DOI: 10.31534/engmod.2026.1.ri.01a
Original scientific paper
Received: 27.01.2025.

An Improved Early Breast Cancer Cells Classification and Prediction Based on a Fuzzy Neural Network Model

Yi Lv(1*), **Perk Lin Chong**(2), **Li Yao Sun**(3)

(1) *School of Mechanical Engineering, Hubei University of Arts and Science, Xiangyang, CHINA*
*e-mail: jiayi_lv@hotmail.com, *corresponding author*

(2) *School of Computing, Engineering & Digital Technologies, Teesside University, Middlesbrough, UNITED KINGDOM*
e-mail: p.chong@tees.ac.uk

(3) *College of Basic Medical Sciences, Liaoning University of Traditional Chinese Medicine, Shenyang, CHINA*
e-mail: 1544299539@qq.com

SUMMARY

Breast cancer is the most common type of cancer among women. Accurately diagnosis requires experienced medical practitioners to determine the nature of the cells. However, given the inherent complexity, there is a potential risk of misdiagnosis. This study proposes an artificial intelligence system that integrates fuzzy reasoning and a neural network to accurately classify cells as benign or malignant. Using the Wisconsin Breast Cancer (Diagnosis) dataset, samples were randomly partitioned into a training set of 400 and a testing set of 169 samples, following a 7:3 ratio. It is worth noting that these samples are correlated with 30 parameters, which can be computationally demanding. To address this issue, the principal component analysis (PCA) technique was employed to eliminate less significant parameters, resulting in a reduced set of only 6 key parameters. The proposed PCA-NF model achieved a test accuracy of 97.63%, with 100% precision, 93.10% recall, and a 96.43% F-measure. The PCA-ANFIS model achieved 95.27% accuracy and 94.12% for both precision and recall. Both models demonstrated reliable discrimination, supported by Matthews correlation coefficients of 94.80% and 90.16% for PCA-NF and PCA-ANFIS, respectively. The research novelty lies in the enhanced ANFIS approach, which provides comparable accuracy to existing artificial intelligence techniques while simplifying the diagnosis process. This user-friendly approach greatly benefits clinical medical experts by enhancing workflow efficiency and effectiveness.

KEY WORDS: *cell classification; fuzzy model; feature extraction; neural network; diagnosis.*

1. INTRODUCTION

Breast cancer is a malignant tumor that develops in the glandular epithelium of the breast. The global incidence of breast cancer has been increasing since the late 1970s. The American Cancer

Institute estimates that approximately 129.7 per 100,000 women are diagnosed with breast cancer each year, and men may also develop the disease. In 2022, there were about 280,000 new breast cancer cases, which was 51,400 more than in 2021 [1, 2]. Breast cancer most commonly occurs in individuals over the age of 45 and ranks second among the top ten cancers, with a mortality rate of 19.4%. Its pathogenesis is be closely related to factors such as heredity, genetic mutation, radiation exposure, lifestyle habits, gender and aging. For early detection of breast cancer, palpation or medical imaging are commonly used for regular screening or monitoring [2, 3]. After the initial diagnosis, tumor markers in medical and molecular tissues must also be confirmed. Histopathological examination is considered the most accurate "gold standard" [2] for breast cancer diagnosis. However, various clinical and technical factors, together with the limited sensitivity and specificity of existing classification and recognition methods [1, 3], require tumor specialists to rely on their extensive professional knowledge and clinical experience to ensure accurate diagnoses. Inexperienced physicians may reach incorrect conclusions, potentially leading to misdiagnoses. Fortunately, if these abnormal cellular tissue characteristics are detected at an early stage of breast cancer, comprehensive treatment such as surgery, radiotherapy and drug treatment can be implemented to prevent cancer progression and metastasis as early as possible, thereby improving patients' quality of life and survival rates [4, 5]. To support efficient medical decision-making, artificial intelligence researchers have developed a predictive system for breast cancer based on existing clinical data. This system assists oncologists in rapidly completing pathological assessments of cells and tissues. Compared with the complexity and high costs of surgical procedures, radiotherapy and chemotherapy techniques, this computer-aided automatic diagnosis system demonstrates greater cost-effectiveness and usability [5-9].

The integration of the learning mechanism of neural networks into fuzzy control systems to create an adaptive system that exhibits human-like perception and cognition has been a topic of research [10]. In such a system, fuzzy logic emulates human logical thinking, while neural networks simulate human conscious associations. By combining these two concepts, the machine acquires a certain level of logical reasoning similar to that of the human brain through continuous learning. When the neural network model is utilized to train randomly generated data, the resulting preliminary diagnosis outcomes are then processed by a fuzzy inference system. This processing enables the generation of optimal membership functions and fuzzy rules, which provide a highly generalized approach to representing the functional relationship between input features and output results. In the process of optimizing and implementing many fuzzy reasoning methods, the adaptive network fuzzy reasoning system (ANFIS) stands out. After performing forward and backward propagation through the network, it uses gradient information to gradually reduce the value of the error function, thus optimizing fuzzy rules, significantly reducing the complexity of fuzzy reasoning, and enhancing the interpretability of the neural network while demonstrating strong predictive performance [10-12]. The schematic diagram describing the hierarchical structure of ANFIS is shown in Figure 1 [10]. This system consists of five layers from I to V, and adopts backpropagation learning to modify the parameters of the input membership functions. It uses the least squares method to determine the curve of the best fits data and finally realizes the Takagi-Sugeno fuzzy inference system [12-14].

Übeyli [12] proposed a solution to the problem of slow training convergence caused by unfixed parameters and a large search space in previous models. This solution includes two mathematical analysis methods, namely, a hybrid algorithm system that combines the forward optimization of the least squares method and the reverse adjustment using gradient information.

The accuracy, specificity and sensitivity of this method are 99.08%, 99.27% and 98.74%, respectively.

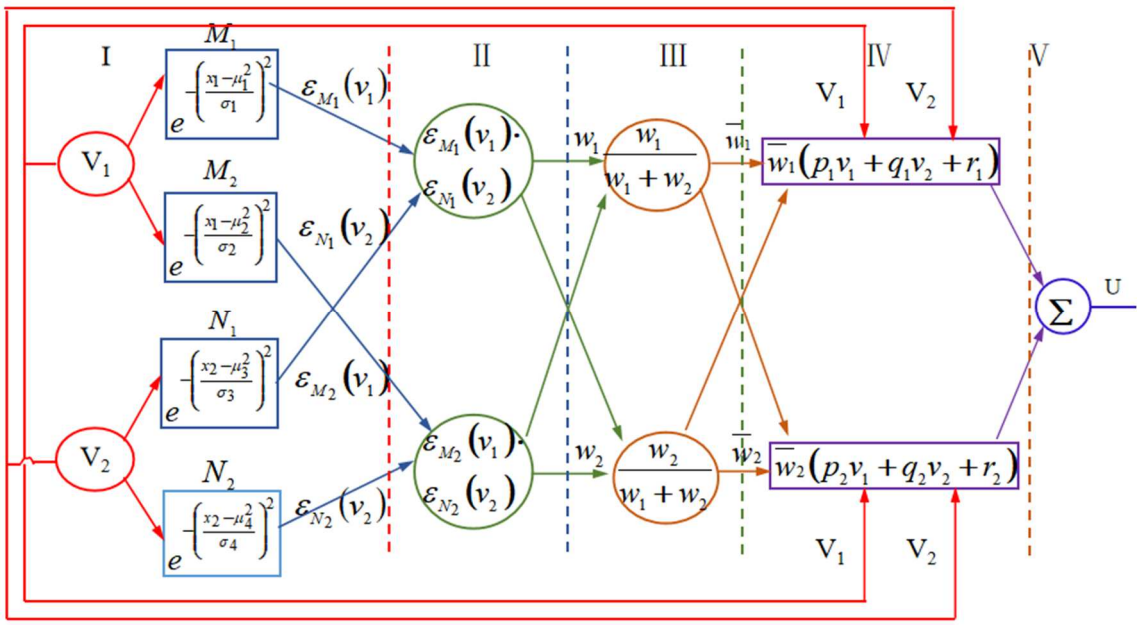

Fig. 1 *Anfis hierarchy description diagram*

Ashraf et al. [14] proposed the concept of information gain in a new fuzzy framework for cancer cell diagnosis. They stated that the information gain is the difference between prior entropy and conditional entropy. Each attribute can be evaluated by calculating the difference. During the training process, the difference is identified through forward propagation, while the parameters are adjusted based on the error in during backpropagation, and the optimal model is obtained through continuous iteration.

Senol and Yildirim [15] compared three types of fuzzy neural networks, in which the propagation rules of the CSFNN network are derived using the analytic equation of a cone. Each structure assigns two membership functions to each input variable, and the output is represented by a linear membership function. Levenberg-Marquardt and least squares algorithms are used to train the model, respectively. The analytical equation can be described as a linear combination of the propagation rules of the MLP network and the Euclidean distance between the center of RBF network and the input variables. When applied to actual clinical data published by the University of Wisconsin Breast Cancer Research Center (WBCD), the hybrid structure demonstrates high prediction performance [16-18].

On the one hand, El Hamdi et al. [17] used a probability vector to represent the virtual population in the CGA algorithm, and employed the SSGA algorithm to identify the best individual in the search space for iterative replacement. When this evolutionary method was applied to the WBCD dataset to determine the optimal fuzzy rules [16-18], they found that its primary advantage is the automatic optimization of parameters, thereby achieving high accuracy.

Zarbakhsh et al. [18] conducted experiments using WBCD datasets to validate the improved performance of the model. Compared with the two experiments conducted before and after applying the association rules method to remove the second attribute of WBCD [16-18], the

accuracy of ANFIS in the second experiment was slightly higher than that in the first experiment. In addition, the optimized cluster radius was determined through the COA module to obtain a fuzzy inference system with the minimum number of rules. At this stage, the accuracy of tumor classification reached 99.26%. These two improvements demonstrate that the number of features and rules provides important support for the performance of the tumor classification system.

Hamdan et al. [19] used the NPI value and survival time as input variables and applied two fuzzy algorithms to define the initial membership function. To reduce the number of rules, they reduced the membership functions in the two improved fuzzy c-means models from the first five to the last three. In the subtractive clustering model, they extracted sixteen rules from three types of data. By continuously adjusting the rules and optimizing the parameters of the learning mechanism, high accuracy was achieved in the breast cancer prognosis model.

The Mamdani algorithm of fuzzy reasoning [20, 21] typically involves designing variable intervals based on experts' experience. The membership functions of all input variables are then defined within these intervals, and various combinations of these functions are generated using specific rules. However, this approach has certain drawbacks. Firstly, it is time-consuming and cumbersome to identify an effective combination of input variables that yields accurate output results. Additionally, as the number of input variables increases, the number of rules grows significantly, making this method suitable only for problems with a small number of variables or membership functions. Moreover, arbitrary discarding fuzzy rules can lead to a lack of clarity in mathematical analysis. Despite the promising results achieved by various AI models on the WBCD dataset, two critical challenges hinder their practical clinical adoption, thereby presenting a clear research gap. First, the "black-box" nature of many high-performing models (e.g., deep neural networks and complex ensembles) limits interpretability, making it difficult for clinicians to understand and trust the diagnostic rationale. Second, the high dimensionality and redundancy among the 30 original features often lead to computationally complex models that are not optimized for efficiency in real-world clinical workflows.

To bridge this gap, the primary objective of this paper is to develop a novel hybrid intelligent diagnostic framework that balances high accuracy with clinical practicality. We propose an enhanced ANFIS approach centered on two key innovations: (1) Feature Simplification: We employ Principal Component Analysis (PCA) to distill the 30 complex histological features into a concise set of six principal components, thereby effectively reducing dimensionality and eliminating redundancy; (2) Interpretable & Efficient Modeling: We integrate this simplified feature set with a fuzzy neural network. Crucially, we introduce a variable membership function allocation strategy within the ANFIS architecture, in which the number of fuzzy sets for each input is weighted according to its importance (i.e., variance contribution), thereby constructing a compact and interpretable rule base without sacrificing model capacity.

The proposed PCA–ANFIS model, alongside a complementary PCA–NF (Neuro-Fuzzy) variant, is designed to achieve competitive diagnostic accuracy comparable to state-of-the-art "black-box" methods. More importantly, our core contribution lies in offering a favorable trade-off: by providing transparent fuzzy logic rules and operating on a reduced feature space, our framework significantly enhances interpretability and computational efficiency. This user-friendly and efficient approach addresses the critical need for trustworthy and deployable AI tools in clinical settings and has the potential to streamline the diagnostic workflow for medical experts.

2. MATERIALS AND METHOD

2.1 DATA FEATURE EXTRACTION

Principal component analysis is a technique for simplifying datasets by extracting important features through the linear transformation of the original variables into a new coordinate system of uncorrelated variables. The transformed variable with the largest variance becomes the first principal component, the second largest becomes the second principal component, and so forth. By retaining lower-order principal components and ignoring the higher-order ones, dimensionality reduction is achieved while preserving features that contribute most to the overall variance of the data. Each principal component represents a linear combination of the original variables, effectively rewriting the original information. In practice, the number of retained components is determined by specifying a threshold for the cumulative proportion of explained variance.

The entire WBCD dataset comprises 32 attributes: the first attribute is a sample identifier, the second is the diagnosis result (B for benign, M for malignant), and the remaining 30 represent histological characteristics of cells [16-18]. These 30 attributes correspond to three statistical measures (mean, standard error and worst/largest value) for each of 10 cytological features, including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension [16-18, 22-27]. This multi-measure representation creates informational redundancy among attributes, thereby increasing computational complexity and complicating the determination of appropriate membership function scopes in fuzzy systems.

2.1.1 MATHEMATICAL FORMULATION OF PCA

To address this redundancy, we employ PCA for dimensionality reduction. Let the feature matrix (excluding identifier and diagnosis columns) be denoted as $X \in R^{n\times m}$, where n=569 is the number of samples and m=30 is the number of original features. After column-wise standardization to zero mean and unit variance, the covariance matrix is computed as:

$$C = \frac{1}{n-1} X^T X \tag{1}$$

Through eigen decomposition $C = Q\Lambda Q^T$, where $Q = [q_1, q_2, \cdots, q_m]$ contains the eigenvectors and $\Lambda = diag(\lambda_1, \lambda_2, \cdots \lambda_m)$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$, the principal components are obtained as:

$$Z = XQ \tag{2}$$

where $Z \in R^{n\times m}$ contains the principal component scores. The cumulative explained variance ratio for the first k components is defined as:

$$R_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \tag{3}$$

We selected k=6 principal components corresponding to $R_6 \approx 85\%$, resulting in the reduced feature matrix $Z_{red} = Z_{:,1:6} \in R^{n\times 6}$. These six uncorrelated variables, ordered by decreasing

variance contribution, serve as the extracted main features for subsequent analysis, as shown in Figure 2.

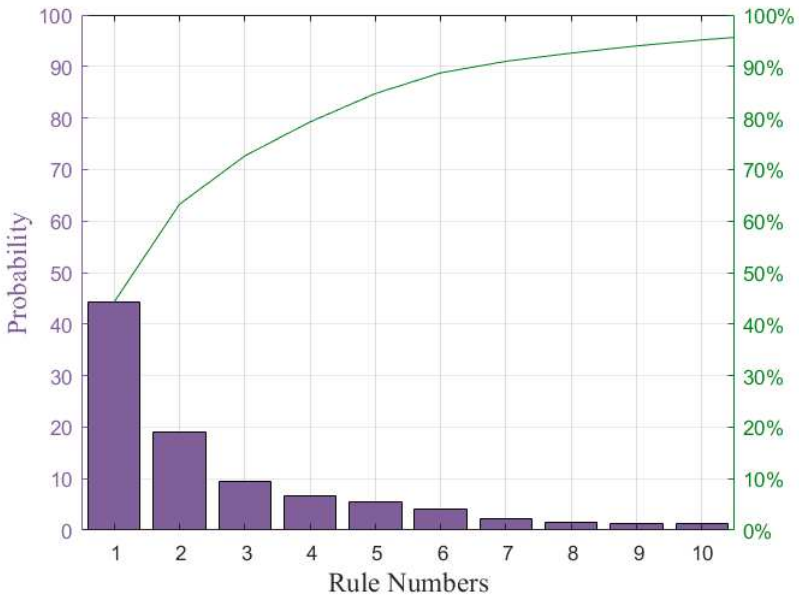

Fig. 2 *Main feature extraction histogram*

2.1.2 MODEL INPUTS AND OUTPUT DEFINITION

(i) Input Variables: The six principal components (PC1-PC6) derived above, denoted as $z_1, z_2, \cdots z_6$, represent the most informative linear combinations of the original 30 attributes.

(ii) Output Definition: Binary classification where output $y=0$ represents benign (B) and $y=1$ represents malignant (M) tumors, directly corresponding to the diagnosis labels in the WBCD dataset.

(iii) Membership Function Allocation: Each principal component z_i is associated with Gaussian membership functions (MFs). The number of membership functions per component was determined through systematic experimentation, as detailed in Section 2.2, with the allocation scheme based on each component's relative importance as indicated by its variance contribution.

After feature extraction, classification and prediction are performed using two complementary fuzzy systems: the ANFIS toolbox in MATLAB (PCA-ANFIS) and a custom fuzzy neural network (PCA-NF).

2.2 THE PCA-ANFIS MODEL: STRUCTURE AND IMPROVEMENTS

A critical aspect of designing the PCA-ANFIS model is the management of the size of the fuzzy rule base, which directly impacts model complexity and generalizability. In fuzzy systems using grid partitioning, rules are generated by enumerating all possible combinations of the membership functions assigned to each input variable. For a system with m input variables, where the i-th variable has q_i membership functions, the total number of rules is given by the product rule:

$$N_{rules} = \prod_{i=1}^{m} q_i \tag{4}$$

To illustrate, consider a hypothetical system with three inputs, each having four MFs. The rule count is $4^3 = 64$. In our work, with six principal components as inputs, a uniform assignment of, for example, three MFs to each would yield $3^6 = 729$ rules, which is excessive and prone to overfitting given our dataset size. To achieve a balance between expressiveness and efficiency, we designed several MF allocation schemes (M2-M6) guided by the principle that features with higher variance contribution from PCA should be allocated more MFs to capture finer data distinctions. For example, M6 was assigned [4, 3, 2, 2, 2, 2] membership functions to principal components PC1 through PC6, respectively. Applying Eq. (4), this yields a total of:

$$N_{rules} = 4 \times 3 \times 2 \times 2 \times 2 \times 2 = 192 \tag{5}$$

Each unique combination of these MFs generates one first-order Takagi-Sugeno rule of the general form:

$$\text{if } PC_1 \text{ is } A_{1j} \text{ and } PC_2 \text{ is } A_{2k} \text{ and ... and } PC_6 \text{ is } A_{6l} \text{ then } y = \alpha_0 + \sum_{i=1}^{6} \alpha_i PC_i \tag{6}$$

where A_{ij} are Gaussian membership functions and α_i the consequent parameters to be learned.

This strategy of variable MF allocation successfully reduces the rule base from a potential 729 rules to a more manageable 192 rules, significantly lowering computational cost while preserving the model's ability to capture key patterns in the data.

2.3 THE PCA-NF MODEL: ARCHITECTURE AND LEARNING

Next, we construct a fuzzy neural network model based on the ANFIS hierarchy shown in Figure 1 to achieve classification and prediction of cancer cell data. In the following section, this model is referred to as PCA-NF. Suppose two variables v_1 and v_2 are input to the fuzzy reasoning system, and an output variable u is obtained. For the first-order fuzzy neural network model based on the Takagi-Sugeno model, two representative rules with fuzzy if-then rules can be described as follows:

The first rule is that when v_1 is M_1 and v_2 is N_1, then $u = p_1 v_1 + q_1 v_2 + r_1$

The second rule is that when v_1 is M_2 and v_2 is N_2, then $u = p_2 v_1 + q_2 v_2 + r_2$

In Figure 1 square nodes indicate that the node parameters are adjustable, whereas circular nodes indicate that the node has no adjustable parameters. For the convenience, the data of the i^{th} node in the n^{th} layer is represented as:

$$O_i^n \left(i = 1,2,3,4,5 \right) \tag{7}$$

where O_i^n is the membership degree of the fuzzy set $A = \left(M_1, M_2, N_1, N_2 \right)$, which determines the degree to a given input x_1 (or x_2) satisfies A. M_i and N_i are language labels associated with node functions.

Layer I: Convert the input variables into the membership of each fuzzy set.

$$O_i^1 = \varepsilon_{M_i}(x_1) \quad i = 1,2 \tag{8}$$

$$O_i^1 = \varepsilon_{N_i}(x_2) \quad i = 1,2 \tag{9}$$

A function $\varepsilon_{A_i}(x)$ with a value range of 0-1 is typically selected to represent the membership function of each variable interval, such as a bell-shaped Gaussian membership functions:

$$\varepsilon_{A_i}(x) = e^{-\left(\frac{x-\mu_i}{\sigma_i}\right)^2} \tag{10}$$

where μ_i and σ_i represent the mean and standard deviation of the Gaussian function, reflecting the center and width of the distribution.

Layer II: Multiply the input signals to obtain the weight coefficient of each rule.

$$w_i = \varepsilon_{M_i}(v_1) \cdot \varepsilon_{N_i}(v_2) \quad i = 1,2 \tag{11}$$

Layer III: Normalize the i^{th} rule corresponding to the i^{th} node.

$$O_i^3 = \overline{w}_i = \frac{w_i}{\sum_j w_j} \quad i = 1,2 \quad j = 2 \tag{12}$$

Layer IV: Calculate the output of each rule as linear combination.

$$O_i^4 = \overline{w}_i f_i = \overline{w}_i (p_i v_1 + q_i v_2 + r_i) \quad i = 1,2 \tag{13}$$

Layer V: Aggregate all rule outputs to obtain the final output value.

$$O_i^5 = \sum_i \overline{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \quad i = 1,2 \tag{14}$$

2.4 MEMBERSHIP FUNCTION SPECIFICATION AND RATIONALE

For the PCA-ANFIS model, Gaussian MFs were chosen for their smoothness, differentiability (essential for gradient-based learning), and widespread applicability in pattern recognition problems. The parameters $\{\mu_i, \sigma_i\}$ for each MF are automatically tuned during the hybrid learning process of ANFIS, which adjusts them to minimize the output error.

In the PCA-NF model, Gaussian MFs are similarly adopted in the fuzzification layer to maintain methodological consistency and facilitate fair comparison. The MF parameters are initialized based on the statistical distribution of each input feature (mean and standard deviation of training data) and are subsequently updated via backpropagation along with the network's connection weights. This enables the MFs to adapt dynamically to the data patterns during training.

The number of MFs per input was determined experimentally (as described in Section 2.2), guided by the cumulative variance contribution of each principal component. This strategy ensures that more influential features are modeled with higher granularity, while less

informative features are represented more coarsely, achieving a balance between model complexity and performance.

3. RESULTS AND DISCUSSION

3.1 DATA SET AND EVALUATING INDICATOR

In the present work, we used the Wisconsin Breast Cancer Diagnostic (WBCD) Dataset, which contains 569 actual tumor sample records. Each record consists of a sample retrieval address, a diagnosis (data set label: "B" indicates benign, "M" indicates malignant) and 30 real-valued input features. According to statistics, 357 cases (62.7%) are benign breast cancer and 212 cases (37.3%) are malignant breast cancer. These feature values are the actual values obtained by measuring the medical digital image of cell biopsy tissue extracted via breast mass puncture. They represent the characteristics of the nuclei present in the image. The criteria for the correct classification of malignant and benign cells are shown in Table 1. In the table, the letters T/F indicate the correctness of the predicted result, signifying whether it aligns with the actual situation. Similarly, the letters M/B indicate whether the predicted result classifies as malignant or benign.

Table 1 *Accuracy criteria for classification of benign and malignant cells [4]*

Actual / *Predicted*	*Malignent*	*Benign*
Malignent	*TM*	*FM*
Benign	*FB*	*TB*

An initial exploratory analysis was conducted on the WDBC dataset, including visual inspection via box plots and consideration of biological plausibility. No extreme values indicative of measurement error were identified; apparent statistical outliers corresponded to genuine pathological features of malignant tumors. Therefore, all 569 samples were retained to preserve the dataset's clinical validity and comparability with prior studies. All features were normalized to the range [0, 1] using Min-Max scaling.

The performance evaluation indicators of the model are as follows [4]:

$$Accuracy(Acc) = \frac{TM + TB}{TM + TB + FM + FB} \tag{15}$$

$$Precision(Pr) = \frac{TM}{TM + FM} \tag{16}$$

$$Recall(Re) = \frac{TM}{TM + FB} \tag{17}$$

$$Specificity(Sp) = \frac{TB}{FM + TB} \times 100\% \tag{18}$$

$$F\text{-}Measure(F_M) = \frac{2 \times Pr \times Re}{Pr + Re} \tag{19}$$

$$MCC = \frac{TM \times TB - FM \times FB}{\sqrt{(TM + FM)(TM + FB)(TB + FM)(TB + FB)}} \quad (20)$$

where *F-Measure* is the comprehensive evaluation index, *MCC* is the Matthews correlation coefficient.

Equations (15)–(20) constitute a set of complementary performance metrics essential for comprehensive evaluation of the binary classification model. Accuracy (Eq. 15) provides an overall measure of correct predictions but can be misleading in imbalanced datasets. Precision (Eq. 16) indicates the reliability of a positive (malignant) prediction, i.e., the proportion of correctly identified malignant cases among all cases predicted as malignant. Recall (Eq. 17), also known as sensitivity, quantifies the model's ability to detect all actual malignant cases, which is critical in medical diagnosis to minimize false negatives. Specificity (Eq. 18) measures the model's performance in correctly identifying benign cases, thereby complementing recall. The F-Measure (Eq. 19) combines precision and recall into a single score and is especially useful when seeking a balance between these two often competing metrics. The Matthews Correlation Coefficient (MCC, Eq. 20) provides a robust measure that considers all four categories of confusion matrix and remains reliable even for imbalanced data. Together, these metrics move beyond a single accuracy score and provide a multi-faceted understanding of the model's classification performance, ensuring that it is not only accurate but also clinically trustworthy-particularly by prioritizing high recall to avoid missing malignant cases while maintaining strong precision and specificity to support effective clinical decision-making.

3.2 EXPERIMENT RESULT

3.2.1 EXPERIMENTAL SETUP AND MODEL CONSTRUCTION

Of the 569 samples in the dataset, 357 are benign and 212 are malignant. When training our model, the samples were randomly divided into training and test sets according to a 7:3 ratio. The number of samples in the training set was 400, and the number in the test set was 169.

First, the principal component analysis was applied to extract the important features of the data, and reduce the 30-dimensional attributes to 6-dimensions. The reduced feature set was then fed into the neural network model and the fuzzy logic model as input variables, with the ultimate goal of performing binary classification of cells in the UCI WBCD dataset. Due to the limited dataset size, a random sub-sampling validation method was employed to estimate the true accuracy, thereby reducing the risk of overfitting caused by an excessive number of rules when modeling the six extracted principal components. Furthermore, key performance indicators of the two proposed models were compared with those reported in existing literature [4, 20-25]. To verify the feasibility of PCA-ANFIS network improvement, we adjusted the number of membership functions and constructed five models, denoted M2 to M6. Simultaneously, by combining principal component analysis with the fuzzy neural network model, model M7 (PCA-NF) was developed. A critical design choice in this network was the number of neurons in the hidden layer, which implicitly determines the model's complexity and, by analogy to fuzzy systems, the effective number of learnable "rules" or pattern templates.

After a series of experiments varying the hidden layer size, performance was observed to saturate beyond a certain point. Specifically, increasing the number of hidden nodes beyond 12 did not yield significant improvements in validation accuracy but resulted in linear increase in

computational cost and the risk of overfitting. Therefore, the architecture was finalized with 12 hidden neurons, achieving an optimal balance between model capacity and generalization ability.

The single output neuron uses a sigmoid activation function, producing a value between 0 and 1, which is thresholded at 0.5 to yield the final binary classification: 0 for Benign and 1 for Malignant.

3.2.2 MODEL PERFORMANCE ANALYSIS

The performance indicators of these models for cancer cell classification prediction are presented in Table 2. A detailed analysis of Table 2 highlights the impact of membership function (MF) allocation on model performance. The six principal features were ranked in descending order by their cumulative contribution to the total variance, with the first feature contributing the most. In models M2 to M6, the number of membership functions per feature ranges from 2 to 4. Models M2, M3, and M4 exhibit relatively lower accuracy and recall on the test set compared to M6. This underperformance can be attributed to two interconnected factors:

(1) Insufficient model capacity: Assigning only two MFs to the most important principal component (as in M2 and M4) results in an overly coarse discretization of that feature's value range. This coarse granularity limits the model's ability to capture subtle yet critical distinctions between benign and malignant cases in the transformed feature space;

(2) Suboptimal rule structure: although M3 assigns three MFs to the first component, the specific combination of MF counts across all six features may not align optimally with the underlying data distribution. As a result, the constructed rule base may be either overly simplistic or structurally misaligned, thereby reducing predictive performance on unseen data.

The case of model M5 is particularly illustrative. It achieves a near-perfect training accuracy of 99.75% but experiences a sharp decline in test accuracy to 89.94%, which is a classic indication of overfitting. The configuration [4, 2, 2, 2, 2, 2], with four MFs on PC1, produces a rule base that is excessively complex relative to the sample size and intrinsic dimensionality of the training data. Consequently, the model tends to memorize dataset-specific noise and idiosyncrasies patterns rather than learning generalizable pathological characteristics.

In contrast, model M6, with its MF allocation of [4, 3, 2, 2, 2, 2], achieves an optimal balance. It attains a test accuracy of 95.27%, with both precision and recall balanced at 94.12%. A direct comparison with the benchmark model M1 (NN-EF) [4], which reported an accuracy of 99.41%, indicates that M6 achieves slightly lower accuracy. However, this comparison must be properly contextualized. The NN-EF model operates in the original, high-dimensional feature space (30 raw features), whereas M6 operates on a parsimonious, information-rich subset consisting of only 6 principal components. This represents a fivefold reduction in feature dimensionality, thereby significantly lowering model complexity and computational cost during inference - an important advantage for clinical deployment.

The complementary performance metrics, with an F-Measure of 94.12% and MCC of 90.16%, further confirm that M6 is a robust and well-calibrated classifier. It successfully avoids the overfitting pitfall observed in M5 while exhibiting greater expressive power and generalization capability than the under-parameterized models (M2-M4). This configuration demonstrates

that a principled, importance-weight allocation of model complexity is essential for achieving a favorable balance amonge performance, interpretability, and efficiency in this diagnostic task.

Table 2 *Comparison of performance indicators with some models*

Model	*Network*	*Data*	*Acc*	*Pr*	*Re*	*Sp*	*FM*	*MCC*
M1[4]	*NN-EF*		*99.41*	*100*	*99.06*	*98.46*	*99.21*	*98.76*
M2	*Number of membership functions=[2,2,2,2,2,2]*	*Training*	*98.00*	*98.63*	*96.00*	*97.30*	*99.20*	*95.73*
		Test	*94.08*	*96.36*	*86.89*	*98.15*	*91.41*	*87.16*
M3	*Number of membership functions=[3,2,2,2,2,2]*	*Training*	*99.25*	*100*	*97.92*	*100*	*98.95*	*98.37*
		Test	*92.90*	*95.45*	*92.64*	*97.03*	*94.23*	*90.13*
M4	*Number of membership functions=[2,2,3,2,2,2]*	*Training*	*98.25*	*99.27*	*95.77*	*97.43*	*99.61*	*96.18*
		Test	*91.72*	*93.44*	*85.07*	*96.08*	*89.06*	*82.65*
M5	*Number of membership functions=[4,2,2,2,2,2]*	*Training*	*99.75*	*100*	*99.32*	*100*	*99.65*	*99.46*
		Test	*89.94*	*86.79*	*82.14*	*84.19*	*93.81*	*77.05*
M6	*Number of membership functions=[4,3,2,2,2,2]*	*Training*	*98.75*	*100*	*96.53*	*100*	*98.23*	*97.30*
		Test	*95.27*	*94.12*	*94.12*	*96.04*	*94.12*	*90.16*
M7	*PCA-NF*	*Training*	*97.25*	*99.32*	*93.15*	*99.59*	*96.35*	*94.25*
		Test	*97.63*	*100*	*93.10*	*100*	*96.43*	*94.80*

The classification and prediction distribution of the M6 model is illustrated in Figure 3, indicating a high degree of agreement between the predicted data and the actual data. The detailed classification of malignant and benign cells for the M6 model is presented in Table 3. The results indicate that the numbers of correctly predicted malignant samples in the training and test samples are 139 and 64, respectively, yielding a total of 203 correctly identified malignant cases. This corresponds to 95.75% of all malignant samples. Furthermore, only five malignant samples in the training set and four in the test set were misclassified as benign.

Meanwhile, the performance indicators of the optimized PCA-NF model (M7) are also presented in Table 4, while its classification and prediction distribution are shown in Figure 4. The predicted outputs exhibit strong consistency with the actual labels in the training data. The confusion matrix in Table 4 provides detailed classification results, indicating that only 10 malignant samples in the training set were incorrectly classified as benign. On the independent test set, the model achieved an accuracy of 97.63% and a recall of 93.10%, demonstrating its strong capability in detecting malignant cases. The comprehensive evaluation metrics-an F-Measure of 96.43% and MCCof 94.80%-further confirm the model's robust and balanced performance.

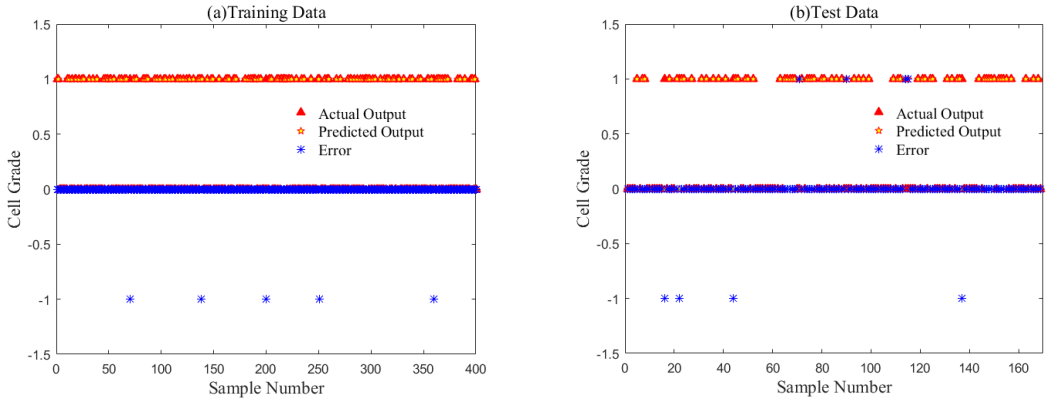

Fig. 3 *Classification and prediction distribution of M6 (PCA-ANFIS) model*

Table 3 *Classification of malignant and benign cells for M6 (PCA-ANFIS) model*

Actual / Predicted	Training Data		Test Data	
	Malignent	*Benign*	*Malignent*	*Benign*
Malignent	*139*	*0*	*64*	*4*
Benign	*5*	*256*	*4*	*97*

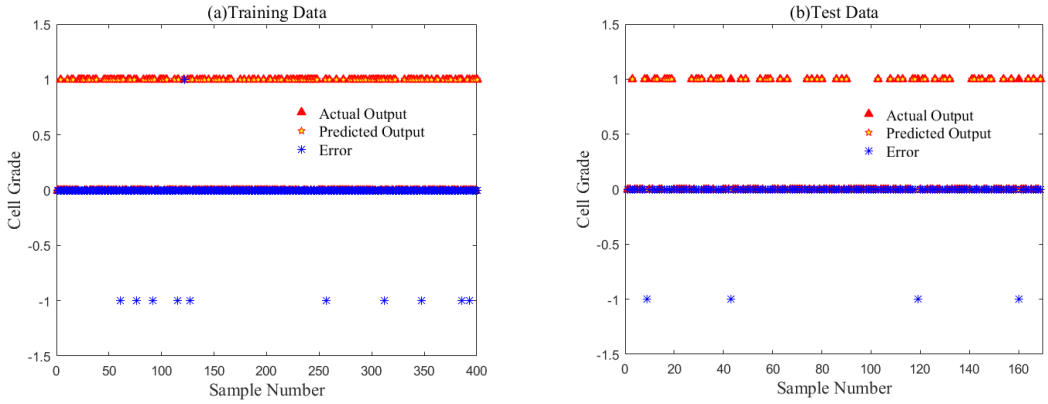

Fig. 4 *Classification and prediction distribution of M7 (PCA-NF) model*

Table 4 *Classification of malignant and benign cells for M7 (PCA-NF) model*

Actual / Predicted	Training Data		Test Data	
	Malignent	*Benign*	*Malignent*	*Benign*
Malignent	*145*	*1*	*54*	*0*
Benign	*10*	*244*	*4*	*111*

3.2.3 COMPARISON WITH EXISTING STUDIES

We contextualize our findings within the broader research landscape by comparing them with state-of-the-art methods published between 2016 to 2022, as summarized in Table 5. This comparison indicates that both our enhanced PCA-NF and PCA-ANFIS models achieve competitive performance relative to contemporary approaches. Specifically, M6 attains a test accuracy of 95.27%, eceeding several reported models in Table 5 (e.g., GAW+BP [24] at 95.00%, SVM Linear [27] at 95.00%), while approaching the performance of the more complex NN-EF model (M1) which reported at 99.41% in [4].

As discussed in Section 4, this level of performance is achieved alongside improved interpretability through the integration of fuzzy logic structures and reduced computational complexity enabled by PCA-based feature dimensionality reduction. Consequently, the proposed models are not only reliable for early breast cancer prediction but also more suitable for practical clinical deployment, where model transparency, reasoning interpretability, and operational efficiency are critical considerations. Overall, the reported metrics are comparable to, and in some cases surpass, those of the benchmark models, demonstrating the effectiveness and practical relevance of the proposed approach.

Table 5 *Comparison of performance indicators with other network models*

Year	*Model*	*Accuracy*
2016[22]	*PSO-KDE*	*98.11%, (2) 96.92%*
2016[22]	*GA-KDE*	*96.87%, (2) 96.19%*
2016[23]	*FW-PHHO-ELM*	*98.76%*
2018[24]	*GAW+BP*	*95.00%*
2018[24]	*IGSAGAW+BP*	*96.30%*
2018[24]	*GAW+3NN*	*92.00%*
2018[24]	*IGSAGAW+3NN*	*95.60%*
2018[24]	*GAW+CSSVM*	*95.20%*
2018[24]	*IGSAGAW+CSSVM*	*95.80%*
2019[25]	*FE-SSAE-SM*	*98.60%*
2019[25]	*SSAE-SM*	*98.25%*
2020[26]	*BCP-T1F*	*96.56%*
2020[26]	*BCP-SVM*	*97.06%*
2022[27]	*SVM RBF*	*96.00%*
2022[27]	*SVM Linear*	*95.00%*
2022[27]	*Random Forest*	*93.00%*
2022[27]	*Xgboost*	*98.00%*
2022[4]	*NN-ET*	*99.40%*
Now	*Our Model: PCA-NF*	*98.75%*
Now	*Our Model: PCA-ANFIS*	*97.25%*

4. CONCLUSION

In this paper, two enhanced hybrid models, namely PCA-NF and PCA-ANFIS, have been proposed to improve the classification of breast cancer cells using the WBCD dataset. By integrating principal component analysis for feature reduction with fuzzy neural architectures, the models achieve a balance between predictive performance and operational simplicity. A critical comparison of our proposed models with the benchmark methods listed in Table 5 reveals distinct trade-offs in terms of accuracy, interpretability, computational cost, and robustness.

Accuracy and Performance: Both PCA–NF (98.75% on the training set / 97.63% on the test set) and PCA–ANFIS (97.25% on the training set / 95.27% on the test set) achieve classification accuracy that is competitive with contemporary machine learning models. For instance, the NN-ET model [4] reports an accuracy of 99.40%, while ensemble methods such as Xgboost [27] achieve 98.00%. Although our models do not surpass the highest reported accuracy (e.g., 99.41% from NN-ET), they remain within the high-performance range (95–99%) typically reported among state-of-the-art methods, thereby confirming their effectiveness for clinical binary classification tasks.

Interpretability and Clinical Usability: This represents a key advantage of our fuzzy-based approaches. Unlike "black-box" models such as SVM-RBF, Random Forest, or deep learning models (e.g., FE-SSAE-SM [25]), both PCA–ANFIS and PCA–NF offer a degree of interpretability through fuzzy rule sets and membership functions. In particular, PCA–ANFIS provides explicit, human-readable rules, which can help clinicians understand the reasoning behind a classification decision. This contrasts with methods such as PSO-KDE [22] or GA-KDE [22], which may offer high accuracy but lack transparent decision logic. In clinical diagnostics, such interpretability can enhance trust and facilitate their integration into decision-support systems.

Computational Cost and Simplicity: The PCA step reduces the feature space from 30 to 6 dimensions, thereby significantly lowering computational complexity for subsequent modeling. Compared to methods that use all 30 features or employ evolutionary optimization (e.g., GAW+BP [24], PSO-KDE [22]), our models are more efficient in both training and inference. PCA–ANFIS, with its rule base reduced to 192 rules, is particularly efficient during the inference phase. In contrast, methods based on ensemble learning or multiple kernel learning may yield slightly higher accuracy but at the cost of greater computational overhead and longer training times, which may pose constraints in real-time or resource-limited clinical settings.

Generalizability and Robustness: The use of PCA not only reduces dimensionality but also decorrelates features, which can improve model generalizability to new data. Our models demonstrated stable performance on the test set, with F-Measure and MCC values exceeding 94%, indicating robustness to potential class imbalance. However, like many models evaluated on the WBCD dataset, their generalizability to external, multi-center clinical data remains to be validated. Methods such as BCP-T1F [26] and BCP-SVM [26] were explicitly designed with robustness in mind; however, our approach, by relying on PCA for noise reduction, also inherently mitigates the impact of feature redundancy and minor data variations.

Limitations and Future Work: The main limitation of our approach lies in its dependency on the quality of PCA transformation; if the retained variance (85%) omits clinically significant but low-variance features, diagnostic sensitivity could be affected. Moreover, while interpretable, the fuzzy rule base in PCA–ANFIS requires expert knowledge to define initial membership functions, which may not be trivial in all clinical contexts. Future work will focus on: (1) validating the models on larger, multi-institutional datasets to assess external generalizability;

(2) exploring automated membership function tuning using metaheuristic algorithms; and (3) extending the framework to multi-class classification tasks involving different cancer subtypes.

In summary, this research contributes a practical, interpretable, and computationally efficient framework for breast cancer cell classification. While it does not surpass the absolute highest accuracy reported in the literature, the PCA–NF and PCA–ANFIS models offer a favorable balance between performance, transparency, and ease of use—attributes that are essential for clinical decision-support tools. The proposed methodology provides a valuable alternative to purely accuracy-driven "black-box" models, emphasizing the importance of interpretability and operational efficiency in medical AI applications.

Acknowledgments: This work was supported by Special Project of Educational Science Planning of Hubei Province (2025ZX089) "Exploration and Practice Research on the STEM+X Interdisciplinary Teaching System Oriented towards Intelligent Manufacturing in the Context of the 'Four New' 2.0", the Higher Education Teaching Reform Research Project of Hubei Province (2024447) and the Hubei Province's New Engineering Discipline Construction Project (XGK03098).

Competing interest: The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

5. REFERENCES

[1] Siegel, R.L., Miller, K.D.H., Fuchs, E., and Jemal, A., Cancer Statistics, *Cancer Journal for Clinicians*, Vol. 72, No. 1, pp. 7-33, 2022. http://dx.doi.org/10.3322/caac.21708

[2] Depciuch, J., Kaznowska, E., Szmuc, K., Zawlik, I., Cholewa, M., Heraud P., and Cebulski, J., Comparing Paraffined and Deparaffinized Breast Cancer Tissue Samples and An Analysis of Raman Spectroscopy and Infrared Methods, *Infrared Physics and Technology*, Vol. 76, pp. 217-226, 2016. http://dx.doi.org/10.1016/j.infrared.2016.02.006

[3] Conti, A., Duggento, A., Indovina, I., Guerrisi, M., and Toschi, N., Radiomics in Breast Cancer Classification and Prediction, *Seminars in Cancer Biology*, Vol. 72, pp. 238-250, 2021. https://doi.org/10.1016/j.semcancer.2020.04.002

[4] Sharma, D., Kumar, R., and Jain, A., Breast Cancer Prediction Based on Neural Networks and Extra Tree Classifier Using Feature Ensemble Learning, *Measurement Sensors*, Vol. 24, pp. 100560, 2022. https://doi.org/10.1016/j.measen.2022.100560

[5] Naji, M.A., Filali, S.E., Aarika, K., Benlahmar, E.H., Abdelouhahid, R.A., and Debauche, O., Machine Learning Algorithms for Breast Cancer Prediction and Diagnosis, *Procedia Computer Science*, Vol. 191, pp. 487-492, 2021.

https://doi.org/10.1016/j.procs.2021.07.062

[6] M. Desai M., and Shah, M., An Anatomization on Breast Cancer Detection and Diagnosis Employing Multi-layer Perceptron Neural Network (MLP) and Convolutional Neural Network (CNN), *Clinical eHealth*, Vol. 4, pp. 1-11, 2021.

https://doi.org/10.1016/j.ceh.2020.11.002

[7] Rane, N., Sunny, J., Kanade R., and Devi, S., Breast Cancer Classification and Prediction Using Machine Learning, *International Journal of Engineering Research and Technology*, Vol. 9, No. 2, pp. 576-580, 2020. https://doi.org/10.17577/IJERTV9IS020280

[8] Al-Massri, R., Al-Astel, Y., Ziadia, H., Mousa, D.K., and Abu-Naser, S.S., Classification Prediction of SBRCTs Cancers Using Artificial Neural Network, *International Journal of Academic Engineering Research*, Vol. 2, No. 11, pp. 1-7, 2018.

http://ijeais.org/wp-content/uploads/2018/11/IJAER181101.pdf

[9] Chen, Y.R., Feng, J., Liu, J., Pang, B.C., Cao D.F., and Li, C., Detection and Classification of Lung Cancer Cells Using Swin Transformer, *Journal of Cancer Therapy*, Vol. 13, No. 7, pp. 464-475, 2022. https://doi.org/10.4236/jct.2022.137041

[10] Kour, H., Manhas, J., and 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, *Artificial Intelligence Review*, Vol. 53, pp. 4651-4706, 2020.

https://doi.org/10.1007/s10462-020-09804-x

[11] Vieira, J., Morgado-Dias, F., Mota, A., Neuro-Fuzzy Systems: A Survey. *WSEAS Transactions on System*, Vol. 3, No. 2, pp. 414-419, 2004.

https://www.researchgate.net/publication/242073375

[12] Übeyli, E.D., Adaptive Neuro-Fuzzy Inference Systems for Automatic Detection of Breast Cancer, *Journal of Medical Systems*, Vol. 33, No. 5, pp. 353-358, 2009a.

https://doi.org/10.1007/s10916-008-9197-x

[13] Jang, J-S.R., ANFIS: Adaptive-Network-Based Fuzzy Inference System, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 23, No. 3, pp. 665-685, 1993.

https://doi.org/10.1109/21.256541

[14] Ashraf, M., Le, K., and Huang, X., Information Gain and Adaptive Neuro-Fuzzy Inference System for Breast Cancer Diagnoses, in *Proceedings of the IEEE 5th International Conference on Computer Sciences and Convergence Information Technology*, *ICCIT 2010,* Seoul, Korea (South), pp. 911-915, Feb. 2010.

https://doi.org/10.1109/ICCIT.2010.5711189

[15] Senol C., and Yildirim, T., Thyroid and Breast Cancer Disease Diagnosis Using Fuzzy-Neural Networks, in *Proceedings of the International Conference on Electrical and Electronics Engineering-ELECO, ELECO 2009*, Bursa, Turkey, pp. 390-393, Dec. 2009.

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5355297

[16] Wolberg, W.H., Nice Street, W., and Mangasarian, O.L., Breast Cancer Wisconsin (Diagnostic) (1995). UCI Machine Learning Repository.

https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

[17] El Hamdi, R., Njah, M., and Chtourou, M., An Evolutionary Neuro-Fuzzy Approach to Breast Cancer Diagnosis, in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, *ICSMC 2010*, Istanbul, Turkey, pp. 142-146, Nov. 2010.

https://doi.org/10.1109/ICSMC.2010.5642219

[18] Zarbakhsh, P., Addeh, A., and Demirel, H., Early Detection of Breast Cancer Using Optimized ANFIS and Features Selection, in *2017 9th International Conference on Computational Intelligence and Communication Networks, CICN 2017*, Girne, Northern Cyprus, pp. 39-42, Mar. 2017. https://doi.org/10.1109/CICN.2017.8319352

[19] Hamdan, H., and Garibaldi, J.M., Automatic Generation of ANFIS Rules in Modelling Breast Cancer Survival, in *2014 International Conference on Computer Assisted System in Health, CASH 2014*, Kuala Lumpur, Malaysia, pp. 12-17, Oct. 2014.

https://doi.org/10.1109/CASH.2014.16

[20] Thani, I., and Kasbe, T., Expert system based on fuzzy rules for diagnosing breast cancer. *Health and Technology*, Vol. 12, No. 2, pp. 473-489, 2022.

https://doi.org/10.1007/s12553-022-00643-0

[21] Dubey, M., and Kumar, S., A model for the diagnosis and prognosis of breast cancer based on fuzzy expert system. *In Mathematical sciences and applications*, pp. 30-36, 2024. CRC Press. https://doi.org/10.1201/9781003451808

[22] Razieh, S., Mehdi, A.S., and Robab, S., Particle Swarm Optimization for Bandwidth Determination and Feature Selection of Kernel Density Estimation Based Classifiers in the Diagnosis of Breast Cancer, *Applied Soft Computing*, Vol. 40, pp. 113-131, 2016.

https://doi.org/10.1016/j.asoc.2015.10.005

[23] Jiang, F., Zhu, Q.N., and Tian, T.H., Breast Cancer Detection Based on Modified Harris Hawks Optimization and Extreme Learning Machine Embedded with Feature Weighting, *Neural Processing Letters*, Vol. 55, pp. 3631-3654, 2023.

https://doi.org/10.1007/s11063-021-10700-w

[24] Liu, N., Qi, E.-S., Xu, M., Gao, B., and Liu, G.-Q., A novel Intelligent Classification Model for Breast Cancer Diagnosis, *Information Processing and Management*, Vol. 56, No. 3, pp. 609-623, 2019. https://doi.org/10.1016/j.ipm.2018.10.014

[25] Kadam, V.J., Jadhav, S.M., and Vijayakumar, K., Breast Cancer Diagnosis Using Feature Ensemble Learning Based on Stacked Sparse Autoencoders and Softmax Regression, *Journal of Medical. Systems*, Vol. 43, No. 8, pp. 1-11, 2019.

https://doi.org/10.1007/s10916-019-1397-z

[26] Khan, F., Khan, M.A., Abbas, S., Athar, A., Siddiqui, S.Y., Khan, A.H., Saeed, M.A., Hussain, H., and Iriguchi, N., Cloud-based Breast Cancer Prediction Empowered with Soft Computing Approaches, *Journal of Healthcare Engineering*, Vol. 2020, No. 1, pp. 1-16, 2020.

https://doi.org/10.1155/2020/8017496

[27] Al-Azzam, N., and Ibrahem, S., Comparing Supervised and Semi-supervised Machine Learning Models on Diagnosing Breast Cancer, *Annals of Medicine and Surgery*, Vol. 62, pp. 53-64, 2021. https://doi.org/10.1016/j.amsu.2020.12.043