Received: 19 November 2022 | Revised: 23 January 2023 | Accepted: 8 February 2023

IET Image Processing

DOI: 10.1049/ipr2.12774

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

A fuzzy rule-based system with decision tree for breast cancer detection

Vedika Gupta[2] | Harshit Gaur[1] | Srishti Vashishtha[1] | Uttirna Das[1] | Vivek Kumar Singh[3] | D. Jude Hemanth[4]

[1]Department of Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi, India

[2]Jindal Global Business School, O.P. Jindal Global University, Sonipat, Haryana, India

[3]Department of Computer Science, Banaras Hindu University, Varanasi, India

[4]Department of Electronics and Communication Engineering, Karunya University, Coimbatore, India

Correspondence
Vivek Kumar Singh, Department of Computer Science, Banaras Hindu University, Varanasi 221005, India.
Email: vivek@bhu.ac.in

D. Jude Hemanth, Department of Electronics and Communication Engineering, Karunya University, Coimbatore, India.
Email: judehemanth@karunya.edu

Abstract

Breast cancer is possibly the deadliest illness in the world and the risks are gradually increasing. One out of eight women has the chance to be detected with breast cancer in their lifetime. The utmost cause for the higher fatality rates is the prolonged prognosis for the detection of breast cancer. The focus of this study is therefore to develop a better fuzzy expert system for the detection of breast cancer using decision tree analysis for deriving the rule base. For this classification problem, the input features of the dataset are converted into human-understandable terms-linguistic variables. The Mamdani Fuzzy Rule-Based system is deployed as the main inference engine and the centroid method for the defuzzification process to convert the final fuzzy score into class labels- benign (not cancerous) or malignant (cancerous). A decision tree algorithm is applied the creating a novel set of 27 fuzzy rules which are fed into FRBS. The investigation is performed on the publicly available Wisconsin Breast Cancer Dataset. The accuracy obtained by the proposed system is about 97%, recall is 99.58% and precision is about 93%. The experiments on this dataset yield higher performance as compared to the state-of-the-art dataset.

1 | INTRODUCTION

Breast Cancer is one of the most extensively seen diseases among women and one of the pervasive variables behind significant deaths among them all over the world. In 2020, there were 2.3 million women diagnosed with breast cancer and 685,000 deaths globally [1]. Around 9.6 million people have died because of prolonged-stage cancer in 2019. A survey conducted in 2019 provided an estimated 2,68,600 newly diagnosed cases of breast cancer among which 41,760 deaths of females were reported (Siegel et al., [2]). Mammography and histopathology are two major and traditional frameworks for executing screening tests for breast cancer. The confidence score of mammograms can deflect from the truly positive results and therefore it cannot be a trustworthy tool in particularly many cases to determine the disease in a big way. A false-negative mammogram can at times literally look normal even though there exists positive breast cancer. Similarly, false-positive mammograms really look abnormal even if there are no chances for positive breast cancer [3]. Subsequently, it is essential to devise machine learning and data mining procedures to make an adequate system for early phase accurate diagnosis of breast cancer malignancy and the need for proper cure methods.

Modern expert knowledge systems and machine learning techniques can be taken into consideration to develop new and faster detection systems. Fuzzy expert systems are applied to those real-world problems whose decision-making process is complex and require human-like thinking to either solve the problem with one of the attributes [4]. This is the reason they are also known as ELSE-IF models. The prediction value falls between 0 and 1 and classifies based on attributes collected from the database. On the other hand, machine learning methods are based on the training of the model with the help of incremental learning and dataset variation. Fuzzy rule-based models are successful in many classification problems- Sentiment Analysis [5], Speech Emotion Recognition [6, 7], Big Data classification [8], and heart disease [9, 10]. There are many machine learning models [11] available to detect cancer of several types like Naive

Bayes [12, 13], Support Vector Machine (SVM) [14, 15], Least square- SVM [16], K-nearest neighbours [17, 18], Decision Tree [19, 20], Convolution Neural networks [21, 22] etc. Data mining techniques are efficient in predicting breast cancer recurrence [23] with the assistance of integrating the classifying algorithms and feature selection algorithms.

1.1 | Motivation and main contributions

The primary motivation of this paper is to classify (whether Benign or Malignant) and detect breast cancer patients by applying a model consisting of the fuzzy rule-based system (FRBS) and decision tree. As a troubleshooting tool, fuzzy systems theory is a formal approach that seeks to address modelling, representation, reasoning, and erroneous information procedures.

The main contributions of this paper are:

(i) Development of a fuzzy model by deploying a Mamdani fuzzy rule-based inference engine for fuzzification of input features of breast cancer.
(ii) Application of decision tree algorithm for the creation of fuzzy rule base of FRBS.
(iii) Development of a novel set of 27 fuzzy rules, identified by a decision tree, for classifying whether the given sample is benign (not cancerous) or malignant (cancerous) and
(iv) Comparison of the proposed fuzzy rule-based approach for breast cancer detection with six state-of-the-art methods.

The rest of the paper is organized as follows: Section 2 describes the state-of-the-art breast cancer classification, while our proposed fuzzy rule-based system and implementation are presented in Section 3. Section 4 is about the experimental setup and results. The overall conclusions and future work are drawn in Section 5.

2 | LITERATURE REVIEW

Expert systems are technologically promising and advanced heuristic rule-based systems can be used for accurate prediction and diagnosing diseases such as diabetes [24]. The performance of the expert systems may be improved with the help of several optimization techniques. Data clustering methods like expectation maximization are effective to achieve higher accuracies with the dataset. Ghasemzadeh et al. [25] have addressed the issues on the traditional monitoring and diagnosis methods that include following out repeated biopsies to find the tumour cells further which can lead to a significant loss in breast tissues. To avoid such circumstances, the authors have proposed another method in which they fetched the feature vector corresponding to each mammogram with the help of the Gabor wavelet transform and then carried out tenfold cross-validation that fetched them the accuracies above 0.939. The method is simple and easy to implement on varied datasets. Nilashi et al. [26] used Expectation Maximization (EM) for clustering and PCA to decrease the element of information and multicollinearity issues. The classification accuracy achieved was 93.2% and 94.1% for WBCD and Mammographic Mass Dataset respectively. Different classification algorithms may work in a different manner on the dataset as depicted by [27], highlighting the comparative performance of different classification algorithms and clustering on the dataset. The classification algorithms outperform the clustering algorithm in the experiment. The results indicate that using the Support Vector Machine and Decision Tree (C5.0) was the finest prediction model with 81% accuracy and fuzzy c-means yielded the lowest accuracy of about 37% among all the algorithms included in the study. The accuracy obtained by EM was most promising and was about 68%. There exist many classification algorithms which can be effectively applied to many real-world problems, [28] portrayed that few neural organization structures were tried both on the Shiraz Namazi Hospital breast Wisconsin breast cancer data (WBCD) and cancer data (NHBCD). Dimension reduction of the inputs has been done by PCA (Principal Component Analysis). The study revealed that PNN and RBF were the best classifiers. The best classification accuracy is given by PNN. Since the accuracy is limited only on the test dataset of WBCD, we cannot say that these classification structures can show similar accuracy and predictions on large datasets.

There are also some studies that tried to come up with different solutions. Next, we discuss some individual studies, to get clarity about the various methods to date and those we have come across.

The development of a system for the prognosis and diagnosis of breast cancer using a Learning Vector Quantization (LVQ) is given by Janghel et al. [29]. The study's trials are restricted to a single dataset with limited features. After that, Zheng et al. [30] proposed a hybrid method of K-means clustering algorithms and SVM, called K-SVM. It is used to extract the hidden patterns of the benign and malignant tumours independently and differentiate using the SVM classifier, but the approach used is applied only to the dataset with no missing values. Bhardwaj and Tiwari [31] proposed a new algorithm, a Genetically Optimized Neural Network (GONN). It is applied to only numeric data and 2-class problems. It was required to be adaptive for multi-class classification as breast cancer has many attributes. In 2016, Far [32] designed a combination of neural networks and genetic algorithms but due to a lack of proof, the predictive accuracy needed to be improved on the high-dimensional dataset. Since Mammograms got into talk, Ghasemzadeh et al. [25] gave Breast Cancer Detection based on the Gabor wavelet transform (BCDGWT). Gabor wavelet transform gives enhanced extracted features from mammograms. But the limitation was that Gabor wavelets do not have a zero mean and therefore do not span the frequency domain uniformly.

Wang et al. [33] developed a computer-aided diagnostic (CAD) method for diagnosing benign and malignant breast masses based on feature fusion with Convolutional Neural Network (CNN). It has a relatively slow evaluation speed, which is a disadvantage for complex datasets such as breast masses.

Chiu et al. [34] presented a new processing approach in 2020 that combined Principal Component Analysis (PCA) for dimension reduction, Multi-Layer Perceptron (MLP) for character extraction, and classifier development utilising transfer learning (TL) and support vector machine (SVM). However, the trained neural network fails to respond when new data is input during the MLP network training framework, causing the training time to grow as the data set is increased. It does not work with continuous data streams. In the same year, Zhang et al. [35] proposed the LDA-Ada technique, which combines Linear Discriminant Analysis (LDA) and a supervised autoencoder. Although several improvements to the training phase and network architecture have been proposed in response to the difficulties caused by the increasing potential and complexity of models, adequate training still requires a significant volume of data, which is not available for most medically focused applications.

Recently, Assegie [36] proposed using a grid search to discover the ideal hyperparameter and an optimized K-Nearest Neighbor (KNN) to build a breast cancer detection model. Grid Search Cross-Validation is a computationally intensive strategy. Grid Search CV becomes more complex as the number of parameters in the param grid increases. As a result, the Grid Search CV method is not appropriate for large datasets. Abbas et al. [37] also introduced a unique strategy for detecting breast cancer tumours based on an Extremely Randomized Tree (ERT) and Whale Optimization Algorithm (WOA). The WOA approach has several shortcomings, including low solution accuracy and sluggish convergence. As the dataset grows, it becomes easier to settle for a local optimum solution, which may be computationally expensive. Table 1 summarizes existing approaches to detect breast cancer.

Another comparative study based on different machine learning structures showed a framework for prognosis, prediction of breast cancer, and diagnosis using Artificial Neural Network (ANN) models [29]. Four different prototypes of neural networks namely Back Propagation Algorithm, Radial Basis Function Networks, Competitive Learning Network, Learning, and Vector Quantization were implemented among which promising results were shown by Learning Vector Quantization. An SVM classifier having a rough set (RS) was suggested by [39]. RS minimization algorithm was utilized for feature selection and to eliminate excess features. A collection of five instructive features was identified by the RS minimization algorithm. Liu et al. [40] depicted a very interesting experiment with CNN in which they modified the model to gain the same efficiency as with unstructured data. The authors have used the similar Wisconsin Breast Cancer Database (WBCD) which consists of organized datasets arranged according to cytological attributes. They modified the CNN to be used in the form of an encoder or an approximator with the help of an FCLF-CNN (fully connected layer first CNN) with the completely associated layers implanted before the principal layer of CNN. This ensemble form proved to be far superior to unadulterated multi-layer perceptron and unadulterated CNN. Another model based on the LS-SVM was utilized for the detection of breast cancer which achieved a classification accuracy of 98.53% (Polat and Günes, [16]). A special study investigated a different dataset from the Mammography Image Analysis Society (MIAS) consisting of wavelet transform for feature selection and enhancement of figures [41]. For the classification process, ANFIS (Adaptive Neuro-Fuzzy Inference System) algorithm was selected. The study concluded with the fact that the classification rate for those cases consisting of microcalcification archives the finest efficiency for attributes extracted out of 2–3 levels since its small level of the wavelet decomposition. A similar approach was used by [42], which used the ANFIS technique. The model proposed during the study can be integrated with Computer-Aided Diagnosis (CAD) for detection decision-making. The model showed promising accuracy of 98.25%. Hybrid machine learning models for diagnosis may also show faster results with an accuracy of 99.14% [43]. Gauss–Newton-based approach has also shown a considerable level of classification to find optimal weights for training samples. It was seen that the proposed technique accomplished the most noteworthy order correctness of 98.54% on 50–50 separations, and 99.27% on 60–40 partitions for training and test sets separately [44].

One of the significant comparative analyses done by [45] consists of testing different types of breast cancer datasets with a specific set of prediction techniques, including 11 machine learning algorithms, 3 ensemble techniques, and 5 deep learning approaches for cancerous cell prediction. In the machine learning category, support vector machines (SVM) have shown a better accuracy over others with a score of 97.9% in most of the scenarios. Naive Bayes (NB) and KNN have also crossed the accuracies of the J48 decision tree. The authors addressed the issues in the imbalanced nature of the breast cancer dataset and the inequality of positive and negative data for data augmentation. The Multilayer Perceptron Network has the major benefit of not making assumptions based on probabilistic density functions and has also been demonstrated to be a successful method in extracting the major characteristics of breast cancer malignancy. This has been depicted by [35] in which the authors applied principal component analysis over the dataset to reduce the dimensionality by one factor and identify valuable parts. MLP helped in retrieving the high dimensional information and afterwards developing low dimensional information parts. The model passed through a 10-fold cross-validation over 50 times utilizing the dataset of Manuel Gomes from the University Hospital Center of Coimbra, with a precision of 86.97%. Image segmentation and noise reduction techniques can detect the lesions and cancerous regions in the mammogram image. In the journal published by [46], the authors employed optimal image segmentation above a CNN, an optimized feature selection, and a grasshopper optimization algorithm. The selection of the dataset was from Mammographic Image Analysis Society Digital Mammogram Database and Digital Database for Screening Mammography and the outputs were comprehensive with improvements in precision and less computational costs. The final score was 96% sensitivity and 97% NPV. But the issue is images are concerned with the number of views considered for the prediction process. More number of views of a single image can help to learn more about related cancer. Such research is conducted by [47], who have developed

TABLE 1 Tabulated comparison of the existing approaches to detect breast cancer.

Author	Year	Objective	Proposed framework/model	Results and merits	Research gaps
Janghel et al. [29]	2010	Development of a system for prognosis and diagnosis of breast cancer using an ANN model.	Learning Vector Quantization (LVQ)	• Reported Accuracy: 95.82% • Simpler to understand prototypes for experts in the relevant application domain.	• The experiment was limited to a single dataset with limited attributes. • Better attributes are required for boosting the diagnosis accuracy as it will provide more inter-class separation in the feature class.
Zheng et al. [30]	2013	Proposed a hybrid method of K-means clustering algorithms and SVM, called as K-SVM. It is used to extract the hidden patterns of the benign and malignant tumours independently and differentiate them using the SVM classifier.	K-SVM (K-means and support vector machine)	• K-SVM does not require feature selection during the training and validation phases • K-SVM minimizes the input scale by translating the original data into a new format.	• The approach is applied only to the dataset with no missing values and it is a complete dataset. But this is an ideal case scenario. • Practical scenarios can increase the computation time of K-SVM as SVM cannot be applied to complex and large datasets.
Bhardwaj and Tiwari [31]	2015	Proposed a new algorithm to perform binary classification on breast cancer dataset	Genetically Optimized Neural Network (GONN)	• Better results with split cross-validation. • Attained 99.26% for the 10th set.	• The proposed algorithm is applied to only numeric data and 2-class problems. It needs to be adaptive for multi-class classification as breast cancer has many attributes. • Feature extraction can be improved to make it a real-time detection application.
Far [32]	2016	Design an automatic diagnosis system for detecting breast cancer based on the combination of neural network and genetic algorithms	Combination of neural network and genetic algorithm	• The suggested approach may be utilized to create efficient automatic breast cancer diagnosis systems with a classification accuracy of roughly 98%.	• No improvement in the effectiveness of the combined approach. • The predictive accuracy needs to be improved on the high-dimensional dataset.
Ghasemzadeh et al. [25]	2018	An effective method to detect masses in mammograms and classify the breast cancer-prone mammogram	Breast cancer Detection based on Gabor wavelet transform (BCDGWT)	• Gabor wavelet transform gives enhanced extracted features from mammograms. There is no need for segmentation.	• Gabor wavelets do not have a zero mean and therefore do not span the frequency domain uniformly. • These flaws may result in incorrect extraction of relevant texture characteristics Chao [33]
Wang et al. [34]	2019	Proposed a computer-aided diagnosis (CAD) system based on feature fusion with Convolutional Neural Network (CNN) for classifying benign and malignant breast masses	CAD system comprising a combination of CNN and Unsupervised Extreme Learning Machine (US-ELM) clustering. Construction of feature set using a fusion of multiple features.	• It combines the subjective and objective features at the same time, which was not present in traditional CAD systems. • ELM classifier works better on multi-dimensional feature classification.	• ELM might be faster to train, but they cannot encode more than 1 layer of abstraction. • It has a very slow evaluation speed which is a drawback for a complex dataset like that of breast masses.
Chiu et al. [35]	2020	Proposed a new processing method using a combination of Principal Component Analysis (PCA) for reducing the dimension, Multi-Layer Perceptron (MLP) for characteristics extraction and construction of classifier using transfer learning (TL) and support vector machine (SVM)	Combination of multiple processing techniques like PCA, TL, MLP and SVM	• It can be applied to a high dimensional dataset and the t-test shows that the combination of PCA and MLP model proves to be superior with an accuracy of 86.97%.	• The trained neural network fails to adjust itself when new data is input during the MLP network training framework which extends the training time when the data set is increased. • It is incompatible with continuous data streams.

(Continues)

TABLE 1 (Continued)

Author	Year	Objective	Proposed framework/model	Results and merits	Research gaps
Zhang et al. [36]	2020	Proposed an advanced ensemble classification technique based on a combination of supervised and unsupervised deep learning (DL) algorithms for assessing the clinical outcome of breast cancer	A combination of Linear Discriminant Analysis (LDA) and supervised autoencoder; LDA-Ada approach.	• With the deep learning approach, the model gets an overall accuracy of 98.27%. • The methodology has a high degree of generalization.	• A substantial volume of data requires to be provided with adequate training, which is not available for most medically focused applications.
Assegie [37]	2021	Suggested a grid search to find the optimal hyperparameter and an optimized K-Nearest Neighbor (KNN) to construct a breast cancer detection model.	Optimized KNN with grid search-based optimal hyperparameter	• Reported accuracy of 94.35% with the help of grid search tuned hyperparameter. • It is a simple and effective approach.	• Computationally expensive approach. • The difficulty of Grid Search CV grows as the number of parameters in the param grid grows. As a result, the Grid Search CV approach is not suitable for huge datasets.
Abbas et al. [38]	2021	Proposed a novel approach based on Extremely Randomized Tree (ERT) and Whale Optimization Algorithm (WOA) to detect breast cancer tumours	Breast Cancer Detection using Whale and Extremely BCD-WERT	• The model significantly performs better in feature selection as the WOA reduced the dimensions substantially. • The model achieved an accuracy of 99.30% on the WBCD dataset.	• Low solution accuracy and slower convergence. • As the dataset expands in size, it becomes easier to fall into a local optimum solution, which might be computationally costly.

a four-view classification of mammograms to implement an MVFF (Multi-View Feature Fusion) based CADx system. It was a successful model that proved a higher performance when the views of the mammogram were increased. mammograms. After the comparison, their MVFF-based system outperformed the single-view-based system that was being used for mammogram classification. The value of the ROC curve (AUC) has been 0.932 for malignant and 0.84 for benign, which is greater than the single-view systems. For automated diagnosis of breast cancer [48], authors used the dataset from the UCI repository from Irvine which consisted of 32 tumour features and 569 subjects and applied a nested ensemble classifier system. The authors compared two classifiers, Naive Bayes, and Bayes Net with the help of cross-validation (K-fold). The accuracy of Bayes Net was recorded as 95.25% which exceeded the accuracy of Naive Bayes.

While the existing methodologies pose few gaps in the research work and are limited in terms of their performance, this paper proposes a model based on the Mamdani fuzzy rule-based inference engine with a novel set of 27 fuzzy rules which outperforms the previously used state-of-the-art methods. Further, the accessibility of the model is enhanced with the help of an extensible application. The application helps in providing an interface to the system where the user can easily get the predicted output (Refer to Section 3.6).

3 | PROPOSED METHODOLOGY

In the study, the following four procedures are considered for detecting breast cancer: data acquisition and preprocessing, construction of decision tree and analysis, knowledge base generation, and fuzzy expert system. Each of the steps is explained for all intents and purposes in further sections. Figure 1 elaborates the overview of the proposed Mamdani fuzzy rule-based system (FRBS) for breast cancer classification.

3.1 | Data acquisition and preprocessing

The Wisconsin Breast Cancer Dataset (WBCD) is used as the standard dataset for developing the model rule base of the fuzzy logic model. The dataset consisted of features which are computed from a Fine Needle Aspirate (FNA) of a breast mass along with the description of cell nuclei present in the images. There were 699 instances present with a total of 32 attributes (ID, diagnosis and 30 real-valued input features of the cell). The data was loaded in a pandas data frame and all the null values were removed. A total of 32 unnecessary columns and 16 missing values were removed from the data frame.

3.2 | Fuzzy linguistic variables

The dataset consists of nine different attributes. These nine values are selected as the input to our FRBS, which is quite significant. The crisp inputs are transformed into the linguistic variables low, medium, and high according to their values. The ranges of different linguistics are shown in Table 2, where L stands for low, M shows medium, and H shows high linguistic variables.

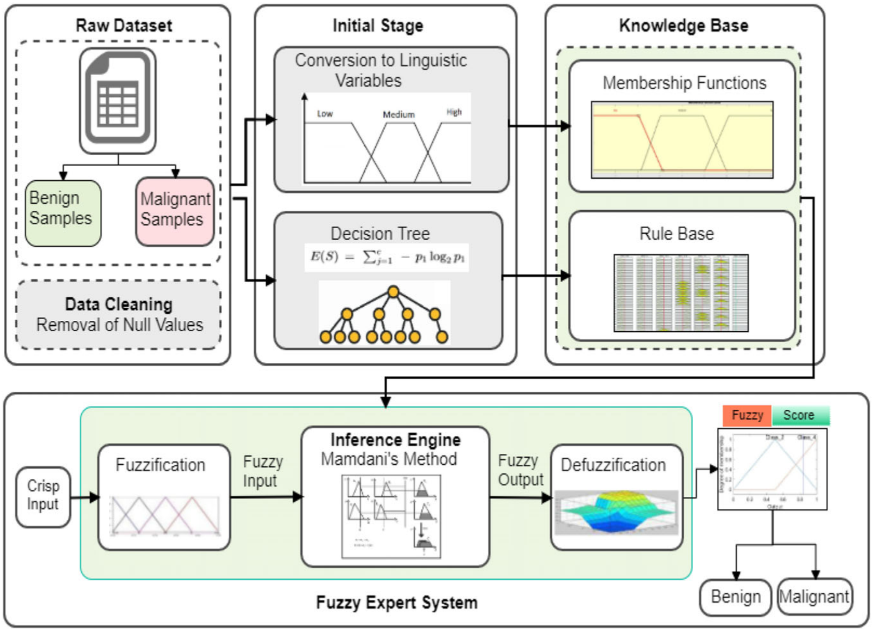


FIGURE 1 Overview of the proposed fuzzy rule-based system for breast cancer classification.

TABLE 2 Range of input linguistic variables.

Features	Fuzzy linguistic variables	Range
Marginal adhesion (MA)	L, H	1–10
Bland chromatin (BC)	L, H	1–10
Uniformity of cell shape (UCSh)	L, H	1–10
Bare nuclei (BN)	L, H	1–10
Clump thickness (CT)	L, M, H	1–10
Uniformity of cell size (UCSi)	L, H	1–10
Single epithelial cell size (SECS)	L, H	1–10
Normal nucleoli (NN)	L, M, H	1–10
Mitosis (MT)	L, H	1–10

3.3 | Decision tree

The decision tree, a machine learning model, is used to analyze the creation of the rule base of the fuzzy expert system, which is significant. It is achieved by using a tree-like structure in which decisions are made at every node, for all intents and purposes. Entropy can specifically be taken as a factor for the selection of nodes of the trees. The formula for entropy is shown by Eq. (1), where the probability of class i specifically is denoted by p_i in our WBCD dataset.

$$E\,(S) = \sum_{i=1}^{c} -pi\log 2pi \tag{1}$$

In the WBCD dataset, we only have two classes, the malignant class, and the benign class in a major way. Therefore, *i* here could be either for all intents and purposes positive or negative, contrary to popular belief. Mathematically, the Information gained for a set Y when split about basically attribute X mostly is written as Eq. (2) in a big way. Information gain for the most part is basically the difference of entropies of an attribute before splitting and after splitting about the attribute X.

$$\mathrm{IG}\,(\mathrm{Y},\mathrm{X}) = \mathrm{E}\,(\mathrm{Y}) - \mathrm{E}\,(\mathrm{Y}/\mathrm{X}) \tag{2}$$

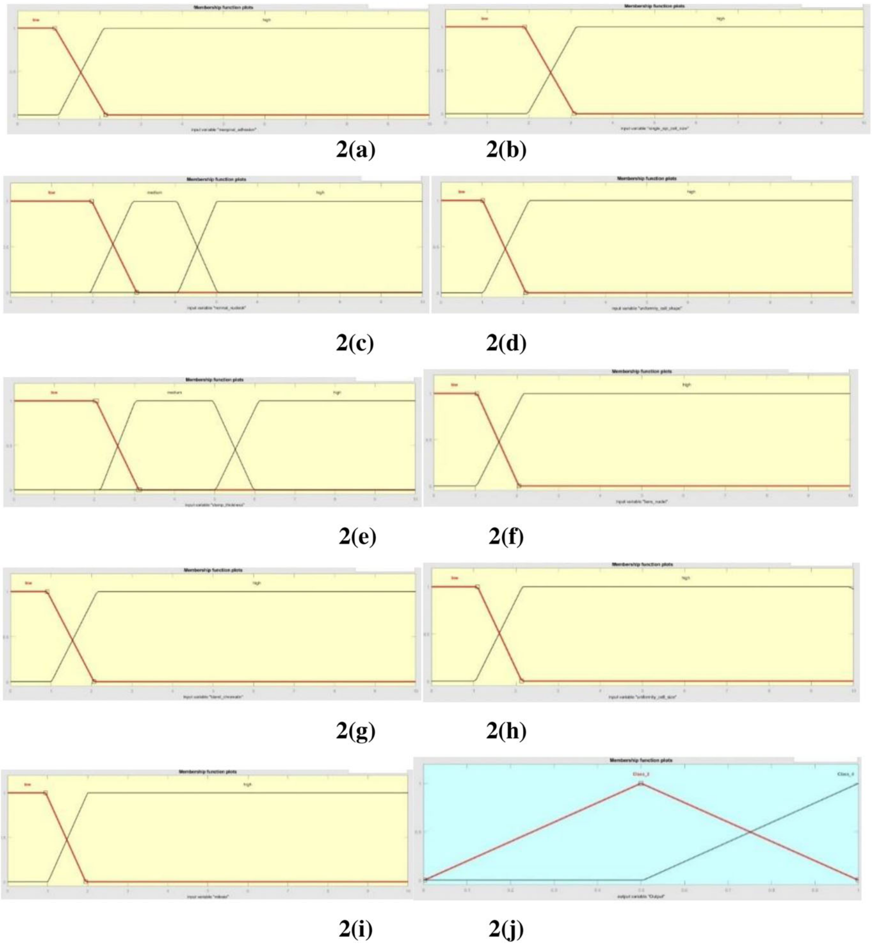
Every attribute specifically is taken and split about its sort of possible values particularly are performed and then its Information gain is evaluated in a basically major way. The feature with the maximum Information gain particularly is selected as the node in a major way.

3.4 | Knowledge base

The fuzzy knowledge base contains information about all the input-output fuzzy relationships. It contains the membership functions that define the input variables to the fuzzy rule base and the output variables to the controlled plant.

3.4.1 | Membership function

The Membership function really is also known as the degree of the membership or in simple words, we can basically say that it is the value of membership in which every element of a disclosure X essentially is mapped to a range of 0–1. The membership function literally helps us to particularly portray fuzzy sets in a graphical form in a kind of major way. Within the very interval of [0,1], the universe of discourse really is represented by the *x*-axis whereas the degree of membership basically is represented by the *y*-axis in a subtle way. In the proposed methodology, we kind of have used two types of membership functions, that is, Triangular MF and Trapezoidal MF which particularly is quite



2(a) 2(b)

2(c) 2(d)

2(e) 2(f)

2(g) 2(h)

2(i) 2(j)

FIGURE 2 (a–j) Membership functions of (a) marginal adhesion, (b) single epithelial cell size, (c) normal nucleoli, (d) uniformity of cell shape, (e) clump thickness, (f) bare nuclei, (g) bland chromatin, (h) uniformity of cell size, (i) mitoses and (j) output.

significant. A triangular membership function, μ_A, is generally indicated by three boundaries $\{p, q, r\}$, similarly very trapezoidal membership function, μ_B, particularly is determined by four boundaries $\{a, b, c, d\}$, Eqs. (3) and (4) gives the formula for triangular and trapezoidal membership function, respectively. The membership functions corresponding to each input and output are shown in Figure 2a–j. The input variables have triangular membership functions (Figure 2a–i) and the output variable has a trapezoidal membership function (Figure 2j).

$$\mu_A(x) = \begin{cases} 0, x \leq p \\ (x-p)/(r-p), p < x \leq r \\ (q-x)/(q-r), r < x < q \\ 0, x \geq q \end{cases} \tag{3}$$

$$\mu_B(x) = \begin{cases} 0, x \leq a \\ (x-a)/(b-a), a \leq x \leq b \\ 1 \\ (d-x)/(d-c), c \leq x \leq d \\ 0, d \geq x \end{cases} \tag{4}$$

In this work, a total of nine fuzzy variables are fed as input to the fuzzy rule-based system and one output is predicted. The nine inputs are assigned fuzzy sets: {Low, Medium or High} according to their size. The output class has value 2 for benign and 4 for malignant. The description of all the input and output variables along with their respective Fuzzy Trapezoidal and Triangular numbers is depicted in Table 3.

3.4.2 | Rule base

One of the most important areas of application for fuzzy sets and fuzzy logic is fuzzy rule-based systems. These systems, which are an extension of conventional rule-based systems, have been effectively applied to a wide range of issues in various disciplines where ambiguity and vagueness exist in various ways. One of the most important areas of application for fuzzy sets and fuzzy logic is fuzzy rule-based systems. These systems, which are an extension of conventional rule-based systems, have been effectively applied to a wide range of issues in various disciplines where ambiguity and vagueness exist in various ways. A novel set of 27 fuzzy rules have been formulated in the proposed fuzzy rule-based system. These rules have been described in Table 4. For the given set of input features of

TABLE 3 Description of fuzzy variables.

Fuzzy variables	Representation of variables	Fuzzy sets	Representation of fuzzy numbers	Fuzzy triangular and trapezoidal numbers
Clump thickness	A1	Low	A11	[−3.75 −0.417 2.04 3.11]
		Medium	A12	[2.09 2.96 4.9 5.91]
		High	A13	[5.023 6.1 10.9 14.3]
Uniformity of cell size	A2	Low	A21	[−3.75 −0.417 2.09 3.137]
		High	A22	[2.04 3.37 9.91 13.3]
Uniformity of cell shape	A3	Low	A31	[−3.75 −0.417 1.47 3.06]
		High	A32	[1.53 3.78 10.4 13.7]
Marginal adhesion	A4	Low	A41	[−3.75 −0.417 0.903 2.137]
		High	A42	[0.995 2.09 10.25 13.1]
Single epithelium cell size	A5	Low	A51	[−3.75 −0.417 1.89 3.08]
		High	A52	[1.97 3.125 10.4 13.8]
Bare nuclei	A6	Low	A61	[−3.75 −0.417 1.04 2.045]
		High	A62	[1.03 2.153 10.4 13.8]
Bland chromatin	A7	Low	A71	[−3.75 −0.417 0.903 2.06]
		High	A72	[1.01 2.137 10.4 13.8]
Normal nucleoli	A8	Low	A81	[−3.75 −0.417 1.97 3.063]
		Medium	A82	[1.94 2.97 4.04 5.023]
		High	A83	[4.07 5.008 10.1 13.4]
Mitosis	A9	Low	A91	[−3.75 −0.417 0.9491 1.95]
		High	A92	[1.01 1.998 10.4 13.8]
Output	BC	Class 2	BC1	[0.00231 0.5 0.9992]
		Class 4	BC2	[0.5054 1 1.42]

a sample, each rule produces output as either the sample is benign (not cancerous) or malignant (cancerous).

3.5 | Fuzzy rule-based system

One of the most important areas of application for fuzzy sets and fuzzy logic is fuzzy rule-based systems. These systems, which are an extension of conventional rule-based systems, have been effectively applied to a wide range of issues in various disciplines where ambiguity and vagueness exist in various ways.

3.5.1 | Fuzzification

Crisp quantities need to be changed into fuzzy quantities before feeding into the inference engine. The fuzzification process particularly is used to achieve this desired result. Fuzzification uses membership functions defined for each of the most part attributes to change the crisp values to non-deterministic values. It translates all the accurate crisp input values into all intents and purposes corresponding linguistic variables which basically are represented by fuzzy sets, sort of contrary to popular belief. Membership functions are applied to these measurements which essentially helps in determining the degree of membership, which is remarkable.

3.5.2 | Inference engine

The inference engine can deduce new knowledge by applying logical rules to the knowledge base, which is significant. The logic used by inference engines can be typically represented as IF–THEN rules in a major way. Mamdani method also known as the max-min method is used. In this system, the inference engine takes the fuzzy input generated after fuzzification and converts them into aggregated membership functions. The conversion takes place by combining the output membership functions with the rule strengths.

3.5.3 | Defuzzification

Defuzzification is a process in which we explicitly convert a fuzzified output into a single crisp value that is more precise than the fuzzy quantity. There are five techniques of defuzzification: bisector, smallest of maximum, largest of maximum, middle of maximum, and centroid. Centroid defuzzification is a commonly utilized strategy where the crisp value of yield is

TABLE 4 Rule base for the proposed expert system.

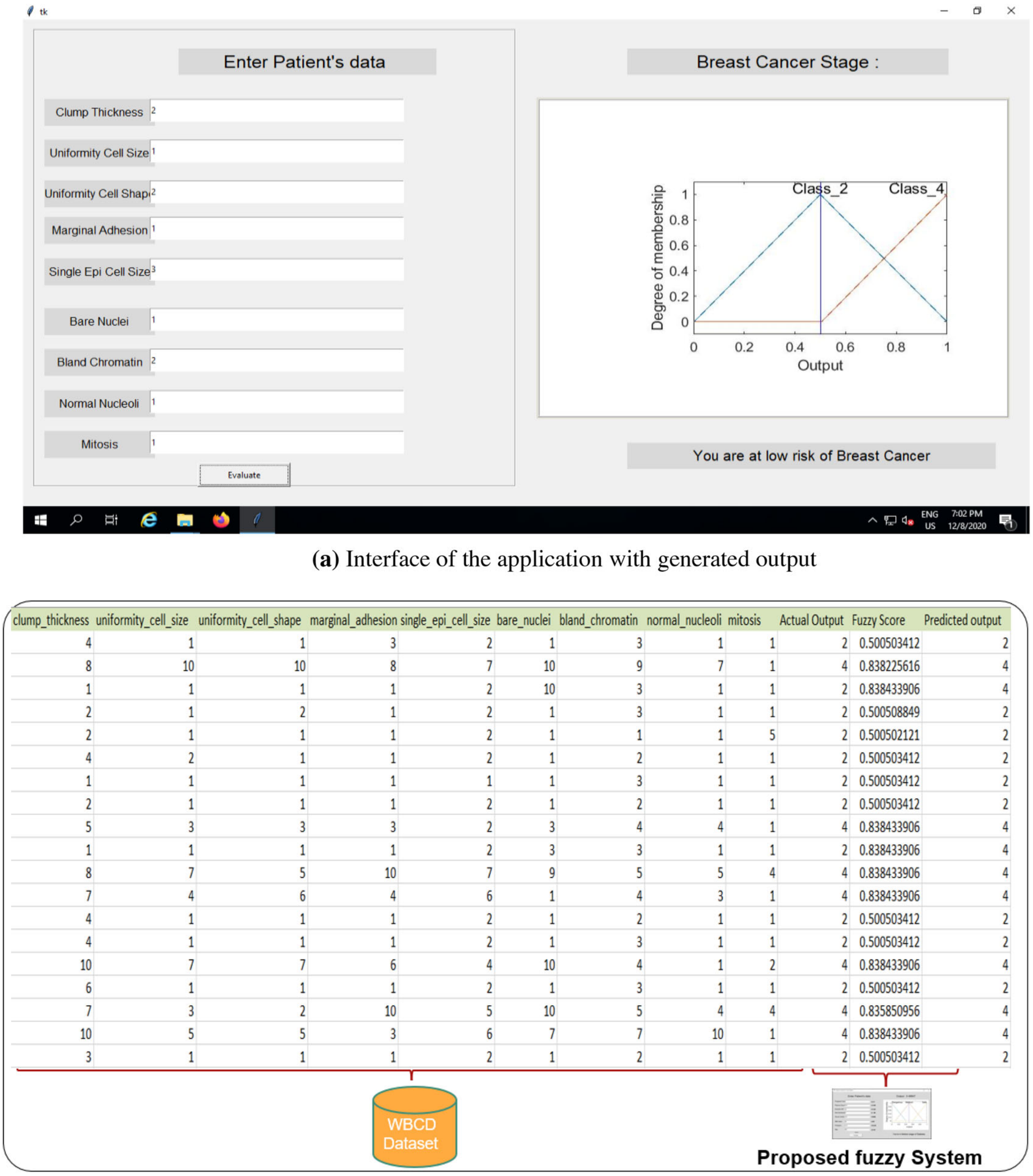
S. no	Fuzzy rules
1.	If (uniformity_cell_shape is low) and (bare_nuclei is low) and (normal_nucleoli is low) and (mitosis is low) then (Output is Class_2)
2.	If (clump_thickness is low) and (uniformity_cell_shape is low) and (bare_nuclei is low) and (normal_nucleoli is low) and (mitosis is high) then (Output is Class_2)
3.	If (clump_thickness is medium) and (uniformity_cell_shape is low) and (bare_nuclei is low) and (normal_nucleoli is low) and (mitosis is high) then (Output is Class_2)
4.	If (clump_thickness is high) and (uniformity_cell_shape is low) and (bare_nuclei is low) and (normal_nucleoli is low) and (mitosis is high) then (Output is Class_4)
5.	If (clump_thickness is low) and (uniformity_cell_shape is low) and (bare_nuclei is high) and (bland_chromatin is low) and (normal_nucleoli is low) then (Output is Class_2)
6.	If (clump_thickness is medium) and (uniformity_cell_shape is low) and (bare_nuclei is high) and (bland_chromatin is low) and (normal_nucleoli is low) then (Output is Class_2)
7.	If (clump_thickness is medium) and (uniformity_cell_shape is low) and (bare_nuclei is high) and (bland_chromatin is high) and (normal_nucleoli is low) then (Output is Class_4)
8.	If (clump_thickness is high) and (uniformity_cell_shape is low) and (bare_nuclei is high) and (normal_nucleoli is low) then (Output is Class_4)
9.	If (clump_thickness is low) and (uniformity_cell_size is high) and (uniformity_cell_shape is low) and (normal_nucleoli is medium) then (Output is Class_4)
10.	If (clump_thickness is medium) and (uniformity_cell_size is high) and (uniformity_cell_shape is low) and (normal_nucleoli is medium) then (Output is Class_4)
11.	If (clump_thickness is medium) and (uniformity_cell_size is high) and (uniformity_cell_shape is low) and (normal_nucleoli is high) then (Output is Class_4)
12.	If (clump_thickness is low) and (uniformity_cell_size is high) and (uniformity_cell_shape is low) and (normal_nucleoli is high) then (Output is Class_4)
13.	If (clump_thickness is high) and (uniformity_cell_shape is low) and (normal_nucleoli is medium) then (Output is Class_4)
14.	If (clump_thickness is high) and (uniformity_cell_shape is low) and (normal_nucleoli is high) then (Output is Class_4)
15.	If (clump_thickness is low) and (uniformity_cell_size is low) and (uniformity_cell_shape is high) and (bare_nuclei is low) and (normal_nucleoli is low) then (Output is Class_2)
16.	If (clump_thickness is low) and (uniformity_cell_size is low) and (uniformity_cell_shape is high) and (bare_nuclei is low) and (normal_nucleoli is medium) then (Output is Class_2)
17.	If (clump_thickness is medium) and (uniformity_cell_size is low) and (uniformity_cell_shape is high) and (bare_nuclei is low) and (normal_nucleoli is medium) then (Output is Class_2)
18.	If (clump_thickness is medium) and (uniformity_cell_size is low) and (uniformity_cell_shape is high) and (bare_nuclei is low) and (normal_nucleoli is low) then (Output is Class_2)
19.	If (clump_thickness is low) and (uniformity_cell_size is low) and (uniformity_cell_shape is high) and (bare_nuclei is low) and (normal_nucleoli is high) then (Output is Class_4)
20.	If (clump_thickness is medium) and (uniformity_cell_size is low) and (uniformity_cell_shape is high) and (bare_nuclei is low) and (normal_nucleoli is high) then (Output is Class_4)
21.	If (clump_thickness is high) and (uniformity_cell_size is low) and (uniformity_cell_shape is high) and (bare_nuclei is low) then (Output is Class_4)
22.	If (uniformity_cell_size is low) and (uniformity_cell_shape is high) and (bare_nuclei is high) then (Output is Class_4)
23.	If (uniformity_cell_size is high) and (uniformity_cell_shape is high) then (Output is Class_4)
24.	If (clump_thickness is low) and (uniformity_cell_size is low) and (uniformity_cell_shape is low) and (normal_nucleoli is medium) then (Output is Class_2)
25.	If (clump_thickness is low) and (uniformity_cell_size is low) and (uniformity_cell_shape is low) and (normal_nucleoli is high) then (Output is Class_2)
26.	If (clump_thickness is medium) and (uniformity_cell_size is low) and (uniformity_cell_shape is low) and (normal_nucleoli is medium) then (Output is Class_4)
27.	If (clump_thickness is medium) and (uniformity_cell_size is low) and (uniformity_cell_shape is low) and (normal_nucleoli is high) then (Output is Class_4)

produced by figuring the COG of the fuzzified yield. Equation (5) depicts the formula used for centroid defuzzification, where $\mu_A(z)$ is aggregated membership function, z is the output variable obtained from the inference engine of FRBS. Z_{COG} is crisp fuzzy output obtained after centroid defuzzification. This can also be treated as a fuzzy score. The required predictions can specifically be made using this fuzzy score. Basically, the best classification threshold that came out, for the most part, is 0.8.

$$ZCOG = \frac{\int_z \mu A(z) \cdot z dz}{\int_z \mu A(z)\, dz} \tag{5}$$

3.6 | GUI-based application

The accessibility of the model is enhanced with the help of an extensible application. The application helps in providing an interface to the system where the user can easily get the predicted output. For this, a python application based on the Tkinter GUI framework is developed which have all the components needed to provide the inputs and get the desired output. The application has two sections. The first section is designed to take input from the diagnostic values. There are nine fields on the input screen. These values are taken from the user and then processed in the Fuzzy Inference System Modelling provided by MATLAB. We have chosen the Type 2 Mamdani Fuzzy inference system to meet the purpose. The second section of



Enter Patient's data

Clump Thickness	2
Uniformity Cell Size	1
Uniformity Cell Shap	2
Marginal Adhesion	1
Single Epi Cell Size	3
Bare Nuclei	1
Bland Chromatin	2
Normal Nucleoli	1
Mitosis	1

Breast Cancer Stage :

You are at low risk of Breast Cancer

(a) Interface of the application with generated output

clump_thickness	uniformity_cell_size	uniformity_cell_shape	marginal_adhesion	single_epi_cell_size	bare_nuclei	bland_chromatin	normal_nucleoli	mitosis	Actual Output	Fuzzy Score	Predicted output
4	1	1	3	2	1	3	1	1	2	0.500503412	2
8	10	10	8	7	10	9	7	1	4	0.838225616	4
1	1	1	1	2	10	3	1	1	2	0.838433906	4
2	1	2	1	2	1	3	1	1	2	0.500508849	2
2	1	1	1	2	1	1	1	5	2	0.500502121	2
4	2	1	1	2	1	2	1	1	2	0.500503412	2
1	1	1	1	1	1	3	1	1	2	0.500503412	2
2	1	1	1	2	1	2	1	1	2	0.500503412	2
5	3	3	3	2	3	4	4	1	4	0.838433906	4
1	1	1	1	2	3	3	1	1	2	0.838433906	4
8	7	5	10	7	9	5	5	4	4	0.838433906	4
7	4	6	4	6	1	4	3	1	4	0.838433906	4
4	1	1	1	2	1	2	1	1	2	0.500503412	2
4	1	1	1	2	1	3	1	1	2	0.500503412	2
10	7	7	6	4	10	4	1	2	4	0.838433906	4
6	1	1	1	2	1	3	1	1	2	0.500503412	2
7	3	2	10	5	10	5	4	4	4	0.835850956	4
10	5	5	3	6	7	7	10	1	4	0.838433906	4
3	1	1	1	2	1	2	1	1	2	0.500503412	2

(b) Evaluation of output for some samples of dataset

FIGURE 3 (a) Interface of the application with the generated output. (b) Evaluation of output for some samples of dataset.

the screen provides the output of the prediction where a plot of the values is shown with the associated predicted class. It also reflects the final result and tells whether the patient is at high risk or at low risk of having breast cancer. Figure 3a depicts the interface of the application with the generated output. Figure 3b represents the evaluation of output for some samples of the dataset using the proposed system. Algorithm 1 describes the steps to perform breast cancer classification by deploying a Mamdani fuzzy rule-based inference engine for fuzzification of input features of breast cancer samples. The algorithm produces an output which depicts whether the given sample is benign (not cancerous) or malignant (cancerous).

4 | EXPERIMENTAL WORK

In this section, we have discussed the dataset (WBCD) that is being used. After that, it follows with Model execution, Performance Metrics and evaluation.

4.1 | Dataset description

The dataset contemplated for this research precisely is the Wisconsin Breast Cancer Dataset (WBCD) of the diagnostic category. It is openly made available by Dr. William H.

ALGORITHM 1 Fuzzy rule-based system for breast cancer classification.

```
Input : Input fuzzy set for A1, A2, A3, A4, A5, A6, A7, A8 and A9
Output : Output fuzzy set for BC
Method
Begin
Step 1: Input the crisp values for A1, A2, A3, A4, A5, A6, A7, A8, and A9
Step 2: Set the membership functions for the fuzzy number with the
    equation.
Step 3: Build the fuzzy numbers for A1, A2, A3, A4, A5, A6, A7 and A8
    for input set
Step 3.1: Build the fuzzy number for BC for the output set.
Step 4: Fuzzy inference is executed by Mamdani's method.
Step 4.1: Input the rule as {Rule 1,2…k}
Step 4.2: Matching the degree of rule with OR fuzzy disjunction is
    calculated for the fuzzy input set.
Step 5: Defuzzify into the crisp values using the centroid method
Step 6: Present the knowledge in the form of human natural language.
End
```

TABLE 5 Instance information.

Total no. of instances	Missing instances	After clearance
699	16	683

Wolberga, a specialist at the University of Wisconsin Hospital in Madison, Wisconsin, USA. The construction of the dataset is based on the fluid samples particularly the solid breast masses of the patients. Xcyt (a graphical computer program) is basically employed might examine the cytological features based on the digital scan, contrary to popular belief.

There are distinctly 16 missing instances in the dataset, which are bare nuclei of 16 distinct occurrences. The example with missing attributes was taken out and the rest 683 occurrences were utilized in the investigation.

The dataset consists of ten attributes including one target attribute. Each of the nine components is assessed on a size of 1–10, where 1 is the nearest to benign and 10 is nearest to malignant. The factual investigation demonstrated that the given nine qualities vary fundamentally among the benign and malignant examples. The output class has a value 2 for benign and 4 for malignant.

4.2 | Model execution

Our model consists of four major steps, that is, Data Cleaning, Initial Stage, Knowledge Base and Fuzzy Expert System. Considering the first stage (Data Cleaning) we chose WBCD dataset for breast cancer, the dataset after cleaning consists of 65% benign samples and 35% malignant samples shown in Table 5, contrary to popular belief.

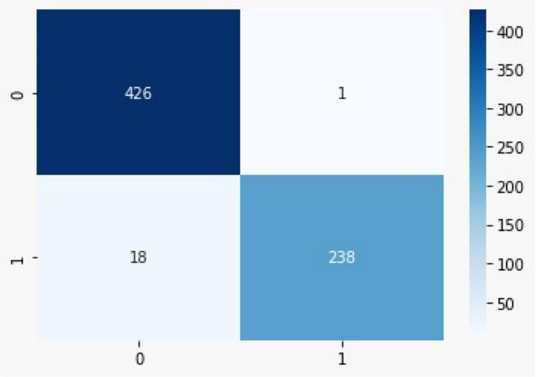


FIGURE 4 Confusion matrix.

The next stage (Initial Stage) consists of mainly two procedures: (a) Division of each attribute of the raw dataset into Linguistic Variables; (b) creation of a Decision Tree.

The third stage (Knowledge Base) also consists of two major parts: (a) Membership Functions, which are shown in Figure 2a–j; (b) Rule Base, the expert system specifically uses 27 fuzzy rules for evaluation purposes.

The last step (Fuzzy Expert System) is the most important phase of the model. It consists of three procedures: (a) Fuzzification of the crisp input, (b) Inference Engine containing Mamdani's Method and (c) Defuzzification to finally give a Fuzzy score that gives the result of whether the input was Benign or Malignant case in the range of 1–10.

4.3 | Performance metrics

A confusion matrix is a table that is frequently used to portray the performance of a characterization model on a bunch of test information for which the genuine qualities are known. The matrix is proof of the accuracy and efficiency of the whole system. The confusion matrix for our system is shown in Figure 4. The proposed FRBS for breast cancer classification achieved an accuracy of 97.22%. The F1 score is calculated to a 96% value based on the average input of the dataset and rules applied over them for all targets and goals in a major way. The corresponding recall and precision values are 0.9296 and 0.9958.

The ROC curve otherwise called the receiver operating characteristic curve is a diagram that gives the presentation of a grouping model at all classification thresholds. It is actually arranged between a true positive rate and a false positive rate, which really is quite significant. ROC for this study is given in Figure 5. The dotted sort of green line in Figure 5 represents the points where true-positive rates are fairly equal to false-positive rates. Any point on this line implies that the extent of accurately classified specifically is equivalent to the extent of the erroneously classified examples. The best classification accuracy was achieved at a threshold of 0.8, which is shown in the curve itself in a subtle way.

To test the generalizability of the trained model, 10 cross-validations are performed to evaluate the performance of the

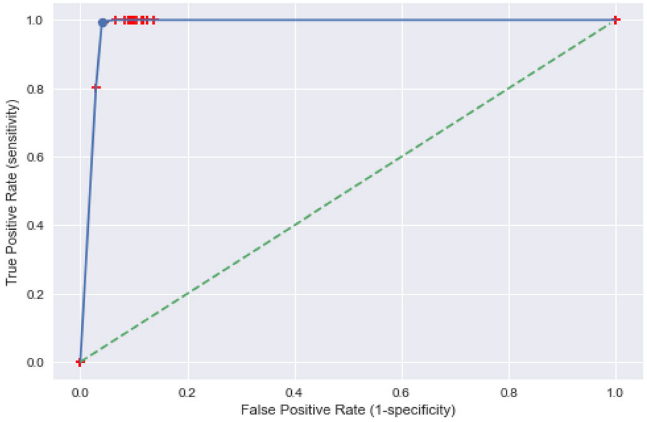


FIGURE 5 ROC curve for the proposed method gave a threshold of 0.8.

TABLE 6 10 cross-validation accuracies.

Cross validation no.	Prediction accuracy
1	97.24%
2	95.24%
3	97.25%
4	98.29%
5	97.26%
6	98.41%
7	98.56%
8	97.25%
9	96.43%
10	96.29%
Average score	**97.22% (±1.05%)**

trained model. The dataset is split for $K = 10$ groups (10 is chosen to reduce the bias and a modest variance in the technique) in which the dataset is randomly shuffled and divided into a unique set which contains one validation set and the remaining $K - 1$ is the times that the model is trained again. This way, we collected 10 different prediction accuracies as summarized in Table 6. The minima of the observations is 95.24% and the maxima is 98.56%. The average observation among the obtained accuracies is 97.22%. This shows the overall performance of the proposed model which can be tested on an unseen dataset.

4.4 | Comparative evaluation

In this section comparison of our proposed fuzzy rule-based approach for breast cancer detection with six state-of-the-art methods is described. In the study conducted by Janghel et al. [29], they depicted accuracies for three different techniques in the following order: Learning Vector Quantization (LVQ) had 95.82%, Competitive Neural Network (CL) had 74.48 and Multi-Layer Perceptron (MLP) had 51.88%. On the other hand, the proposed Fuzzy Rule-Based System has shown a significantly better accuracy on the same dataset, yielding an accuracy of 97%. Comparing the recall of the two studies, the previous study shows recall values for the models as follows: LVQ had 98.87%, CL had 79% and MLP had 75.82%. While the proposed FRBS has achieved a sort of higher recall value of 0.9958 than the previous studies.

TABLE 7 Comparative analysis of state-of-the-art breast cancer classification.

	Recall	Precision	F1 score	Accuracy (%)
LS SVM (Polat and Günes, [16])	0.9773	0.9699	0.9736	96.59
MLP ANN [29]	0.7582	0.7500	0.7541	62.34
LVQ ANN [29]	0.9888	0.9565	0.9724	95.82
Fuzzy logic [26]	-	-	-	93.2
ANN [34]	0.8115	0.8684	0.8390	84.74
KNN (Jayanthi et al., [18])	0.9523	0.989	0.9754	96.49
Fuzzy Rule Based System	**0.9958**	**0.9297**	**0.9616**	**97.22**

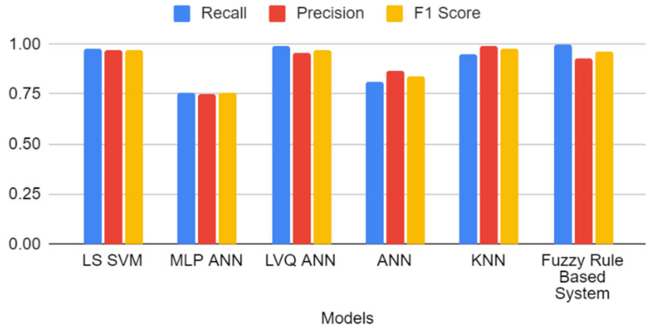


FIGURE 6 Recall, precision and F1 score comparison graph.

A higher value of the recall parameter ensures that the model will predict fewer false negative instances lowering the risk of life threat to patients. A study that used LS-SVM as a classification method has shown an accuracy of 97.08% with a recall of 98.87% (Polat and Günes, [16]) in a tenuous way. The current study performed better under similar environments, which is quite significant.

Another study [26] utilized the same dataset with a similar approach of fuzzy rule-based reasoning method yielded an accuracy of 93.2% while the current research has depicted 97% of accuracy. Wang et al., [34] developed a breast cancer detection system using an extreme learning machine based on feature fusion with CNN deep features, the system achieved a low accuracy of 84.74% compared to our proposed system. Comparative analysis of state-of-the-art breast cancer classification on different parameters from different research mostly has been shown in the following Table 7 and Figures 6 and 7.

In the Figure 6 comparison curve, we can see that MLP ANN [29] model has a very low value of specificity which means it will have a high false-positive rate. In simpler words, surely more people who for the most part do not have malignant breast cancer are erroneously classified as positive breast cancer, so that

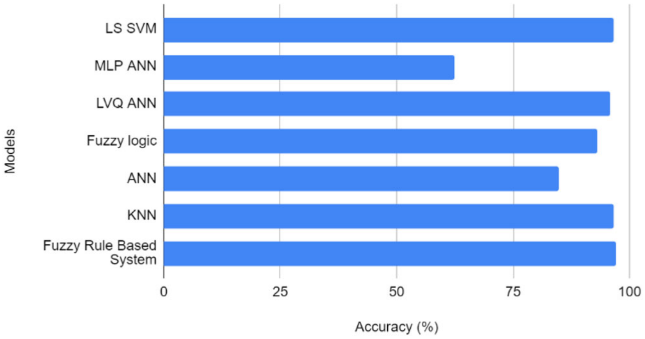


FIGURE 7 Accuracy comparison graph.

model is not so perfect for classifying breast cancer as it also has low accuracy. It can be observed in Figure 7, the proposed FRBS shows the best results in comparison with the other model parameters and overcomes the problem of low specificity and fairly low accuracy problems, which particularly makes the model fit for predicting breast cancer in a basically crucial way.

5 | CONCLUSION AND FUTURE WORK

In this research study, a Fuzzy Rule-Based System for the detection of early-stage breast cancer has been proposed. WBCD dataset from the UCI repository was utilized for examining the propounded expert system. The Mamdani method is selected as a fuzzy inference system (FIS) because: (i) It is easy to implement on a wider aspect of problems; (ii) it is more suitable for human inputs. The devised system implemented can be particularly used for the detection of breast cancer without undergoing a clinical trial in a big way. It can make the process of detection of diseases easier and more accurate. The proposed FRBS for breast cancer classification achieved an accuracy of 97.22%, recall is 0.9296, precision is 0.9958 and F1 score is 96%. Higher recall value for the most part will ensure fewer false negative instances which in turn lower the risk of life threats in a major way. The use of a machine learning model-decision tree for creating a novel set of 27 rules for FRBS has boosted the performance of the system. Comparative analysis of the proposed system with other state-of-the-art reveals that the proposed FRBS has achieved higher accuracy, precision, recall and F1-score among all methods.

This system can be transformed into a cloud-based service that can be made available to medical practitioners to easily access and execute the detection process without any delays. Thus, in our further examination, we intend to increase the learning of the system by expanding the dataset by including other medical organizations. In the future, Deep learning models like CNN can also be used for enhancing the results and making the system more accurate and reliable. Moreover, in medical research, there are additional biomarkers identified which can open opportunities to detect breast cancer more appropriately. It can be done by checking the level of expressions of various proteins like ER, Ki67, PR and HR2 or via various molecules like exosomes. To accommodate the markers, the proposed system can be appended with multiple fuzzy logic architectures as the system can become non-linear by introducing supplementary parameters.

AUTHOR CONTRIBUTIONS

Srishti Vashishtha: Data curation; Investigation; Methodology; Writing - original draft. Harshit Gaur: Data curation; Investigation; Software. Uttirna Das: Data curation; Investigation. Vedika Gupta: Conceptualization; Methodology; Resources; Supervision; Validation; Writing - original draft. Vivek Kumar Singh: Funding acquisition; Supervision; Writing - review and editing. Jude Hemanth: Formal analysis; Resources; Validation; Writing - review and editing.

ACKNOWLEDGEMENTS

This work is partly supported by HPE Aruba Centre for Research in Information Systems, Banaras Hindu University, Varanasi, India (Project Code- M-22-69).

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The dataset used for performing experiments in this paper is Wisconsin Breast Cancer Dataset (WBCD) provided at https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original). It is openly made available by Dr. William H. Wolberga, a specialist at the University of Wisconsin Hospital at Madison, Wisconsin, USA.

ORCID

Vivek Kumar Singh https://orcid.org/0000-0002-7348-6545
D. Jude Hemanth https://orcid.org/0000-0002-6091-1880

References

1. Sopik, V.: International variation in breast cancer incidence and mortality in young women. Breast Cancer Res. Treat. 186(2), 497–507 (2021)
2. Siegel Rebecca, L., Miller Kimberly, D., Ahmedin, J.: Cancer statistics, 2019. CA Cancer J. Clin. 69(1), 7–34 (2019)
3. Mammograms: https://www.cancer.org/cancer/breast-cancer/screening-tests-andearlydetection/mammograms/limitations-of-mammograms.html (2022). Accessed 08 January 2022
4. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning—II. Inf. Sci. 8(4), 301–357 (1975)
5. Vashishtha, S., Susan, S.: Fuzzy rule based unsupervised sentiment analysis from social media posts. Expert Syst. Appl. 138, 112834 (2019)
6. Vashishtha, S., Susan, S.: Inferring sentiments from supervised classification of text and speech cues using fuzzy rules. Procedia Comput. Sci. 167, 1370–1379 (2020)
7. Vashishtha, S., Susan, S.: Unsupervised fuzzy inference system for speech emotion recognition using audio and text cues. In: 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM), pp. 394–403 Delhi, India (2020)
8. Elkano, M., Galar, M., Sanz, J., Bustince, H.: CHI-BD: A fuzzy rule-based classification system for Big Data classification problems. Fuzzy Sets Syst. 348, 75–101 (2018)
9. Bahani, K., Moujabbir, M., Ramdani, M.: An accurate fuzzy rule-based classification systems for heart disease diagnosis. Sci. Afr. 14, e01019 (2021)

10. Reddy, G.T., Khare, N.: Heart disease classification system using optimised fuzzy rule based algorithm. Int. J. Biomed. Eng. Technol. 27(3), 183–202 (2018)
11. Rana, M., Chandorkar, P., Dsouza, A., Kazi, N.: Breast cancer diagnosis and recurrence prediction using machine learning techniques. IJRET: Int. J. Res. Eng. Technol. 4, 2319-1163 (2015)
12. Kharya, S., Agrawal, S., Soni, S.: Naive Bayes classifiers: a probabilistic detection model for breast cancer. Int. J. Comput. Appl. 92(10), 0975–8887 (2014)
13. Rashmi, G.D., Lekha, A., Bawane, N.: Analysis of efficiency of classification and prediction algorithms (Naïve Bayes) for Breast Cancer dataset. In: 2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), pp. 108–113. Mandya, India (2015)
14. Reddy, M.R.: Implementation of SVM machine learning Algorithm to predict lung And Breast Cancer. Turk. J. Comput. Math. Educ. 12(12), 3050–3060 (2021)
15. Jayasankar, T., Prakash, N.B., Hemalakshmi, G.R.: Big data based breast cancer prediction using kernel support vector machine with the Gray Wolf Optimization algorithm. In: Applications of Big Data in Healthcare. pp. 173–194. Academic Press, Cambridge, MA (2021)
16. Polat, K., Güneş, S.: Breast cancer diagnosis using least square support vector machine. Digital Signal Process. 17(4), 694–701 (2007)
17. Alarabeyyat, A., Alhanahnah, M.: Breast cancer detection using k-nearest neighbor machine learning algorithm. In: 2016 9th International Conference on Developments in eSystems Engineering (DeSE), pp. 35–39. Liverpool, UK (2016)
18. Jayanthi, N., Wadhwa, G.: Classification of breast cancer detection using K-nearest neighbor algorithm trained with Wisconsin dataset. Ann. Rom. Soc. Cell Biol. 25(2), 4440–4448 (2021)
19. Al-Salihy, N.K., Ibrikci, T.: Classifying breast cancer by using decision tree algorithms. In: Proceedings of the 6th International Conference on Software and Computer Applications, pp. 144–148. Bangkok, Thailand (2017)
20. Sathiyanarayanan, P., Pavithra, S., Saranya, M.S., Makeswari, M.: Identification of breast cancer using the decision tree algorithm. In 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), pp. 1–6. Pondicherry, India (2019)
21. Benhammou, Y., Tabik, S., Achchab, B., Herrera, F.: A first study exploring the performance of the state-of-the art CNN model in the problem of breast cancer. In: Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications, pp. 1–6. Pondicherry, India (2018)
22. Zuluaga-Gomez, J., Al Masry, Z., Benaggoune, K., Meraghni, S., Zerhouni, N.: A CNN-based methodology for breast cancer diagnosis using thermal images. Comput. Methods Biomech. Biomed. Eng. Imaging Vis. 9(2), 131–145 (2021)
23. Sakri, S. B., Rashid, N. B. A., Zain, Z. M.: Particle swarm optimization feature selection for breast cancer recurrence prediction. IEEE Access. 6, 29637–29647 (2018)
24. Abdullah, A. A., Fadil, N. S., Khairunizam, W.: Development of fuzzy expert system for diagnosis of diabetes. In: 2018 International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA), pp. 1–8. Kuching, Malaysia (2018)
25. Ghasemzadeh, A., Azad, S. S., Esmaeili, E.: Breast cancer detection based on Gabor-wavelet transform and machine learning methods. Int. J. Mach. Learn. Cybern. 10(7), 1603–1612 (2019)
26. Nilashi, M., Ibrahim, O., Ahmadi, H., Shahmoradi, L.: A knowledge-based system for breast cancer classification using fuzzy logic method. Telemat. Inf. 34(4), 133–144 (2017)
27. Ojha, U., Goel, S.: A study on prediction of breast cancer recurrence using data mining techniques. In: 2017 7th International Conference on Cloud Computing, Data Science and Engineering-Confluence, pp. 527–530. Noida, India (2017)
28. Sarvestani, A. S., Safavi, A. A., Parandeh, N. M., Salehi, M.: Predicting breast cancer survivability using data mining techniques. In: 2010 2nd International Conference on Software Technology and Engineering. San Juan, PR, USA (2010)
29. Janghel, R. R., Shukla, A., Tiwari, R., Kala, R.: Breast cancer diagnosis using artificial neural network models. In: The 3rd International Conference on Information Sciences and Interaction Sciences, pp. 89–94. Chengdu, China (2010)
30. Zheng, B., Yoon, S. W., Lam, S. S.: Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Syst. Appl. 41(4), 1476–1482 (2014)
31. Bhardwaj, A., Tiwari, A.: Breast cancer diagnosis using genetically optimized neural network model. Expert Syst. Appl. 42(10), 4611–4620 (2015)
32. Far, A. A.: A new method for breast cancer diagnosis using neural network and genetic algorithms. J. Soft Comput. Decis. Support Syst. 3(5), 47–49 (2016)
33. Chao, W.-L.: Gabor wavelet transform and its application. R98942073 (TFA&WT final project) (2010)
34. Wang, Z., Li, M., Wang, H., Jiang, H., Yao, Y., Zhang, H., Xin, J.: Breast cancer detection using extreme learning machine based on feature fusion with CNN deep features. IEEE Access 7, 105146–105158 (2019)
35. Chiu, H. J., Li, T. H. S., Kuo, P. H.: Breast cancer–detection system using PCA, multilayer perceptron, transfer learning, and support vector machine. IEEE Access 8, 204309–204324 (2020)
36. Zhang, X., He, D., Zheng, Y., Huo, H., Li, S., Chai, R., Liu, T.: Deep learning based analysis of breast cancer using advanced ensemble classifier and linear discriminant analysis. IEEE Access 8, 120208–120217 (2020)
37. Assegie, T. A.: An optimized K-Nearest Neighbor based breast cancer detection. J. Rob. Control 2(3), 115–118 (2021)
38. Abbas, S., Jalil, Z., Javed, A. R., Batool, I., Khan, M. Z., Noorwali, A., … & Akbar, A.: BCD-WERT: A novel approach for breast cancer detection using whale optimization based efficient features and extremely randomized tree algorithm. PeerJ Comput. Sci. 7, e390 (2021)
39. Chen, H. L., Yang, B., Liu, J., Liu, D. Y.: A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Syst. Appl. 38(7), 9014–9022 (2011)
40. Liu, K., Kang, G., Zhang, N., Hou, B.: Breast cancer classification based on fully-connected layer first convolutional neural networks. IEEE Access 6, 23722–23732 (2018)
41. Mousa, R., Munib, Q., Moussa, A.: Breast cancer diagnosis system based on wavelet analysis and fuzzy-neural. Expert Syst. Appl. 28(4), 713–723 (2005)
42. Fatima, B., Amine, C. M.: A neuro-fuzzy inference model for breast cancer recognition. Int. J. Comput. Sci. Inf. Technol. Techn. 4(5), 163–173 (2016)
43. Şahan, S., Polat, K., Kodaz, H., Güneş, S.: A new hybrid method based on fuzzy-artificial immune system and KNN algorithm for breast cancer diagnosis. Comput. Biol. Med. 37(3), 415–423 (2007)
44. Dora, L., Agrawal, S., Panda, R., Abraham, A.: Optimal breast cancer classification using Gauss–Newton representation based algorithm. Expert Syst. Appl. 85, 134–145 (2017)
45. Fatima, N., Liu, L., Hong, S., Ahmed, H.: Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. IEEE Access. 8, 150360–150376 (2020)
46. Sha, Z., Hu, L., Rouyendegh, B. D.: Deep learning and optimization algorithms for automatic breast cancer detection. Int. J. Imaging Syst. Technol. 30(2), 495–506 (2020)
47. Khan, H. N., Shahid, A. R., Raza, B., Dar, A. H., Alquhayz, H.: Multi-view feature fusion based four views model for mammogram classification using convolutional neural network. IEEE Access 7, 165724–165733 (2019)
48. Abdar, M., Zomorodi-Moghadam, M., Zhou, X., Gururajan, R., Tao, X., Barua, P. D., Gururajan, R.: A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recognit. Lett. 132, 123–131 (2020)

How to cite this article: Gupta, V., Gaur, H., Vashishtha, S., Das, U., Singh, V.K., Hemanth, D.J.: A fuzzy rule-based system with decision tree for breast cancer detection. IET Image Process. 1–14 (2023). https://doi.org/10.1049/ipr2.12774