Received: February 9, 2025; **Accepted:** June 8, 2025; **Published:** July 1, 2025.
DOI. 10.30473/coam.2025.73974.1295
Summer-Autumn (2025) Volume 10, Issue 2, (51-73)
Research Article

Open Access

Control and Optimization in Applied Mathematics - COAM

Deep Learning-Based Decision Fusion for Breast Cancer Classification Using Multi-Source Medical Data

Mohammad Zahaby[1], **Mostafa Boroumandzadeh**[1], **Iman Makhdoom**[2]

[1] Department of Computer engineering and information technology, Payame Noor University, Tehran, Iran.
[2]Department of Statistics, Payame Noor University, Tehran, Iran.

Correspondence:
Mohammad Zahaby
E-mail:
zahaby@pnu.ac.ir

How to Cite
Zahaby, M., Boroumandzadeh, M., Makhdoom, I. (2025). "Deep learning-based decision fusion for breast cancer classification using multi-source medical data", Control and Optimization in Applied Mathematics, 10(2): 51-73, doi: 10.30473/coam.2025.73974.1295.

Abstract. Breast cancer is one of the most prevalent cancers among women and remains a leading cause of cancer-related mortality. Mammography is the primary imaging modality for the early detection of breast tumors. Providing timely and highly accurate diagnoses is a top priority for physicians and healthcare providers in the management of critical illnesses. This paper presents a Medical Decision Support System (MDSS) that utilizes Yager's rule of combination to classify and diagnose breast cancer patients by integrating information from multiple data sources. Medical text reports (MTR) and key feature vectors extracted from electronic health records (EHR) were reduced using Principal Component Analysis (PCA) and then classified using Convolutional Neural Networks (CNN), Multi-Layer Perceptrons (MLP), and Support Vector Machines (SVM). Medical images were preprocessed and classified using a U-Net model. A novel decision fusion algorithm, called weighted Yager, was introduced to determine the Breast Imaging-Reporting and Data System (BI-RADS) categories, taking into account the accuracy of each class in each classifier as evidence. The performance of the proposed system was evaluated based on standard metrics including accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1-score. The proposed system achieved the highest accuracy of 96.23%, outperforming individual classifiers (CNN: 86.37%, MLP: 92.11%, SVM: 87.92%, U-Net: 92.97%, and Yager: 93.49%). The weighted Yager fusion method yielded the best performance with an accuracy of 96.23%, sensitivity of 98.80%, specificity of 85.90%, PPV of 86.21%, NPV of 97.82%, and F1-score of 85.87%. These findings demonstrate that integrating decisions from multiple classifiers significantly improves diagnostic accuracy and robustness.

Keywords. Medical decision support system, Text mining, BI-RADS, Deep learning.

MSC. 68U35; 68T05.

1 Introduction

Breast cancer is the most common type of cancer and a leading cause of cancer-related deaths among women. Despite significant medical advances, it remains a major challenge in healthcare and treatment [38]. This underscores the need for more accurate methods to diagnose and classify breast cancer, facilitating faster identification and treatment to improve patient outcomes. A key aspect of effective follow-up and treatment is the standardization of medical reports, which aids in the timely diagnosis and staging of the disease. To address this, the American College of Radiology (ACR) introduced BI-RADS, a system for standardizing mammography reports [7, 37]. BI-RADS categorizes findings into seven levels: 0 (incomplete), 1 (negative), 2 (benign findings), 3 (probably benign), 4 (suspicious abnormality), 5 (highly suggestive of malignancy), and 6 (biopsy-proven malignancy).

Deep learning and artificial intelligence are increasingly utilized to support physicians in evidence-based decision-making by analyzing complex, multidimensional data. A key aspect of this process is decision fusion, which combines information from multiple sources to address conflicting or uncertain data, ultimately improving the accuracy and reliability of the decision-making process [12]. Among the various decision fusion techniques, Yager's method stands out due to its ability to effectively manage such uncertainties and contradictions, providing a robust framework for enhancing the capabilities of medical systems [42]. Furthermore, results from clinical trials demonstrate that integrating data and decisions leads to more accurate diagnoses and improved treatment outcomes for breast cancer [32].

This article investigates a novel decision support system designed using Yager's rule of combination to achieve accurate and effective classification of breast cancer. The proposed system leverages clinical data, medical reports, and imaging analysis through deep learning algorithms. Its innovative approach to data combination enhances classification accuracy while improving performance in handling data variations and conflicting evidence. This is due to its flexibility in learning and decision-making processes. The developed decision support system offers a powerful tool for medical professionals, enabling faster and more effective treatment decisions, and contributing significantly to improved clinical outcomes for patients.

2 Related Works

Breast cancer diagnosis has been widely studied, with significant research focused on the integration of medical imaging, clinical reports, and patient history. Several studies have proposed various approaches for improving breast cancer classification, often relying on either medical image analysis or text-based clinical reports. However, these approaches tend to have notable limitations, which our proposed system aims to address. In recent years, numerous studies have focused on improving breast cancer diagnosis and treatment by utilizing advanced computational techniques. A significant portion of this research has concentrated on medical image analysis and text mining from clinical reports. However, many of these studies have limitations in terms of the data sources used, the integration of decision fusion techniques, and the generalization of results across diverse patient populations.

Esmaeili et al. [18] developed a clinical decision support system (CDSS) using data mining techniques to interpret mammography reports. This approach relied solely on the text data from the reports,

which limited its ability to incorporate visual data from mammograms. Furthermore, the system did not employ decision fusion, which could have improved the robustness and accuracy of the classification. Boumaraf et al. [10] proposed a system that utilized a genetic algorithm (GA) for BI-RADS classification based solely on mammography images. While their method achieved reasonable performance, it overlooked the valuable contextual information provided by patient history and clinical reports. Additionally, the absence of decision fusion techniques in their model hindered its ability to handle conflicting evidence from different data sources. Borkowski et al. [8] employed deep learning techniques for the automatic classification of background parenchymal enhancement (BPE) in breast MRI images. Although their method provided accurate results for image-based analysis, it ignored other critical sources of information such as medical text and patient history, which could have enriched the decision-making process. Zhang et al. [47] focused on analyzing clinical text data using various deep learning approaches. While their method successfully extracted features from the reports, it did not incorporate imaging data, and lacked decision fusion techniques to reconcile information from multiple sources. As a result, the system may have been less robust in handling the complexity and variability of breast cancer data.

Jesneck et al. [26] evaluates various classification algorithms applied to two breast cancer datasets, focusing on decision fusion methods that optimize clinically significant performance measures, such as the area under the curve (AUC) and partial area under the curve (pAUC). The findings suggest that decision fusion can outperform traditional machine learning techniques in certain scenarios. Manali et al. [30] proposes a three-parallel-channel artificial intelligence-based system that combines support vector machines (SVM) and convolutional neural networks (CNN) through decision fusion. The approach aims to enhance system performance in classifying mammogram images. Yan et al. [43] proposed a fusion network for classifying benign and malignant breast cancer cases by integrating multimodal data. This approach leverages diverse data sources to improve diagnostic accuracy. Fogliatto et al. [19] proposes a method for feature selection and classification of breast cancer cases into benign or malignant categories by deriving a feature importance index, contributing to the development of decision support systems in breast cancer detection.

Destrempes et al. [17] evaluated different combinations of 13 features derived from shear wave elasticity (SWE), statistical metrics, and spectral tissue dispersion from ultrasound images, combined with BI-RADS classifications, using the random forest algorithm. This approach did not consider the text of medical reports. The most relevant studies [4, 5, 11, 13, 17, 20, 22, 23, 34, 39] center on BI-RADS determination for breast cancer diagnosis, predominantly using either medical images or clinical text alone. In contrast, this article proposes a novel decision support system that integrates hospital information systems (HIS), medical text, and imaging data to train classifiers and determine BI-RADS categories. By applying a new decision fusion method based on Yager's rule of combination, this approach yields superior results, promising significant improvements in the breast cancer treatment process.

Zahaby and Makhdoom [44] proposed a decision support system that combines various classifiers, such as CNN, Decision Tree, multi-class SVM, and XGBoost, using weighted ensemble learning with majority voting to determine BIRADS values. This approach aims to improve diagnostic accuracy by integrating multiple models, although it may struggle with imbalances in classifier performance. The weighted ensemble learning method aggregates classifier outputs based on predefined weights, which can sometimes lead to suboptimal decision fusion when individual classifier strengths vary significantly. In contrast, our method, based on Weighted Yager, achieves superior performance by more effectively handling the imbalance in classifier results and offering enhanced robustness in breast cancer diagnosis.

Sharifonnasabi and Makhdoom [21] compared various deep learning and machine learning algorithms for breast cancer diagnosis and found that CNN achieved the highest accuracy. This study explored the effectiveness of several machine learning algorithms, including deep learning approaches such as multilayer perceptron (MLP) and convolutional neural networks (CNN), as well as traditional classifiers like decision tree (DT), Naïve Bayesian (NB), support vector machine (SVM), K-nearest neighbors (KNN), and eXtreme Gradient Boosting (XGBoost). Their research, conducted using the Breast Cancer Wisconsin Diagnostic dataset, evaluated classifier performance based on metrics such as confusion matrix, accuracy, and precision. The results of their study revealed that CNN outperformed all other models.

Zahaby et al. [46] investigates the impact of multimodal data integration in medical decision support systems. Their work highlights the importance of combining data from multiple sources, such as imaging and patient history, to improve diagnostic accuracy. However, the study emphasizes that while multimodal data integration is promising, there are challenges related to data standardization and the handling of conflicting evidence from diverse data sources.

Boroumandzadeh et al. [9] explores the use of machine learning algorithms for breast cancer diagnosis, focusing on feature selection and classification accuracy. The authors use clinical and text data but limit their focus to a single data source at a time, with employing decision fusion for two methods. While the study achieves promising results, the lack of decision fusion prevents it from leveraging the full potential of multimodal data, particularly in cases where data from different sources may conflict or provide incomplete information.

Furthermore, while many studies have attempted to apply deep learning techniques, such as CNNs for image analysis, they often overlook the importance of integrating textual information from medical reports. This is a critical gap, as clinical reports provide essential context that can significantly enhance diagnostic decisions.

Our proposed system builds upon these previous efforts by integrating textual data, imaging data, and patient history using an innovative decision fusion method based on Yager's rule of combination. This approach addresses the limitations observed in previous studies by allowing for more robust and accurate classifications, even in the presence of contradictory or incomplete information.

Table 1 summarizes the key characteristics and limitations of the most relevant studies in breast cancer diagnosis. As shown, while prior studies predominantly rely on single data sources (text or images), our proposed system benefits from multi-modal integration and decision fusion, which enhances its diagnostic performance.

Table 1: Comparison of related studies in breast cancer diagnosis.

Study	Data Sources	Methodology	Decision Fusion	Limitations
Esmaeili et al. [18]	Text Data (Mammography Reports)	Data Mining	No	Only text-based analysis, no image data
Boumaraf et al. [10]	Image Data (Mammography Images)	Genetic Algorithm (GA)	No	No integration of patient history or clinical text
Borkowski et al. [8]	Image Data (MRI)	Deep Convolutional Neural Network (CNN)	No	No use of clinical text or patient history
Zhang et al. [47]	Text Data (Clinical Reports)	Deep Learning	No	Excludes imaging data, lacks decision fusion
Proposed System	Text, Image Data, HIS	Deep Learning, Weighted Yager's Decision Fusion	Yes	Multi-source integration, improved diagnostic accuracy

3 Material and Methods

This paper presents a novel framework for a decision support system for breast cancer diagnosis, encompassing dataset preparation, preprocessing, classification, decision fusion, and result evaluation to optimize BI-RADS classification using an improved Yager method. The segmentation process of the proposed decision support system is illustrated in Figure 1. Initially, the medical report text for each individual in the dataset is processed and transformed into a vector representation using the Word2Vec algorithm [45]. Alongside features extracted from mammography reports, additional features derived from patients' electronic health records (EHR) are incorporated. These features are combined at the data level and, following preprocessing in the second stage, are used to train the CNN, SVM, and MLP classifiers in the subsequent stage. Additionally, mammography images undergo normalization and noise removal before being used to train the U-Net classifier. In the classification stage, these models are employed to predict BI-RADS values. The predictions are then aggregated in the decision fusion stage using the proposed improved Yager method. Finally, in the evaluation stage, the system's results and overall performance are reviewed and assessed.

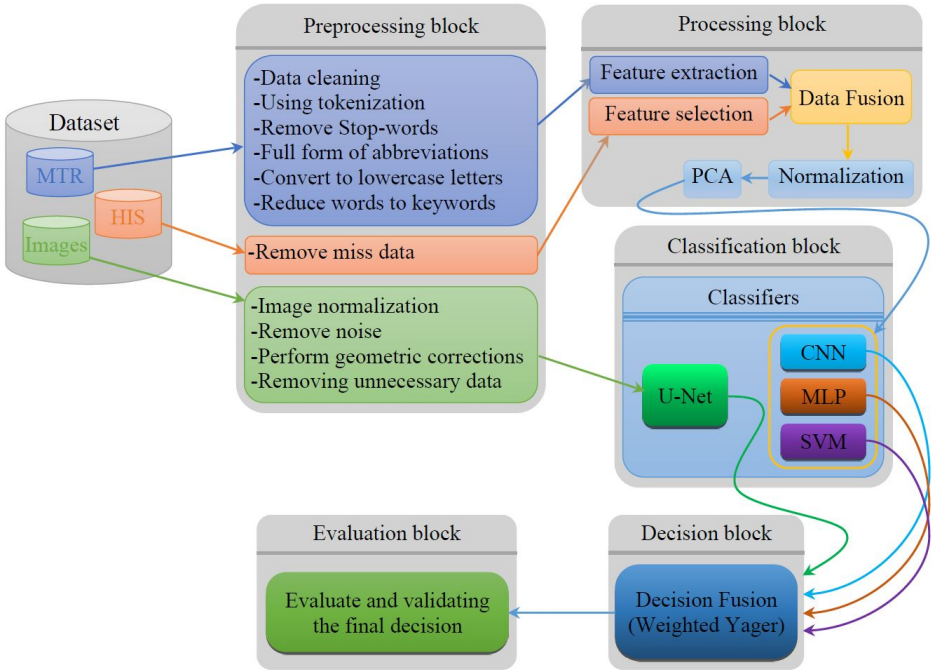

Figure 1: Segmentation of the proposed DSS.

3.1 Dataset

The dataset used in this research includes two main sources; Mammography images and reports, as well as the patient's electronic file, which were extracted from PACS and patient records, respectively. These datasets were obtained from the information available in Shahidzadeh Hospital Medical Training Center

in Behbahan City in the period of 2020 to 2022, which includes mammography reports and electronic files of 400 patients, and since the information of some patients was incomplete and had missing data, finally, only the information of 250 patients who had complete information was used. According to Equation (1), which is known as Cochran's formula [16] and was presented by William Cochran, to achieve an accuracy with an error of 5% in performing calculations, at least 138 samples will be needed, and considering that in this research, the sample size is 250 cases, it indicates the appropriateness of the sample size for conducting this research with an error accuracy of 5%.

$$n = \frac{z^2 p(1-p)}{d^2}. \tag{1}$$

In Equation (1), p represents the estimated proportion of breast cancer in the population. The parameter z is the $Z-score$ corresponding to the desired confidence level, which is set to 1.96 for a 95% confidence interval, thus $z^2 = 3.8416$. The term d denotes the margin of error, chosen as 0.05 to reflect acceptable precision in the estimation. Finally, n indicates the required minimum sample size, which was calculated to be 138 using these parameter values.

Some of the key features and elements extracted from the medical text reports and keywords are: density, asymmetry, distribution pattern, shape, size, and history of breast surgery. Also, some of the characteristics checked by medical experts and patient records (related to HIS) were: menopause, breast-feeding history, sports activities, pregnancy history, marital status, age, family history of cancer, etc.

3.2 Preprocessing and Processing

In this research, preprocessing was applied to the textual data from mammography reports. Key medical descriptors were extracted from the mammography text reports, including breast density, asymmetry, architectural distortion, distribution patterns, lesion size, prior breast surgeries, and lesion shape. These elements were identified and standardized during preprocessing to ensure consistency and enhance the quality of text-based feature extraction. This involved comprehensive cleaning steps, including removing impurities, standardizing characters, and correcting grammatical and spelling errors. Tokenization was used to break the data into processable units. Stopwords were removed, and abbreviated phrases were expanded to their full forms. The text was converted to lowercase, and words were stemmed or lemmatized to reduce them to their root forms. These steps improved the effectiveness of subsequent modeling. Feature extraction was performed using Word2Vec embeddings combined with min-max normalization [33], transforming each report into a numerical vector.

Additionally, structured clinical data was extracted from the Hospital Information System (HIS) data included 30 features related to patient history and clinical context. These features encompassed elements such as breastfeeding history, physical activity levels, pregnancy history, marital status, patient age, and family history of cancer. Each feature was rated by seven expert physicians on a scale from 1 to 10 based on its perceived clinical relevance to breast cancer diagnosis. The average scores were used to select the top 24 most relevant features for further analysis and classification.

To reduce dimensionality, Principal Component Analysis (PCA) was applied. PCA was chosen to eliminate multicollinearity among features, reduce computational cost, and preserve the most informative variance in the data. This step enhanced model performance and helped prevent overfitting.

To enhance the quality of diagnosis and classification, the vectors extracted from medical text reports were combined with selected features from the Hospital Information System (HIS) at the data level and processed using CNN, MLP, and SVM classification algorithms. Additionally, due to the challenges posed by mammography images—such as noise, intensity variations, and non-uniform contrasts—preprocessing was applied. The first step in image processing involved normalizing the images by adjusting their contrast and brightness, ensuring homogeneity across the dataset. To remove noise and perform post-processing tasks, a Gaussian filter was applied, which preserves the most important edges and details. Geometric corrections were made to the mammography images, followed by advanced metadata segmentation algorithms that enhanced edge differentiation and eliminated unnecessary data. This process strengthened feature extraction by providing a cleaner and more distinct surface for pattern recognition. These preprocessing steps allow for more accurate analysis of mammography images, leading to reliable results when using the U-Net deep learning technique.

3.3 Classification

In this section, the vectors generated in the previous step are passed to the CNN, MLP, and SVM classification algorithms to predict the BI-RADS score in patient reports. Additionally, mammography images are processed by the U-Net classification algorithm to determine the corresponding BI-RADS value.

Machine learning algorithms are essential for extracting knowledge from data and typically operate within reasonable computational times for specific problems [3]. Convolutional neural networks (CNNs) [31] are a class of deep learning models commonly applied to image, speech, and text analysis in machine learning. In this study, CNNs are used for BI-RADS classification, as they can capture complex relationships between data variables and handle noisy data effectively. A convolution operation is performed on the input, followed by pooling layers where sampling is applied to reduce dimensionality and prevent overfitting [41].

During the backpropagation phase, the parameter θ is updated by minimizing the error. The $ReLU$ activation function is used in both the first and second convolutional layers, while the output layer employs the softmax function, and the loss function used is the mean squared error. The $Adam$ optimization algorithm [24], an adaptive learning rate optimizer, is also utilized.

The U-Net architecture is a convolutional neural network originally designed for medical image segmentation at the University of Freiburg's Computer Science Department [36]. U-Net has been modified and extended for tasks requiring fewer training images, achieving more accurate segmentation [29]. This algorithm allows for high-speed processing and learning with reduced reliance on complex or expensive hardware. It can operate efficiently with smaller datasets, improving accuracy by extracting complex features [35]. In this study, the U-Net architecture is used for BI-RADS classification and detection.

Support Vector Machines (SVM) are used to find the optimal separating hyperplane that maximizes the margin between two classes. In this paper, the Radial Basis Function (RBF) kernel is employed [14], and once the model is trained, the possible BI-RADS class values for each sample are determined. A one-vs-all approach is utilized for BI-RADS classification, as there are seven possible classes. This approach involves using seven distinct SVMs, with each SVM making a decision for each sample, and the sample is assigned to the class corresponding to the highest probability [40].

The multilayer perceptron (MLP) model consists of input, hidden, and output layers. The input layer assigns a neuron to each input variable, while the hidden layer performs the main computational tasks of the network. In this study, the output layer comprises seven neurons, which are used to detect BI-RADS, along with the $softmax$ activation function. In this research, seven neurons were used for the output layer to detect BI-RADS along with the $softmax$ activation function. The primary computational power of the MLP comes from the hidden units between the input and output layers. Data flows forward through the network, similar to a feed-forward structure. The neurons are trained using a backpropagation algorithm. MLPs are composed of neurons, each referred to as a perceptron. A perceptron takes n features as input $ip = \{ip_1,\ ip_2, \ldots,\ ip_n\}$, and each feature is assigned a weight, and the weights assigned to the features must be a value be numerical [1]. In this article, the multilayer perceptron model is used for BI-RADS classification and diagnosis.

3.4 Decision Fusion

In this paper, we propose a framework for predicting and determining BI-RADS by combining the decisions of various classifier methods along with the assigned weight for each class, based on Yager's rule of combination. The models derived from CNN, MLP, SVM, and U-Net are integrated to aid decision-making in determining BI-RADS. Yager introduced and formulated an efficient method that also accounts for conflicts between evidences. To address this issue, Yager defined a new function, q, known as the probability mass allocation function. According to Equation (2), the probability mass allocation value can be greater than or equal to zero, indicating the possibility of conflict between the evidence [2].

$$q(\phi) \geq 0. \tag{2}$$

Yager considers only one weight for each evidence, which can reduce accuracy, especially in datasets where the class distribution is not uniform or when evidence lacks sufficient accuracy to distinguish a class. However, in the proposed method, a weight is assigned to each class within each piece of evidence. Let $\vec{O}_i(J)$ represent the estimation of classifier i for class j, and $\vec{A}_i(J)$ denote the weight of classifier i for class j. The mass values of the function $m_i(J)$ are then defined in Equation eq3. Additionally, the value $\vec{A}_i(J)$, which represents the accuracy of evidence i for class j, is computed based on the trained model and is given by (4).

$$m_i(J) = \ \vec{A}_i(J)\ \circ\ \vec{O}_i(J), \tag{3}$$

$$\vec{A}_i(J) = \ \frac{TP_i^j + TN_i^j}{TP_i^j + TN_i^j + \ FP_i^j + FN_i^j} = \ \text{Accuracy}_i^j. \tag{4}$$

According to Equation (5), the contradiction between evidences is classified in a set $\vec{\Omega}_i(J)$ which is equal to $\vec{\Omega}_i(J) = \left\{\omega_i^1, \ldots, \omega_i^j\right\}$. The combination of evidence decision in the proposed method is also calculated by Equations (6)-(8).

$$\vec{\Omega}_i(J) = 1 - \vec{A}_i(J), \tag{5}$$

$$q(J^j) = \sum_{\cap J_i^j = J^j} \left[m_1^j(J_1^j) \times m_2^j(J_2^j) \times \ldots \times m_i^j(J_i^j) + \omega_i^j \times m_i^j\right], \tag{6}$$

$$q(\phi)^j = \sum_{\cap m_i^j = \phi} m_i^j, \tag{7}$$

$$m_i(J^j) = \frac{q(J^j)}{1 - q\,(\phi)^j}. \tag{8}$$

In the proposed method, a weight is assigned to each class within each evidence, giving more influence to evidence that has better accuracy in distinguishing a certain class. The weight assigned to each class of evidence is based on the accuracy obtained from each classifier for that particular class during the classification phase. Finally, based on (8), a weight is calculated for each class, and the class with the highest value is selected as the decision integration output of the proposed system.

The procedural framework of the proposed method is presented in Algorithm 2, providing a systematic overview of its key algorithmic steps.

Algorithm 2 The algorithm of the proposed method.

Inputs: Models, $\vec{O}_i(J)$, $\vec{A}_i(J)$

Output: Predicted $m_i(J^j)$

$m_i(J) = \vec{A}_i(J) \circ \vec{O}_i(J)$.

$\vec{\Omega}_i(J) = 1 - \vec{A}_i(J)$.

for j **in** $\vec{O}_i(J)$ **do**

  for i **in** $Models$ **do**

    $q(\phi)^j = \sum_{\cap m_i^j = \phi} m_i^j$.

    Calculating $q(J^j)$ using (6)

    $m_i(J^j) = \frac{q(J^j)}{1-q(\phi)^j}$.

  end for

end for

Return $m_i(J^j)$.

4 Results

In this study, we encountered the inherent imbalance in breast cancer datasets, which poses challenges for effective classification. To address this, we employed various data balancing techniques, including the Synthetic Minority Over-sampling Technique (SMOTE) and under-sampling of the majority class. These approaches are well-recognized in the literature for their effectiveness in dealing with class imbalance [6, 15]. Using SMOTE, we synthetically generated new instances for the minority class by interpolating between existing instances. This method enhances the representation of the minority class and aids in mitigating imbalance. Additionally, we implemented under-sampling to reduce the size of the majority

class, ensuring a more balanced distribution of instances across classes. Given the balanced dataset achieved through these preprocessing methods, we deemed accuracy to be a suitable and fair metric for determining classifier weights in Yager's rule. Accuracy provides a comprehensive measure of the overall performance of the classifier, especially in scenarios where the dataset has been balanced [25]. While we acknowledge that metrics such as recall, precision, or F1-score are crucial for imbalanced datasets, in our context, the preprocessing steps allowed us to use accuracy reliably.

Zahaby et al. [45], systematically analyzed the impact of multi-source data (EHR + text vs. text alone). The findings confirmed that EHR data complements textual information and improved outcomes. This aligns with the current study's objective of leveraging multimodal data for improved performance.

Figure 2 illustrates the accuracies of the CNN, MLP, and SVM classifiers, which utilized text mining and HIS to detect BI-RADS. Since a large number of features are obtained through text mining, and some features extracted from HIS do not significantly contribute to the classification, PCA was applied to identify the most suitable and relevant features [27]. Features with dimensions ranging from 110 to 200 were selected and classified using CNN, MLP, and SVM. It was observed that as the dimensionality increased, the classification accuracy also increased, but the accuracy began to decrease at dimensions greater than 160. Several studies, including [28], have shown that the quality of word2vec deteriorates with increasing dimensionality, which leads to a reduction in accuracy. Ultimately, the maximum accuracy for the CNN, MLP, and SVM classifiers at 160 dimensions was 86.37%, 92.11%, and 87.92%, respectively. This value was chosen as the base dimension for subsequent calculations and processes.

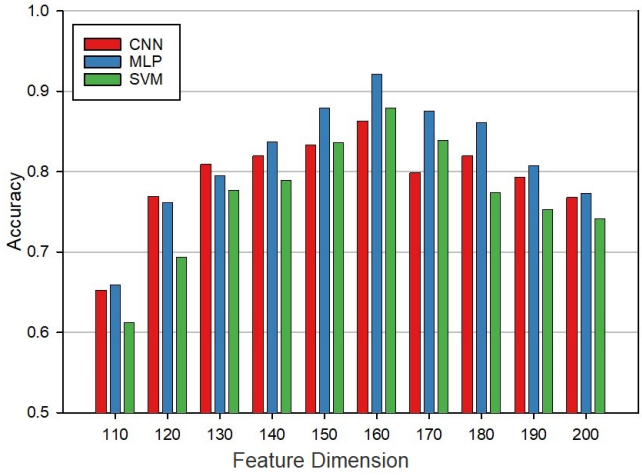

Figure 2: The variation of accuracy with the number of features.

Figures 3 and 4 present a comparison of the evaluation parameters of the proposed Decision Support System (DSS) with other methods and classifiers. Figure 3 shows the evaluation parameters of accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1-measure for all classifiers used in this research, along with the Yager method and the proposed DSS.

The best accuracy for the proposed DSS was 96.23%, compared to 86.37%, 92.11%, 87.92%, 92.97%, and 93.49% for CNN, MLP, SVM, U-Net, and the Yager method, respectively.

The highest sensitivity for the proposed DSS was 85.90%, whereas CNN, MLP, SVM, U-Net, and Yager achieved sensitivity values of 78.39%, 83.55%, 77.97%, 81.76%, and 84.89%, respectively. The

best specificity for the proposed DSS was 97.80%, while CNN, MLP, SVM, U-Net, and Yager recorded specificity values of 85.76%, 90.48%, 86.15%, 92.54%, and 92.91%, respectively.

The highest PPV for the proposed DSS was 86.21%, whereas the PPV values for CNN, MLP, SVM, U-Net, and Yager were 79.71%, 83.49%, 76.46%, 83.87%, and 84.62%, respectively. The best NPV value for the proposed DSS was 97.82%, compared to 85.33%, 91.02%, 85.27%, 91.35%, and 92.48% for CNN, MLP, SVM, U-Net, and Yager, respectively. Finally, the highest F1-measure for the proposed DSS was 85.87%, while CNN, MLP, SVM, U-Net, and Yager had F1-measure values of 77.28%, 78.93%, 75.21%, 81.17%, and 82.25%, respectively.

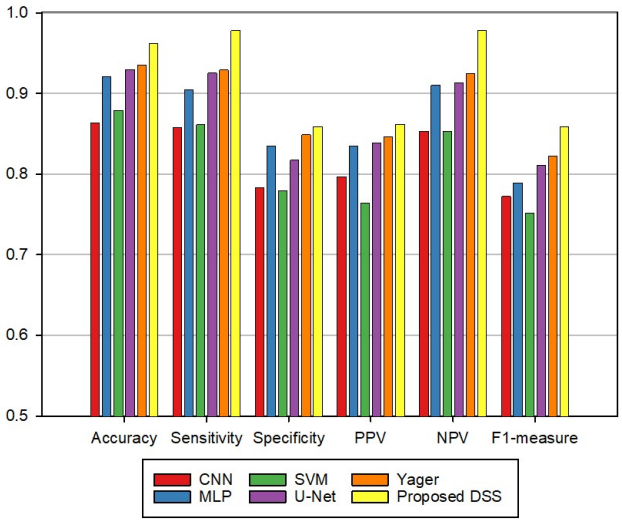

Figure 3: Comparison of evaluation parameters of the proposed DSS with other classifiers.

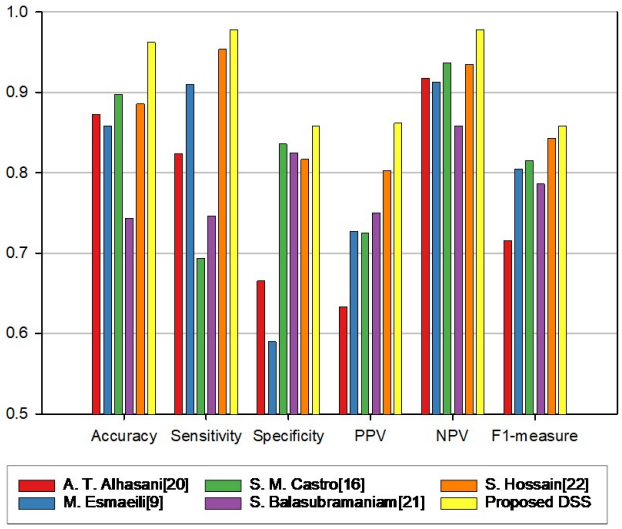

Figure 4: Comparison of evaluation parameters of the proposed DSS with other methods.

Figure 4 presents the parameters of accuracy, specificity, sensitivity, PPV, NPV, and F1-measure for different methods, all using the same dataset employed in this study. The results demonstrate that the proposed Decision Support System (DSS) outperforms the other methods. By examining the evaluation metrics, it is evident that the proposed method, which employs decision fusion to enhance accuracy, performs better than similar methods used for identifying BI-RADS classes. This improvement in diagnostic performance ultimately contributes to more effective patient treatment and follow-up care.

To further validate the performance of our proposed method, we compared it with Yager's original decision fusion approach as well as other commonly used classifiers. As illustrated in Figure 3, our method consistently achieved superior results across all evaluation metrics. This improvement is mainly attributed to the introduction of a weighted fusion mechanism, where the contribution of each classifier is adjusted based on its classification accuracy, unlike Yager's original method which assumes equal importance for all sources of evidence.

Table 2: Confusion matrix of proposed decision support system.

Confusion Matrix						
32	0	2	1	0	1	0
0	33	1	0	1	3	2
1	1	25	2	0	0	1
0	0	1	32	1	1	2
0	1	2	0	26	1	1
1	1	0	1	0	37	0
1	1	0	1	3	1	29

Class	Sensitivity	Specificity	PPV	NPV	F1-Measure	Accuracy
BI-RADS 0	76.67%	97.73%	82.14%	96.85%	79.31%	95.20%
BI-RADS 1	95.65%	96.57%	86.27%	98.99%	90.72%	96.40%
BI-RADS 2	88.57%	97.21%	83.78%	98.12%	86.11%	96.00%
BI-RADS 3	81.82%	97.81%	78.26%	98.24%	80.00%	96.40%
BI-RADS 4	94.74%	97.64%	87.80%	99.04%	91.14%	97.20%
BI-RADS 5	86.05%	99.03%	94.87%	97.16%	90.24%	96.80%
BI-RADS 6	77.78%	98.60%	90.32%	96.35%	83.58%	95.60%

Table 2 displays the confusion matrix of the proposed system for BI-RADS classification, along with the evaluation metrics of sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1-measure, and accuracy. Most disease classes were detected with an accuracy exceeding 95%. The average sensitivity value is 85.90%, with the lowest sensitivity for BI-RADS 0 and the highest sensitivity for BI-RADS 1 (95.65%).

The specificity value for healthy individuals is 96.57%, demonstrating the system's high performance in detecting healthy people. The average specificity value is 97.80%, with the minimum value for BI-RADS 1 and the maximum value for BI-RADS 5, which reaches 99.03%. These values indicate that the proposed method performs well in terms of specificity.

The average positive predictive value (PPV) is 86.21%, with the highest PPV value of 94.87% for BI-RADS 5. The negative predictive value (NPV) for healthy individuals is 98.99%, showcasing the proposed method's strong performance. The maximum NPV value is 99.04% (BI-RADS 4), and the minimum value is 96.35% (BI-RADS 6).

The average F1-measure value is 85.87%, with the maximum value of 91.14% for BI-RADS 4 and the minimum value of 79.31% for BI-RADS 0. These results indicate that the proposed method provides a good detection rate.

The overall accuracy, or the system's ability to correctly diagnose both healthy and sick individuals, is 96.23% on average, with minimum and maximum accuracy values of 95.20% and 97.20%, respectively.

In conclusion, by analyzing these evaluation metrics, it is evident that the proposed Decision Support System (DSS) performs effectively in detecting BI-RADS classes, which aids in disease diagnosis

and the determination of appropriate treatment methods. Since multi-evidence decision fusion was employed, the proposed method improved the detection performance of BI-RADS.

5 Discussion

The proposed method was examined and implemented on a desktop computer with the following specifications: Intel® Core™ i7-4790 CPU @ 3.60 GHz, 16 GB DDR RAM (2x8 GB), 2 GB GT 730 graphics card, and a 256 GB SSD hard drive with an additional 1 TB SATA hard drive. For implementation, Python 3.8.7 was used within the Visual Studio Code environment.

5.1 Evaluation Parameters

K-fold cross-validation was employed to assess the quality and validity of the results. The data was first divided into 10 subsets ($K = 10$), and for each subset, the system was trained based on the proposed framework, with the average evaluation criteria being reported.

The confusion matrix is one of the evaluation criteria for classifiers [46]. It is an $N \times N$ matrix, where N represents the number of classes; in this case, there are 7 BI-RADS classes. The diagonal of the matrix contains the number of correct predictions (True Positives), while the off-diagonal elements represent false detections. In binary classification models, which only detect the presence or absence of a disease, the confusion matrix includes concepts such as True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). This study aims to classify patients into the 7 BI-RADS categories. Here, TP_i refers to the True Positive value for class i, which indicates cases where both the true class and predicted class are i. The value for TP_i is calculated using Equation (9). False Positives (FP) are represented by FP_i, where the true class is i, but the predicted class is different. The value for FP_i is calculated using Equation (10). Additionally, FN_i, representing False Negatives, denotes cases where the predicted class is i, but the true class is different from i. FN_i is computed with Equation (11). Finally, TN_i is the True Negative value, indicating cases where the true class is not i, and the predicted class is also not i. The value for TN_i is calculated using (12).

$$TP_i = C_{ii}, \qquad i = 0, 1, \ldots, 6, \tag{9}$$

$$FP_i = \sum_{i \neq j = 0}^{6} C_{ij}, \qquad i = 0, 1, \ldots, 6, \tag{10}$$

$$FN_i = \sum_{i \neq j = 0}^{6} C_{ji}, \qquad i = 0, 1, \ldots, 6, \tag{11}$$

$$TN_i = \sum_{i \neq j = 0}^{6} \sum_{i \neq k = 0}^{6} C_{jk}, \qquad i = 0, 1, \ldots, 6. \tag{12}$$

Using these values, other important parameters such as accuracy, specificity, sensitivity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), and the F1-score can be calculated. These metrics

are computed using Equations (13)-(18), where TP, TN, FP, and FN represent the averages of TP_i, TN_i, FP_i, and FN_i, respectively [28].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{13}$$

$$\text{Specificity} = \frac{TN}{TN + FP}, \tag{14}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \tag{15}$$

$$PPV = \frac{TP}{TP + FP}, \tag{16}$$

$$NPV = \frac{TN}{TN + FN}, \tag{17}$$

$$F1 - \text{measure} = \frac{2 \times PPV \times \text{Sensitivity}}{PPV + \text{Sensitivity}}. \tag{18}$$

5.2 Evaluation of Methodologies

In this segment, we analyze the methods employed in this study from the perspective of statistical hypothesis testing. Initially, each method was executed 30 times under identical conditions, and the accuracy for each iteration was recorded. Table 3 presents the cumulative and mean execution times for all methods, including their respective results.

The proposed model exhibits a longer runtime compared to some other classifiers, as shown in Table 3, this increase in execution time is justified by its superior diagnostic performance. In clinical practice, particularly in the context of cancer diagnosis, the emphasis is primarily on accuracy and reliability rather than speed. Since cancer detection does not usually require real-time processing, a slightly longer processing time is acceptable if it results in a more accurate and trustworthy diagnosis. Therefore, the runtime of the proposed method is not expected to hinder its practical applicability in clinical environments.

Subsequently, an ANOVA test, conducted using SPSS version 25, was performed to compare the six methods. The results of the analysis are presented in Tables 3 and 4.

Table 3: Descriptives statistics.

	N	Runtime (ms)		Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
		Overall	Average				Lower Bound	Upper Bound		
CNN	30	325821	10861	.757257142933	.0192775351727	.0035195802891	.750058792998	.764455492868	.7314285710	.8171428570
MLP	30	16699	557	.841980952333	.0131712977633	.0024047389655	.837062708919	.846899195747	.8125714290	.8662857140
SVM	30	13909	464	.826057142833	.0154861905440	.0028273786303	.820274504249	.831839781417	.7954285710	.8514285710
U-Net	30	76281	2543	.870857142833	.0205341753768	.0037490103512	.863189555734	.878524729932	.8022857140	.8982857140
Yager	30	30664	1022	.902929482433	.0073922379998	.0013496318343	.900169175400	.905689789467	.8891428570	.9154285710
Proposed DSS	30	231870	7729	.911042385433	.0049901840045	.0009110787818	.909179020102	.912905750764	.8971428570	.9165714290

Table 4: ANOVA.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.484	5	.097	453.019	.000
Within Groups	.037	174	.000		
Total	.523	179			

To compare the accuracy of the mentioned methods, the ANOVA test using $Fisher's\ F$ statistic was employed, testing the hypothesis outlined in Equation (19) as follows:

$$\begin{cases} H_0: \ \mu_{CNN} = \mu_{DecisionTree} = \mu_{MLF} = \mu_{SVM} = \mu_{XGboost} = \mu_{ProposedDSS} \\ H_A: \ \text{At least one of the means is different from the others.} \end{cases} \tag{19}$$

Based on the findings presented in Table 4, the p-value $(Sig: 0.000)$ is clearly below the predetermined significance level of $\alpha = 0.05$. Therefore, the null hypothesis is rejected, indicating that at least one of the means significantly deviates from the others. To further investigate this discrepancy, a post hoc test was conducted. As shown by the p-values highlighted in yellow in Tables 5 and 6, all values associated with the proposed DSS are less than 0.05, which results in the rejection of the null hypotheses. This suggests a significant difference between the mean of the proposed DSS and the means of the other classifiers. Consequently, post hoc tests (Tukey's HSD and LSD) reveal a significant difference between the mean of the proposed method and the means of the other classifiers, but no significant difference between the proposed DSS (weighted Yager) and Yager's method. Furthermore, Figure 5 visually demonstrates the superiority of the proposed DSS method over its counterparts.

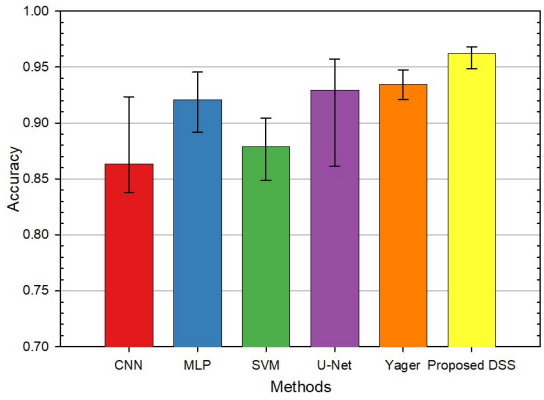

Figure 5: Comparison of accuracy of proposed DSS with other classifiers.

Table 5: Multiple comparisons for tukey HSD.

(I) Classifier	(J) Classifier	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
CNN	MLP	-.0847238094000*	0.003780253	0	-0.095617527	-0.073830092
	SVM	-.0687999999000*	0.003780253	0	-0.079693718	-0.057906282
	U-Net	-.1135999999000*	0.003780253	0	-0.124493718	-0.102706282
	Yager	-.1456723395000*	0.003780253	0	-0.156566057	-0.134778622
	Proposed DSS	-.1537852425000*	0.003780253	0	-0.16467896	-0.142891525
MLP	CNN	.0847238094000*	0.003780253	0	0.073830092	0.095617527
	SVM	.0159238095000*	0.003780253	0.001	0.005030092	0.026817527
	U-Net	-.0288761905000*	0.003780253	0	-0.039769908	-0.017982473
	Yager	-.0609485301000*	0.003780253	0	-0.071842248	-0.050054812
	Proposed DSS	-.0690614331000*	0.003780253	0	-0.079955151	-0.058167715
SVM	CNN	.0687999999000*	0.003780253	0	0.057906282	0.079693718
	MLP	-.0159238095000*	0.003780253	0.001	-0.026817527	-0.005030092
	U-Net	-.0448000000000*	0.003780253	0	-0.055693718	-0.033906282
	Yager	-.0768723396000*	0.003780253	0	-0.087766057	-0.065978622
	Proposed DSS	-.0849852426000*	0.003780253	0	-0.09587896	-0.074091525
U-Net	CNN	.1135999999000*	0.003780253	0	0.102706282	0.124493718
	MLP	.0288761905000*	0.003780253	0	0.017982473	0.039769908
	SVM	.0448000000000*	0.003780253	0	0.033906282	0.055693718
	Yager	-.0320723396000*	0.003780253	0	-0.042966057	-0.021178622
	Proposed DSS	-.0401852426000*	0.003780253	0	-0.05107896	-0.029291525
Yager	CNN	.1456723395000*	0.003780253	0	0.134778622	0.156566057
	MLP	.0609485301000*	0.003780253	0	0.050054812	0.071842248
	SVM	.0768723396000*	0.003780253	0	0.065978622	0.087766057
	U-Net	.0320723396000*	0.003780253	0	0.021178622	0.042966057
	Proposed DSS	-0.008112903	0.003780253	0.002	-0.019006621	0.002780815
Proposed DSS	CNN	.1537852425000*	0.003780253	0	0.142891525	0.16467896
	MLP	.0690614331000*	0.003780253	0	0.058167715	0.079955151
	SVM	.0849852426000*	0.003780253	0	0.074091525	0.09587896
	U-Net	.0401852426000*	0.003780253	0	0.029291525	0.05107896
	Yager	0.008112903	0.003780253	0.002	-0.002780815	0.019006621

*. The mean difference is significant at the 0.05 level.

5.3 Practical Challenges and Limitations

Although the proposed decision fusion method demonstrates significant improvements in diagnostic accuracy, there are several practical challenges and limitations that need to be addressed for its successful application in clinical practice:

- **Computational Cost:** Deep learning models such as CNN and U-Net are computationally intensive, which could limit their real-time use in clinical settings, especially when processing large

Table 6: Multiple comparisons for tukey LSD.

(I)Classifier	(J) Classifier	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
CNN	MLP	-.0847238094000*	0.003780253	0	-0.092184862	-0.077262757
	SVM	-.0687999999000*	0.003780253	0	-0.076261053	-0.061338947
	U-Net	-.1135999999000*	0.003780253	0	-0.121061053	-0.106138947
	Yager	-.1456723395000*	0.003780253	0	-0.153133392	-0.138211287
	Proposed DSS	-.1537852425000*	0.003780253	0	-0.161246295	-0.14632419
MLP	CNN	.0847238094000*	0.003780253	0	0.077262757	0.092184862
	SVM	.0159238095000*	0.003780253	0	0.008462757	0.023384862
	U-Net	-.0288761905000*	0.003780253	0	-0.036337243	-0.021415138
	Yager	-.0609485301000*	0.003780253	0	-0.068409583	-0.053487477
	Proposed DSS	-.0690614331000*	0.003780253	0	-0.076522486	-0.06160038
SVM	CNN	.0687999999000*	0.003780253	0	0.061338947	0.076261053
	MLP	-.0159238095000*	0.003780253	0	-0.023384862	-0.008462757
	U-Net	-.0448000000000*	0.003780253	0	-0.052261053	-0.037338947
	Yager	-.0768723396000*	0.003780253	0	-0.084333392	-0.069411287
	Proposed DSS	-.0849852426000*	0.003780253	0	-0.092446295	-0.07752419
U-Net	CNN	.1135999999000*	0.003780253	0	0.106138947	0.121061053
	MLP	.0288761905000*	0.003780253	0	0.021415138	0.036337243
	SVM	.0448000000000*	0.003780253	0	0.037338947	0.052261053
	Yager	-.0320723396000*	0.003780253	0	-0.039533392	-0.024611287
	Proposed DSS	-.0401852426000*	0.003780253	0	-0.047646295	-0.03272419
Yager	CNN	.1456723395000*	0.003780253	0	0.138211287	0.153133392
	MLP	.0609485301000*	0.003780253	0	0.053487477	0.068409583
	SVM	.0768723396000*	0.003780253	0	0.069411287	0.084333392
	U-Net	.0320723396000*	0.003780253	0	0.024611287	0.039533392
	Proposed DSS	-.0081129030000*	0.003780253	0.033	-0.015573956	-0.00065185
Proposed DSS	CNN	.1537852425000*	0.003780253	0	0.14632419	0.161246295
	MLP	.0690614331000*	0.003780253	0	0.06160038	0.076522486
	SVM	.0849852426000*	0.003780253	0	0.07752419	0.092446295
	U-Net	.0401852426000*	0.003780253	0	0.03272419	0.047646295
	Yager	.0081129030000*	0.003780253	0.033	0.00065185	0.015573956

*. The mean difference is significant at the 0.05 level.

datasets. To optimize runtime, techniques such as model pruning, parallel processing, or cloud-based solutions are necessary to improve system performance in high-throughput environments.

- **Data Quality and Completeness:** The system's performance heavily depends on the quality and completeness of input data, including mammography reports, clinical records, and medical images. Incomplete or erroneous data may lead to inaccurate predictions, posing a major challenge for real-world applications.

- **Integration with Hospital Information Systems (HIS) and Electronic Health Records (EHR):** Our system's integration with HIS and EHR presents practical challenges related to data privacy,

security, and interoperability. Different healthcare institutions may use various HIS and EHR platforms, which could make it difficult to develop a universal solution that works seamlessly across all hospitals.

- **System Complexity:** The fusion of multiple classifiers improves accuracy but also increases system complexity. This added complexity could pose challenges for system maintenance, updates, and ease of interpretability. For clinical adoption, ensuring the system's outputs are understandable and actionable by healthcare professionals is essential.

6 Conclusions

The American College of Radiology (ACR) introduced the Breast Imaging Reporting and Data System (BI-RADS) to standardize mammogram reporting and improve patient care. This standard aims to help patients prioritize treatment progress more accurately based on their condition. However, certain challenges remain, such as disagreements among clinicians regarding BI-RADS results, which may hinder the determination of precise treatment strategies based on these values. To address these challenges, this paper proposes a hybrid model that integrates unstructured data (medical reports) with structured data from Hospital Information Systems (HIS) and medical images to create a biological model for medical data analysis. Mammography reports were converted into vectors using Word2Vec, following text processing. Essential features were selected using Principal Component Analysis (PCA). The classification of BI-RADS features was performed using CNN, MLP, SVM, and U-Net classifiers. These classification outputs were then combined using the proposed method to generate BI-RADS classification results. To evaluate the system, K-fold cross-validation was employed with $K = 10$. Metrics such as specificity, sensitivity, positive predictive value (PPV), negative predictive value (NPV), and F1-measure were calculated. The maximum values achieved for these metrics in the proposed method were 85.90%, 97.80%, 86.21%, 97.82%, and 85.87%, respectively, with an overall diagnostic accuracy of 96.23%. The proposed decision support system (DSS) converts medical textual records into vectors, utilizes HIS features derived from medical literature, and processes medical images for BI-RADS diagnosis. By synthesizing these various evidences, the DSS aids physicians in making informed decisions. Consequently, the system enhances the evaluation of patient treatment progress, potentially saving many lives by facilitating timely and accurate medical interventions. Although this study demonstrates promising results, there are several potential areas for future research and development. First, further research could focus on optimizing the system's computational efficiency to enable real-time clinical applications. This may include reducing model complexity and utilizing hardware accelerations like GPUs. Second, expanding the dataset to include data from different healthcare institutions, as well as incorporating additional data types such as genetic and demographic data, would help improve the generalizability and robustness of the system. Third, enhancing the explainability and interpretability of the decision fusion process is essential for ensuring the system's adoption in clinical environments. Finally, investigating the real-time integration of the system into clinical workflows would allow for continuous patient monitoring and decision support, improving patient care in practice.

Declarations

Availability of Supporting Data

All data generated or analyzed during this study are included in this published paper.

Funding

The authors conducted this research without any funding, grants, or support.

Competing Interests

The authors declare that they have no competing interests relevant to the content of this paper.

Authors' Contributions

All authors contributed equally to the design of the study, data analysis, and writing of the manuscript, and share equal responsibility for the content of the paper.

References

[1] Ahmed, S. (2023). "A software framework for predicting the maize yield using modified multi-layer perceptron", *Sustainability*, 15(4), 3017, doi:https://doi.org/10.3390/su15043017.

[2] Akram, M., Shahzadi, G. (2021). "A hybrid decision-making model under q-rung orthopair fuzzy Yager aggregation operators", *Granular Computing*, 6(4), 763-777, doi:https://doi.org/10.1007/s41066-020-00229-z.

[3] Alesheykh, R. (2016). "Comparative analysis of machine learning algorithms with optimization purposes", *Control and Optimization in Applied Mathematics*, 1(2), 63-75.

[4] Alhasani, A.T., Alkattan, H., Subhi, A.A., El-Kenawy, El.S.M., Eid, M.M. (2023). "A comparative analysis of methods for detecting and diagnosing breast cancer based on data mining", *Journal of Artificial Intelligence and Metaheuristics*, 4(2), 08-17, doi:https://doi.org/10.54216/JAIM.040201.

[5] Balasubramaniam, S., Velmurugan, Y., Jaganathan, D., Dhanasekaran, S.J.D. (2023). "A modified LeNet CNN for breast cancer diagnosis in ultrasound images", *Diagnostics*, 13(17), 2746, doi:https://doi.org/10.3390/diagnostics13172746.

[6] Batista, G.E., Prati, R.C., Monard, M.C. (2004). "A study of the behavior of several methods for balancing machine learning training data", *ACM SIGKDD Explorations Newsletter*, 6(1), 20-29, doi:https://doi.org/10.1145/1007730.1007735.

[7] Boyer, B., Canale, S., Arfi-Rouche, J., Monzani, Q., Khaled, W., Balleyguier, C. (2013). "Variability and errors when applying the BIRADS mammography classification", *European Journal of Radiology*, 82(3), 388-397, doi:https://doi.org/10.1016/j.ejrad.2012.02.005.

[8] Borkowski, K., Rossi, C., Ciritsis, A., Marcon, M., Hejduk, P., Stieb, S., Boss, A., Berger, N. (2020). "Fully automatic classification of breast MRI background parenchymal enhancement using

a transfer learning approach", *Medicine*, 99(29): p e21243, doi:https://doi.org/10.1097/MD.0000000000021243.

[9] Boroumandzadeh, M., Parvinnia, E., Boostani, R., Sefidbakht, S. (2021). "A decision support system framework based on text mining and decision fusion techniques to classify breast cancer patients", *Control and Optimization in Applied Mathematics*, 6(1), 11-29, doi:https://doi.org/10.30473/coam.2021.60533.1175.

[10] Boumaraf, S., Liu, X., Ferkous, C., Ma, X. (2020). "A new computer-aided diagnosis system with modified genetic feature selection for BI-RADS classification of breast masses in mammograms", *BioMed Research International*, 2020, doi:https://doi.org/10.1155/2020/7695207.

[11] Bozkurt, S., Gimenez, F., Burnside, E.S., Gulkesen, K.H., Rubin, D.L. (2016). "Using automatically extracted information from mammography reports for decision-support", *Journal of Biomedical Informatics*, 62, 224-231, doi:https://doi.org/10.1016/j.jbi.2016.07.001.

[12] Carlsson, C., Brunelli, M., Mezei, J. (2012). "Decision making with a fuzzy ontology", *Soft Computing*, 16, 1143-1152, doi:https://doi.org/10.1007/s00500-011-0789-x.

[13] Castro, S.M., Tseytlin, E., Medvedeva, O., Mitchell, K., Visweswaran, Sh., Bekhuis, T., Jacobson, R.S. (2017). "Automated annotation and classification of BI-RADS assessment from radiology reports", *Journal of Biomedical Informatics*, 69, 177-187, doi:https://doi.org/10.1016/j.jbi.2017.04.011.

[14] Chang, C.-C., Lin, C.-J. (2011). "LIBSVM: A library for support vector machines", *ACM Transactions on Interactive Intelligent Systems*, 2(3), 27, doi:https://doi.org/10.1145/1961189.1961199.

[15] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. (2002). "Smote: Synthetic minority over-sampling technique", *Journal of Artificial Intelligence Research*, 16, 321-357, doi:https://doi.org/10.1613/jair.953.

[16] Cochran, W.G. (1977). "Sampling techniques", 3rd Edition, *John Wiley & Sons*, New York.

[17] Destrempes, F., Trop, I., Allard, L., Chayer, B., Garcia-Duitama, J., El Khoury, M., Lalonde, L., Cloutier, G. (2020). "Added value of quantitative ultrasound and machine learning in BI-RADS 4–5 assessment of solid breast lesions", *Ultrasound in Medicine & Biology*, 46(2), 436-444, doi:https://doi.org/10.1016/j.ultrasmedbio.2019.10.024.

[18] Esmaeili, M., Ayyoubzadeh, S.M., Ahmadinejad, N., Ghazisaeedi, M., Nahvijou, A., Maghooli, K. (2020). "A decision support system for mammography reports interpretation", *Health Information Science and Systems*, 8(1), 17, doi:https://doi.org/10.1007/s13755-020-00109-5.

[19] Fogliatto, F.S., Anzanello, M.J., Soares, F., Brust-Renck, P.G. (2019). "Decision support for breast cancer detection: Classification improvement through feature selection", *Cancer Control*, 26(1), doi:https://doi.org/10.1177/1073274819876598.

[20] Gao, H., Aiello Bowles, E.J., Carrell, D., Buist, D.S.M. (2015). "Using natural language processing to extract mammographic findings", *Journal of Biomedical Informatics*, 54, 77-84, doi:https://doi.org/10.1016/j.jbi.2015.01.010.

[21] Ghazalnaz Sharifonnasabi, F., Makhdoom, I. (2022). "Comparison of deep learning and machine learning algorithms to diagnose and predict breast cancer", In: Ullah, A., Anwar, S., Calandra, D., Di Fuccio, R. (eds) *Proceedings of International Conference on Information Technology and Applications, ICITA, Lecture Notes in Networks and Systems*, 839. Springer, Singapore, doi:https://doi.org/10.1007/978-981-99-8324-7_4.

[22] Gupta, A., Banerjee, I., Rubin, D.L. (2018). "Automatic information extraction from unstructured mammography reports using distributed semantics", *Journal of Biomedical Informatics*, 78, 78-86, doi:https://doi.org/10.1016/j.jbi.2017.12.016.

[23] Hossaina, Sh., Azamb, S., Montahac, S., Karimb, A., Sultana Chowaa, S., Mondola, Ch., Hasana, Md.Z., Jonkman, M. (2023). "Automated breast tumor ultrasound image segmentation with hybrid UNet and classification using fine-tuned CNN model", *Heliyon*, 9(11), e21369, doi:https://doi.org/10.1016/j.heliyon.2023.e21369.

[24] Jais, I.K.M., Ismail, A.R., Nisa, S.Q. (2019). "Adam optimization algorithm for wide and deep neural network", *Knowledge Engineering and Data Science*, 2(1), 41-46, doi:https://doi.org/10.17977/um018v2i12019p41-46.

[25] Japkowicz, N., Stephen, S. (2002). "The class imbalance problem: A systematic study", *Intelligent Data Analysis*, 6(5), 429-449, doi:https://doi.org/10.3233/IDA-2002-6504.

[26] Jesneck, J.L., Nolte, L.W., Baker, J.A., Floyd, C.E., Lo, J.Y. (2006). "Optimized approach to decision fusion of heterogeneous data for breast cancer diagnosis", *Medical Physics*, 33(8), 2945-2954, doi:https://doi.org/10.1118/1.2208934.

[27] Jia, W., Sun, M., Lian, J., Hou, S. (2022). "Feature dimensionality reduction: A review", *Complex & Intelligent Systems*, 8(3), 2663-2693, doi:https://doi.org/10.1007/s40747-021-00637-x.

[28] Li, B., Drozd, A., Guo, Y., Liu, T., Matsuoka, S., Du, X. (2019). "Scaling Word2Vec on big corpus", *Data Science and Engineering*, 4(2), 157-175, doi:https://doi.org/10.1007/s41019-019-0096-6.

[29] Long, J., Shelhamer, E., Darrell, T. (2015). "Fully convolutional networks for semantic segmentation", in: *2015 IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA. 3431-3440, doi:https://doi.org/10.1109/CVPR.2015.7298965.

[30] Manalı, D., Demirel, H., Eleyan, A. (2024). "Deep learning based breast cancer detection using decision fusion", *Computers*, 13(11), 294, doi:https://doi.org/10.3390/computers13110294.

[31] Nal Kalchbrenner, E.G., Blunsom, Ph. (2014). "A convolutional neural network for modelling sentences", *arXiv*, doi:https://doi.org/10.48550/arXiv.1404.2188.

[32] Oh, S., Lee, M.S., Zhang, B.T. (2011). "Ensemble learning with active example selection for imbalanced biomedical data classification", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2), 316-325, doi:https://doi.org/10.1109/TCBB.2010.96.

[33] Patro, S.G.K., Kumar Sahu, K. (2015). "Normalization: A preprocessing stage", *arXiv*, abs/1503.06462, doi:https://doi.org/10.48550/arXiv.1503.06462.

[34] Percha, B., Nassif, H., Lipson, J., Burnside, E., Rubin, D. (2012). "Automatic classification of mammography reports by BI-RADS breast tissue composition class", *Journal of the American Medical Informatics Association*, 19(5), 913-916, doi:https://doi.org/10.1136/amiajnl-2011-000607.

[35] Punn, N.S., Agarwal, S. (2022). "Modality specific U-Net variants for biomedical image segmentation: a survey", *Artificial Intelligence Review*, 55(7), 5845-5889, doi:https://doi.org/10.1007/s10462-022-10152-1.

[36] Ronneberger, O., Fischer, P., Brox, T. (2015). "U-Net: Convolutional networks for biomedical image segmentation", *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Cham*, 234-241, Springer International Publishing, doi:https://doi.org/10.1007/978-3-319-24574-4_28.

[37] Sefidbakht, S., Jalli, R., Izadpanah, E. (2015). "Adherence of academic radiologists in a non-English speaking imaging center to the BI-RADS standards of reporting breast MRI", *Journal of Clinical Imaging Science*, 5, 66, doi:https://doi.org/10.4103/2156-7514.172970.

[38] Siegel, R.L., Miller, K.D., Wagle, N.S., Jemal, A.J.C.C.J.C. (2023). "Cancer Statistics", *CA: A Cancer Journal for Clinicians*, 73(1), 17-48, doi:https://doi.org/10.3322/caac.21763.

[39] Sippo, D.A., Warden, G.I., Andriole, K.P., Lacson, R., Ikuta, I., Birdwell, R.L., Khorasani, R. (2013). "Automated extraction of BI-RADS final assessment categories from radiology reports with natural language processing", *Journal of Digital Imaging*, 26(5), 989-94, doi:https://doi.org/10.1007/s10278-013-9616-5.

[40] Tang, F., Adam, L., Si, B. (2018). "Group feature selection with multiclass support vector machine", *Neurocomputing,* 317, 42-49, doi:https://doi.org/10.1016/j.neucom.2018.07.012.

[41] Xu, Q., Zhang, M., Gu, Z., Pan, G. (2019). "Overfitting remedy by sparsifying regularization on fully-connected layers of CNNs", *Neurocomputing*, 328, 69-74, doi:https://doi.org/10.1016/j.neucom.2018.03.080.

[42] Yager, R.R. (1988). "On ordered weighted averaging aggregation operators in multicriteria decisionmaking", *IEEE Transactions on Systems, Man, and Cybernetics*, 18(1), 183-190, doi:https://doi.org/10.1109/21.87068.

[43] Yan, R., Zhang, F., Rao, X., Lv, Zh., Li, J., Zhang, L., Liang, Sh., Li, Y., Ren, F., Zheng Ch., Liang, J. (2021). "Richer fusion network for breast cancer classification based on multimodal data", *BMC Medical Informatics and Decision Making,* 21c, 1-15, doi:https://doi.org/10.1186/s12911-020-01340-6.

[44] Zahaby, M., Makhdoom, I. (2025). "Enhanced decision support system for breast cancer diagnosis with weighted ensemble learning methods", *Journal of Data Science and Modeling*, 71-102, doi:https://doi.org/10.22054/jdsm.2025.82414.1055.

[45] Zahaby, M., Shiri, M.E., Haj Seyed Javadi, H., Broumandzadeh, M. (2024). "Automatic classification of BI-RADS in mammography reports using data fusion", *Armaghane Danesh*, 29(3), 365-85, doi:http://dx.doi.org/10.61186/armaghanj.29.3.365.

[46] Zahaby, M., Shiri, M.E., Haj Seyed Javadi, H., Boroumandzadeh, M. (2025). "Decision support system for improving breast cancer diagnosis using ensemble learning", *Computing and Informatics*, 44(1), 124-150, doi:https://doi.org/10.31577/cai_2025_1_124.

[47] Zhang, X. et al. (2019). "Extracting comprehensive clinical information for breast cancer using deep learning methods", *International Journal of Medical Informatics*, 132, 103985, doi:https://doi.org/10.1016/j.ijmedinf.2019.103985.