International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 16s,2025
https://theaspd.com/index.php

Optimized Data Mining Approach For Breast Cancer Prediction Using Fuzzy Theory And Machine Learning

Parul Bhatnagar[1], Dr. Bhupendra Kumar[2]
[1]Research Scholar, School of Computer Science & Applications, IIMT University Meerut, Uttar Pradesh, India, Parul0710@gmail.com, 0009-0007-4431-4270
[2]Professor, School of Computer Science & Applications, IIMT University Meerut, Uttar Pradesh, India, Singhbhupender231@gmail.com, 0000-0001-9281-3655

Abstract
Breast Cancer (BC) continues to be a major health challenge for women around the globe due to factors such as delayed diagnosis and difficulties distinguishing between benign and malignant tumors. Mammography lacks accuracy and often requires interpretatively invasive procedures that can lead to complications. The current research aims to develop a hybrid BC prediction model based on fuzzy logic and ensemble Machine Learning (ML) techniques, utilizing the WDBC dataset comprising 569 samples. Treated under the ensemble approach, the proposed methodology consists of parallel processes of fuzzy rule-based risk evaluation and ML classification to compute a fitness score for definitive selection. The model's interpretability and accuracy were improved through fuzzy rule generation alongside feature selection using PCA. The experimental findings showed that the proposed Fuzzy + Ensemble model surpassed the individual classifiers' performance, attaining a precision of 93%, recall of 91%, accuracy of 99% and 92% in the F1-score. These values indicate a substantial improvement over conventional methods, such as Decision Tree (86%) and Naïve Bayes (84%). The confusion matrix further confirmed a low false positive and false negative rate. The research signifies a notable advancement in diagnostic decision-making by combining interpretability with predictive power, making it a viable framework for real-time, intelligent BC prognosis systems in clinical environments.
Keywords: *Breast Cancer Prediction, Fuzzy Logic, Ensemble Learning, Machine Learning Classification, Medical Data Mining.*

1. INTRODUCTION

Cancer is a life-threatening ailment primarily brought about by environmental influences and factors that cause mutations in genes responsible for critical cell regulatory proteins [1]. These changes can interfere with the cells' normal processes, resulting in tumor formation due to uncontrolled cell growth. From any area of the human body, cancer can arise and metastasize, which is why its early detection is needed to enhance the chances of survival [2, 3]. Finding cancer in the early stages, especially before it becomes malignant, improves management procedures significantly and halts progression to other organs [4]. BC is high among cancers and ranks second in women for cancer-related deaths, behind only lung cancer [5,6]. It is a true global health challenge, affecting millions of women every year. Despite medical breakthroughs in early diagnosis and treatment, it continues to remain an enormous burden in the health care system [7]. The one of the BC can be marked by different symptoms, which start subtly before becoming more pronounced as time goes on [8]. Some of the most general symptoms are a lump within the breast, unusual pain in the breast or armpit not associated with the menstrual cycle and a change in the skin color to red or pitting that resembles an orange peel [9,10]. Early attendance to the noticed symptoms has the potential to receive non-invasive treatment and timely medical intervention, which hinders the progression of the ailment [11]. In general, BC can be classified into two major groups: benign and malignant tumors [12]. Benign tumors refer to growths that are non-cancerous and do not propagate to other regions of the body which greatly reduces the health risk [13]. Unlike benign tumors, cancerous tumors are severely risky because they can infiltrate the tissues around them and spread to other divisions of the body. Cancerous tumors have the potential to be successfully treated if identified at an early stage and medicated without delay [14,15]. The traditional diagnosis of BC involved invasive procedures and utilized breast biopsy as one of the most prevalent methods [16]. Technological innovations better diagnostic methodologies that are non-invasive and predictive, increasing their accuracy while decreasing stress on the patient [17]. The application of ML in disease diagnosis has revolutionized the field by enabling the construction of predictive models for the assessment of risk factors that make BC diagnosis more effective [18][19].

Among the ML algorithms that were found effective in prediction and breast cancer classification are Support Vector Machines (SVM), Decision Trees (DT), Artificial Neural Networks (ANN) and K-nearest neighbor (K-NN) [20]. These sophisticated techniques use a data-driven approach to predict BC by identifying complex patterns and correlations among various clinical parameters [21]. The use of ML and data mining, along with medical informatics, has remarkably increased clinical decision-making quality and speed of decision-making. The use of computational techniques, data analysis, and predictive modeling has enabled clinicians to process medical data effectively and make appropriate decisions on patient care [22]. The use of fuzzy theory with ML has also enhanced the BC predictive models, minimized diagnostic complications, and enhanced the possibility of early diagnosis [23]. Despite the advancement of predictive models, BC diagnosis is still hampered by numerous challenges. Medical data are typically described in terms of uncertainty, ambiguity, and factors that cannot be accurately classified and described by the majority of classifier models with correct diagnosis [24-26]. The consequences of high false positive and false negative rates can lead to unnecessarily complicated treatments being done or missed detections, which can have a catastrophic impact on the health of the patient [27,28]. Moreover, most of the medical datasets are imbalanced as the malignant cases are significantly fewer in quantity than the benign ones and this leads to biased models developed. Interpretability of the AI-based diagnosis systems is also a problem for medical experts to understand the output. Apart from this, fuzzy logic and data mining together pose a humongous algorithmic burden that requires optimized algorithm design. A solution that combines the strengths of fuzzy logic and data mining's prediction ability is required to provide the flexibility and interpretability that such problems entail. The motivation behind the application of fuzzy logic and data mining to predict BC is to enhance reliability and enable interpretability by enabling machines to reason and simulate uncertainty in data logic. In contrast to conventional models that demand numerical measures of accuracy, fuzzy logic supports classification in terms of linguistic variables, enhancing the results' understandability to health practitioners. This research aims to combine fuzzy logic and data mining to enhance classification accuracy using enhanced fuzzy membership functions based on tumor characteristics and validate the model using the Wisconsin Diagnostic Breast Cancer (WDBC) database. It also aims to improve intelligibility by developing justifiable fuzzy rule explanations and examining the clinical relevance of advanced automated diagnostic systems designed to optimize the patient's health and diagnostic outcomes. The key objectives of this research are:

- To develop a hybrid BC prediction model that integrates fuzzy logic with ML to enhance classification accuracy and interpretability.
- To design a parallel processing framework that utilizes a fuzzy rule-based system for risk assessment alongside ML classifiers for improved decision-making.
- To apply ensemble learning techniques to combine multiple model predictions and determine an optimal fitness score for reliable BC recurrence prediction.
- To compare the proposed hybrid approach with traditional diagnostic methods to determine its efficiency in early BC detection and recurrence prediction.

The paper is structured as follows: in section 1 introduction of the topic is provided, and then in section 2, the literature review examines existing research on BC prediction, highlighting strengths and limitations. In section 3, the methodology outlines the hybrid model, including data preprocessing, feature selection, fuzzy rule generation, and classification.

2. LITERATURE REVIEW

In this section, existing research on BC prediction is examined and summarized which are done by several authors by highlighting further research gaps. Recently, in 2025, Nassih and Berrado [29] presented a novel PRIM-based BC classification system that made use of XGBoost, Logistic Regression (LR), and Random Forest (RF). On the Wisconsin dataset, LR achieved 94.1% accuracy vs 96.8% and 85.4% recall versus 94.2% for the PRIM-based architecture. Previously, in 2024, Ashika et al. [30] developed a new hybrid technique (Neutrosophic Set Theory (NST) and ML) that enhanced BC detection by integrating NST and ML techniques. To emphasize the effectiveness of NS in improving diagnostic reliability for the WDBC dataset, N-AdaBoost models achieved remarkable outcomes with 99.12% accuracy and 100% precision. Similarly, to perform the same task in 2024, Mohammed and Muhammad [31] designed a

neural network–fuzzy logic AI hybrid intelligent expert system to aid in BC detection by offering a potential method for the problem of ambiguity. An accuracy of 96.77% in correctly diagnosing patients with mild, moderate, or severe BC was shown by the system's performance. Earlier, in 2023, Nemade and Fegade [32] used the WDBC dataset to test six ML classification methods (DT, KNN, SVM, RF, Naïve Bayesian (NB), and LR) with ensemble approaches. After comparing all of the methods, it was determined that the DT classifier using the Gini index had the best accuracy (97%), while the LR classifier had the maximum area under the curve (0.996) and the best accuracy 97% for XGBoost, among the ensemble approaches. In the same year (2023), Atban et al. [33] suggested a hybrid model for histopathology picture binary classification that combines ML techniques with metaheuristic algorithms and pre-trained transfer learning methods. Results showed that the suggested method produced a 97.75% F-score for features derived from ResNet18-EO. Moreover, in 2023, Gupta et al. [34] improved the fuzzy expert system's rule-based derivation using DT analysis to aid in the identification of breast cancer. Using WBCD, the Mamdani Fuzzy Rule-Based system was used to turn the final fuzzy score into class labels beneficial. Results revealed that with a recall of 99.58% and a precision of 93%, the suggested approach achieves an accuracy of around 97%. Previously, in the same field of research, Nasir et al. (2022) [35] suggested a strategy for fine-tuning the neural network to use AlexNet for training BC photos by extracting characteristics from the images. To identify healthy and cancerous breast tissue, researchers modified AlexNet's first and final three layers in the suggested model. The suggested model attained a greater accuracy of 98.44% during training and 98.1% during testing, respectively. In the same year, Idris et al. (2022) [36] suggested FID3-AF, a classification method for detecting breast cancer, which used the fuzzy-ID3 algorithm with an association function implemented. When it came to BC categorization, the data showed that FID3-AF held its own. When evaluated on the WDBC dataset, the FID3-AF attained a 94.517% accuracy rate. Similarly, in 2022, Altameem et al. [37] used a variety of deep convolutional neural network (CNN) models as foundational classifiers across four mammography imaging datasets, including 1145 equal numbers of benign, normal, and cancerous images. The suggested method made use of an ensemble strategy that ranked the underlying classification methods using fuzzy logic based on the Gompertz function. The proposed Inception V4 ensemble model's accuracy was 99.32% when using the fuzzy rank-based Gompertz function. Meanwhile, in 2022, Chidambaram et al. [38] suggested a new method for improving the accuracy of input data categorization using a hybrid neuro-fuzzy classifier (HNFC). Before using the neural network to evaluate performance, fuzzy logic was used on the input dataset. In comparison to the other two methods, the suggested method achieved a classification accuracy of 86.2% on the BC dataset. In the past (2021), Chinniyan and Subramani [39] researched a new method for quantitatively inferring BC tissues utilizing ANN and NB classifiers. Using k-fold cross-validation, the suggested approach was assessed. The Wisconsin Original BC dataset was utilized from the UCI Repository to assess the effectiveness of the proposed model. The findings showed that the suggested method achieved the highest accuracy rates for ANNs (98.0%) and NB Classifiers (95.2%).

Despite significant advancements in BC prediction, several research gaps remain. First, most studies [29, 30] focus on WDBC, lacking evaluation across diverse datasets for generalizability. Second, while hybrid AI systems [31, 33] show promise, integration with real-time clinical data and multi-modal imaging remains unexplored. Lastly, deep learning models [35, 37] achieve high accuracy but lack interpretability, necessitating research into explainable AI for clinical trust.

3. RESEARCH METHODOLOGY

The proposed methodology integrates fuzzy logic and ML to predict malignant body part recurrence. It utilizes the WDBC dataset, where recurring patients are transferred to a separate dataset for better tracking. Preprocessing ensures data quality through normalization, cleaning, and handling missing values. Feature extraction and selection refine the dataset, which is then processed in two parallel paths: a fuzzy rule-based system analyzing age and risk factors and ML models trained on the data. A hybrid decision-making approach combines both methods and ensemble learning to optimize performance by selecting the best classifier combination. The final fitness score determines the likelihood of recurrence, enhancing predictive accuracy. The proposed methodology involves preprocessing the dataset to prepare the data for analysis, as shown in Figure 1.

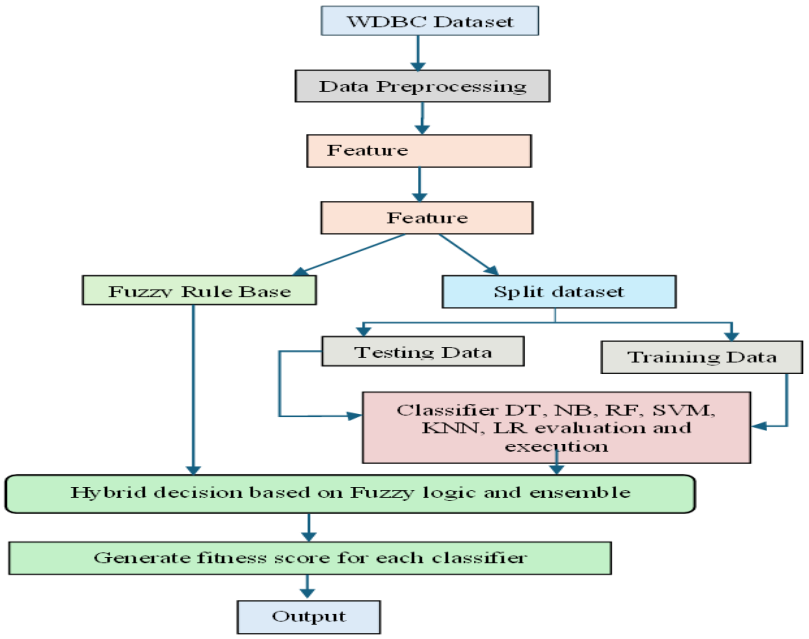


Figure 1: Flowchart of the proposed methodology

3.1 Dataset description

The WDBC Dataset [40] is used in this research and comprises a total of 569 patient cases defined by 30 numeric attributes. These attributes are obtained from FNA biopsy images and have valuable information about tumor classification as Malignant (M) or Benign (B) classes. Out of the 569 instances, 357 are benign, and 212 are malignant. The major tumor characteristics like texture, smoothness, compactness, concavity, radius and symmetry are presented, and this renders the dataset vital in tumor behavior understanding as well as enhancing diagnostic accuracy. Results from this dataset are vital in the diagnosis of malignant and benign tumors, therefore enabling better clinical decision-making.

3.2 Feature extraction and selection

Feature extraction and selection are critical steps to advance classification performance and reduce computational complexity after performing data pre-processing. Feature extraction from the WDBC dataset focuses on morphological characteristics obtained from FNA biopsy images, such as texture, symmetry, and fractal dimensions. Dimensionality reduction technique known as Principal Component Analysis (PCA) is applied to retain only the most informative attributes using the formula below:

$$I(X;Y) = \sum_i p(x_i)\log\frac{p(x_i)}{p(x_i|y)} \quad (1)$$

Where $p(x_i)$ is the probability of a feature (x_i) occurring and $p(x_i|y)$ is its conditional probability given the class (y). This step ensures that redundant or irrelevant features do not negatively impact model performance, leading to faster and more precise BC classification.

3.3 Fuzzy rule base

The fuzzy rule base serves as the core decision-making framework in our methodology, addressing uncertainties and imprecise patterns in medical data. Given the complexity and variability of cancer diagnosis, crisp classification models often struggle with borderline cases. Fuzzy logic provides a robust alternative by incorporating linguistic variables and membership functions to classify patients into different risk categories. For instance, hormone receptor status, tumor size, and lymph node involvement are transformed into fuzzy sets (e.g., "small," "medium," or "large" tumor size). Expert-defined fuzzy rules, combined with ML-generated patterns, facilitate an interpretable and flexible classification approach. The integration of a fuzzy inference system (FIS) ensures that the system adapts dynamically to new patient data, improving prediction accuracy while maintaining clinical interpretability [41].

- **Fuzzy membership function**

Let $\mu_A(x)$ be the membership function for an input x. Membership function for tumor size x [42]:

$$\mu_A(x) = \begin{cases} 0 & x \le a \\ \frac{x-a}{x-b}, & a < b < x \\ 1, & x \le b \end{cases} \quad (2)$$

Where a and b are defined as threshold values for classifying tumor sizes as small, medium, or large.

- **Fuzzy rule generation**

Fuzzy logic is applied to model uncertainty in medical diagnosis. A FIS is developed with input variables (e.g., tumor size hormone receptor status) and an output variable (cancer classification). The output is derived using a Mamdani inference system with the centroid defuzzification method:

$$y = \frac{\sum_{i=1}^{n}(\mu_i(x).w_i)}{\sum_{i=1}^{n}(\mu_i(x))} \quad (3)$$

Where $\mu_i(x)$ is the degree of membership for rule (i), and w_i is the weight assigned to that rule. Let $x_1, x_2, \ldots, x_n$ be the input features, and y be the output cancer classification (Benign, Malignant). The fuzzy rules are defined using **IF-THEN** statements.

3.4 ML-based classification

ML classifiers play a pivotal role in predicting BC outcomes and tumor malignancy. A comparative analysis of different classifiers DT, NB, RF, SVM, KNN, and LR is conducted to determine the most effective model using the following functions:

- **DT:** Decision boundaries are created based on recursive splitting, represented as [43]:

$$H(X) = -\sum p(x) \log_2 p(x) \quad (4)$$

where $H(X)$ is the entropy function determining information gain for splitting at attribute X.

NB Classifier: A probabilistic model based on Bayes' theorem:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \quad (5)$$

This classifier is well-suited for the Dataset, where tumor characteristics are categorical.

- **NB:** It takes input information and determines which class has the greatest posterior probability [44]. The classification rule is given by:

$$P(C_k \mid X) = \frac{P(C_k)\prod_{i=1}^{n}P(x_i|C_k)}{P(X)} \quad (6)$$

Where x_i represents individual features in the feature set X.

- **RF:** An ensemble learning technique that generates multiple DTs and aggregates results [45]. The prediction is given by:

$$\hat{y} = \text{mode}(h_1(x), h_2(x), \ldots, h_n(x)) \quad (7)$$

where $h_i(x)$ is the decision of individual trees in the ensemble.

- **SVM:** This model finds the hyperplane that maximizes the margin between malignant and benign classes, formulated as [46]:

$$\max_{w,b}\left(\frac{1}{\|w\|}\right), \quad \text{subject to } y_i(w.x_i + b) \ge 1 \quad (8)$$

where w is the weight vector, x_i represents input data points and y_i is the corresponding class label.

- **KNN:** A majority of the label among the neighboring samples is used by this non-parametric approach to classify a new sample [47]. The distance metric is typically the Euclidean distance:

$$d(x,y) = \sqrt{\sum(x_i - y_i)^2} \quad (9)$$

Each classifier is evaluated based on precision, recall, F1-score, and AUC-ROC values to assess performance. By combining these models with the fuzzy rule-based system, the hybrid decision framework ensures robust predictions with reduced bias and variance, leading to more reliable patient classification.

- **LR:** It is a statistical model used for binary classification, predicting the probability of a sample belonging to a class using the sigmoid function [48]:

$$P(Y = 1 \mid X) = \frac{1}{1+e^{-z}} \quad (10)$$

Where $z = \beta_0 + \sum_{i=1}^{n}\beta_i x_i$, and $P(Y = 1 \mid X)$ is the probability of the positive class (e.g., malignant tumor), β_0 is the intercept, and β_i represents the coefficients (weights) for the feature x_i.

3.5 Proposed algorithm

Step 1: Dataset Management

Define:

- D (WDBC Dataset)

For each patient record $P_i \in D$:

If P_i Matches a record (based on patient ID or medical history similarity).

Step 2: Data Preprocessing

Data Cleaning: Remove inconsistencies and fill in missing values using imputation:

$$(x'_{ij}) = \begin{cases} x_{ij}, & \text{if } x_{ij} \text{ is available} \\ \mu_j, & \text{if } x_{ij} \text{ is missing} \end{cases}$$

where x'_{ij} is the cleaned value, and μ_j is the meaning of feature j.

Normalization (Min-Max Scaling):

$$x^*_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_{ij})}$$

wherex^*_{ij} is the normalized feature value.

Step 3: Feature Extraction

Define the feature transformation function $\phi(x)$:

$$X' = \phi(X)$$

Where X' is the extracted feature set.

Step 4: Feature Selection

Apply feature selection methods such as mutual information, PCA, or correlation-based filtering:

$$F' = \{f_i \mid \text{Relevance}(fi) > \theta\}$$

Where F′ is the selected feature set, and θ is the selection threshold.

Step 5: Parallel Processing Paths

Fuzzy Rule Base Path:

- Define fuzzy sets $A_1, A_2, \ldots, A_n$for risk factors (e.g., Age, Medical History).
- Define fuzzy membership functions $\mu_{A_i}(x)$ using a Gaussian function.
- Compute fuzzy risk score R:

$$R = \sum_{i=1}^{n} w_i . \mu_{A_i}(x)$$

where w_i is the weight for each risk factor.

ML Path:

- Split dataset into training (80%) and testing (20%):

$$X_{train}, X_{test}, y_{train}, y_{test} = \text{Split}(X, y, 0.8)$$

Step 6: Model Training

Train multiple classifiers M on training data:

$M_i = \text{Train}(X_{train}, y_{train}, \theta)$,

Where θ represents model hyperparameters.

Step 7: Hybrid Decision-Making

Combine fuzzy risk score R and model predictions $M_i(X)$:

$$P = \alpha R + \sum_{i=1}^{k} \beta_i M_i(X)$$

where α and β_i are weighting factors for fuzzy logic and ML classifiers.

Step 8: Ensemble Learning & Fitness Score Determination

- Compute weighted ensemble prediction:

$$P_{ensemble} = \sum_{i=1}^{k} w_i M_i(X)$$

- Compute fitness score:

$$S_{fitness} = \frac{1}{k} \sum_{i=1}^{k} \text{Accuarcy}(M_i)$$

where $S_{fitness}$ evaluates the ensemble's overall performance.

Step 9: Final Prediction

- Compute final recurrence probability:

$$P_{final} = \lambda P_{ensemble} + (1 - \lambda)R$$

where λ is a balancing factor.

- Classify as malignant if $P_{final} > \delta$, where δ is a decision threshold.

4. RESULTS AND DISCUSSION

In this section, the results of the research are presented in detail with appropriate tables, graphs and metrices.

4.1 Evaluation Metrices

Evaluation matrices help to compute the efficiency of the proposed framework. Evaluation matrices give a simpler means of experimental results interpretation and standardizing how different methods or approaches can be compared.

- **Accuracy:** Accuracy refers to the ratio of predicted correct results, both TP as well as TN, to the number of total predictions made. The formula below is used to compute accuracy:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

- **Precision:** It measures the ratio of correct positive predictions to all anticipated positive outcomes, that is, to both TP aa well as FP.

$$\text{Precision(P)} = \frac{TP}{TP+FP} \quad (12)$$

- **Recall:** It can be referred to as sensitivity that measures the ratio of TP correctly anticipated to the sum of actual positive cases, i.e. True Positive (TP) as well as False Negative (FN) cases.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (13)$$

- **F1-score:** This measure calculates the harmonic average of the recall and accuracy values. It can be calculated by using:

$$F1 - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \quad (14)$$

4.2 Result Analysis

The results obtained from the research or experiment must be systematically analyzed, interpreted, and evaluated. The following is a detailed description of the approach's systematic outcome analysis:

- **Dataset Distribution**

Figure 2 shows the connections between the target parameter diagnosis (malignant, blue, and benign, orange, and smoothness, respectively) and three BC data characteristics (radius, texture, and smoothness). According to the diagonal subplots, which display the feature distributions using kernel density estimations, benign tumors often have lower radius and texture means than malignant ones. Malignant cases cluster towards higher values in both radius mean and texture mean, whereas benign instances are more concentrated around lower values; the scatter plots in the off-diagonal locations show pairwise connections between the characteristics. Since smoothness does not significantly differentiate between the two conditions, it may not be very useful as a diagnostic tool in and of itself. The graph shows that when it comes to this dataset, the means of radius and texture are better at differentiating between benign and malignant tumors.

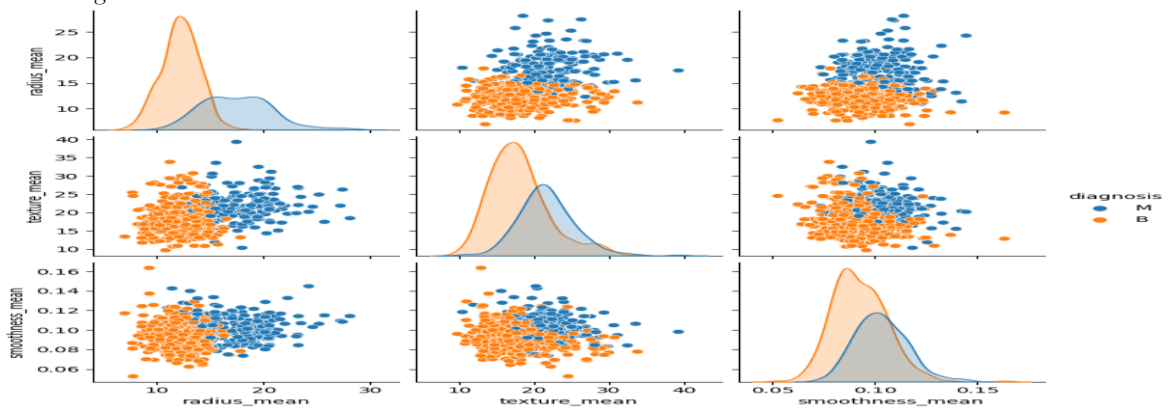
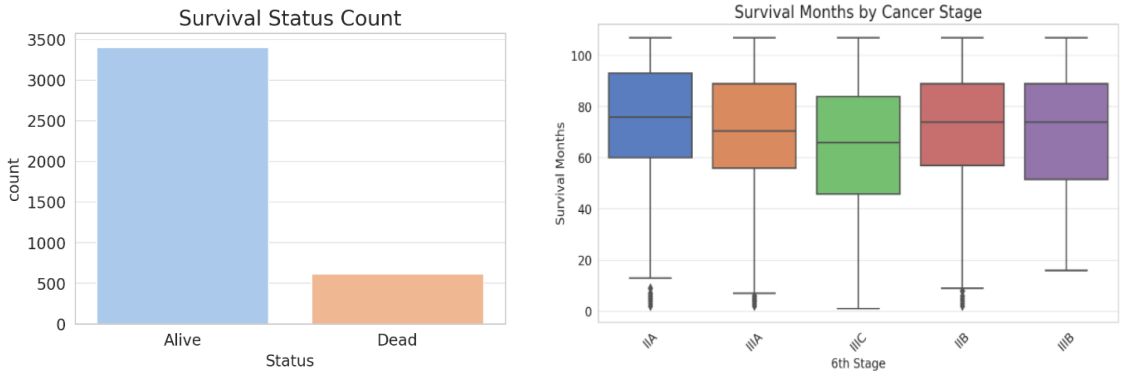


Figure 2: Data Distribution Graph

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 16s,2025
https://theaspd.com/index.php

- **Data Pre-processing Representation**

Figure 3 illustrates key survival data essential for preprocessing. In Figure 4 (a), the survival status is clearly imbalanced, with over 3,300 patients alive and around 600 deceased, highlighting the need for techniques to address class imbalance and prevent model bias. Figure 4 (b) shows survival months by cancer stage, where median survival ranges from approximately 65 to 80 months. The presence of outliers and variability across stages suggests the importance of handling skewness and extreme values to ensure robust model training.



(a) (b)

Figure 3: Survival Data Representation (a) Status Count (b) Months by Cancer Steps

Figure 4 displays the mean values of significant numerical features, facilitating comprehension of data distribution before preprocessing. Scale differences between features, for example, survival months being considerably higher, underscore the importance of normalization or standardization. The process provides fair model training by avoiding feature dominance by higher-magnitude features.

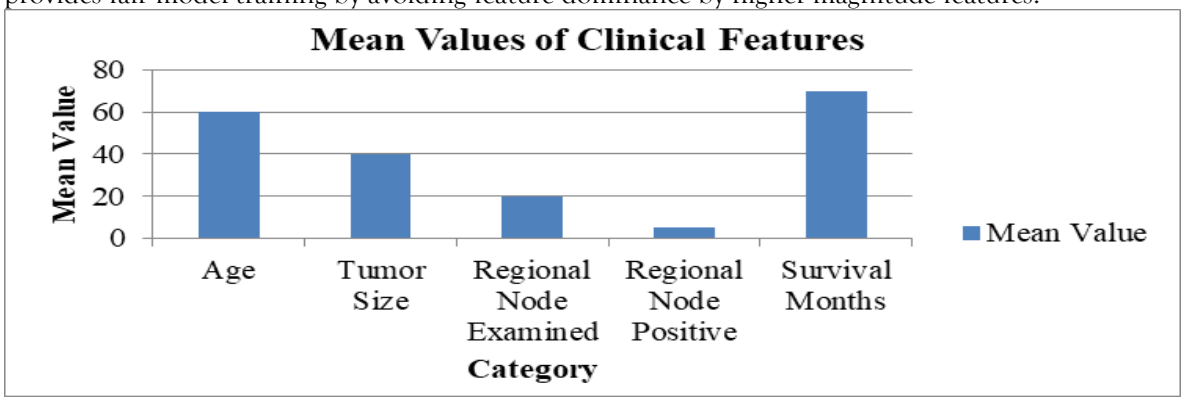


Figure 4: Mean of Numerical Features

Figure 5 illustrates the standard deviation of numerical features, representing data dispersion. Elevated variability in survival months and tumor size reflects a wide range of values, where normalization is necessary to provide model stability. Feature standardization using features with different scales avoids ML model bias, leading to improved predictive precision and interpretability.

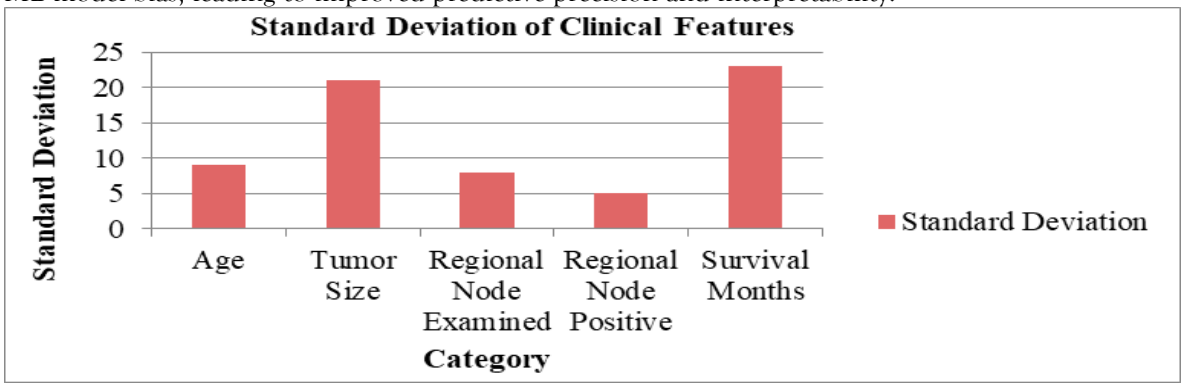


Figure 5: Standard Deviation of Numerical Features

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 16s,2025
https://theaspd.com/index.php

- **Feature Extraction**

Feature extraction is an important part of dimensionality reduction because it makes it easier to turn high-dimensional data into a more useful and compact picture. The correlation heatmap shows the relationships between features in a breast cancer dataset, with values ranging from minus one to plus one, as shown in Figure 6. Strong positive correlations are seen among the means of radius, perimeter, and area, indicating they are closely related. Diagnosis is positively correlated with features like radius mean, and concave points mean, highlighting their significance in predicting malignancy. The means of features like compactness, concavity, and concave points are highly correlated with one another. This visualization helps identify key features and potential redundancies for effective feature extraction in modeling.

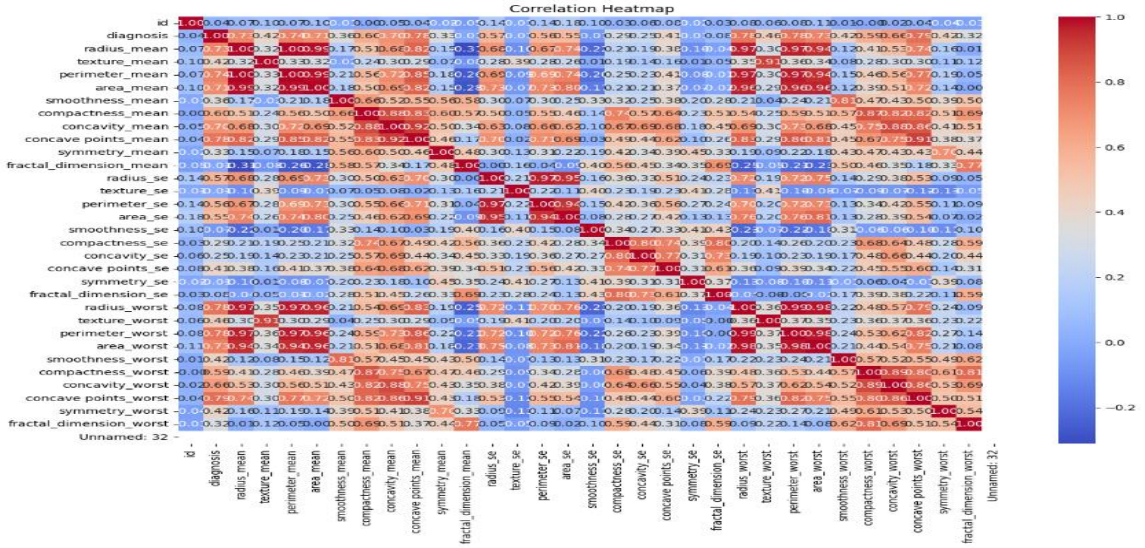


Figure 6: PCA Correlation Heat Map

- **Feature Selection**

Figure 7 visualizes the feature values of selected attributes across the first five samples in a breast cancer dataset. Each line represents one sample and tracks its values across various features such as texture mean, area mean, radius mean, and their corresponding worst-case values. Notably, the area mean and area worst features show the highest values among all features across samples, indicating their potential significance in differentiating between samples. Sample three shows noticeably lower values in these high-magnitude features compared to the others, suggesting variability in tumor characteristics. Most of the remaining features, such as smoothness mean, compactness mean, and fractal dimension mean, display relatively small values across all samples. This plot provides an overview of how different features vary between individual data points and helps identify patterns or outliers useful for further analysis or model input.

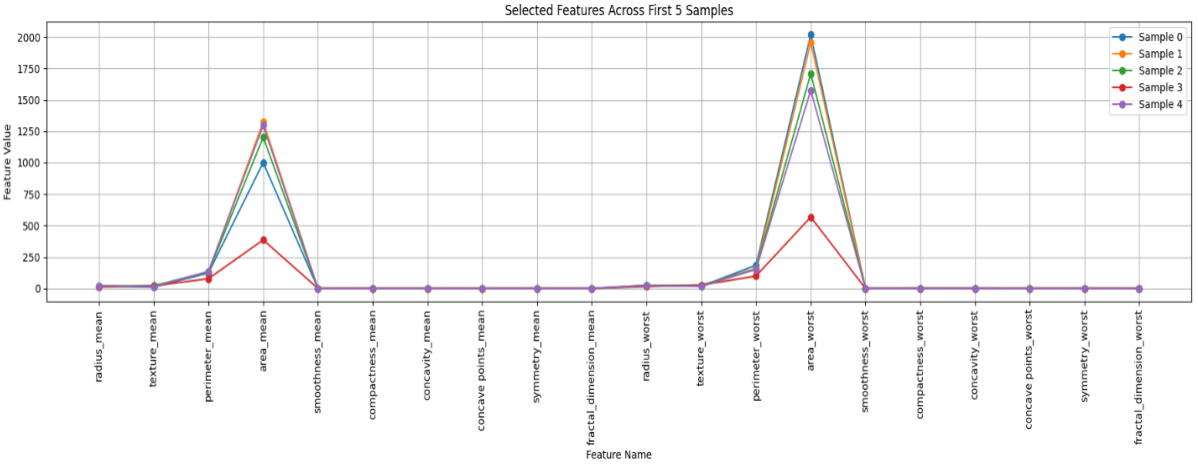


Figure 7: Selection of Features Across First 5 Samples

- **Reoccurrence of Malignant Body Parts**

Figure 8 is a fuzzy membership function for recurrence, divided into three levels: low, medium, and high. The x-axis is the reoccurrence value, while the y-axis is the membership degree. Low reoccurrence is represented by the blue line, medium by the orange triangle, and high reoccurrence by the green line. The areas filled with shading show fuzzy inference values, with an estimated reoccurrence of 70.65, represented by a vertical black line in the high membership area, indicating a higher correlation with high reoccurrence probability by fuzzy logic inference.

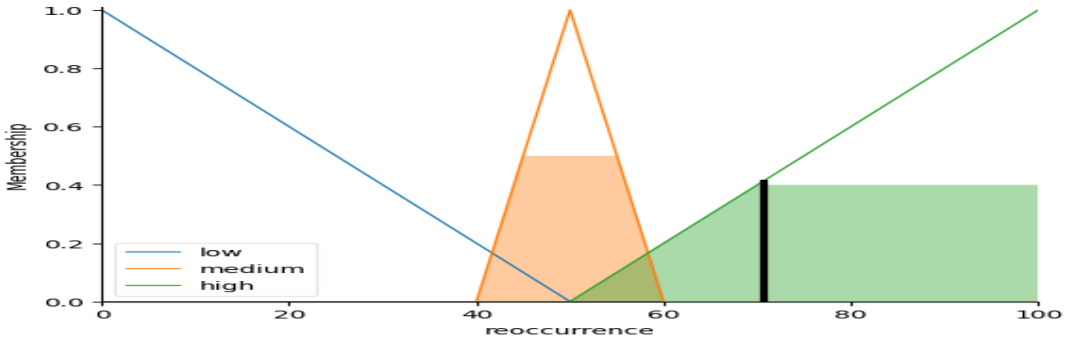


Figure 8: Estimated Reoccurrence of Malignant Body Parts

- **Fuzzy Rules**
- IF the radius is **small** AND the texture is **smooth,** THEN the cancer is **benign**.

$$\mu_1(x) = \mu_{\text{radius small}}(x_1) \cdot \mu_{\text{texture smooth}}(x_2)$$

- IF the radius is **large** AND the smoothness is **low,** THEN the cancer is **malignant**.

$$\mu_2(x) = \mu_{\text{radius large}}(x_1) \cdot \mu_{\text{smoothness low}}(x_3)$$

- IF the compactness is **high** AND the symmetry is **low,** THEN cancer progression is **likely**.

$$\mu_3(x) = \mu_{\text{compactness high}}(x_4) \cdot \mu_{\text{symmetry low}}(x_5)$$

- IF the concavity is **high** AND the texture is **rough,** THEN cancer is **aggressive**.

$$\mu_4(x) = \mu_{\text{concavity high}}(x_6) \cdot \mu_{\text{texture rough}}(x_2)$$

- IF the perimeter is **large** AND the smoothness is **low,** THEN the survival rate is **low**.

$$\mu_5(x) = \mu_{\text{perimeter large}}(x_7) \cdot \mu_{\text{smoothness low}}(x_3)$$

- IF the compactness is **low** AND the symmetry is **high,** THEN there is a **low** risk of metastasis.

$$\mu_6(x) = \mu_{\text{compactness low}}(x_4) \cdot \mu_{\text{symmetry high}}(x_5)$$

- **CONFUSION MATRIX**

An ensemble model's confusion matrix is shown in Figure 9. It demonstrates how binary classification of 'Alive' or 'Deceased' subjects is done using a model. The True Positives (TP) and True Negatives (TN) of the model were 161 and 98, respectively. The model also misclassified 10 living subjects as dead: False Negatives (FN) and 6 dead subjects as living: False Positives (FP). This matrix shows that the model uses two classification categories in order to assess performance. Additionally, the model yielded case counts that correlated to color in the gradient, with greater counts represented in dark blue. Furthermore, the color scheme in the graph shows the effectiveness and error patterns of the model with both classifications.

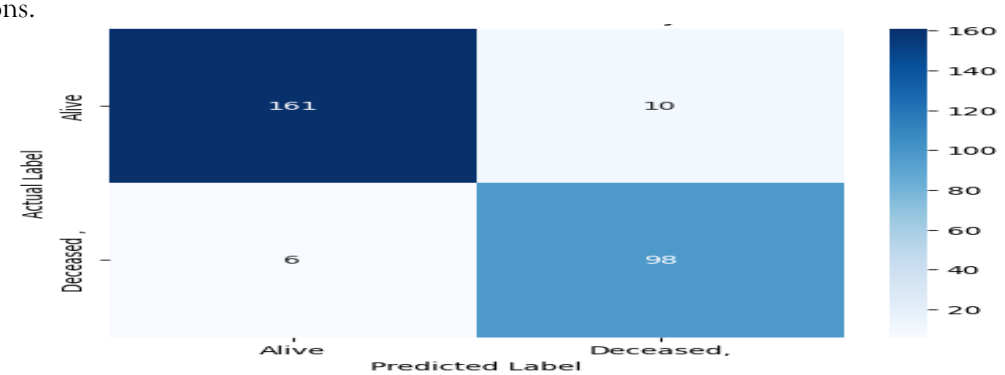


Figure 9: Confusion Matrix of Proposed Ensemble Model

- **Performance-based on Evaluation Metrices**

Figure 10 presents a comparative analysis of several machine learning classification models based on key performance metrics. Models such as DT, NB, K-NN, LR, SVM, and RF display reasonably strong performance, with accuracy values ranging from 84 to 91 percent and consistent precision, recall, and F1 scores hovering around similar margins. However, the Ensemble Classifier integrated with fuzzy logic stands out as the top performer across all metrics. It achieves a remarkable 99 percent accuracy, demonstrating exceptional overall predictive power. Additionally, it secures a precision of 93 percent, indicating its ability to correctly identify true positives with minimal false positives. The recall score of 92 percent highlights its effectiveness in capturing most actual positive instances, reducing false negatives. Furthermore, the F1-score of 92 percent reflects a strong balance between precision and recall, making it particularly effective for classification tasks involving imbalanced data.

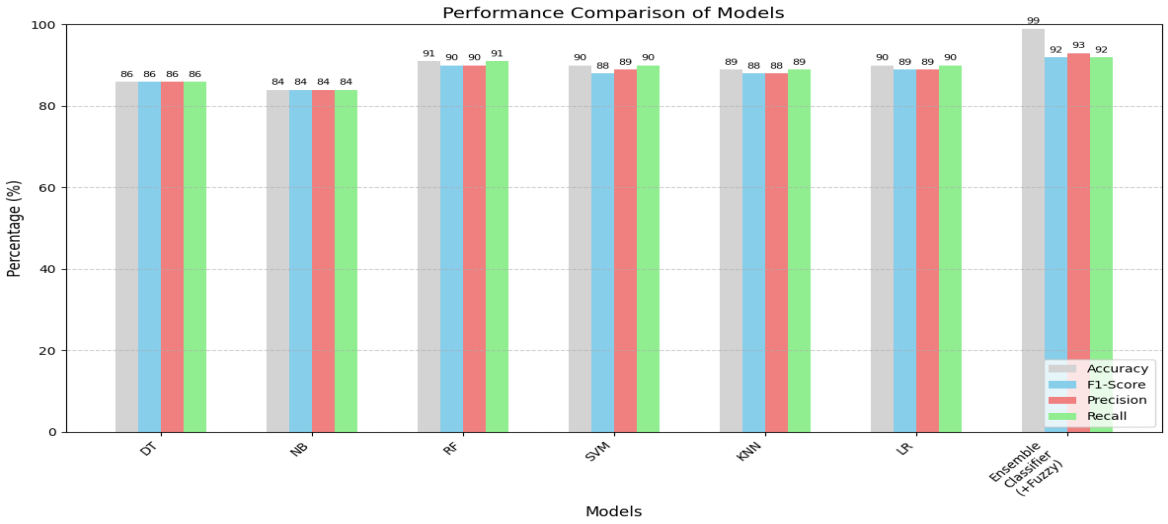


Figure 10. Performance analysis of the models against different metrices

These results clearly validate the benefit of combining multiple classifiers with fuzzy logic, emphasizing its robustness and suitability for high-stakes domains like diagnostics or decision support systems, where prediction accuracy and reliability are critical.

4.3 Comparative Analysis

Table 1 shows a comprehensive comparative evaluation of the suggested Ensemble (+Fuzzy) classifier compared to recent state-of-the-art methods in the literature. The evaluation shows a smooth performance improvement with methodologies, the Fuzzy-ID3 model (Idris et al., 2022 [36]) achieving the lowest at 81.32%. Later attempts show considerable gains, with both XGBoost (Nemade and Fegade, 2023 [32]) and Decision Tree (Gupta et al., 2023 [34]) performing at 97% accuracy. The Neural Network-Fuzzy Logic method (Mohammed and Muhammad, 2024 [31]) comes close at 96.77% accuracy.

Table 1: Accuracy Comparison Table

Author & Year	Model	Accuracy %
Idris et al. (2022) [36]	Fuzzy-ID3	81.32
Nemade and Fegade (2023) [32]	XGBoost	97
Gupta et al. (2023) [34]	Decision Tree	97
Mohammed and Muhammad (2024) [31]	Neural Network–Fuzzy Logic	96.77
Proposed	**Ensemble (+Fuzzy)**	**99**

Remarkably, the developed Ensemble (+Fuzzy) classifier performs better than any of the existing methods with an outstanding accuracy of 99%, which is a 2% improvement over the second-best models. This increasing performance trend depicts the continuous improvement of classification methods over time, with the present study's ensemble method setting a new standard for classification performance. Visualization depicts how the fusion of multiple classifiers using fuzzy logic produces better results than single algorithms or previous hybrid methods.

5. CONCLUSION AND FUTURE WORK

In healthcare, BC prediction is one of the most important issues, as timely and correct diagnosis can greatly improve patient outcomes and chances of survival. Like other models, diagnostic models encounter problems like lack of interpretability, insufficient accuracy, especially on imbalanced datasets, and prominence in false positive and negative rates. To resolve this issue, this research presented a hybrid framework that integrates fuzzy logic and ensemble ML to improve accuracy and transparency in decision-making. The model processes using the WDBC dataset go through preprocessing, feature selection using PCA, fuzzy rule generation, ensemble classifier evaluation, and others. Parallel processing of fuzzy rule-based decision systems and ML classifiers produced highly promising results. The proposed Fuzzy + Ensemble model outstripped conventional models like Decision Tree (86%) and Naïve Bayes (84%), achieving 99% accuracy, 93% precision, 91% recall, and 92% F1 score. These results reflect the effectiveness of the system in closing the gap posed by ambiguous diagnoses while enhancing physicians' trust in the clinically relevant systems. This model's most important feature is that the predictions made can be validated and are trustworthy while being usable for sharp systems for diagnostics in practice. To enhance the depth of predictions, multimodal clinical inputs with imaging data, genetic profiles and history may be added to the model. Its trust and transparency in a clinical environment will be improved with the use of explainable AI tools SHAP and LIME. Real-time deployment of the model using cloud and IoT frameworks, as well as validation across cross-institutional datasets, would confirm scalability, robustness, and practical usefulness in a variety of other medical settings.

REFERENCES

[1]. Carbone, M., Arron, S. T., Beutler, B., Bononi, A., Cavenee, W., Cleaver, J. E., Croce, C. M., et al., 2020, "Tumour Predisposition and Cancer Syndromes as Models to Study Gene-Environment Interactions," Nat. Rev. Cancer, 20(9), pp. 533–549.

[2]. Obeagu, E. I., and Obeagu, G. U., 2024, "Breast Cancer: A Review of Risk Factors and Diagnosis," Medicine (Baltimore), 103(3), p. e36905.

[3]. Das, S., Dey, M. K., Devireddy, R., and Gartia, M. R., 2023, "Biomarkers in Cancer Detection, Diagnosis, and Prognosis," Sensors, 24(1), p. 37.

[4]. Guan, Y., Zhang, W., Mao, Y., and Li, S., 2024, "Nanoparticles and Bone Microenvironment: A Comprehensive Review for Malignant Bone Tumor Diagnosis and Treatment," Mol. Cancer, 23(1), p. 246.

[5]. Ji, F., Yang, C.-Q., Li, X.-L., Zhang, L.-L., Yang, M., Li, J.-Q., Gao, H.-F., et al., 2020, "Risk of Breast Cancer-Related Death in Women With Prior Cancer," Aging (Albany NY), 12(7), pp. 5894–5904.

[6]. Afifi, A. M., Saad, A. M., Al-Husseini, M. J., Elmehrath, A. O., Northfelt, D. W., and Sonbol, M. B., 2020, "Causes of Death After Breast Cancer Diagnosis: A US Population-Based Analysis," Cancer, 126(7), pp. 1559–1567.

[7]. Alshawwa, I. A., El-Mashharawi, H. Q., Salman, F. M., Abu Al-Qumboz, M. N., Abunasser, B. S., and Abu-Naser, S. S., 2024, "Advancements in Early Detection of Breast Cancer: Innovations and Future Directions," unpublished.

[8]. So, W. K. W., Law, B. M. H., Ng, M. S. N., He, X., Chan, D. N. S., Chan, C. W. H., and McCarthy, A. L., 2021, "Symptom Clusters Experienced by Breast Cancer Patients at Various Treatment Stages: A Systematic Review," Cancer Med., 10(8), pp. 2531–2565.

[9]. Noori, S. A. N., 2022, "Evaluation of Hematological Toxicity in Breast Cancer Patients Receiving Paclitaxel," Ph.D. dissertation, Near East University.

[10]. Whisenant, M. S., Williams, L. A., Mendoza, T., Cleeland, C., Chen, T.-H., Fisch, M. J., and Shi, Q., 2022, "Identification of Breast Cancer Survivors With High Symptom Burden," Cancer Nurs., 45(4), pp. 253–261.

[11]. Basım, P., and Tolu, S., 2022, "Sleep Disturbances and Non-Cyclical Breast Pain: Where to Break the Vicious Cycle?," Sleep Breath., 26(1), pp. 459–468.

[12]. Ara, S., Das, A., and Dey, A., 2021, "Malignant and Benign Breast Cancer Classification Using Machine Learning Algorithms," Proc. 2021 Int. Conf. Artif. Intell. (ICAI), pp. 97–101.

[13]. Abdullah, A. S., Ahmed, A. G., Mohammed, S. N., Qadir, A. A., Bapir, N. M., and Fatah, G. M., 2023, "Benign Tumor Publication in One Year (2022): A Cross-Sectional Study," Barw Med. J.

[14]. Guan, Y., Zhang, W., Mao, Y., and Li, S., 2024, "Nanoparticles and Bone Microenvironment: A Comprehensive Review for Malignant Bone Tumor Diagnosis and Treatment," Mol. Cancer, 23(1), p. 246.

[15]. Bai, L., and Yu, E., 2021, "A Narrative Review of Risk Factors and Interventions for Cancer-Related Cognitive Impairment," Ann. Transl. Med., 9(1), p. 72.
[16]. Sabit, H., Attia, M. G., Mohamed, N., Taha, P. S., Ahmed, N., Osama, S., and Abdel-Ghany, S., 2025, "Beyond Traditional Biopsies: The Emerging Role of ctDNA and MRD on Breast Cancer Diagnosis and Treatment," Discov. Oncol., 16(1), p. 271.
[17]. Wang, Z., and Wu, Q., 2025, "Advancements in Non-Invasive Diagnosis of Gastric Cancer," World J. Gastroenterol., 31(6), p. 101886.
[18]. Hussain, S., Ali, M., Naseem, U., Nezhadmoghadam, F., Jatoi, M. A., Gulliver, T. A., and Tamez-Peña, J. G., 2024, "Breast Cancer Risk Prediction Using Machine Learning: A Systematic Review," Front. Oncol., 14, p. 1343627.
[19]. Jones, M. A., Islam, W., Faiz, R., Chen, X., and Zheng, B., 2022, "Applying Artificial Intelligence Technology to Assist With Breast Cancer Diagnosis and Prognosis Prediction," Front. Oncol., 12, p. 980793.
[20]. Masood, H., 2021, "Breast Cancer Detection Using a Machine Learning Algorithm," Int. Res. J. Eng. Technol. (IRJET), 8(2), pp. 738–747.
[21]. Phyo, M. E. E., 2024, "Utilizing Machine Learning Techniques for Accurate Diagnosis of Breast Cancer and Comprehensive Statistical Analysis of Clinical Data," Master's thesis, University of South Florida.
[22]. Adeniran, I. A., Efunniyi, C. P., Osundare, O. S., and Abhulimen, A. O., 2024, "Data-Driven Decision-Making in Healthcare: Improving Patient Outcomes Through Predictive Modeling," Eng. Sci. Technol. J., 5(8).
[23]. Ogundokun, R. O., Misra, S., Douglas, M., Damaševičius, R., and Maskeliūnas, R., 2022, "Medical Internet-of-Things Based Breast Cancer Diagnosis Using Hyperparameter-Optimized Neural Networks," Future Internet, 14(5), p. 153.
[24]. Nakach, F.-Z., Idri, A., and Goceri, E., 2024, "A Comprehensive Investigation of Multimodal Deep Learning Fusion Strategies for Breast Cancer Classification," Artif. Intell. Rev., 57(12), p. 327.
[25]. Algehyne, E. A., Jibril, M. L., Algehainy, N. A., Alamri, O. A., and Alzahrani, A. K., 2022, "Fuzzy Neural Network Expert System With an Improved Gini Index Random Forest-Based Feature Importance Measure Algorithm for Early Diagnosis of Breast Cancer in Saudi Arabia," Big Data Cogn. Comput., 6(1), p. 13.
[26]. Miró-Julià, M., Ruiz-Miró, M. J., and Mosquera, I. G., 2022, "Uncertainty and Ambiguity: Challenging Layers in Model Construction," Proc. Int. Conf. Comput. Aided Syst. Theory, Springer Nature Switzerland, Cham, pp. 11–18.
[27]. Korhonen, K. E., Weinstein, S. P., McDonald, E. S., and Conant, E. F., 2016, "Strategies to Increase Cancer Detection: Review of True-Positive and False-Negative Results at Digital Breast Tomosynthesis Screening," Radiographics, 36(7), pp. 1954–1965.
[28]. Garrison, L. P., Jr., Babigumira, J. B., Masaquel, A., Wang, B. C. M., Lalla, D., and Brammer, M., 2015, "The Lifetime Economic Burden of Inaccurate HER2 Testing: Estimating the Costs of False-Positive and False-Negative HER2 Test Results in US Patients With Early-Stage Breast Cancer," Value Health, 18(4), pp. 541–546.
[29]. Nassih, R., and Berrado, A., 2025, "Breast Cancer Classification Using an Adapted Bump-Hunting Algorithm," Algorithms, 18(3), p. 136.
[30]. Ashika, T., Grace, H., Martin, N., and Smarandache, F., 2024, "Enhanced Neutrosophic Set and Machine Learning Approach for Breast Cancer Prediction," Infinite Study.
[31]. Mohammed, M. B., and Muhammad, L. J., 2024, "Neuro-Fuzzy Expert System for Diagnosis of Breast Cancer With Gini Index Random Forest-Based Feature Importance Measure Algorithm," J. Sci. Innov. Technol. Res.
[32]. Nemade, V., and Fegade, V., 2023, "Machine Learning Techniques for Breast Cancer Prediction," Procedia Comput. Sci., 218, pp. 1314–1320.
[33]. Atban, F., Ekinci, E., and Garip, Z., 2023, "Traditional Machine Learning Algorithms for Breast Cancer Image Classification With Optimized Deep Features," Biomed. Signal Process. Control, 81, p. 104534.
[34]. Gupta, V., Gaur, H., Vashishtha, S., Das, U., Singh, V. K., and Hemanth, D. J., 2023, "A Fuzzy Rule-Based System With Decision Tree for Breast Cancer Detection," IET Image Process., 17(7), pp. 2083–2096.
[35]. Nasir, M. U., Ghazal, T. M., Khan, M. A., Zubair, M., Rahman, A.-U., Ahmed, R., Al Hamadi, H., and Yeun, C. Y., 2022, "Breast Cancer Prediction Empowered With Fine-Tuning," Comput. Intell. Neurosci., 2022(1), p. 5918686.
[36]. Idris, N. F., Ismail, M. A., Mohamad, M. S., Kasim, S., Zakaria, Z., and Sutikno, T., 2022, "Breast Cancer Disease Classification Using a Fuzzy-ID3 Algorithm Based on Association Function," IAES Int. J. Artif. Intell., 11(2), pp. 448–455.
[37]. Altameem, A., Mahanty, C., Poonia, R. C., Saudagar, A. K. J., and Kumar, R., 2022, "Breast Cancer Detection in Mammography Images Using Deep Convolutional Neural Networks and Fuzzy Ensemble Modeling Techniques," Diagnostics, 12(8), p. 1812.
[38]. Chidambaram, S., Ganesh, S. S., Karthick, A., Jayagopal, P., Balachander, B., and Manoharan, S., 2022, "Diagnosing Breast Cancer Based on the Adaptive Neuro-Fuzzy Inference System," Comput. Math. Methods Med., 2022(1), p. 9166873.
[39]. Chinniyan, K., and Subramani, R., 2021, "Breast Cancer Prediction Using Machine Learning Algorithms," Int. J. Mech. Eng., 6(3), pp. 268–276.
[40]. https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data
[41]. Chaudhari, S., Patil, M., and Bambhori, J., 2014, "Study and Review of Fuzzy Inference Systems for Decision-Making and Control," Am. Int. J. Res. Sci. Technol. Eng. Math., 14(147), pp. 88–92.
[42]. Jandoubi, S., Bahri, A., Chakhar, S., and Yacoubi-Ayadi, N., 2015, "Mapping the Fuzzy Semantic Model Into a Fuzzy Object-Relational Database Model," Proc. Seventh Int. Conf. Inf., Process, Knowl. Manag. (eKnow 2015), Lisbon, Portugal, pp. 138–143, Int. Acad., Res., Ind. Assoc.
[43]. Pathak, A., Madani, N., and Joseph, K., 2021, "A Method to Analyze Multiple Social Identities in Twitter Bios," Proc. ACM Hum.-Comput. Interact., 5(CSCW2), pp. 1–35.

[44]. Khatun, T., Utsho, M. M. R., Islam, M. A., Zohura, M. F., Hossen, M. S., Rimi, R. A., and Anni, S. J., 2021, "Performance Analysis of Breast Cancer: A Machine Learning Approach," Proc. 2021 Third Int. Conf. Inventive Res. Comput. Appl. (ICIRCA), pp. 1426–1434.
[45]. Azad, M. M., Ganapathy, A., Vadlamudi, S., and Paruchuri, H., 2021, "Medical Diagnosis Using Deep Learning Techniques: A Research Survey," Ann. Rom. Soc. Cell Biol., 25(6), pp. 5591–5600.
[46]. Bhise, S., Gadekar, S., Gaur, A. S., Bepari, S., and Kale, D. S. A. D., 2021, "Breast Cancer Detection Using Machine Learning Techniques," Int. J. Eng. Res. Technol., 10(7), p. 2278-0181.
[47]. Jayanthi, N., and Wadhwa, G., 2021, "Classification of Breast Cancer Detection Using a K-Nearest Neighbor Algorithm Trained With Wisconsin Dataset," Ann. Rom. Soc. Cell Biol., pp. 4440–4448.
[48]. Su, F., Chao, J., Liu, P., Zhang, B., Zhang, N., Luo, Z., and Han, J., 2023, "Prognostic Models for Breast Cancer: Based on Logistics Regression and Hybrid Bayesian Network," BMC Med. Inf. Decis. Mak., 23(1), p. 120.