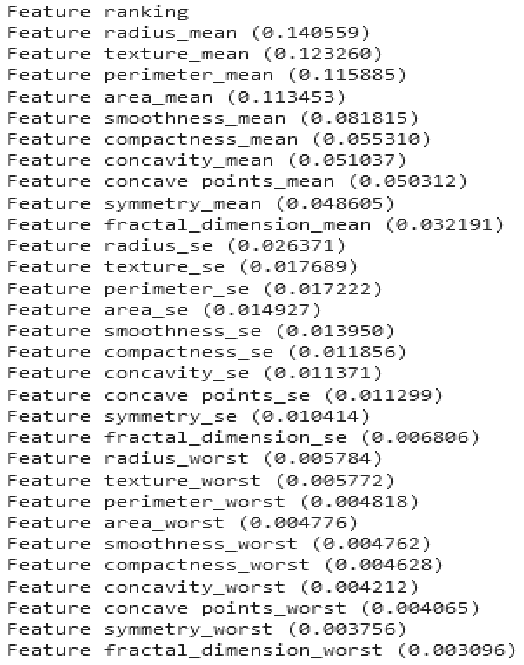
Breast cancer dataset was collected from UCI public database (*University of California*) [69]. The dataset is called the diagnostic wisconsin breast cancer database and it has 32 features and 569 instances. In this phase, the dataset went through the cleaning process where missing, noisy, and irrelevant data were eliminated. There are no missing values in the dataset as shown in Figure 4.



```
radius_mean                    0
texture_mean                   0
perimeter_mean                 0
area_mean                      0
smoothness_mean                0
compactness_mean               0
concavity_mean                 0
concave points_mean            0
symmetry_mean                  0
fractal_dimension_mean         0
radius_se                      0
texture_se                     0
perimeter_se                   0
area_se                        0
smoothness_se                  0
compactness_se                 0
concavity_se                   0
concave points_se              0
symmetry_se                    0
fractal_dimension_se           0
radius_worst                   0
texture_worst                  0
perimeter_worst                0
area_worst                     0
smoothness_worst               0
compactness_worst              0
concavity_worst                0
concave points_worst           0
symmetry_worst                 0
fractal_dimension_worst        0
dtype: int64
```

Figure 4. Statistics of the missing values of the datasets features.

The dataset has 32 features and 569 instances, but not all of them are useful to build the system and they may reduce the generalization ability and overall accuracy of the system. An Improved Gini Index Random Forest-Based Feature Importance Measure Algorithm was used as a feature selection technique that assigns a score to all input features based on how useful and important they are for predicting breast cancer in Saudi Arabia. Figure 5 shows the importance measure scores of the features of the dataset and Figure 6 shows the graphical presentation of the feature ranking scores.



```
Feature ranking
Feature radius_mean (0.140559)
Feature texture_mean (0.123260)
Feature perimeter_mean (0.115885)
Feature area_mean (0.113453)
Feature smoothness_mean (0.081815)
Feature compactness_mean (0.055310)
Feature concavity_mean (0.051037)
Feature concave points_mean (0.050312)
Feature symmetry_mean (0.048605)
Feature fractal_dimension_mean (0.032191)
Feature radius_se (0.026371)
Feature texture_se (0.017689)
Feature perimeter_se (0.017222)
Feature area_se (0.014927)
Feature smoothness_se (0.013950)
Feature compactness_se (0.011856)
Feature concavity_se (0.011371)
Feature concave points_se (0.011299)
Feature symmetry_se (0.010414)
Feature fractal_dimension_se (0.006806)
Feature radius_worst (0.005784)
Feature texture_worst (0.005772)
Feature perimeter_worst (0.004818)
Feature area_worst (0.004776)
Feature smoothness_worst (0.004762)
Feature compactness_worst (0.004628)
Feature concavity_worst (0.004212)
Feature concave points_worst (0.004065)
Feature symmetry_worst (0.003756)
Feature fractal_dimension_worst (0.003096)
```

Figure 5. Feature ranking scores.

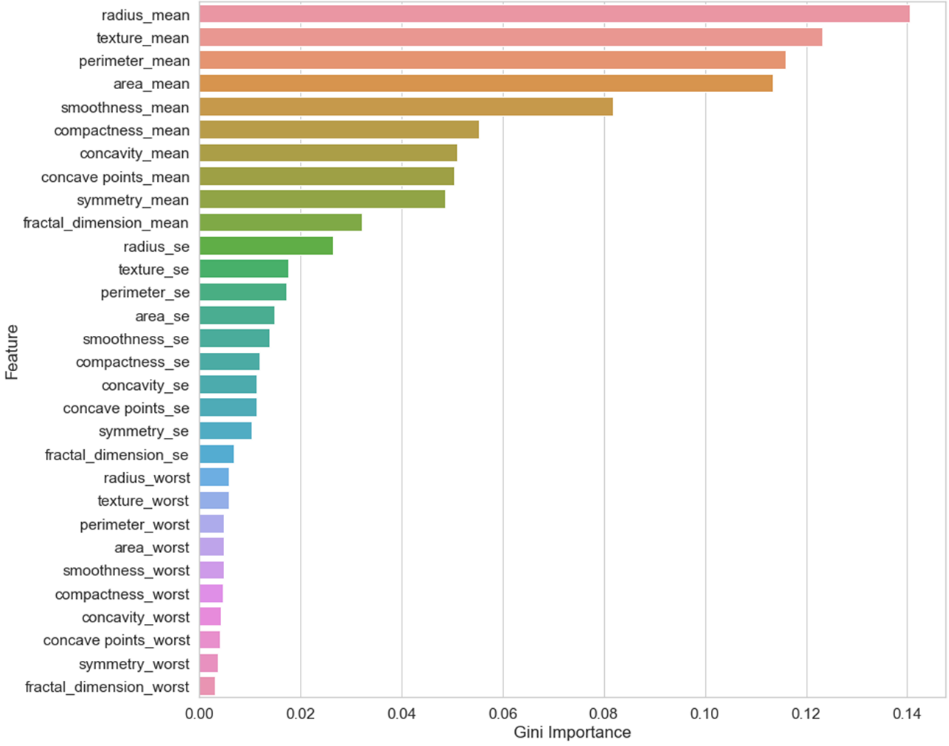


Figure 6. Graphical presentation of feature ranking scores.

Based on the importance measure scores of the features of the dataset, we choose the five fitness features of the dataset for the development of the fuzzy neural network-based expert system to reduce a heavier burden on the overlay of the network nodes of the system. This would help to build a simpler and more reliable system. The five fittest features of the dataset include radius mean, texture mean, perimeter mean, area mean, and smoothness mean. Table 1 shows the unit and value ranges of the features of the dataset.

Table 1. Unit and value ranges of the features.

SN	Features	Units	Range
1	RadiusMean	mm	6.981–28.11
2	TextureMean	dimensionless	9.71–39.28
3	PerimeterMean	mm	43.79–188.5
4	AreaMean	µm^2	143.5–2501
5	SmoothnessMean	mm	0.05263–0.1634
6	The Diagnosis of Breast Tissues	malignant (0), benign (1)	0,1

1 µm (micrometer) = 1 × 10^{-6} m; 1 mm (millimeter) = 1 × 10^{-3} m.

4.1.1. Classification Models with full features

Five machine algorithms, which include logistic regression, support vector machine, k-nearest neighbor, random forest, and gaussian naïve bayes learning algorithms, were used to develop classification models with full features of the dataset. The models were evaluated with 10-folds cross-validation techniques. The result of the experimental performance evaluation of classification models has been shown in Table 2 and Figure 7, respectively. Random forest, in terms of accuracy, and specificity random-forest–forest-based model achieved the highest accuracy of 76.20% and 93.40% specificity, respectively. While in terms of sensitivity, the gaussian naïve bayes based model has the highest sensitivity of 87.12%.

Table 2. Performance evaluation result of classification models with full features.

Model	Accuracy (%)	Specificity (%)	Sensitivity (%)
Logistic regression	66.30	74.24	36.80
Support vector machine	75.00	85.20	56.00
K-nearest neighbor	66.10	75.40	67.19
Random forest	76.20	93.40	71.16
Gaussian naïve bayes	70.00	78.70	87.12

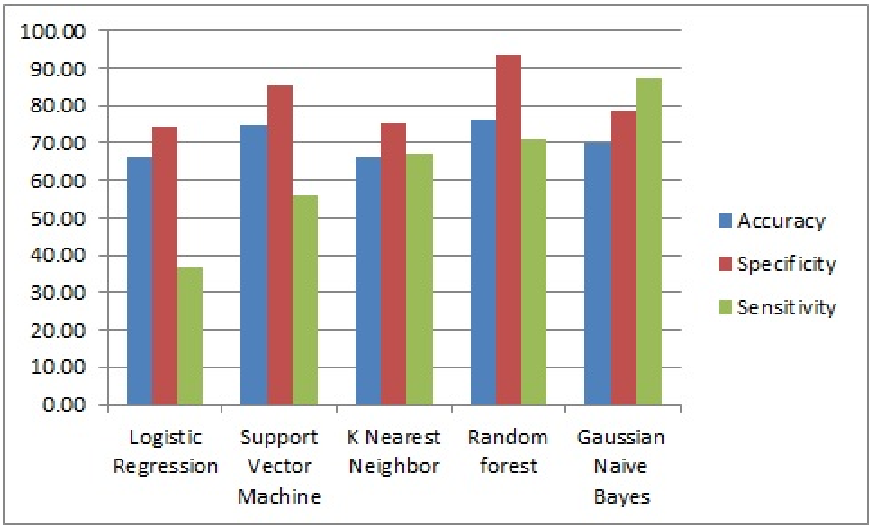


Figure 7. Chart presentation of performance evaluation result of classification models with full features.

4.1.2. Classification Models with Selected Fittest Features

Logistic regression, support vector machine, k-nearest neighbor, random forest, and gaussian naïve bayes learning algorithms were used to develop classification models with only five fittest selected features of the dataset. The models were evaluated with 10-folds cross-validation techniques. The result of the experimental performance evaluation of classification models has been shown in Table 3 and Figure 8, respectively. In terms of accuracy, random forest-based also achieved the highest accuracy of 95.61%, while in terms of specificity, the gaussian naïve bayes based model achieved the highest specificity of 98.34% and, in terms of sensitivity, the k-nearest neighbor-based model achieved the highest sensitivity of 98.11%.

Table 3. Performance evaluation of classification models with five fittest selected features.

Model	Accuracy (%)	Specificity (%)	Sensitivity (%)
Logistic Regression	95.60	96.90	90.11
Support vector machine	95.60	97.45	78.51
K-nearest neighbor	94.74	98.00	98.11
Random forest	95.61	93.56	88.45
Gaussian naïve bayes	93.86	98.34	87.18

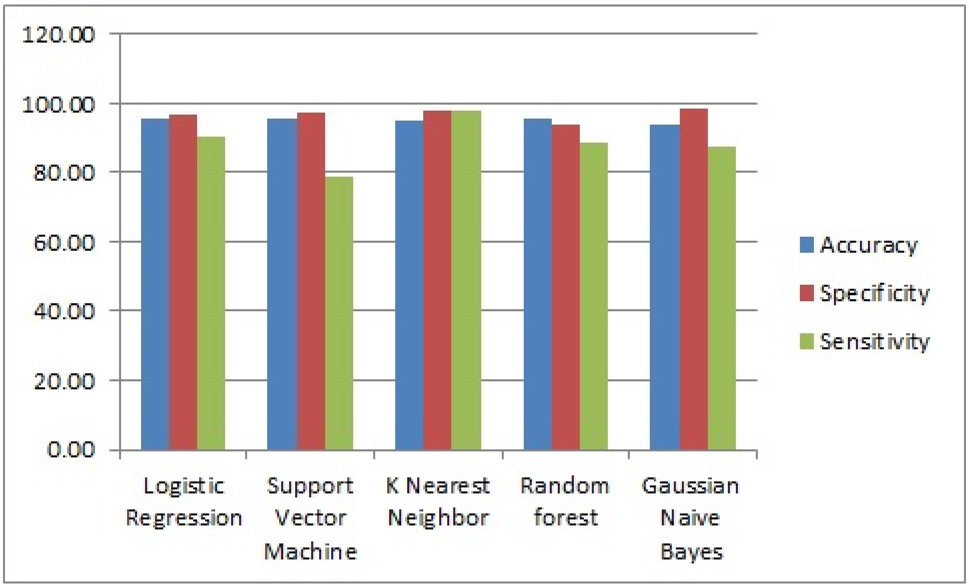


Figure 8. Chart presentation of performance evaluation result of classification models with selected fittest features.

4.1.3. Comparative Analysis of Performance Evaluation of the Classification Models

In this study, we used three performance evaluation metrics which include accuracy, sensitivity, and sensitivity. Hence, the experimental results of the classification models with full features and classification models with selected fittest features, respectively, are shown in Table 4.

The results of the comparative analysis of the type of performance evaluation of classification models of both classification models with full features and classification models with five selected fittest features show that the models with five selected fittest features outperformed their counterparts with full features in terms of accuracy, sensitivity, and sensitivity, as shown in Figures 9–11, respectively.

Table 4. Performance evaluation of classification models with full features against that of the models with selected fittest features.

Model.	Accuracy of Models with Full Features (%)	Accuracy of Models with Selected Fittest Features (%)	Specificity of Models with Full Features (%)	Specificity of Models with Selected Fittest Features (%)	Sensitivity of Models with Full Features (%)	Sensitivity of Models with Selected Fittest Features (%)
Logistic regression	66.30	95.61	74.24	96.90	36.80	90.11
Support vector machine	75.00	95.60	85.20	97.45	56.00	78.51
K-nearest neighbor	66.10	94.74	75.40	98.00	67.19	98.11
Random forest	76.20	95.61	93.40	93.56	71.16	88.45
Gaussian naïve bayes	70.00	93.86	78.70	98.34	87.12	87.18

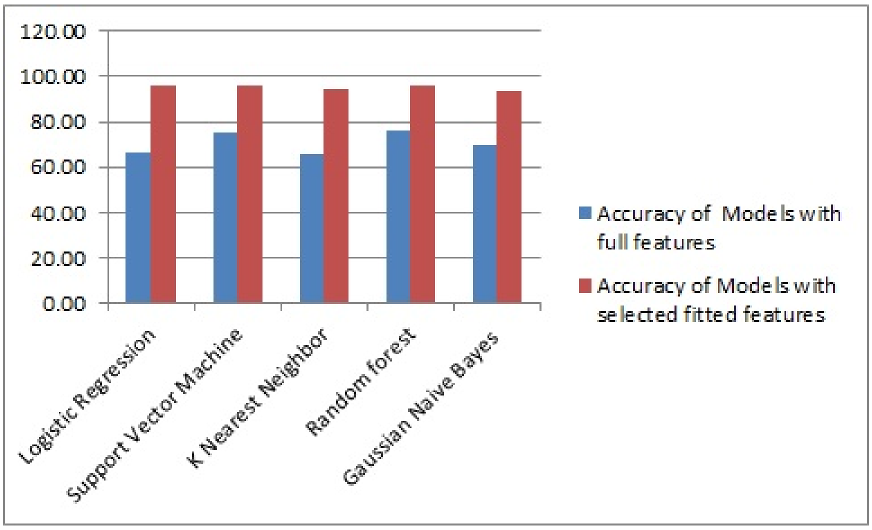


Figure 9. Chart presentation of the accuracy of the classification models with full features against that of the models with selected fittest features.

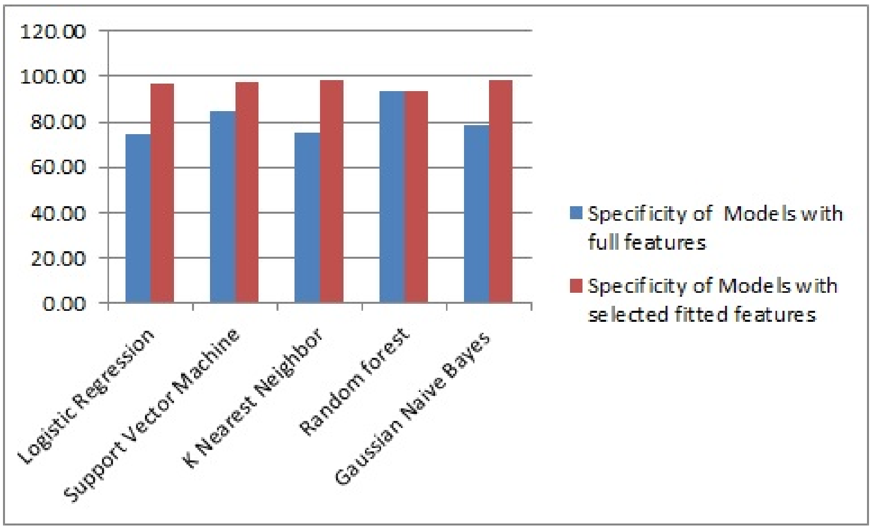


Figure 10. Chart presentation of the specificity of the classification models with full features against that of the models with selected fittest features.

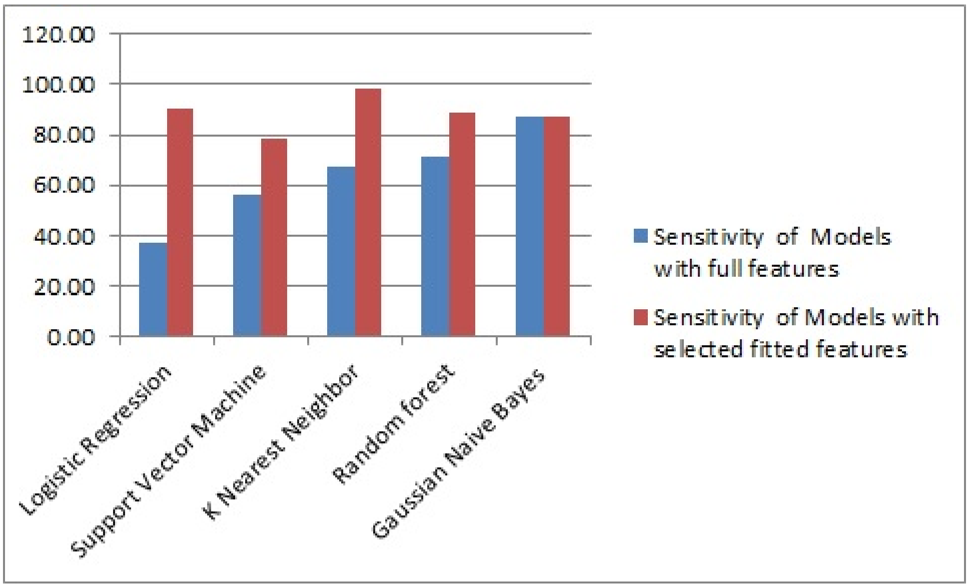


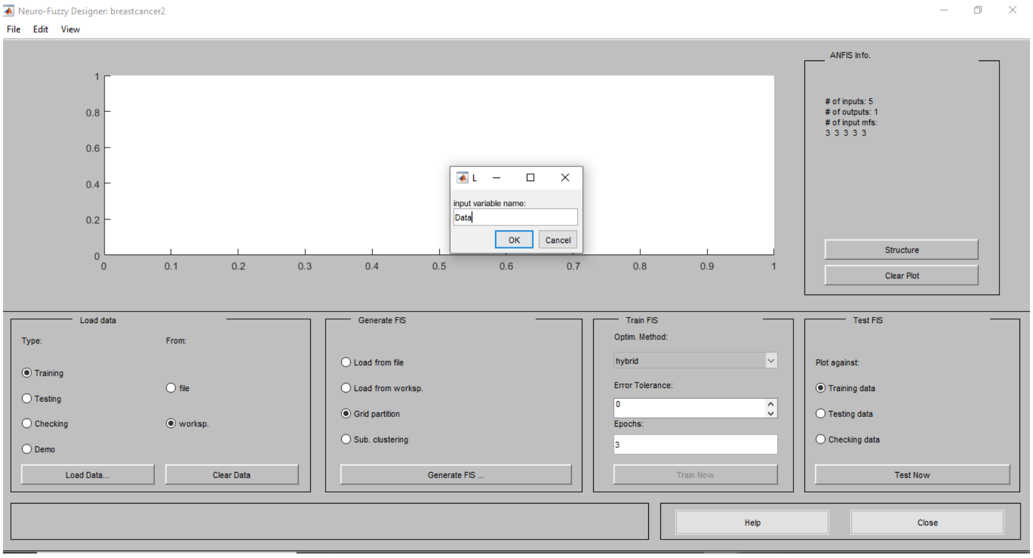
Figure 11. Chart presentation of the sensitivity of the classification models with full features against that of the models with selected fittest features.

4.2. Selection of the Size of Training Dataset

The dataset has been split into 80% training and 20% testing sets to avoid over flitting and under flitting of the system, respectively.

4.3. Rule Generation from Dataset

As the dataset has been split into 80% training and 20% testing sets to avoid over flitting and under flitting of the system, respectively. The dataset was used to generate the rules that would be used for the diagnosis of breast cancer. In this study, MALTAB neuro-fuzzy designer toolbox was used to generate the rules. The dataset was saved to MATLAB workspace and loaded it on the toolbox as shown in Figures 12 and 13 shows the trend of the trained dataset.



input variable name: Data

# of inputs: 5
# of outputs: 1
# of input mfs:
3 3 3 3 3

Type: ● Training ○ Testing ○ Checking ○ Demo

From: ○ file ● worksp.

○ Load from file ○ Load from worksp. ● Grid partition ○ Sub. clustering

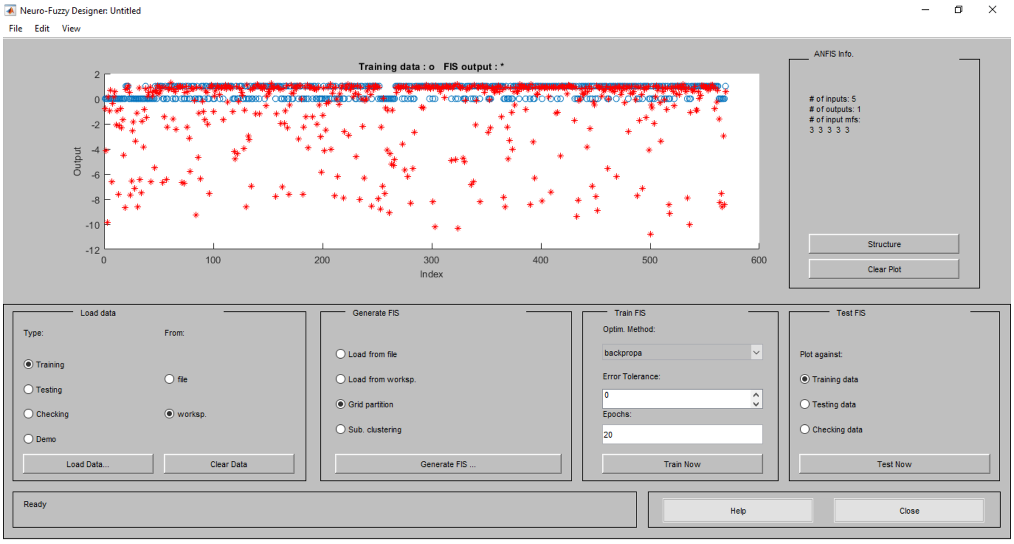
Optim. Method: hybrid

Error Tolerance: 0

Epochs: 3

Plot against: ● Training data ○ Testing data ○ Checking data

Figure 12. Loading of dataset.



Neuro-Fuzzy Designer: Untitled

Type: ● Training ○ Testing ○ Checking ○ Demo

From: ○ file ● worksp.

○ Load from file ○ Load from worksp. ● Grid partition ○ Sub. clustering

Optim. Method: backpropa

Error Tolerance: 0

Epochs: 20

Plot against: ● Training data ○ Testing data ○ Checking data

# of inputs: 5
# of outputs: 1
# of input mfs:
3 3 3 3 3

Ready

Figure 13. Trend of trained dataset.

4.4. Rule Optimization

In this step, the fuzzy rules being generated from the selected training dataset are redefined using the neural network technique. Therefore, five network neural layers are to be used for redefining the fuzzy rules and each layer defines the inference mechanism steps in the fuzzy neural network-based expert system.

4.4.1. Input Layer

The five fittest features are the inputs of the system, which would have a linear function to represent the current state of the system. Therefore, the inputs of the system include radius mean, texture mean, perimeter mean, area mean, and smoothness mean, and the membership of each input would be determined in the next layer.

4.4.2. Fuzzification Layer

The five fittest selected features of the dataset which are the input of the system, which include radius mean, texture mean, perimeter mean, area mean, and smoothness mean, are to be fuzzified in this layer. The inputs are the neurons of this layer which are the scalar values of the input features received from the input layer; therefore, their corresponding fuzzy values are to be determined by using the membership function. Fuzzification is the process of converting the scalar values of the inputs of the system to their corresponding fuzzy values and the process is carried out with fuzzy logic [70,71]. Fuzzy logic is a multivalued logic unlike traditional logic where the value is either 0 or 1, but fuzzy logic has an infinite number of values between 0 and 1 [70]. In fuzzy theory, A fuzzy set *B* in Y is defined as a set of ordered pairs =, ( ), ∈ a where μ_B *(x)* is called the membership function of set *B*.

$$M_B(y) : \rightarrow \{0,\ 1\},\ where\ \mu\beta(y) = 1\ is\ totally\ in\ \ B; \tag{11}$$

$$M_B(y) = 0\ if\ \ y\ is\ not\ is\ B; \tag{12}$$

$$0\ <\ \ \mu\beta(y) < 1\ if\ \ y\ is\ party\ in\ \ B. \tag{13}$$

The fuzzy logic allows for having possible options of choices. For any element *y* of the set Y, the member function of *μB(u)* is equal to the degree that *y* is an element of the set Y [31,51]. Therefore, the infinite value between 0 and 1 is considered and called the order of membership function. The membership function expresses the ambiguity and uncertainty by overlapping values to avoid the problem of sharp boundaries of values [19]. Therefore, there are many membership function distributions; however, for the input features of the system, the trapezoid membership function distribution is used, while for the output feature of the system, triangle membership function distribution is used. The trapezoidal membership function distribution is represented as *Trapezoidal* (x; a, b, c, d). The membership function values at $x = a$, $x = b$, $x = c$, and $x = d$ are set equal to 0.0, 1.0, 1.0, and 0.0, respectively. The trapezoidal membership function expressed in Equation (6) below:

$$trapezoid\ (\ x;a,\ b,\ c,\ d) = \max\left(\min\left(\frac{x-a}{b-a}\ ,\ 1, \frac{d-x}{d-c}\right), 0\right) \tag{14}$$

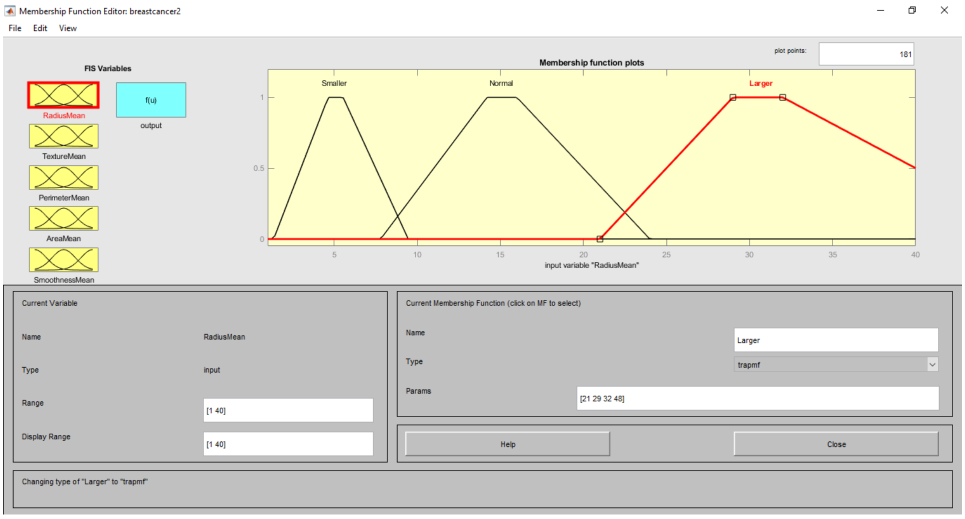
The triangular membership function is donated by *Triangle (x; a, b, c)*. The membership function values at $x = a$, $x = b$, and $x = c$ are set equal to 0.0, 1.0, and 0.0, respectively. The triangular membership function expressed in Equation (7) below:

$$triangle\ (\ x;a,\ b,\ c) = \max\left(\min\left(\frac{x-a}{b-a}\ ,\ 1, \frac{c-x}{c-b}\right), 0\right) \tag{15}$$

Table 5 above shows the input and output feature ranges with linguistic terms and membership function of the system. The trapezoid membership function distribution is used for inputs and output of the system was calculated and visualized in MATLAB Fuzzy Logic Toolbox. Figures 14–18 show the visualized membership functions of the linguistic variables of radius mean, texture mean, perimeter mean, area mean, and smoothness mean input features of the system, respectively, and Figure 19 shows the visualized membership functions of the linguistic variables output feature.

Table 5. Input and output features, ranges on linguistic terms, and membership function.

Input (Features)	Range	Linguistic Term	Membership Function
RadiusMean	<10 7–28 >21	Smaller Normal Larger	Trapezoid
TextureMean	<15 10–30 >25	Smaller Normal Larger	Trapezoid
PerimeterMean	<40 30–180 >160	Smaller Normal Larger	Trapezoid
AreaMean	<150 130–2200 >2300	Smaller Normal Larger	Trapezoid
SmoothnessMean	<0.04 0.06–0.14 >0.12	Smaller Normal Larger	Trapezoid
Diagnosis	<1 1–2 >2	Mild Healthy Severe	Triangle



Current Variable

Name RadiusMean

Type input

Range [1 40]

Display Range [1 40]

Current Membership Function (click on MF to select)

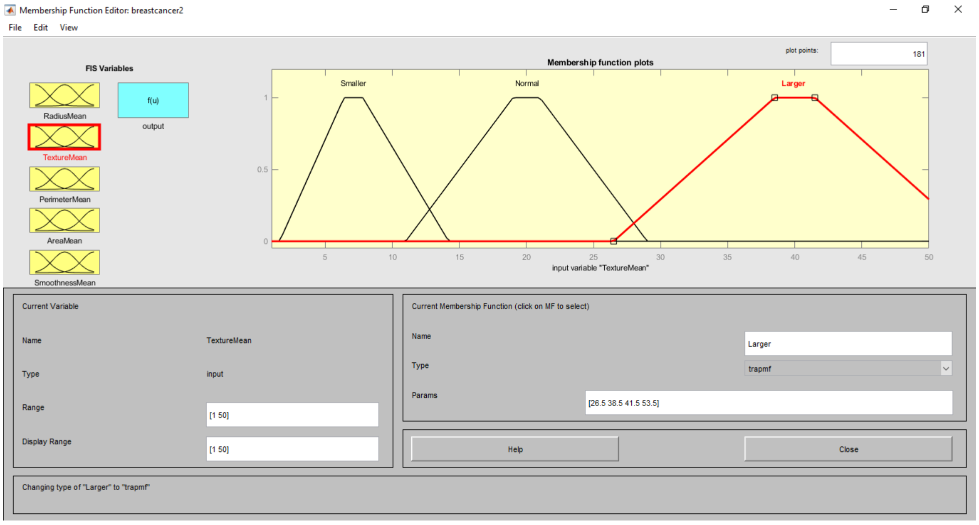
Name Larger

Type trapmf

Params [21 29 32 48]

Changing type of "Larger" to "trapmf"

Figure 14. Membership functions of the linguistic variables of the radius mean input.



Current Variable

Name TextureMean

Type input

Range [1 50]

Display Range [1 50]

Current Membership Function (click on MF to select)

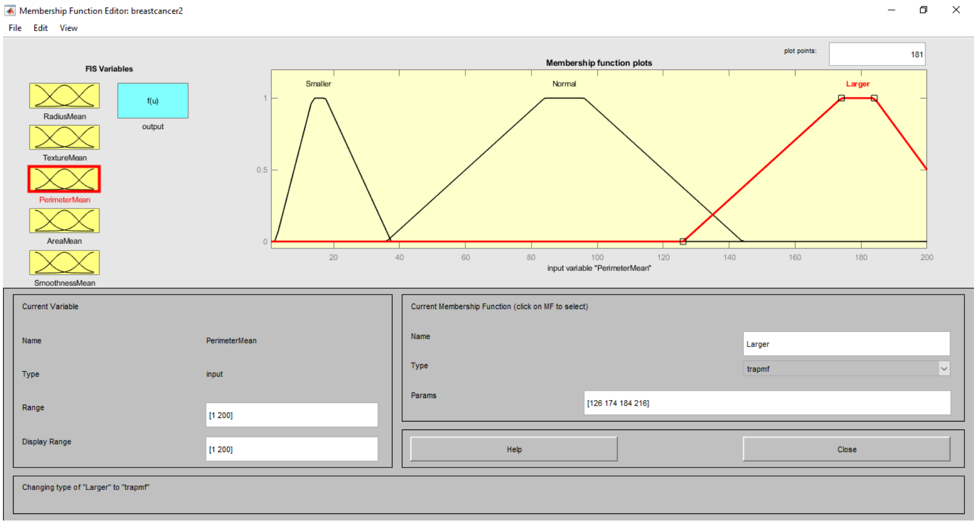
Name Larger

Type trapmf

Params [26.5 38.5 41.5 53.5]

Changing type of "Larger" to "trapmf"

Figure 15. Membership functions of the linguistic variables of texture mean input.



Current Variable

Name PerimeterMean

Type input

Range [1 200]

Display Range [1 200]

Current Membership Function (click on MF to select)

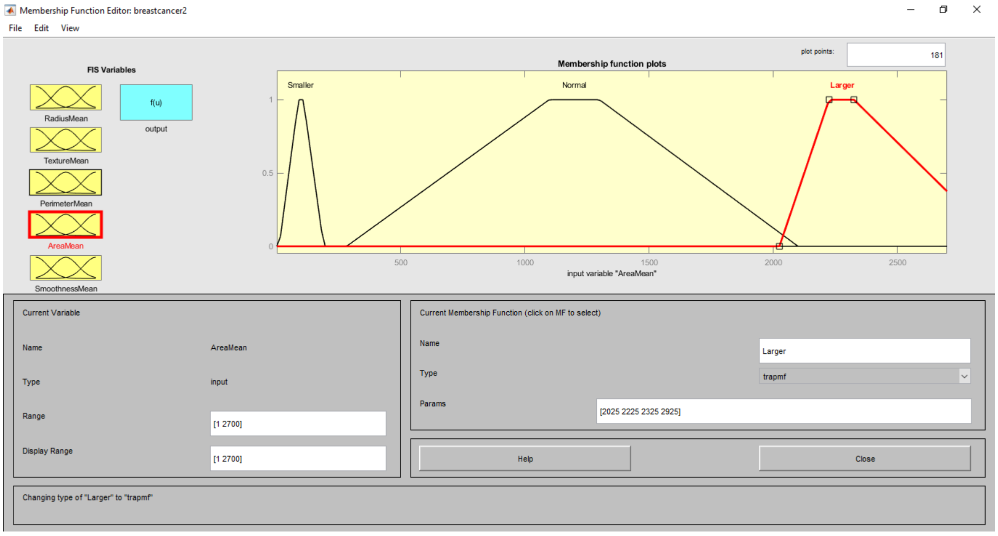
Name Larger

Type trapmf

Params [126 174 184 216]

Changing type of "Larger" to "trapmf"

Figure 16. Membership functions of the linguistic variables of perimeter mean input.



plot points: 181

Current Variable

Name AreaMean

Type input

Range [1 2700]

Display Range [1 2700]

Current Membership Function (click on MF to select)

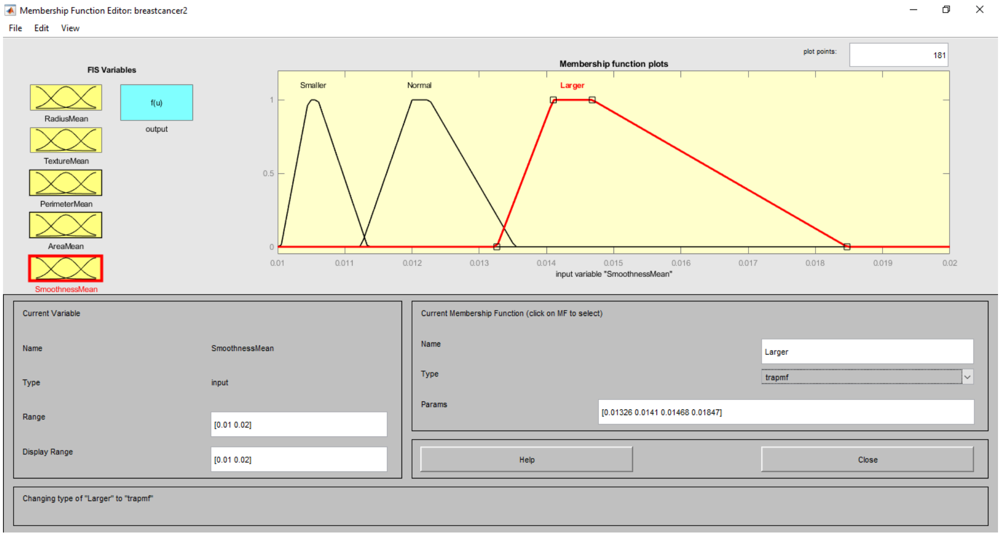
Name Larger

Type trapmf

Params [2025 2225 2325 2925]

Changing type of "Larger" to "trapmf"

Figure 17. Membership functions of the linguistic variables of area mean input.



plot points: 181

Current Variable

Name SmoothnessMean

Type input

Range [0.01 0.02]

Display Range [0.01 0.02]

Current Membership Function (click on MF to select)

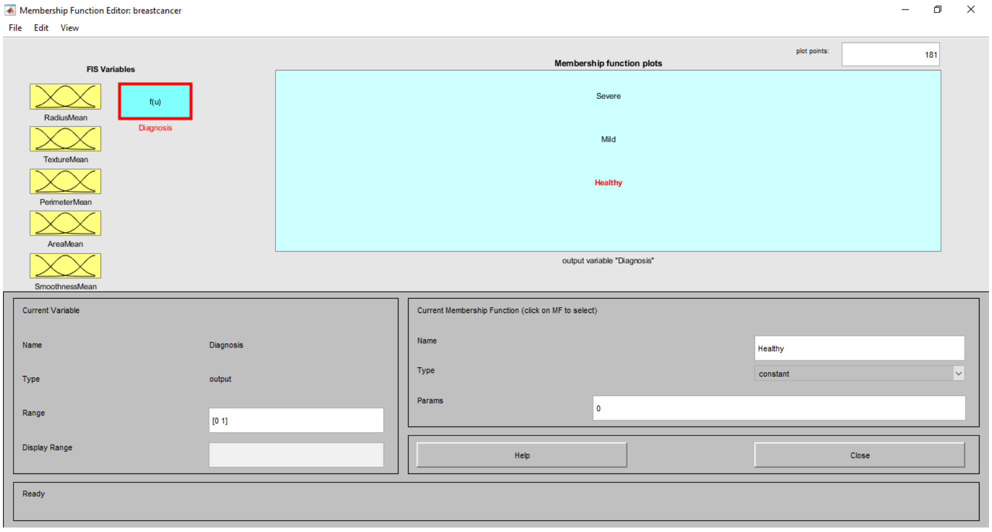
Name Larger

Type trapmf

Params [0.01326 0.0141 0.01468 0.01847]

Changing type of "Larger" to "trapmf"

Figure 18. Membership functions of the linguistic variables of smoothness mean input.



plot points: 181

Current Variable

Name Diagnosis

Type output

Range [0 1]

Display Range

Current Membership Function (click on MF to select)

Name Healthy

Type constant

Params 0

Ready

Figure 19. Membership functions of the linguistic variables output feature.

4.4.3. Fuzzy Rules Layers

The fuzzy rules are generated in this layer called the knowledge base of predicator, because it contains the possible combination of all fuzzified values of system inputs with performed logical AND operations. In fuzzy logic, the universal set Y:Y→[0,1] is called the universe of discourse, or simply the universe. The implication Y→[0,1] is the abbreviation for the IF-THEN rule: —IF y is in Y, THEN its MF μY(y) is in [0,1].‖, where μY(x) is the membership function of y. The universe Y may contain either discrete or continuous values.

Therefore, the neurons in this layer represent the fuzzy rules of the system and the numbers of the neurons are equal to the number of fuzzy rules. The main function of the neurons in this layer is to perform logical AND operations by the product of their respective input features of the system and pass the output to the next layer. Below are the same with the sample of fuzzy rules:

i. (RadiusMean == Larger) & (TextureMean == Normal)) & (PerimeterMean == Normal)) & (AreaMean == Normal) & (SmoothnessMean == Small) => (Diagnosis == Mild)

ii. (RadiusMean == Normal) & (TextureMean == Normal)) & (PerimeterMean == Normal)) & (AreaMean == Normal)) & (SmoothnessMean == Normal) => (Diagnosis == Healthy)

iii. (RadiusMean == Larger) & (TextureMean == Larger)) & (PerimeterMean == Larger)) & (AreaMean == Larger)) & (SmoothnessMean == Larger) = > (Diagnosis == Severe)

iv. (RadiusMean ==Small) & (TextureMean == Normal)) & (PerimeterMean == Normal)) & (AreaMean == Normal) & (SmoothnessMean == Normal) => (Diagnosis == Healthy)

v. (RadiusMean == Lager) & (TextureMean == Small)) & (PerimeterMean == Normal)) & (AreaMean == Small)) & (SmoothnessMean == Larger) => (Diagnosis == MIld)

vi. (RadiusMean == Normal) & (TextureMean == Small)) & (PerimeterMean == Normal)) & (AreaMean == Small) & (SmoothnessMean == Normal) => (Diagnosis == Healthy)

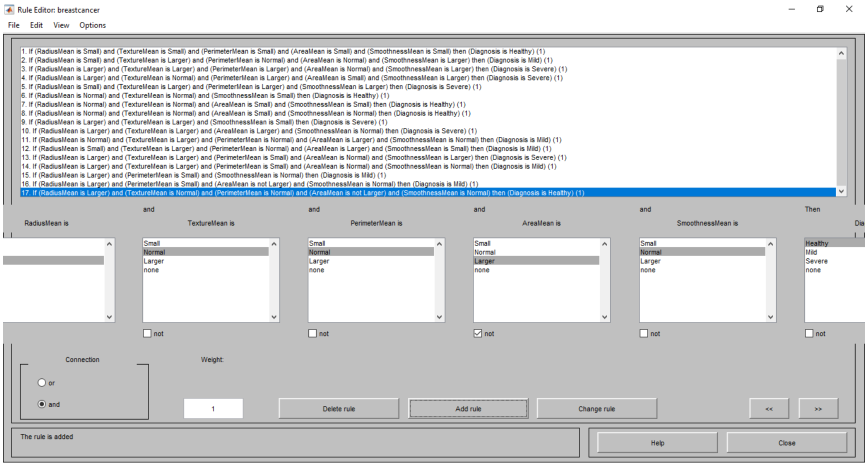
vii. (RadiusMean == Small) & (TextureMean == Larger)) & (PerimeterMean == Normal)) & (AreaMean == Small)) & (SmoothnessMean == Larger) => (Diagnosis == Mild)

viii. (RadiusMean == Normal) & (TextureMean == Small)) & (PerimeterMean == Normal)) & (AreaMean == Normal) & (SmoothnessMean == Normal) => (Diagnosis == Healthy)

ix. (RadiusMean == Larger) & (TextureMean == Larger)) & (PerimeterMean == Larger)) & (AreaMean == Larger)) & (SmoothnessMean == Larger) => (Diagnosis == Severe)

4.4.4. Inference Layer

In this layer, the neurons identify the firing rules corresponding to the set of input features of the system. Therefore, the neurons represent the output membership sets and each neuron combines all its input features AND operation. Therefore, the output of each neuron shows the stretcher of its input rules. Figure 20 shows how each neuron combines all its input features AND operation.


Rule Editor: breastcancer

1. If (RadiusMean is Small) and (TextureMean is Small) and (PerimeterMean is Small) and (AreaMean is Small) and (SmoothnessMean is Small) then (Diagnosis is Healthy) (1)
2. If (RadiusMean is Small) and (TextureMean is Larger) and (PerimeterMean is Normal) and (AreaMean is Normal) and (SmoothnessMean is Larger) then (Diagnosis is Mild) (1)
3. If (RadiusMean is Larger) and (TextureMean is Larger) and (PerimeterMean is Larger) and (AreaMean is Normal) and (SmoothnessMean is Larger) then (Diagnosis is Severe) (1)
4. If (RadiusMean is Larger) and (TextureMean is Normal) and (PerimeterMean is Larger) and (AreaMean is Small) and (SmoothnessMean is Larger) then (Diagnosis is Severe) (1)
5. If (RadiusMean is Small) and (TextureMean is Larger) and (PerimeterMean is Larger) and (SmoothnessMean is Larger) then (Diagnosis is Severe) (1)
6. If (RadiusMean is Normal) and (TextureMean is Normal) and (SmoothnessMean is Small) then (Diagnosis is Healthy) (1)
7. If (RadiusMean is Normal) and (TextureMean is Normal) and (AreaMean is Small) and (SmoothnessMean is Small) then (Diagnosis is Healthy) (1)
8. If (RadiusMean is Normal) and (TextureMean is Normal) and (AreaMean is Small) and (SmoothnessMean is Normal) then (Diagnosis is Healthy) (1)
9. If (RadiusMean is Larger) and (TextureMean is Larger) and (SmoothnessMean is Small) then (Diagnosis is Severe) (1)
10. If (RadiusMean is Larger) and (TextureMean is Larger) and (AreaMean is Larger) and (SmoothnessMean is Normal) then (Diagnosis is Severe) (1)
11. If (RadiusMean is Normal) and (TextureMean is Larger) and (PerimeterMean is Normal) and (AreaMean is Larger) and (SmoothnessMean is Normal) then (Diagnosis is Mild) (1)
12. If (RadiusMean is Small) and (TextureMean is Larger) and (PerimeterMean is Normal) and (AreaMean is Larger) and (SmoothnessMean is Small) then (Diagnosis is Mild) (1)
13. If (RadiusMean is Larger) and (TextureMean is Larger) and (PerimeterMean is Small) and (AreaMean is Normal) and (SmoothnessMean is Larger) then (Diagnosis is Severe) (1)
14. If (RadiusMean is Larger) and (TextureMean is Larger) and (PerimeterMean is Small) and (AreaMean is Normal) and (SmoothnessMean is Normal) then (Diagnosis is Mild) (1)
15. If (RadiusMean is Larger) and (PerimeterMean is Small) and (SmoothnessMean is Normal) then (Diagnosis is Mild) (1)
16. If (RadiusMean is Larger) and (PerimeterMean is Small) and (AreaMean is not Larger) and (SmoothnessMean is Normal) then (Diagnosis is Mild) (1)
17. If (RadiusMean is Larger) and (TextureMean is Normal) and (PerimeterMean is Normal) and (AreaMean is not Larger) and (SmoothnessMean is Normal) then (Diagnosis is Healthy) (1)

RadiusMean is: (selected)
TextureMean is: Normal
PerimeterMean is: Normal
AreaMean is: Larger, ☑ not
SmoothnessMean is: Normal
Then Diagnosis is: Healthy

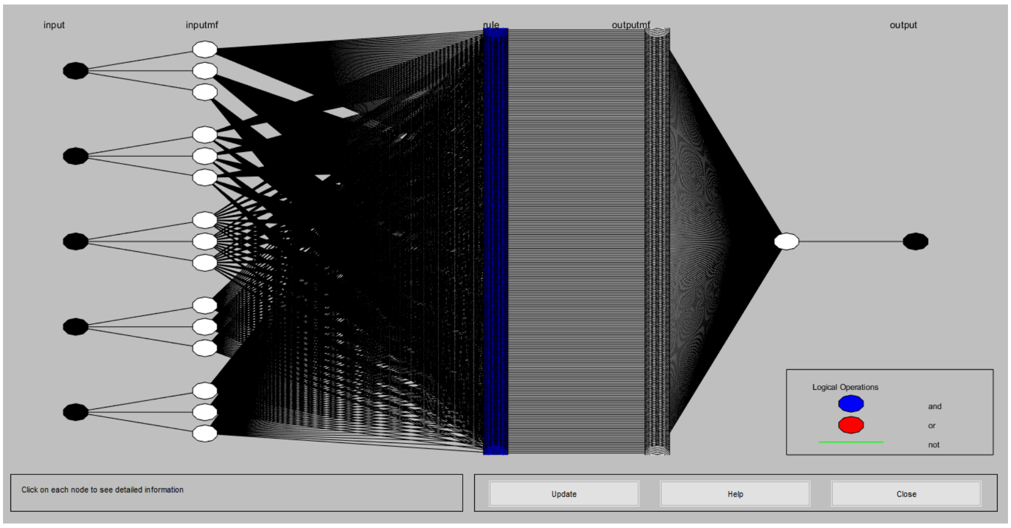
Connection: ○ or ● and
Weight: 1

The rule is added

Figure 20. Neuron combines all its input features AND operation.

4.4.5. Defuzzification Layer

As shown in Figure 21, this layer has only one neuron that represents the final out of the system, as it takes input from the output fuzzy rules and combines them into single fuzzy rules. Therefore, all the fuzzified outputs that have been shown in the figure are defuzzified into corresponding crisp values. Back propagation learning is used to rule optimization to adjust the parameters of the membership functions optimally. Therefore, the training data of input and output are compared with the system output and the errors are propagated backwards through the network from output layer to input layer, with the aim of modifying the membership functions of the neurons in layer 2.

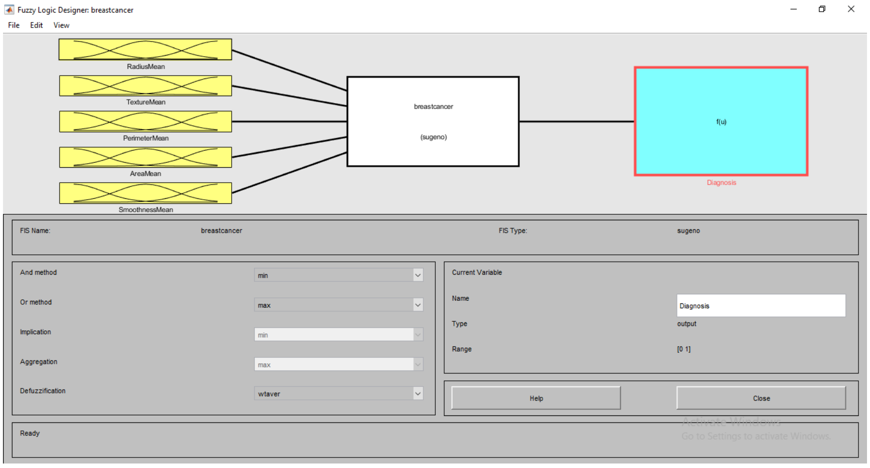


Click on each node to see detailed information

Figure 21. Fuzzy Neural Network Expert System for Early Diagnosis of Breast Cancer in Saudi Arabia.

A centroid is employed in this work for defuzzification, called the center of area or center of gravity, where *z* is the output feature, and (*z*) is the membership function of the aggregated fuzzy set A referring to *z*. The centroid method defuzzifies the system's diagnosis result's undefined values, which is the output of the system to crisp values. Figure 22 shows that the defuzzification layer of the system.

$$Z_{COA} = \frac{\int_Z^{\cdot} \mu A\,(z) \cdot zdz}{\int_z^{\cdot} \mu A\,(z)} \tag{16}$$



FIS Name: breastcancer FIS Type: sugeno

And method: min
Or method: max
Implication: min
Aggregation: max
Defuzzification: wtaver

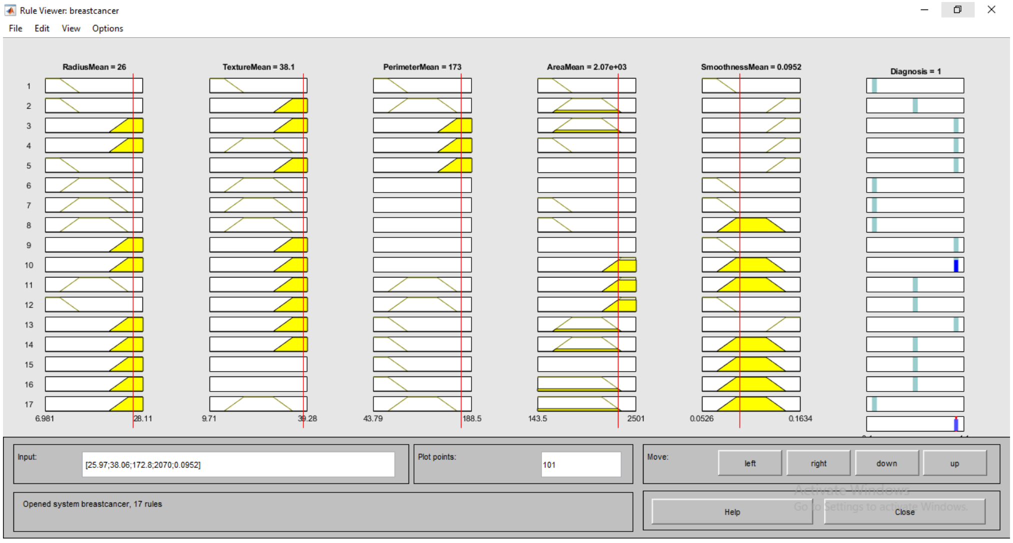
Current Variable
Name: Diagnosis
Type: output
Range: [0 1]

Ready

Figure 22. Defuzzification layer.

5. Performance Evaluation of Fuzzy Neural Network Expert System with an Improved Gini Index Random Forest-Based Feature Importance Measure Algorithm

The fuzzy neural network expert system with an improved gini index random forest-based feature importance measure algorithm for early diagnosis of breast cancer in Saudi Arabia has been developed and carefully evaluated to determine its performance against other systems that have been developed with other AI approaches. The system has been used to predict breast cancer on, randomly selected, 300 instances from the wisconsin diagnostic breast cancer dataset, considering only the five fittest features which have the highest random forest gini based importance measure scores. The features include radius mean, texture mean, perimeter mean, area mean, and smoothness mean. Figures 23–25 show how the system predicts and diagnoses breast cancer on the healthy, mild, and severe linguistic variables basis.

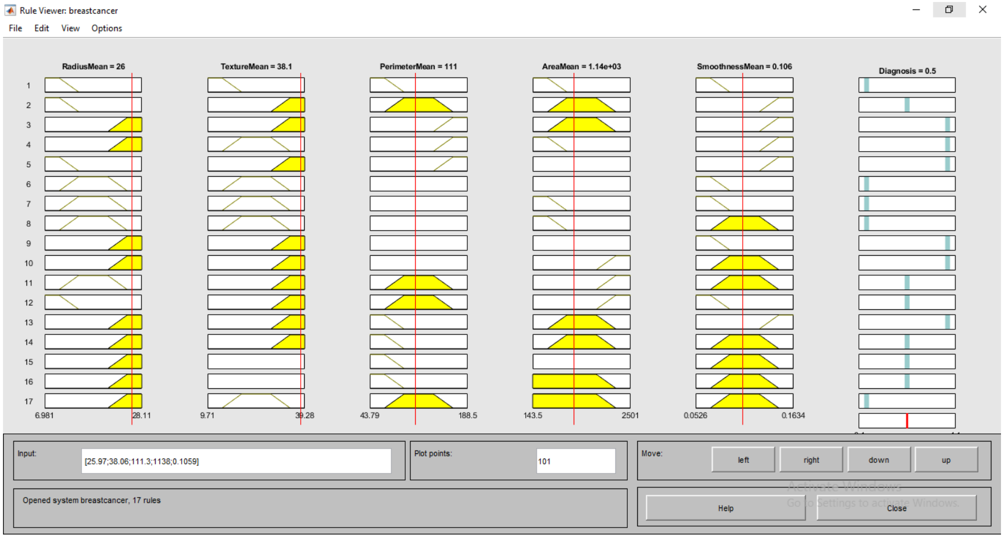


Input: [25.97;38.06;172.8;2070;0.0952]

Plot points: 101

Opened system breastcancer, 17 rules

Figure 23. Diagnosis for severe cases.

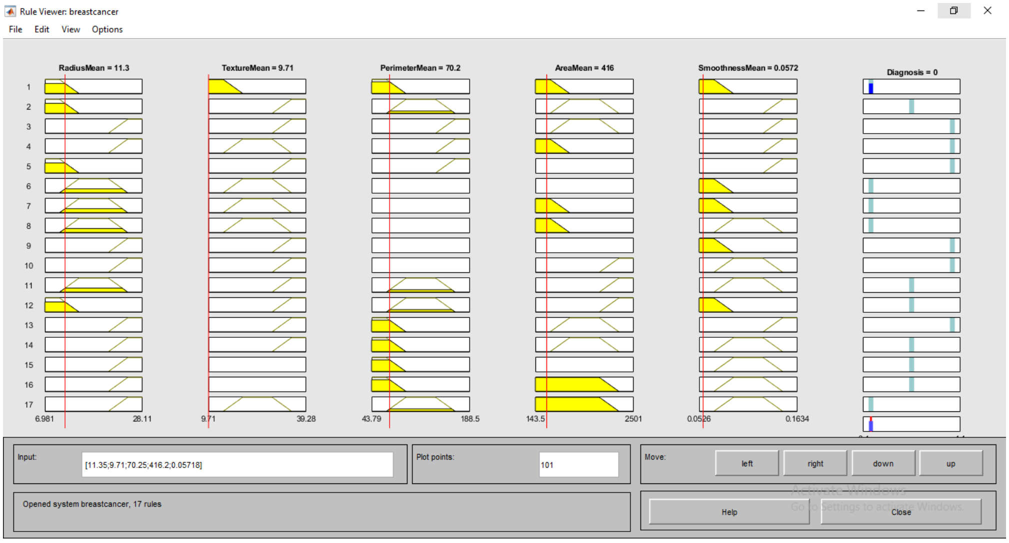


Input: [25.97;38.06;111.3;1138;0.1059]

Plot points: 101

Opened system breastcancer, 17 rules

Figure 24. Diagnosis for mild cases.



Input: [11.35;9.71;70.25;416.2;0.05718]

Plot points: 101

Move:

Opened system breastcancer, 17 rules

Figure 25. Diagnosis for healthy cases.

Table 6 shows the results of system diagnosis of breast cancer based on the healthy, mild, and severe cases, which were 300 randomly selected instances from the wisconsin diagnostic breast cancer dataset.

Table 6. Results of system for diagnosis of breast cancer.

Class	No. of Instances	No. of Corrected Diagnosis	No. of Wrong Diagnosis
Healthy	131	130	1
Mild	80	80	0
Severe	89	88	1
Grand Total	300	298	2

The system was used to diagnose the breast cancer based on five fittest features selected with an improved gini index random forest-based feature importance measure algorithm and the following performance evaluation techniques were used to determine the efficiency of the system:

iv. Accuracy: It is used to determine the efficiency of the system in terms of the percentage of suspected breast cancer patients that were able to predict correctly against the total number of available instances and it is defined using the formula:

$$\frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

v. Sensitivity: It is used to determine the efficiency of the system in terms of the percentage of mild and severe cases of breast cancer patients that is able to predict correctly against the total number of positive cases of the instances and it is defined using the formula:

$$\frac{TP}{TP + FP} \tag{18}$$

vi. Specificity: It is used to determine the efficiency of the system in terms of the percentage of healthy cases of breast cancer patients that is able to predict correctly against the total number of negative cases of the instances and it is defined using the formula:

$$\frac{TP}{TP + FN} \tag{19}$$

Note: TP stands for True Positive, FP stands for False Positive, TN stands for True Negative, FN stands for False Negative.

Table 7 shows the results of the performance evaluation result of the expert system and Figure 26 shows the chart presentation of the results. The system achieved 99.33% of accuracy for the ability to correctly diagnose healthy, mild, and severe cases of breast cancer patients, while for the ability to diagnose mild and severe cases of breast cancer patients correctly, the system achieved 99.41% sensitivity. In addition, likewise for the ability to correctly diagnose healthy cases of breast cancer patients, the system achieved 99.24% specificity.

Table 7. Performance evaluation result.

Accuracy (%)	Specificity (%)	Sensitivity (%)
99.33	99.24	99.41

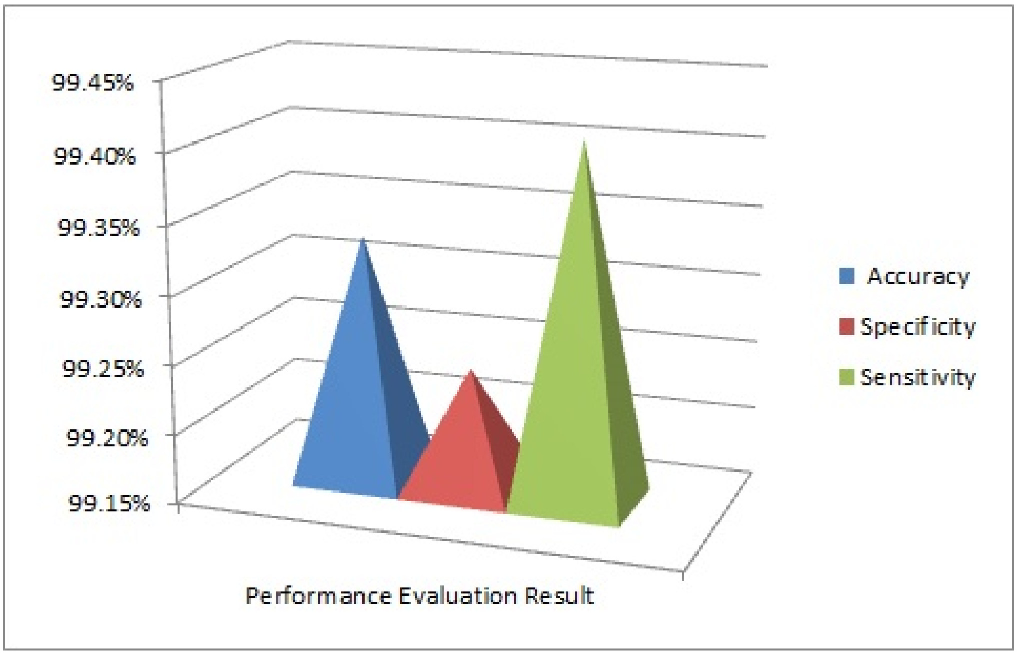


Figure 26. Chart presentation of performance evaluation result.

The result of the performance evaluation obtained from the fuzzy neural network expert system with an improved gini index random forest-based feature importance measure algorithm in this work has been compared with previous works that used fuzzy neural network and other applied artificial intelligence techniques on the same dataset for the diagnosis of breast cancer using the same dataset instances. Table 8 shows the comparison between the proposed method and existing works using the WBCD dataset.

Based on the comparison between the proposed method and existing works using the WBCD dataset, the proposed method of this work stands to be the best with the achievement of 99.33%, 99.41%, and 99.24% for accuracy, sensitivity, and specificity, respectively.

Table 8. Comparison between the proposed method and existing works using WBCD dataset.

Reference	Technique	Accuracy (%)	Sensitivity (%)	Specificity (%)
Proposed Method	Fuzzy neural network expert system with an improved gini index random forest-based feature importance measure algorithm	99.33	99.41	99.24
Ref. [36]	Automatic fuzzy database definition	94.53%	NA	NA
Ref. [3]	A semantic rule-based approach	NA	89.00	81.00
Ref. [51]	fuzzy method for pre-diagnosis of breast cancer from the fine needle aspirate analysis	NA	85.43	98.59
Ref. [39]	Breast cancer diagnosis using GA feature selection and Rotation Forest	96.00	94.67	97.69
Ref. [12]	Normalized neural networks for breast cancer classification	99.27	NA	NA
Ref. [38]	artificial metaplasticity neural network	99.26	NA	NA
Ref. [72]	An expert system for detection of breast cancer based on association rules and neural network (for four inputs)	95.60	NA	NA
Ref. [72]	An expert system for detection of breast cancer based on association rules and neural network (for eight inputs)	97.40	NA	NA
Ref. [36]	Particle swarm optimization feature selection for breast cancer recurrence prediction (naïve bayes)	81.30	63.20	86.90
Ref. [36]	Particle swarm optimization feature selection for breast cancer recurrence prediction (Reftree)	80.00	36.80	93.89
Ref. [36]	particle swarm optimization feature selection for breast cancer recurrence prediction (K-nearest neighbors (IBK))	75.00	42.10	85.29

Therefore, this work addressed all the research questions that we have assumed in the Introduction. In this work, we have developed an expert system that addressed the uncertainty often associated with diagnosis of breast cancer with a fuzzy logic technique. The heavier burden on the overlay of the network nodes of the fuzzy neural network system were addressed by leveraging the feature selection technique. We have found out the five fittest features of the diagnostic wisconsin breast cancer database, among 32 features of the dataset, by leveraging the improved gain index random forest-based feature importance measure algorithm. We have also justified why the five fittest features of the diagnostic wisconsin breast cancer database are more important than using an all feature dataset, where we have built evaluated machine learning models with logistic regression, support vector machine, k-nearest neighbor, random forest, and Gaussian naïve Bayes algorithms with both full features and selected five fittest features selected, respectively, evaluated them, and compared the results of the evaluation.

6. Statistical Testing

In order to be assured the results obtained from the performance evaluation of the fuzzy neural network expert system with an improved gini index random forest-based feature importance measure algorithm are truly accurate and not produced by chance, the statistical z-test was performed using R package. The null hypothesis assumes that there is no significant accuracy achieved by fuzzy neural network expert system with an improved gini index random forest-based feature importance measure algorithm for diagnosis of breast cancer, while the alternative hypothesis assumes that there is significant accuracy achieved by fuzzy neural network expert system with an improved gini index random forest-based feature importance measure algorithm for diagnosis of breast cancer. Using a record of the healthy, mild, and severe cases of breast cancer on the 300 randomly selected instances of the wisconsin diagnostic breast cancer dataset, the null hypothesis that there is no significant accuracy achieved by fuzzy neural network expert system with an improved gini index random forest-based feature importance measure algorithm for diagnosis of breast cancer is rejected, since the p-value obtained from the test, p-value < 0.0001, is less than the significance level, 0.05. That is, the proportion of the correctly diagnosed breast cancer cases by the system is greater than the proportion of incorrectly diagnosed cases. Thus, there is significant accuracy achieved by the fuzzy neural network expert system with an improved gini index random forest-based feature importance measure algorithm for diagnosis of breast cancer.

7. Conclusions

Breast cancer is one of the multifactorial genetic disorders or diseases that is likely associated with the effect of multiple genes in the combination of environmental factors and lifestyle. It is one of the common malignancies among females in Saudi Arabia and is ranked as the one most prevalent and the number two killer disease in Saudi Arabia. However, the conventional clinical diagnosis process of diseases is often associated with uncertainty and ambiguity due to complexity and fuzziness in the course of diagnosis of most of the deadly diseases, such as breast cancer, coronary artery diseases, diabetes, and waterborne diseases, among others. Hence, in this study, a fuzzy neural network expert system with an improved gini index random forest-based feature importance measure algorithm has been developed to address the challenges which include the heavier burden on the overlay of the network nodes of fuzzy neural network-based expert system due to many insignificant features that are used to predict or diagnose the disease and the uncertainty and ambiguity often associated with diagnostic decision making of breast cancer. The fuzzy neural network expert system with an improved gini index random forest-based feature importance measure algorithm achieved 99.33% accuracy for the ability to correctly diagnose healthy, mild, and severe cases of breast cancer patients, while for the ability to diagnose mild and severe cases of breast cancer patients correctly, the system achieved 99.41% sensitivity and, likewise for the ability to correctly diagnose healthy cases of breast cancer patients, the system achieved 99.24% specificity. Moreover, based on the comparison of the system with previous works that used fuzzy neural network or other applied artificial intelligence techniques on the same dataset for predicting breast cancer, the system stands to be the best with the achievement of 99.33%, 99.41%, and 99.24% for accuracy, sensitivity, and specificity, respectively. Z-test statistical testing was also conducted, and the testing result shows that there is significant accuracy achieved by the system for diagnosis of breast cancer. In the future, the method used in this work would be used in some other domains to further ascertain its accuracy.

Limitation and Future Direction

Fuzzy neural network expert system with an improved gini index random forest-based feature importance measure algorithm for early diagnosis of breast cancer in Saudi Arabia has been developed in this work and the system proved to be very efficient in addressing the challenges of the heavier burden on the overlay of the network nodes of fuzzy neural network-based expert system. This is due to many insignificant features that are used to predict or diagnose the disease and uncertainty and ambiguity often associated with diagnostic decision making of breast cancer. However, in future research, there is a need to apply the method used in this work in other domains, such as agriculture, mining, education, and learning. In the future research, fuzzy type 2 instead of type 1 would also be leveraged and a new feature selection would be proposed to make the system more robust and efficient.

Author Contributions: Conceptualization, E.A.A. and M.L.J. methodology, A.K.A. and M.L.J.; software, A.K.A. and M.L.J.; validation, E.A.A., A.K.A. and N.A.A.; formal analysis, O.A.A., M.L.J. and E.A.A.; investigation, N.A.A.; resources, O.A.A. and N.A.A.; data curation, A.K.A. and O.A.A.; writing—original draft preparation, M.L.J., E.A.A. and A.K.A.; writing—review and editing, O.A.A. and N.A.A.; visualization, M.L.J. and E.A.A.; supervision, E.A.A. and M.L.J.; project administration, N.A.A. and E.A.A.; funding acquisition E.A.A. All authors have read and agreed to the published version of the manuscript.

Funding: The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number (0021-1442-S).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The study used the data that are publicly available online (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)), accessed on 4 December 2021) for research purposes. No participants' personal information (e.g., name or address) was included in this study.

Acknowledgments: The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia and University of Tabuk, Tabuk71491, Saudi Arabia as well.

Conflicts of Interest: The authors declare no conflict of interest of any type to anyone.

References

1. Song, H. Detectability of Breast Tumors in Excised Breast Tissues of Total Mastectomy by IR-UWB-Radar-Based Breast Cancer Detector. *IEEE Trans. Biomed. Eng.* **2019**, *66*, 2296–2305. [CrossRef] [PubMed]
2. Alharthi, H. Healthcare predictive analytics: An overview with a focus on Saudi Arabia. *J. Infect. Public Health* **2018**, *11*, 749–756. [CrossRef] [PubMed]
3. Oyelade, O.N.; Obiniyi, A.A.; Junaidu, S.B.; Adewuyi, S.A. ST-ONCODIAG: A semantic rule-base approach to diagnosing breast cancer base on Wisconsin datasets. *Inform. Med. Unlocked* **2018**, *10*, 117–125. [CrossRef]
4. Reis, S. Automated Classification of Breast Cancer Stroma Maturity From Histological Images. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 2344–2352. [CrossRef] [PubMed]
5. Idris, N.F.; Ismail, M.A. Breast cancer disease classification using fuzzy-ID3 algorithm with FUZZYDBD method: Automatic fuzzy database definition. *PeerJ Comput. Sci.* **2021**, *7*, e427. [CrossRef] [PubMed]
6. Aibe, N. Results of a nationwide survey on Japanese clinical practice in breast-conserving radiotherapy for breast cancer. *J. Radiat. Res.* **2019**, *60*, 142–149. [CrossRef]
7. Sitaula, C.; Aryal, S. Fusion of whole and part features for the classification of histopathological image of breast tissue. *Health Inf. Sci. Syst.* **2020**, *8*, 38. [CrossRef]
8. Alanazi, M.; Parine, N.R.; Shaik, J.P.; al Naeem, A.; Aldhaian, S. Targeted sequencing of crucial cancer causing genes of breast cancer in Saudi patients. *Saudi J. Biol. Sci.* **2020**, *27*, 2651–2659. [CrossRef]
9. Assiri, A.S.; Nazir, S.; Velastin, S.A. Breast Tumor Classification Using an Ensemble Machine Learning Method. *J. Imaging* **2020**, *6*, 39. [CrossRef]
10. L-Abad, A.M.A. A Semantic Social Network Service for Educating Saudi Breast Cancer Patients. In Proceedings of the 2009 Ninth IEEE International Conference on Advanced Learning Technologies, Riga, Latvia, 15–17 July 2009; pp. 81–82. [CrossRef]
11. Fu, B.; Liu, P.; Lin, J.; Deng, L.; Hu, K.; Zheng, H. Predicting Invasive Disease-Free Survival for Early Stage Breast Cancer Patients Using Follow-Up Clinical Data. *IEEE Trans. Biomed. Eng.* **2019**, *66*, 2053–2064. [CrossRef]
12. Alickovic, E.; Subasi, A. Normalized Neural Networks for Breast Cancer Classification. In *CMBEBIH 2019*; Badnjevic, A., Škrbić, R., Gurbeta, P.L., Eds.; IFMBE Proceedings Springer: Cham, Switzerland, 2020. [CrossRef]
13. Rawan, S.; Manal, A. Real time data analysis and visualization for the breast cancer disease. *Period. Eng. Nat. Sci.* **2019**, *7*, 395–407.
14. Feng, X. Accurate Prediction of Neoadjuvant Chemotherapy Pathological Complete Remission (pCR) for the Four Sub-Types of Breast Cancer. *IEEE Access* **2019**, *7*, 134697–134706. [CrossRef]
15. Alshammari, F.D. Breast cancer genetic susceptibility: With focus in Saudi Arabia. *J. Oncol. Sci.* **2019**, *5*, 6–12. [CrossRef]
16. Almutlaq, B.; Almuazzi, R.F.; Almuhayfir, A.A. Breast cancer in Saudi Arabia and its possible risk factors. *J. Cancer Policy* **2017**, *12*, 83–89. [CrossRef]
17. Alsharif, F.H.; Mazanec, S.R. The use of complementary and alternative medicine among women with breast cancer in Saudi Arabia. *Appl. Nurs. Res.* **2019**, *48*, 75–80. [CrossRef]
18. Al-Gaithy, Z.K.; Yaghmoor, B.E.; Koumu, M.I. Trends of mastectomy and breast-conserving surgery and related factors in female breast cancer patients treated at King Abdulaziz University Hospital, Jeddah, Saudi Arabia, 2009–2017: A retrospective cohort study. *Ann. Med.* **2019**, *41*, 47–52. [CrossRef]

19. Muhammad, L.J.; Garba, E.J.; Oye, N.D.; Wajiga, G.M.; Garko, A.B. Fuzzy rule-driven data mining framework for knowledge acquisition for expert system. In *Translational Bioinformatics in Healthcare and Medicine*; Elsevier: Amsterdam, The Netherlands; Academic Press: New Delhi, India, 2021; pp. 201–214.
20. Muhammad, L.J.; Jibrin, M.B.; Yahaya, B.Z.; Jibrin, I.A.M.B.; Ahmad, A.; Amshi, J.M. An Improved C4.5 Algorithm using Principle of Equivalent of Infinitesimal and Arithmetic Mean Best Selection Attribute for Large Dataset. In Proceedings of the 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, 29–30 October 2020; pp. 006–010.
21. Pertiwi, D.A.; Daniawan, B.; Gunawan, Y. Analysis And Design of Decision Support System in Major Assignment at Buddhi High School Using AHP and SAW Methods. *Tech-E* **2019**, *3*, 13–21. [CrossRef]
22. Muhammad, L.J.; Ahmad, A.H.; Ibrahim, A.M.; Mansir, A.; Bature, B.; Jamila, M.A. Performance Evaluation of Classification Data Mining Algorithms On Coronary Artery Disease Dataset. In Proceedings of the IEEE 9th International Conference on Computer and Knowledge Engineering (ICCKE 2019), Ferdowsi University of, Mashhad, Mashhad, Iraq, 24–25 October 2019.
23. Muhammad, L.J.; Garba, A.; Abba, G. Security Challenges for Building Knowledge Based Economy in Nigeria. *Int. J. Secur. Its Appl.* **2015**, *9*, 13–21. [CrossRef]
24. Tchiera, F.; Alharbia, A. Fuzzy Relational Model and Genetic Algorithms for Early Detection and Diagnosis of Breast Cancer in Saudi Arabia. *Filomat* **2016**, *30*, 547–556. [CrossRef]
25. Ishaq, F.S.; Muhammad, L.J.; Yahaya, Y.Z. Fuzzy-Based Expert System for Diagnosis of Diabetes Mellitus. *Int. J. Adv. Sci. Technol.* **2020**, *136*, 39–50. [CrossRef]
26. Huang, G.B.; Zhu, Q.Y.; Siew, C.K. Extreme learning machine: Theory and applications. *Neurocomputing* **2006**, *70*, 489–501. [CrossRef]
27. Han, H.; Guo, X.; Yu, H. Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest. In Proceedings of the 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 26–28 August 2016; pp. 219–224. [CrossRef]
28. Haruna, A.A.; Jung, L.T.; Arputharaj, V.; Muhammad, L.J. Incentive-Scheduling Algorithms to Provide Green Computational Data Center. *SN Comput. Sci.* **2021**, *2*, 252. [CrossRef]
29. Muhammad, L.J.; Algehyne, E.A. Fuzzy based expert system for diagnosis of coronary artery disease in Nigeria. *Health Technol.* **2021**, *11*, 319–329. [CrossRef]
30. Das, S.; Ghosh, P.K.; Kar, S. Hypertension diagnosis: A comparative study using fuzzy expert system and neuro fuzzy system. In Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Hyderabad, India, 7–10 July 2013. [CrossRef]
31. Kour, H.; Manhas, R.; Sharma, V. Usage and implementation of neuro-fuzzy systems for classification and prediction in the diagnosis of different types of medical disorders: A decade review. *Artif. Intell. Rev.* **2020**, *53*, 4651–4706. [CrossRef]
32. González-Pérez, B.; Núñez, C.; Sánchez, J.L.; Valverde, G.; Velasco, J.M. Expert System to Model and Forecast Time Series of Epidemiological Counts with Applications to COVID-19. *Mathematics* **2021**, *9*, 1485. [CrossRef]
33. Park, K.; Chen, W.; Chekmareva, M.A.; Foran, D.J.; Desai, J.P. Electromechanical Coupling Factor of Breast Tissue as a Biomarker for Breast Cancer. *IEEE Trans. Biomed. Eng.* **2018**, *65*, 96–103. [CrossRef] [PubMed]
34. Hussain, S.; Muhammad, L.J.; Ishaq, F.S.; Yakubu, A.; Mohammed, I.A. Performance Evaluation of Various Data Mining Algorithms on Road Traffic Accident Dataset. In *Information and Communication Technology for Intelligent Systems. Smart Innovation, Systems and Technologies*; Satapathy, S., Joshi, A., Eds.; Springer Nature: Singapore, 2019.
35. Ubeyli, E.D. Adaptive neuro-fuzzy inference systems for automatic detection of breast cancer. *J. Med. Syst.* **2009**, *33*, 353–358. [CrossRef]
36. Sakri, S.B.; Rashid, N.B.A.; Zain, Z.M. Particle Swarm Optimization Feature Selection for Breast Cancer Recurrence Prediction. *IEEE Access* **2018**, *6*, 29637–29647. [CrossRef]
37. Ahmad, A. Evaluation of Modified Categorical Data Fuzzy Clustering Algorithm on the Wisconsin Breast Cancer Dataset. *Scientifica* **2016**, *2016*, 4273813. [CrossRef]
38. Marcano-Cedeño, A.; Marin-de-la-Barcena, A.; Jimenez-Trillo, J.; Piñuela, J.A.; Andina, D. Artificial metaplasticity neural network applied to credit scoring. *Int. J. Neural. Syst.* **2011**, *21*, 311–317. [CrossRef] [PubMed]
39. Aličković, E.; Subasi, A. Breast cancer diagnosis using GA feature selection and Rotation Forest. *Neural Comput. Appl.* **2017**, *28*, 753–763. [CrossRef]
40. Nembrini, S.; König, I.R.; Wright, M.N. The revival of the Gini importance? *Bioinformatics* **2018**, *34*, 3711–3718. [CrossRef] [PubMed]
41. Haruna, A.A.; Muhammad, L.J.; Yahaya, B.Z. An Improved C4.5 Data Mining Driven Algorithm for the Diagnosis of Coronary Artery Disease. In Proceedings of the International Conference on Digitization (ICD), Sharjah, United Arab Emirates, 18–19 November 2019; pp. 48–52.
42. Menze, B.H.; Kelm, B.M.; Masuch, R. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinform.* **2009**, *10*, 213. [CrossRef] [PubMed]
43. Muhammad, L.J. Deep Learning Models for Predicting COVID-19 Using Chest X-Ray Images. In *Trends and Advancements of Image Processing and Its Applications*; Johri, P., Diván, M.J., Khanam, R., Marciszack, M., Will, A., Eds.; EAI/Springer Innovations in Communication and Computing; Springer: Cham, Switzerland, 2022. [CrossRef]

44. Muhammad, L.J.; Algehyne, E.A.; Usman, S.S. Predictive Supervised Machine Learning Models for Diabetes Mellitus. *SN Comput. Sci.* **2020**, *1*, 240. [CrossRef] [PubMed]
45. Kaur, G.; Kaushik, A.; Sharma, S. Cooking Is Creating Emotion: A Study on Hinglish Sentiments of Youtube Cookery Channels Using Semi-Supervised Approach. *Big Data Cogn. Comput.* **2019**, *3*, 37. [CrossRef]
46. Shah, S.R.; Kaushik, A.; Sharma, S.; Shah, J. Opinion-Mining on Marglish and Devanagari Comments of YouTube Cookery Channels Using Parametric and Non-Parametric Learning Models. *Big Data Cogn. Comput.* **2020**, *4*, 3. [CrossRef]
47. Sarumi, O.A.; Aouedi, O. Potential of Deep Learning Algorithms in Mitigating the Spread of COVID-19. In *Understanding COVID-19: The Role of Computational Intelligence*; Nayak, J., Naik, B., Abraham, A., Eds.; Springer: Manhattan, NY, USA, 2021.
48. Govinda, K.; Singla, K.; Jain, K. Fuzzy based uncertainty modeling of Cancer Diagnosis System. In Proceedings of the 2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, Palladam, India, 7–8 December 2017; pp. 740–743. [CrossRef]
49. Meesad, P.; Yes, G.G. A hybrid intelligent system for medical diagnosis. In Proceedings of the International Joint Conference on Neural Networks, Proceedings (Cat. No.01CH37222), Washington, DC, USA, 15–19 July 2001; pp. 2558–2563. [CrossRef]
50. Salah, B.; Alshraideh, M.; Beidas, R.; Hayajneh, F. Skin cancer recognition by using a neuro-fuzzy system. *Cancer Inform.* **2011**, *10*, CIN-S5950. [CrossRef]
51. Sizilio, G.R.; Leite, C.R.; Guerreiro, A.M. Fuzzy method for prediagnosis of breast cancer from the Fine Needle Aspirate analysis. *BioMed. Eng. OnLine* **2012**, *11*, 83. [CrossRef]
52. Nadia, G.; Bilal, R.; Ebrahem, A.A. The dynamics of fractional order Hepatitis B virus model with asymptomatic carriers. *Alex. Eng. J.* **2021**, *60*, 3945–3955.
53. Aldrich, C. Process Variable Importance Analysis by Use of Random Forests in a Shapley Regression Framework. *Minerals* **2020**, *10*, 420. [CrossRef]
54. Ebrahem, A.A.; Din, R. On global dynamics of COVID-19 by using SQIR type model under non-linear saturated incidence rate. *Alex. Eng. J.* **2021**, *60*, 393–399.
55. Kanagarathinam, K.; Algehyne, E.A.; Sekar, K. Analysis of 'earlyR'epidemic model and time series model for prediction of COVID-19 registered cases. *Mater. Today Proc.* **2020**, *10*, 2214–7853.
56. Park, H.; Kwon, H. Improved Gini-Index Algorithm to Correct Feature-Selection Bias in Text Classification. *IEICE Trans. Inf. Syst.* **2011**, *94*, 855–865. [CrossRef]
57. Strobl, C.; Boulesteix, A.L.; Zeileis, A. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinform.* **2007**, *8*, 25. [CrossRef] [PubMed]
58. Rani, L.N.; Defit, S. Determination of Student Subjects in Higher Education Using Hybrid Data Mining Method with the K-Means Algorithm and FP Growth. *Int. J. Artif. Intell. Res.* **2021**, *5*, 91–101. [CrossRef]
59. Shang, W.; Dong, H.; Zhuo, H. A Novel Feature selection algorithm for text categorization. *Expert Syst. Appl.* **2007**, *33*, 1–5. [CrossRef]
60. Ebrahem, A.A.; Ibrahim, M. Fractal-Fractional Order Mathematical Vaccine Model of COVID-19 under non-singular kernel. *Chaos Solitons Fractals* **2021**, *148*, 111–150.
61. Cassidy, A.P.; Deviney, F.A. Calculating feature importance in data streams with concept drift using Online Random Forest. In Proceedings of the 2014 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 27–30 October 2014; pp. 23–28. [CrossRef]
62. Uddin, M.T.; Uddiny, M.A. A guided random forest based feature selection approach for activity recognition. In Proceedings of the 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), Dhaka, Bangladesh, 21–23 May 2015; pp. 1–6. [CrossRef]
63. Chen, R.A.; Dewi, C.; Huang, S.W. Selecting critical features for data classification based on machine learning methods. *J. Big Data* **2020**, *7*, 52. [CrossRef]
64. Sethi, K.; Jaiswal, V.; Ansari, M.D. Machine Learning Based Support System for Students to Select Stream (Subject). *Recent Adv. Comput. Sci. Commun.* **2020**, *13*, 336–344. [CrossRef]
65. Bhargava, N.; Sharma, G.; Bhargava, R.; Mathuria, M. Decision tree analysis on J48 algorithm for data mining. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **2013**, *3*, 1114–1119.
66. Alsayed, O.; Rahim, M.S.M.; Bidewi, I.A. Selection of the Right Undergraduate Major by Students Using Supervised Learning Techniques. *Appl. Sci.* **2021**, *11*, 10639. [CrossRef]
67. Hjerpe, A. *Computing Random Forests Variable Importance Measures (VIM) on Mixed Continuous and Categorical Data*; KTH Royal Institute of Technology School of Computer Science and Communication: Stockholm, Sweden, 2016.
68. Keles, A.; Yavuz, A.U. Expert system based on neuro-fuzzy rules for diagnosis breast cancer. *Expert Syst. Appl.* **2011**, *38*, 5719–5726. [CrossRef]
69. University of California Irvine Machine Learning Repository. Available online: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic) (accessed on 10 March 2021).
70. Dilip, K. *Soft Computing: Fundamentals and Applications*; NAROSA: New Delhi, India, 2013; pp. 103–121.

71. Nilashi, M.; Ibrahim, O.; Ahmadi, H.; Shahmoradi, L. A Knowledge-Based System for Breast Cancer Classification Using Fuzzy Logic Method. *Telemat. Inform.* **2017**, *34*, 133–144. [CrossRef]
72. Karabatak, M.M.; Ince, M.C. An expert system for detection of breast cancer based on association rules and neural network. *Expert Syst. Appl.* **2009**, *36*, 3465–3469. [CrossRef]