Journal of Data Science and Modeling, Vol. 2, No. 2, 71-102, June 2024

Research Manuscript

Enhanced Decision Support System for Breast Cancer Diagnosis with Weighted Ensemble Learning Methods

Mohammad Zahaby*[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

Recieved: 13/10/2024 Accepted: 09/03/2025

Abstract: Breast cancer (BC) is one of the leading causes of death in women worldwide, and early diagnosis can save many lives. The Breast Imaging Reporting and Data System (BIRADS), developed by the American College of Radiology (ACR), is a standard method used in diagnosis. However, physicians face challenges in determining BIRADS values, and many factors have not been considered in past diagnostic methods. This article presents a novel decision support system (DSS). In the proposed DSS, c-mean clustering is used to determine the molecular subtype for patients lacking this value, combining mammography report processing with hospital information systems (HIS) from electronic files. Several classifiers, including convolutional neural networks (CNN), decision tree (DT), multi-level fuzzy min-max neural network (MLF), multi-class support vector machine (SVM), and XGBoost, are trained to determine the BIRADS value. These classifier outputs are then combined using weighted ensemble learning with the majority voting algorithm. This assists physicians in the early diagnosis of BC. Results are evaluated using accuracy, specificity, sensitivity, positive predictive value (PPV), negative predictive value (NPV), and F1-score. The obtained values are 97.94%, 98.79%, 92.08%, 92.34%, 98.80%, and 92.19%, respectively.

Keywords: weighted ensemble learning, combined machine learning, decision support system, breast cancer diagnosis, BIRADS

Mathematics Subject Classification (2010): 68T05, 62C20.

*Corresponding Author: zahaby@pnu.ac.ir

1. Introduction

Cancer is now one of the top causes of death globally. In developed nations, it is the second leading cause of death after heart disease, while in less developed regions, it ranks third. Cancer results in more fatalities than diseases like tuberculosis, AIDS, and malaria combined Balakumar and *et al.* (2016). Without effective prevention, the next decade could see over 85 million deaths from cancer worldwide Balakumar and *et al.* (2016). Presently, cancer is responsible for 12% of global deaths Bray and *et al.* (2018). Among women, breast cancer is one of the most frequently diagnosed cancers. Statistics show that 19.9% of cancer-related deaths in women are linked to breast cancer U.S. Cancer Statistics Working Group (2024). The World Health Organization (WHO) estimates that between one in eight to one in ten women will be diagnosed with breast cancer in their lifetime Isfahani and *et al.* (2020). Early detection is a key factor in the successful treatment of the disease, as identifying breast cancer in its early stages significantly improves the likelihood of recovery and survival Dehghan and *et al.* (2018); Ginsburg and *et al.* (2020). Medical decision support systems (MDSS), developed through collaboration between physicians and engineers, are designed to assist healthcare providers in making informed medical decisions Alaa and *et al.* (2016); Mazo and *et al.* (2020); Sim and *et al.* (2017). Today, many medical facilities have recognized the value of MDSS in managing breast cancer Mazo and *et al.* (2020). Research suggests that these systems, by providing visualized patient data, enable doctors to quickly access the necessary information to determine the most suitable treatment Park and *et al.* (2021). Mammography reports are one of the important inputs used by MDSS to help diagnose and treat breast cancer Esmaeili and *et al.* (2020).

Radiologists rely on the BIRADS classification system, developed by the American College of Radiology, to interpret mammography findings and describe them in medical reports Magny and *et al.* (2023). This system is recognized as one of the most reliable methods for evaluating and assessing the risk of breast lesions through mammography Farrokh and *et al.* (2019). The BIRADS system is divided into seven levels, ranging from 0 to 6, with each level offering a specific interpretation of the mammogram results Vanderheyden and Xie (2020). Despite various medical decision support systems (MDSS) being introduced to assist in diagnosing cancer patients using electronic health record data, no MDSS has yet been developed to classify breast cancer patients by combining information from mammography reports, electronic patient records (HIS), and molecular subtypes. Preliminary studies indicate a gap in this area Castro and *et al.* (2017); Gao and *et al.* (2015); Gupta and *et al.* (2018); Nassif and *et al.* (2012); Percha and *et al.* (2012); Dorothy and *et al.* (2013); Zhang and *et al.* (2019).

This article is divided into four sections. Section 1 provides an introduction to the article's context, outlining the advancements and limitations in the field of study, along with a brief overview of its objectives. Section 2 presents a proposed decision support system, explaining its various stages and components in detail. All necessary information for understanding the system's operation is also provided. In Section 3, the results from the proposed system are evaluated and analyzed. Finally, Section 4 discusses the system and offers conclusions about its practicality and effectiveness.

2. Literature Review

In 2012, Percha et al. Percha and *et al.* (2012) processed medical reports and categorized them according to BIRADS, though their focus was limited to breast tissue. Also in 2012, Nassif et al. Nassif and *et al.* (2012) extracted BIRADS features from clinical texts and compared them with manual reports, but no BIRADS grading was applied. In 2013, Sippo et al. Dorothy and *et al.* (2013) automated the extraction of BIRADS features from radiology reports using the BIRADS Observation Kit and natural language processing (NLP). In 2015, Gao et al. Gao and *et al.* (2015) used NLP to extract information from unstructured mammography reports, but their approach was confined to diagnosing four types of breast complications, relying solely on medical reports. In 2016, Bozkurt et al. Bozkurt and *et al.* (2016) introduced a decision support system based on NLP to diagnose malignancies from BIRADS reports and radiology texts. In 2017, Castro et al. Castro and *et al.* (2017) presented a rule-based NLP approach for classifying radiology reports, though it utilized only one type of textual data. In 2017, Gupta et al. Gupta and *et al.* (2018) proposed a method using parse tree structures and semantics to convert mammography reports into structured data, utilizing only medical reports. Esmaili et al., in 2020, Esmaeili and *et al.* (2020) introduced a decision support system to assist doctors in interpreting mammography text reports and developed a model capable of predicting when a biopsy might be necessary. In 2022, Achilonu et al. Achilonu and *et al.* (2022) developed a rule-based NLP algorithm that extracted key breast cancer parameters from pathology reports, focusing specifically on molecular subtypes, using only molecular subtype reports. Higa, in 2018, Higa (2018) employed artificial neural networks and decision tree classification methods to predict breast cancer using clinical data. In 2019, Zhang et al. Zhang and *et al.* (2019) applied deep learning techniques to extract clinical information related to breast cancer, though their approach was notably complex. In 2023, Spaeth et al. Spaeth and *et al.* (2023) introduced a breast cancer diagnostic model that incorporates key clinical factors,

including family history and polygenic risk, enabling the exclusion of moderate factors to enhance diagnostic accuracy.

Based on previous research, mammography reports, health information system (HIS) data, and molecular subtypes have typically been utilized independently for BIRADS diagnosis. In this study, however, we combine electronic health record information and molecular subtypes with mammography reports to assess the impact of this additional information on BIRADS diagnosis. The aim of this research is to develop a decision support system that predicts BIRADS values and molecular subtypes. To achieve this, mammography reports are initially processed using natural language processing (NLP) and transformed into vectors with word2vec Guo and *et al.* (2019). Additionally, 15 features from the electronic health records of patients are extracted, which include 2 numerical and 13 nominal variables that are then combined with the vectors from the mammography reports. The unsupervised c-mean method is employed to cluster the molecular subtypes of the samples, assigning molecular subgroup values to each cluster's data. For classification and determination of BIRADS, several algorithms are utilized, including convolutional neural networks (CNN), decision trees (DT), multi-level fuzzy min-max neural networks (MLF), multi-class support vector machines (SVM), and XGBoost. The predicted BIRADS values from each classifier are then used as base learners, which are combined through weighted ensemble learning using a majority voting algorithm to improve prediction accuracy.

3. The Proposed Method

This paper presents a novel BIRADS diagnosis prediction model as part of the proposed decision support system (DSS). The dataset used comprises two main sources: mammography reports and electronic patient records, extracted from the hospital information system (HIS). The dataset consists of 250 mammography images, accompanied by their reports and electronic medical records from Shahidzadeh Hospital Medical Training Center in Behbahan, Iran, covering the years 2020 to 2022. The mammography text reports contain 210 distinct features, while the electronic records contribute 15 features. Table 1 outlines 2 numerical features, and Table 2 lists 13 nominal features. In total, 225 features are collected for each patient, combining both mammography report data and electronic record information.

Additionally, Table 3 provides details on the distribution of the 250 patients across the various BIRADS categories.

Figure 1 illustrates the different stages of the proposed approach, organized into five phases. In the first phase, the dataset is gathered, consisting of mammography

Table 1: Numerical features extracted from HIS

	Variable name	Variable description	Healthy people (n=17) Standard deviation ± mean	Patients (n=233) Standard deviation ± mean
1	Size	Lesion size	5.41±5.59	6.29±4.71
2	Age	Age of clients/patients	53.52± 11.48	43.89±32.11

reports and HIS data for each individual. Since mammography reports contain unstructured text, they were processed and transformed into vectors using natural language processing (NLP) techniques. In the second phase, important features from the HIS data were identified through consultation with a physician. In the third phase, because the dataset only includes BIRADS classifications, it was necessary to determine the molecular subtype classes. Using the c-means algorithm, an unsupervised clustering technique, the data were grouped into four clusters as outlined in Table 4. Each cluster was then assigned a value corresponding to its respective molecular subtype. In the fourth phase, multiple models were trained to predict BIRADS values, including convolutional neural networks (CNN), decision tree (DT), multi-level fuzzy min-max neural network (MLF), multi-class support vector machine (SVM), and XGboost. In the final phase, the predicted BIRADS values from these models were combined using a weighted ensemble learning method with majority voting, and the results were validated with evaluation parameters.

3.1 The first phase: Dataset

Our dataset consists of two primary sources: mammography reports and electronic patient records (a subset from the HIS). The research process began with the collection of medical data from Shahidzadeh Hospital Medical Training Center in Behbahan, Iran, covering the period from 2020 to 2022. This dataset initially included the mammography reports and electronic records of 400 patients. However, due to incomplete data for some patients, only the information of 250 patients with complete records was ultimately used for the study.

3.2 The second phase: Processing

3.2.1 Convert MTR to vector

Figure 2 shows the components of the proposed method for classifying medical reports and the process of extracting a vector from a mammography report. It is important to mention that this figure highlights only the text processing workflow.

During preprocessing, mammography reports were stemmed using the NLTK library Loper and *et al.* (2002). Prepositions and punctuation marks were re-

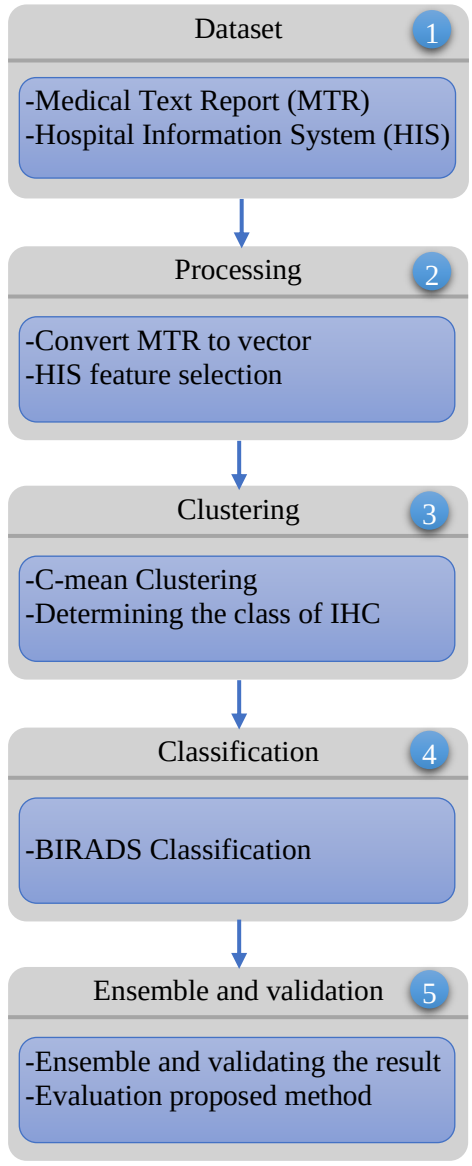

Figure 1: Phasing of the proposed method

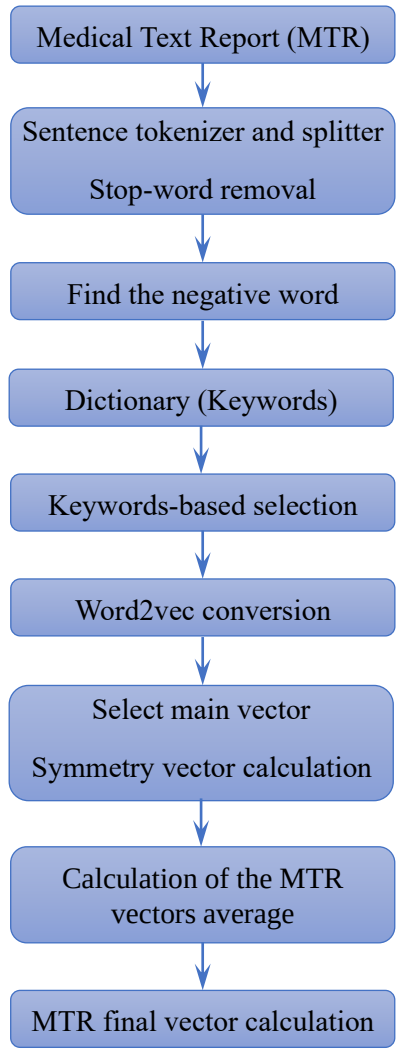

Figure 2: Converting a report to a vector

Table 2: Nominal features extracted from HIS

	Variable name	Variable description	Healthy people Qty (No.=17)	Patients Qty (No.=233)
1	Breast secretion	Presence/absence of abnormal breast discharge	No=5 Yes=12	No=136 Yes=97
2	Side	Left, right, or Bilateral (both sides of the chest)	Left=5 Right=8 Bilateral=4	Left=83 Right=108 Bilateral=42
3	Pain	History of pain in the breast area	No=6 Yes=11	No=86 Yes=147
4	Pregnancy	Presence/absence of pregnancy history	No=7 Yes=10	No=40 Yes=193
5	Disease	Presence/absence of disease history	No=12 Yes=5	No=121 Yes=112
6	Breastfeeding	Presence/absence of a history of the Breastfeeding	No=9 Yes=8	No=72 Yes=161
7	Shape	Breast shape with three states: oval, round and irregular, which can be different based on genetics, age, weight, and hormone level.	Oval=3 Round=6 Irregular=8	Oval=34 Round=47 Irregular=152
8	Menstruation	Presence/absence of regular menstruation according to age	No=5 Yes=12	No=37 Yes=196
9	Birth control pills	Taking/not taking birth control pills	No=13 Yes=4	No=142 Yes=91
10	Heredity	Inheritance was divided into three groups. People who have no family history of cancer. People with a history of other cancers and people with a family history of breast cancer	No=8 Yes (Breast)=3 Yes (Others)=6	No=44 Yes (Breast)=49 Yes (Others)=140
11	Marital status	Presence/absence of marriage history	Single=2 Married=15	Single=40 Married=193
12	Related features	Presence/absence of the following as related features in the patient's records: skin thickening, skin shrinkage, nipple shrinkage, structural distortion, axillary adenopathy, and calcium masses.	Skin thickening=3 Skin retraction=4 Nipple retraction=5 Architectural distortion=2 Axillary adenopathy=2 Calcification=1	Skin thickening=48 Skin retraction=65 Nipple retraction=21 Architectural distortion=26 Axillary adenopathy=31 Calcification=42
13	Menopause	Entering/not entering the menopause period	No=8 Yes=9	No=186 Yes=47

moved, except for negations. For instance, in the sentence "No tangible mass in the breast or axillary is seen," the negative term is retained. Numbers, both integer and decimal, were converted to their corresponding text form. To maintain local dependencies, a bigram collection of word pairs was created based on mutual information. To enhance the accuracy of word embeddings, bigrams that appeared less than 50 times were discarded, while those occurring more than 1,000 times were treated as single words.

Next, key terms were extracted using a predefined dictionary. If a negation appeared in a sentence, its meaning was inverted, or the corresponding vector was reversed. For example, in the previous sentence, "tangible mass" could indicate the absence of breast cancer. If this phrase is in the dictionary, its meaning is

Table 3: Patients distribution according to BIRADS class

Class	Number of patients
BIRADS 0	9
BIRADS 1	17
BIRADS 2	24
BIRADS 3	21
BIRADS 4	78
BIRADS 5	69
BIRADS 6	32
Total	250

flipped, and if no opposite term is found, the Word2vec vector is reversed. To reduce ambiguities and enhance the semantic precision of the reports, domain ontology was applied in the text processing phase. A lexical crawler Banerjee and *et al.* (2019) was employed to identify derived terms sharing a common root with predefined terms, which were then mapped to controlled terms (key terms). Along with the dictionary, commonly available terms (CLEVER) Banerjee and *et al.* (2019) were also used to help identify clinical contexts and map them appropriately.

After merging the key terms with those obtained from CLEVER, a total of 260 key terms were created. These terms serve two main purposes: (a) to shorten reports through mapping, and (b) to assist in generating text-aware vectors. An unsupervised method was utilized to create word embeddings with the Word2vec model Guo and *et al.* (2019). To train Word2vec, the Skip-gram technique was used with a vector length of 210 and a window size of 8. Each report was represented using the selected key terms, and the mean of all corresponding word vectors was computed to obtain the final report vector, as shown in Equation 3.1.

After merging key terms from CLEVER, a total of 260 key terms were obtained, primarily serving two purposes: (a) reducing report complexity through mapping and (b) aiding in the generation of text-aware vectors. An unsupervised method was employed to generate word embeddings using the Word2vec model Guo and *et al.* (2019). To train Word2vec, Skipgram with vector length 210 and window width 8 was used. In each report, selected key terms were used to describe that text. The final text representation was obtained by averaging all the extracted vectors. Each report vector was calculated based on equation 3.1.

$$V_{MTR} = \frac{1}{N} \sum_{i=1}^{N} V_{W_i} \tag{3.1}$$

Here, V_{MTR} is the report's vector, N is the number of words selected from the report, and $V_(W_i)$ is the vector of each word obtained from Word2vec.

3.2.2 HIS feature selection

Breast cancer specialists were consulted to identify the most relevant features for diagnosing breast cancer based on data obtained from the hospital information system (HIS). HIS data were extracted from the picture archiving and communication system (PACS) and the electronic files of patients at Shahidzadeh Hospital Medical Training Center in Behbahan, Iran from 2020 to 2022. The electronic records encompass medical documents, images, and reports stored in PACS. HIS is a comprehensive information system that covers various aspects of hospital operations, including financial, patient health, legal, and administrative services. The database incorporates information from the PACS system used in medical training centers.

3.3 The third phase: Clustering

Breast cancer is one of the heterogeneous diseases characterized by a variety of molecular subtypes, which are classified based on receptor and immunochemical status. Key receptors involved include the estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), the proliferation marker Ki67, and the epidermal growth factor receptor (EGFR). Breast cancer can be categorized into four primary molecular subtypes: luminal A, luminal B, HER2, and basal-like molecular class (BLBC). Each subtype is associated with different rates of recurrence and survival, which are critical factors in determining the appropriate treatment strategies Kao and *et al.* (2011).

3.3.1 c-mean clustering

To identify molecular subtypes, patients need to undergo an invasive biopsy procedure to collect breast tissue. In this study, only 52 out of the 250 samples contained molecular subtype features. Given the importance of molecular subtypes in breast cancer progression, c-mean clustering was used to assign subtypes to samples without this information. This approach enables the system to be trained to accurately identify the molecular subtype for patients in the early stages of the disease who have not yet undergone a biopsy. Initially, all patients were organized into four clusters using c-mean clustering, based on the characteristics obtained in the second phase. Once the clustering process was completed, molecular subtypes were assigned to each cluster according to the values of the cluster centers. In the c-mean method, the samples are divided into c clusters, where c (the number of

Table 4: Molecular subtypes and immunophenotype

Molecular subtypes	Immunophenotype
BLBC	ER-, PR-, HER2- (triple negative), CK5/6+, and/or EGFR+
HER2	ER-, PR-, HER2+, CK5/6±
Luminal A	ER+ and/or PR+, HER2-, CK5/ 6±, and Ki67 <14%
Luminal B	ER+ and/ or PR+, CK5/ 6±, HER2+, or Ki67≥14%; or PR < 20%

molecular subtypes) is predetermined. The objective function is represented by equation 3.2.

$$J = argmin \left(\sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^{m} \left\| x_i - c_j \right\|^2 \right) \tag{3.2}$$

In equation 3.2, m is a real number greater than 1, typically set to 2. Here, n represents the number of samples, c denotes the cluster centre, u indicates the degree of membership, and x refers to the sample. To minimize the value of j, the membership degree and cluster centres are updated in each iteration using equations 3.3 and 3.4, respectively Bezdek and *et al.* (1984),Davtalab and *et al.* (2013).

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \tag{3.3}$$

$$c_j = \frac{\sum_{i=1}^{n} u_{ij}^{m} . x_i}{\sum_{i=1}^{n} u_{ij}^{m}} \tag{3.4}$$

The clustering process followed these key criteria: (1) Only individuals diagnosed with breast cancer were included, and (2) each cluster was assigned to one of four molecular subtypes based on the immunohistochemical results obtained after surgery or biopsy, as outlined by the 13th St. Gallen International Breast Cancer Conference in 2013 Kao and *et al.* (2011). The relationship between molecular subtypes and immunophenotypes is presented in Table 4.

Thus, a logical connection between BIRADS classification and molecular subtypes was established. With a classifier, molecular subtypes can now be identified along with their associated probabilities using BIRADS data.

3.4 The Fourth Phase: Classification

3.4.1 Convolutional Neural Network (CNN)

Machine learning algorithms perform well within practical computational limits, enabling valuable data-driven insights Alesheykh (2016). Convolutional Neural

Networks (CNNs) play a crucial role in machine learning and are widely applied in image, speech, and text processing Kalchbrenner and Blunsom (2014). In this study, CNN is used as a classifier for BIRADS detection, showcasing its capability to capture complex relationships between variables while effectively handling noisy data. The input is processed through convolution operations, followed by pooling layers to decrease dimensionality and reduce the risk of overfitting Xu and *et al.* (2019). During backpropagation, the parameters are optimized through error minimization. ReLU is generally applied as the activation function in the first and second convolution layers, while the softmax function is used in the output layer. The loss function utilized is mean squared error, and the optimization algorithm used is Adam Jais and *et al.* (2019), which is recognized for its adaptive learning rate.

3.4.2 Decision Tree (DT)

Decision tree learning is a widely used supervised machine learning technique for classification and regression tasks. In these tree structures, the leaves represent class labels, while the branches signify combinations of features that lead to those labels Hastie and Tibshirani (2001); Provost and Fawcett (2013); Piryonesi and El-Diraby (2020). Decision trees are constructed by minimizing entropy, a measure of uncertainty in data Hastie and Tibshirani (2001); Piryonesi and El-Diraby (2020). While early decision tree models were limited to discrete variables, modern algorithms can handle both discrete and continuous variables Piryonesi and El-Diraby (2020); Wu and *et al.* (2008). The purpose of a decision tree is to predict the value of a target variable based on input measurements. A key advantage of decision trees is their simplicity and interpretability, making them widely used Piryonesi and El-Diraby (2020); Wu and *et al.* (2008); Piryonesi and El-Diraby (2018). However, they also have limitations, such as a lack of robustness and suboptimal accuracy Wu and *et al.* (2008). In this study, decision trees are also applied to classification and BIRADS detection.

3.4.3 Multi-Level Fuzzy Min-Max Neural Network (MLF)

MLF is an improved version of the Fuzzy Min-Max Neural Network Davtalab and *et al.* (2013), which employs "hyper-boxes" to classify samples. A hyper-box is an n-dimensional structure defined by a minimum point, a maximum point, and a membership function, with each hyper-box representing a distinct class. During training, hyper-boxes are dynamically created and adjusted as new samples are introduced. The definition of a hyper-box is given by Equation 3.5:

$$B_j = \{X, V_j, W_j, f\,(X, V_j, W_j)\,\forall X \in I^n\} \tag{3.5}$$

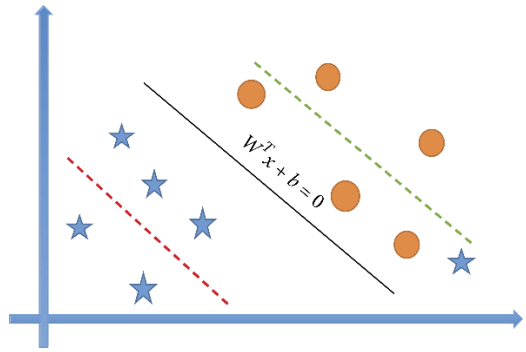

Figure 3: SVM hyperplane

V_j and W_j represent the upper and lower boundaries of a hyper-box, respectively. X refers to an individual sample, while n indicates the number of dimensions in the feature vectors. The sizes of these hyper-boxes are controlled by Equation 3.6:

$$\forall_{i=1...D} \left(max \left(w_b^i, x^i \right) - min \left(v_b^i, x^i \right) \right) \leq \Theta \tag{3.6}$$

Equation 3.6 defines the expansion coefficient, symbolized by Θ. The algorithm consists of three layers: an input layer, a hyper-box layer, and an output layer, where class predictions are made Davtalab and *et al.* (2013). In this study, the MLF was used to train the system for effective recognition of BIRADS features.

3.4.4 Multi-class Support Vector Machine (SVM) Algorithm

The Support Vector Machine (SVM) algorithm aims to find the optimal hyperplane that maximizes the margin between two classes. This separating hyperplane is mathematically expressed through Equation 3.7 Vishwanathan and Murty (2002):

$$W^T x + b = 0 \tag{3.7}$$

In this context, x represents the input vector containing the feature data, b refers to the bias term, W is the weight vector that defines the distance between the hyperplane and the data points, and W^T stands for the transpose of W. Determining the optimal hyperplane involves evaluating several potential hyperplanes that can separate the class labels. The algorithm selects the one that maximizes the margin, meaning it is the furthest from the data points, as shown in Figure 3.

In this study, the Radial Basis Function (RBF) kernel was used, and after model extraction Chang and Lin (2011), class probabilities were determined based on

BIRADS. Data normalization was performed using the standard deviation method Hafemeister and Satija (2019). Seven Support Vector Machines (SVMs) were employed, each corresponding to a different BIRADS category. As shown in Table 5, seven SVMs evaluated each sample. The fourth SVM yielded the highest probability, indicating that the sample belongs to the fourth class ("Probably benign") Magny and *et al.* (2023).

Table 5: SVM values

SVM 1	SVM 2	SVM 3	SVM 4	SVM 5	SVM 6	SVM 7
0.01	0.04	0.02	0.76	0.06	0.05	0.06

3.4.5 XGBoost

XGBoost is a distributed and scalable machine learning library that leverages Gradient Boosted Decision Trees (GBDT) to enhance performance in various tasks, prioritizing both speed and efficiency. Gradient Boosting is a technique commonly used for solving regression and classification problems by combining several weak models into a single, strong predictive model through an iterative process Piryonesi and El-Diraby (2020). As part of the ensemble learning category, this method consistently outperforms simpler algorithms like decision trees or bagging techniques such as Random Forest. However, the effectiveness of Gradient Boosting can depend on the nature of the input data Piryonesi and El-Diraby (2020). In this study, XGBoost was also utilized for classification purposes, specifically for detecting BIRADS categories.

3.5 The Fifth Phase: Ensemble and Validation

In the fifth phase, as depicted in Figure 4, when a patient is referred to the treatment system, the process begins with gathering information from medical text reports (MTR), specifically mammography reports in this study, along with the patient's electronic health records from the HIS system. After performing data fusion, text processing, and clustering, the BIRADS values are predicted using various base learners, including CNN, DT, MLF, SVM, and XGBoost. These predictions are then combined using weighted ensemble learning to generate the final output.

3.5.1 Weighted Ensemble Learning

Weighted ensemble learning is a machine learning strategy where multiple models, often called "base learners," are trained to address the same problem and then

combined to improve performance in tasks like classification or regression. Each model's contribution is weighted, and the predictions are aggregated to achieve better accuracy, reduce errors, and enhance generalization. Compared to a single model, ensemble methods provide stronger and more reliable predictions García-Pedrajas and *et al.* (2005); Chen and *et al.* (2022).

A common and straightforward approach to ensemble learning is majority voting Dimitriadou and *et al.* (2001); Wang and *et al.* (2013), where the class of an object is determined by the majority decision of the individual classifiers. The decision of the t-th classifier for class j is denoted as $d_{(t,j)} \in \{0, 1\}$, where t represents the index of the classifier ($t = 1, 2, 3, ..., T$) and j refers to the class ($j = 1, 2, 3, ..., C$). In this context, T stands for the number of outputs from the base classifiers, and C indicates the total number of classes. If the t-th classifier selects class j, $d_{(t,j)}$ is assigned a value of 1; otherwise, it is set to 0. The weighted ensemble decision for class k, as computed by Equation 3.8, is determined through a majority voting process.

$$\sum_{t=1}^{T} d_{t,k} = \max_{j} \sum_{t=1}^{T} w_t d_{t,j} \tag{3.8}$$

Here, w_t represents the weight of the t-th classifier, calculated as the average accuracy of that base learner. Weighted ensemble learning is applied here to predict BIRADS values using the majority voting approach. The performance of this method is then compared to that of the individual base classification algorithms.

3.5.2 Validation

The BIRADS results generated by the base learners, as well as the final output from the weighted ensemble learning using the majority voting method, are validated using evaluation metrics derived from the confusion matrix. These metrics include accuracy, specificity, sensitivity, positive predictive value (PPV), negative predictive value (NPV), and F1-score. To assess the performance, K-fold cross-validation was employed with K set to 10. The calculations for these metrics are provided in Equations 4.13 through 4.18.

4. Analysis and Evaluation of Results

A computer with the following specifications was utilized to implement this plan:

Processor: Intel(R) Core(TM) i7-4790 CPU @ 3.60GHz
Installed memory (RAM): 2 × 8 GB DDR RAM
VGA: GT 730 2GB

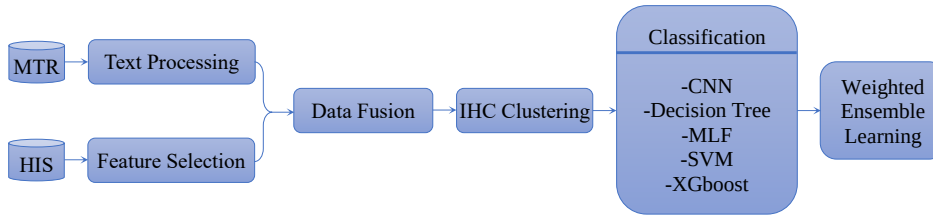

Figure 4: Roadmap of the proposed method

HDD: 256GB SSD + 1TB SATA

The research was conducted using the Microsoft Windows 10 64-bit operating system, with Python 3.8.7 employed for programming within the Visual Studio Code environment.

4.1 Evaluation Parameters

As shown in Table 6, the confusion matrix is used as one of the evaluation metrics for classifiers. It is an $N \times N$ square matrix, where N represents the number of classes—in this case, 7 classes for BIRADS. The main diagonal of the matrix indicates the number of correct detections, while the other entries represent incorrect detections.

Table 6: Confusion matrix Hafemeister and Satija (2019)

		Original/actual values		
		Original Class 1	...	Original Class j
Predicted values	Predicted Class 1	Class 1, which is correctly recognized as class 1	...	Class j, which is mistakenly recognized as class 1
	⋮	⋮	⋮	⋮
	Predicted Class j	Class 1, which is mistakenly recognized as class j	...	Class j, which is correctly recognized as class j

In binary classification models, where only the presence or absence of a disease is diagnosed, the confusion matrix includes terms such as true positive (TP), true negative (TN), false positive (FP), and false negative (FN). However, for BIRADS diagnosis, which involves seven classes, these concepts are extended.

In this context, TP_i represents the true positive value for the i-th class, referring to cases where both the actual and predicted class are i. TP_i is calculated using Equation 4.9. Similarly, FP_i represents the false positive value for the i-th class,

indicating cases where the actual class is not i, but the predicted class is i. FP_i is determined using Equation 4.10.

On the other hand, FN_i represents the false negative value for the i-th class, which occurs when the actual class is i, but the predicted class is not i. FN_i is calculated using Equation 4.11. Finally, TN_i denotes the true negative value for the i-th class, representing cases where neither the actual class nor the predicted class is i. TN_i is obtained using Equation 4.12.

$$TP_i = C_{ii} \qquad i = 0, 1, \cdots, 6 \tag{4.9}$$

$$FP_i = \sum_{i \neq j = 0}^{6} C_{ij} \qquad i = 0, 1, \cdots, 6 \tag{4.10}$$

$$FN_i = \sum_{i \neq j = 0}^{6} C_{ji} \qquad i = 0, 1, \cdots, 6 \tag{4.11}$$

$$TN_i = \sum_{i \neq j = 0}^{6} \sum_{i \neq k = 0}^{6} C_{jk} \qquad i = 0, 1, \cdots, 6 \tag{4.12}$$

The parameters including accuracy, specificity, sensitivity, positive predictive value (PPV), negative predictive value (NPV), and F1-score are calculated using Equations 4.13 to 4.18, respectively Shahabi and Hassanpour (2016); Tharwat (2021).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.13}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4.14}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{4.15}$$

$$PPV = \frac{TP}{TP + FP} \tag{4.16}$$

$$NPV = \frac{TN}{TN + FN} \tag{4.17}$$

$$F1 - measure = \frac{2 \times PPV \times Sensitivity}{PPV + Sensitivity} \tag{4.18}$$

where TP, TN, FP, and FN denote as mean of TP_i, TN_i, FP_i, and FN_i respectively.

Table 7: Methods runtime

	Runtime (ms)	
	Overall	Average
CNN	349765	11659
Decision Tree	5292	176
MLF	17581	586
SVM	9711	324
XGboost	15033	501
Proposed DSS	5859	195

Table 8: Descriptives statistics

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
CNN	30	.845371429	.0121918513	.0022259173	.840818916	.849923941	.8068571	.8674286
Decision Tree	30	.885409524	.0101376684	.0018508766	.881624056	.889194991	.8617143	.8982857
MLF	30	.839619048	.0201697697	.0036824793	.832087532	.847150563	.8034286	.8822857
SVM	30	.888723810	.0146771464	.0026796680	.883243273	.894204346	.8525714	.9120000
XGboost	30	.853676190	.0105664302	.0019291574	.849730621	.857621760	.8365714	.8777143
Proposed DSS	30	.898780952	.0076145471	.0013902197	.895937634	.901624271	.8788571	.9108571

4.2 Assessment of Methods

In this section, the methods discussed in this article were evaluated using statistical hypothesis testing. Each method was executed 30 times under identical conditions, and the accuracy of all methods was recorded for each run. The overall and average execution times for each method are shown in Table 7.

Subsequently, an ANOVA test was performed using SPSS version 25 to compare the six methods. The results of this analysis are presented in Tables 8 and 9.

To compare the accuracy of the methods, the ANOVA test with Fisher's F statistic was applied, with the hypothesis defined in Equation 4.19, as follows:

$$\begin{cases} H_0: & \mu_{CNN} = \mu_{Decision\ Tree} = \mu_{MLF} = \mu_{SVM} = \mu_{XGboost} = \mu_{Proposed\ DSS} \\ H_A: & At\ least\ one\ of\ the\ means\ is\ different\ from\ the\ others. \end{cases} \tag{4.19}$$

Based on the results in Table 9, the p-value ($Sig = 0.000$) is less than the significance level ($\alpha = 0.05$), leading to the rejection of the null hypothesis. This indicates that at least one of the means differs from the others.

To identify the source of this difference, post hoc tests were conducted. As shown in Tables 10 and 11, the p-values associated with the proposed DSS, highlighted in yellow, are all less than 0.05. This result leads to the rejection of the null hypothesis in all cases, confirming a significant difference between the mean

Table 9: ANOVA result

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.096	5	.019	110.441	.000
Within Groups	.030	174	.000		
Total	.126	179			

of the proposed DSS and the other five methods.

The post hoc tests (Tukey's HSD and LSD) further validate the presence of a significant difference between the proposed DSS and the other methods.

Figure 5 also illustrates the superior performance of the proposed DSS method compared to the other approaches.

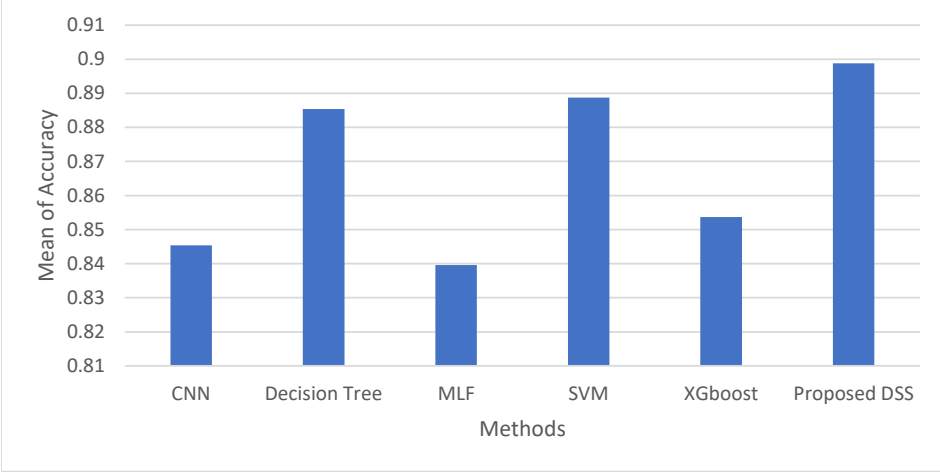

Figure 5: Mean of accuracy for all methos

4.3 Results

Figures 6 through 11 illustrate the accuracy, specificity, PPV, NPV, sensitivity, and F1-measure for various methods, including Convolutional Neural Network (CNN), Decision Tree, Multi-Level Fuzzy Min-Max Neural Network (MLF), Support Vector Machine (SVM), XGBoost, and the proposed Decision Support System (DSS) for BIRADS detection, using only text mining.

It is evident that as the dimensions of the resulting vector increase, the classification accuracy improves. However, beyond 160 dimensions, there is a noticeable decline in accuracy. This behavior aligns with findings from previous studies (e.g.,

Table 10: Multiple Comparisons for Tukey HSD

(I) Methods	(J) Methods	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
CNN	Decision Tree	-.0400380952	.0034047068	.000	-.049849586	-.030226604
	MLF	.0057523809	.0034047068	.541	-.004059110	.015563872
	SVM	-.0433523810	.0034047068	.000	-.053163872	-.033540890
	XGboost	-.0083047619	.0034047068	.149	-.018116253	.001506729
	Proposed DSS	-.0534095239	.0034047068	.000	-.063221015	-.043598033
Decision Tree	CNN	.0400380952	.0034047068	.000	.030226604	.049849586
	MLF	.0457904761	.0034047068	.000	.035978985	.055601967
	SVM	-.0033142858	.0034047068	.926	-.013125776	.006497205
	XGboost	.0317333332	.0034047068	.000	.021921843	.041544824
	Proposed DSS	-.0133714288	.0034047068	.002	-.023182919	-.003559938
MLF	CNN	-.0057523809	.0034047068	.541	-.015563872	.004059110
	Decision Tree	-.0457904761	.0034047068	.000	-.055601967	-.035978985
	SVM	-.0491047619	.0034047068	.000	-.058916253	-.039293271
	XGboost	-.0140571429	.0034047068	.001	-.023868634	-.004245652
	Proposed DSS	-.0591619049	.0034047068	.000	-.068973396	-.049350414
SVM	CNN	.0433523810	.0034047068	.000	.033540890	.053163872
	Decision Tree	.0033142858	.0034047068	.926	-.006497205	.013125776
	MLF	.0491047619	.0034047068	.000	.039293271	.058916253
	XGboost	.0350476190	.0034047068	.000	.025236128	.044859110
	Proposed DSS	-.0100571430	.0034047068	.041	-.019868634	-.000245652
XGboost	CNN	.0083047619	.0034047068	.149	-.001506729	.018116253
	Decision Tree	-.0317333332	.0034047068	.000	-.041544824	-.021921843
	MLF	.0140571429	.0034047068	.001	.004245652	.023868634
	SVM	-.0350476190	.0034047068	.000	-.044859110	-.025236128
	Proposed DSS	-.0451047620	.0034047068	.000	-.054916253	-.035293271
Proposed DSS	CNN	.0534095239	.0034047068	.000	.043598033	.063221015
	Decision Tree	.0133714288	.0034047068	.002	.003559938	.023182919
	MLF	.0591619049	.0034047068	.000	.049350414	.068973396
	SVM	.0100571430	.0034047068	.041	.000245652	.019868634
	XGboost	.0451047620	.0034047068	.000	.035293271	.054916253

Table 11: Multiple Comparisons for Tukey LSD

(I) Methods	(J) Methods	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
CNN	Decision Tree	-.0400380952	.0034047068	.000	-.046757936	-.033318254
	MLF	.0057523809	.0034047068	.093	-.000967460	.012472222
	SVM	-.0433523810	.0034047068	.000	-.050072222	-.036632540
	XGboost	-.0083047619	.0034047068	.016	-.015024603	-.001584921
	Proposed DSS	-.0534095239	.0034047068	.000	-.060129365	-.046689683
Decision Tree	CNN	.0400380952	.0034047068	.000	.033318254	.046757936
	MLF	.0457904761	.0034047068	.000	.039070635	.052510317
	SVM	-.0033142858	.0034047068	.332	-.010034127	.003405555
	XGboost	.0317333332	.0034047068	.000	.025013493	.038453174
	Proposed DSS	-.0133714288	.0034047068	.000	-.020091269	-.006651588
MLF	CNN	-.0057523809	.0034047068	.093	-.012472222	.000967460
	Decision Tree	-.0457904761	.0034047068	.000	-.052510317	-.039070635
	SVM	-.0491047619	.0034047068	.000	-.055824603	-.042384921
	XGboost	-.0140571429	.0034047068	.000	-.020776984	-.007337302
	Proposed DSS	-.0591619049	.0034047068	.000	-.065881746	-.052442064
SVM	CNN	.0433523810	.0034047068	.000	.036632540	.050072222
	Decision Tree	.0033142858	.0034047068	.332	-.003405555	.010034127
	MLF	.0491047619	.0034047068	.000	.042384921	.055824603
	XGboost	.0350476190	.0034047068	.000	.028327778	.041767460
	Proposed DSS	-.0100571430	.0034047068	.004	-.016776984	-.003337302
XGboost	CNN	.0083047619	.0034047068	.016	.001584921	.015024603
	Decision Tree	-.0317333332	.0034047068	.000	-.038453174	-.025013493
	MLF	.0140571429	.0034047068	.000	.007337302	.020776984
	SVM	-.0350476190	.0034047068	.000	-.041767460	-.028327778
	Proposed DSS	-.0451047620	.0034047068	.000	-.051824603	-.038384921
Proposed DSS	CNN	.0534095239	.0034047068	.000	.046689683	.060129365
	Decision Tree	.0133714288	.0034047068	.000	.006651588	.020091269
	MLF	.0591619049	.0034047068	.000	.052442064	.065881746
	SVM	.0100571430	.0034047068	.004	.003337302	.016776984
	XGboost	.0451047620	.0034047068	.000	.038384921	.051824603

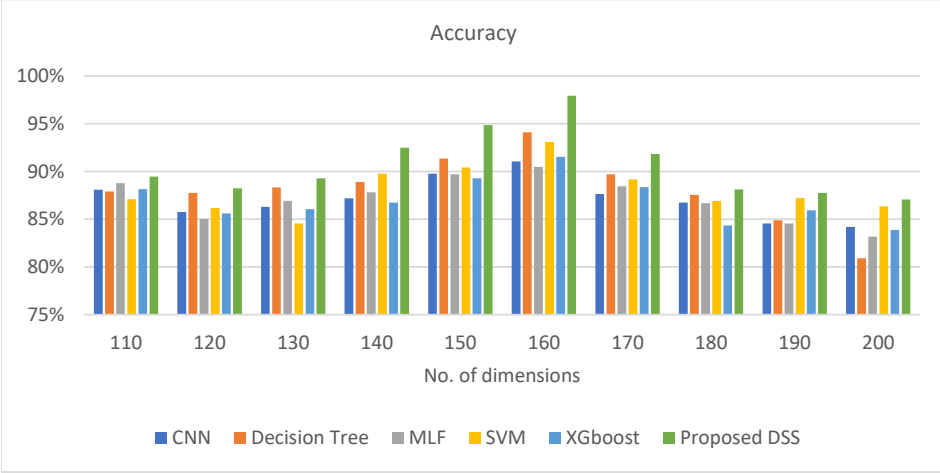

Figure 6: Variations of accuracy with the change of dimensions in the vector resulting from word2vec

Bofang and *et al.* (2019)), which suggest that increasing dimensions can negatively affect the quality of word2vec, and consequently, the accuracy. This issue was further examined by adjusting the dimensions, and 160 dimensions were ultimately selected for further processing, as they produced the best results.

Figure 6 shows the variation in accuracy for all classifiers used in this research across dimensions ranging from 110 to 200. The proposed DSS achieved the highest accuracy of 97.94% when using a dimension of 160. At the same dimension, the accuracies for CNN, Decision Tree (DT), Multi-Level Fuzzy Min-Max Neural Network (MLF), Support Vector Machine (SVM), and XGBoost were 91.06%, 94.11%, 90.49%, 93.09%, and 91.54%, respectively.

Figure 7 illustrates the variation in specificity for all classifiers across the specified dimensions. The proposed decision support system (DSS) achieved the highest specificity at 98.79% in dimension 160. In the same dimension, the specificity for CNN, Decision Tree (DT), MultiLevel Fuzzy Min-Max Neural Network (MLF), Support Vector Machine (SVM), and XGboost were 96.79%, 97.83%, 96.38%, 97.67%, and 97.34%, respectively.

Figure 8 shows the variation in sensitivity for all classifiers within the mentioned dimensions. The proposed DSS reached its highest sensitivity at 92.08% in dimension 160. In the same dimension, the sensitivity for CNN, DT, MLF, SVM, and XGBoost was 84.62%, 84.49%, 80.45%, 87.45%, and 87.28%, respectively.

Figure 9 presents the variation in positive predicted value (PPV) across the specified dimensions. The best PPV for the proposed DSS occurred in dimension

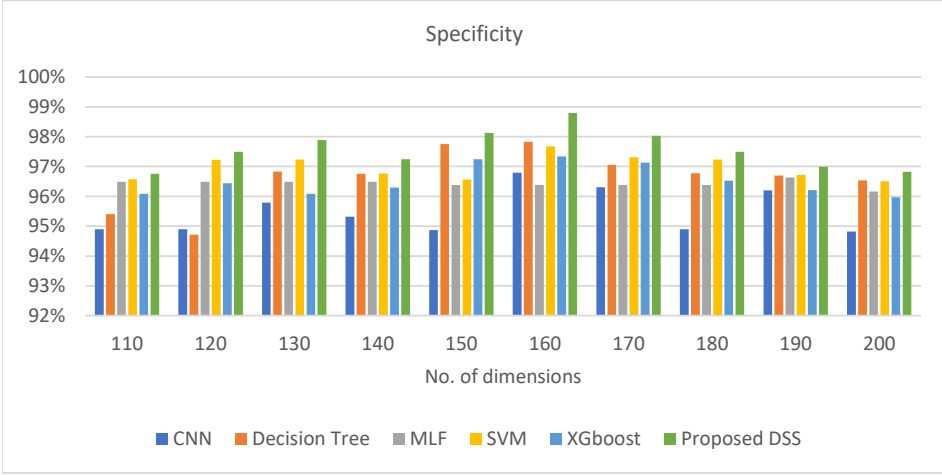

Figure 7: Variations of specificity with the change of dimensions in the vector resulting from word2vec

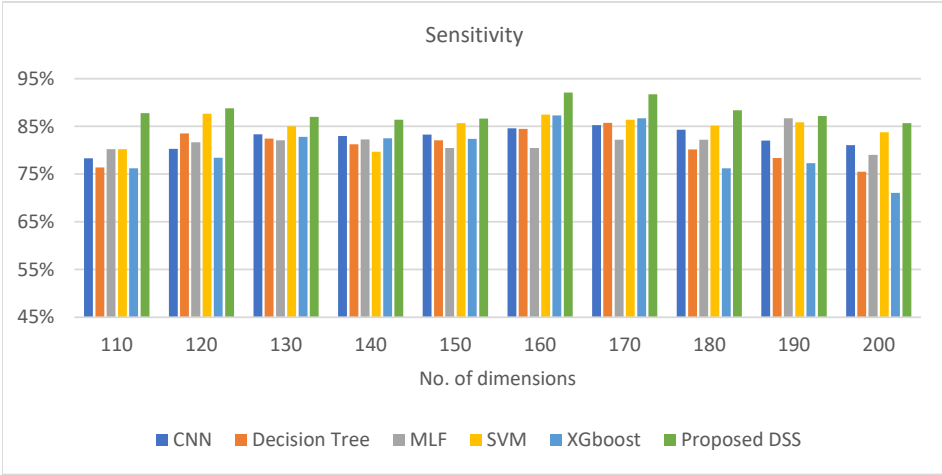

Figure 8: Variations of sensitivity with the change of dimensions in the vector resulting from word2vec

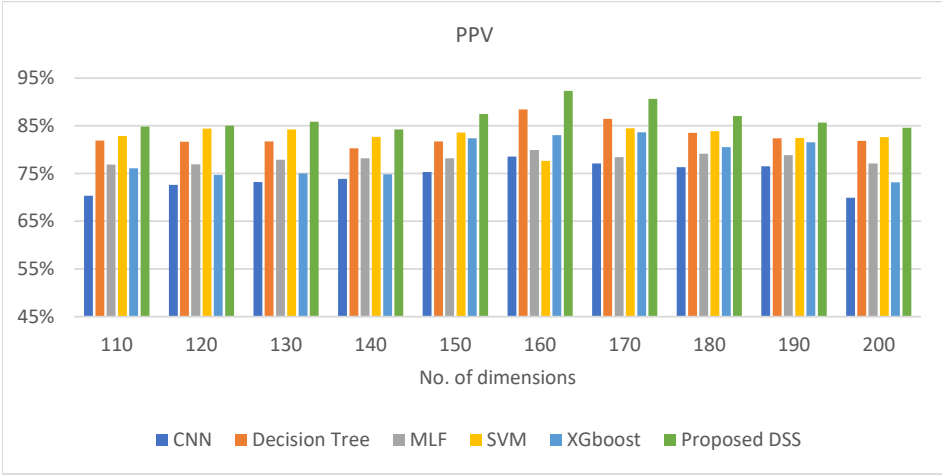

Figure 9: Variations of PPV with the change of dimensions in the vector resulting from word2vec

Table 12: Confusion matrix of proposed DSS

Confusion Matrix						
20	0	1	0	1	0	1
1	20	0	0	0	1	0
0	1	34	0	0	1	1
1	0	0	36	1	0	0
0	1	1	0	49	0	0
1	0	0	1	1	44	1
1	0	0	1	0	0	27

ID Class	Sensitivity	Specificity	PPV	NPV	F1 Measure	Accuracy
Class1	84.62%	98.66%	88.00%	98.22%	86.27%	97.20%
Class2	90.91%	99.12%	90.91%	99.12%	90.91%	98.40%
Class3	94.44%	98.60%	91.89%	99.06%	93.15%	98.00%
Class4	94.74%	99.06%	94.74%	99.05%	94.74%	98.40%
Class5	94.23%	98.99%	96.08%	98.49%	95.15%	98.00%
Class6	95.65%	98.04%	91.67%	99.01%	93.62%	97.60%
Class7	90.00%	99.09%	93.10%	98.64%	91.53%	98.00%

160 with a value of 92.34%. In that same dimension, the PPV for CNN, DT, MLF, SVM, and XGboost was 78.53%, 88.43%, 79.91%, 77.68%, and 83.07%, respectively.

Figure 10 displays the variation in negative predicted value (NPV) for all classifiers in the given dimensions. The proposed DSS showed the highest NPV at 98.80% in dimension 160. In the same dimension, the NPV for CNN, DT, MLF, SVM, and XGboost was 95.86%, 97.87%, 97.40%, 97.26%, and 97.54%, respectively.

Figure 11 presents the variation in the F1-measure for all classifiers across the specified dimensions. The proposed Decision Support System (DSS) achieved the highest F1-measure of 92.19% at a dimension of 160. At the same dimension, the F1-measures for CNN, Decision Tree (DT), Multi-Level Fuzzy Min-Max Neural Network (MLF), Support Vector Machine (SVM), and XGBoost were 78.24%, 88.63%, 82.79%, 87.97%, and 86.10%, respectively.

Table 12 summarizes the sensitivity, specificity, positive predictive value (PPV),

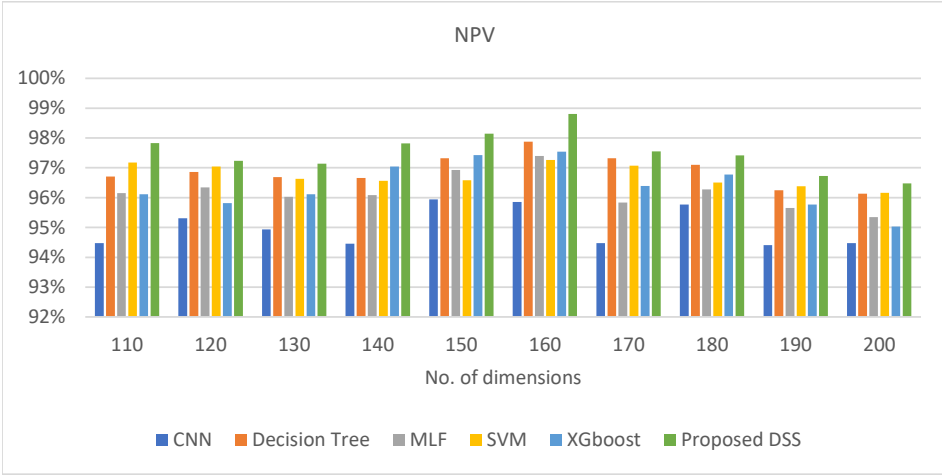

Figure 10: Variations of NPV with the change of dimensions in the vector resulting from word2vec

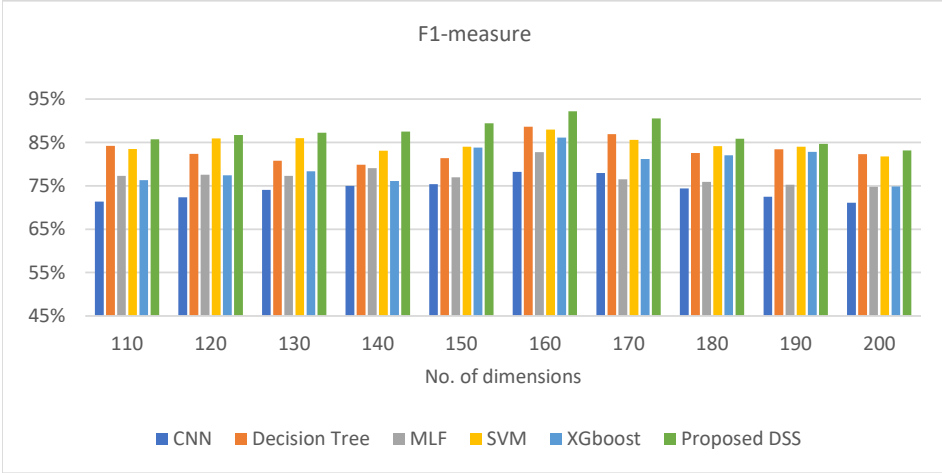

Figure 11: Variations of f1-measure with the change of dimensions in the vector resulting from word2vec

negative predictive value (NPV), F1-measure, and accuracy for BIRADS classification using the proposed DSS. The values for classes one through seven correspond to BIRADS levels zero through six. Most disease classes were diagnosed with an accuracy exceeding 98%. The disease class corresponding to BIRADS = 5 (the sixth class) exhibited the highest sensitivity at 95.65%.

The specificity for healthy individuals was 99.12%, demonstrating strong performance in correctly identifying healthy cases. The average specificity was 98.79%, with the lowest value observed for the third class and the highest for the seventh class (99.09%). These results indicate that the proposed method excels in specificity.

The average PPV value was 92.34%, with the highest PPV of 96.08% observed in the fifth class.

The NPV for healthy individuals was 99.12%, confirming the proposed method's strong ability to correctly identify healthy cases. The highest NPV value, 99.06%, was found in the third class, while the lowest was 98.22% in the first class.

The average F1-measure was 92.19%, with the maximum value reaching 95.15% in the fifth class, and the minimum at 86.27% in the first class, reflecting a solid detection rate across all classes.

The average accuracy of the proposed method in differentiating between sick and healthy cases was 97.94%. The accuracy ranged from a minimum of 97.20% to a maximum of 98.40%, demonstrating the method's consistency in classification.

In conclusion, these evaluation metrics demonstrate that the proposed method performs effectively in detecting BIRADS classes, significantly aiding in disease diagnosis and the determination of appropriate treatment strategies. The integration of HIS (Hospital Information System) values alongside text processing results further enhances BIRADS detection performance, contributing to more accurate and reliable outcomes.

Table 13: Comparison between proposed DSS (Weighted Ensemble Learning) and old method (Ensemble Learning)

Metric	Proposed DSS	Old Method
Accuracy	97.94%	89.35%
Specificity	98.79%	65.87%
Sensitivity	97.08%	87.41%
Positive Predicted Value (PPV)	92.34%	75.19%
Negative Predicted Value (NPV)	98.80%	93.81%
F1-Measure	92.19%	72.69%

Based on the metrics presented in Table 13, it is clear that the proposed method

significantly outperforms the old method in almost all evaluation criteria:

1. **Accuracy**: The proposed method achieves 97.94%, demonstrating a substantial improvement over the old method's 89.35%. This indicates better overall performance in correctly classifying results. 2. **Specificity**: The proposed method's specificity is 98.79%, compared to the old method's 65.98%. This indicates that the new method is much more effective in correctly identifying negative cases. 3. **Sensitivity**: The proposed method achieves a sensitivity of 92.08%, which is slightly higher than the old method's 87.41%. This indicates a better ability to identify positive cases correctly. 4. **Positive Predicted Value (PPV)**: The proposed method has a PPV of 92.34%, significantly outperforming the old method's 75.19%, indicating higher reliability in predicting positive cases. 5. **Negative Predicted Value (NPV)**: The proposed method achieves 98.80%, while the old method scores 93.81%, highlighting superior accuracy in predicting negative cases. 6. **F1-Measure**: The F1-measure of the proposed method is 92.19%, compared to the old method's 72.69%. This metric balances precision and recall, underscoring the better-rounded performance of the proposed method.

5. Discussion and Conclusion

The American College of Radiology (ACR) introduced the BIRADS system to standardize mammography reports, significantly contributing to the uniformity of reports and improving the consistency of treatment planning. This standardization allows for more precise prioritization of treatment progress. However, this approach also has limitations, such as varying interpretations among physicians when determining BIRADS values. To address this issue, this study proposes leveraging information from electronic health records (EHRs). By combining unstructured data (mammography reports) with structured data (electronic records from the Hospital Information System or HIS), a hybrid approach is used to enhance the accuracy and reliability of BIRADS classification.

After preprocessing the mammography reports, keywords were transformed into vectors using Word2vec, with the average vectors of the keywords representing each text. This process resulted in a 210-dimensional vector for each report. Additionally, 15 features were selected from the patients' electronic health records, which included 2 numerical variables and 13 nominal variables. These features were combined with the 210-dimensional vectors from the mammography reports, yielding a total of 225 features for classification. The BIRADS classes were determined using various algorithms such as CNN, DT, MLF, SVM, and XGboost. The estimated BIRADS classes were then aggregated using weighted ensemble learning with a majority voting algorithm. The performance of the model was evaluated

using several metrics, including sensitivity, specificity, PPV, NPV, F1-measure, and accuracy. The highest evaluation scores for BIRADS estimation were 95.65% for sensitivity, 99.12% for specificity, 96.08% for PPV, 99.12% for NPV, 95.15% for F1-measure, and 98.40% for accuracy.

The accuracy of BIRADS detection using the proposed method is 97.94%. This decision support system (DSS) enhances the physician's ability to make more informed decisions by integrating data from both the Hospital Information System (HIS) and medical text reports. Compared to similar approaches, this method improves the detection of diseases, the assessment of a patient's health status, and the determination of the severity of the condition. As a result, physicians can more accurately tailor the treatment plans for individual patients.

In this study, data fusion was employed to enhance accuracy. For future research, it is advised to employ weighted base learners to boost system efficiency. Additionally, since mammography images provide valuable insights to physicians, it is suggested to explore decision fusion, image analysis, and deep learning integration techniques. This approach could lead to more precise estimations of disease severity, ultimately aiding physicians in making better-informed treatment decisions.

Compliance with Ethical Standards

In this study, there were no conflicts of interest, and none of the authors had any direct contact with the patients. Additionally, the authors did not have access to the patients' identities.

References

Balakumar, P., K. Maung-U, and G. Jagadeesh (2016), Prevalence and prevention of cardiovascular disease and diabetes mellitus, *Pharmacological research,* **113**, 600-609.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018), Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA: a cancer journal for clinicians,* **68(6)**, 394-424.

U.S. Cancer Statistics Working Group (2024), U.S. Cancer Statistics Data Visualizations Tool, *U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute; https://www.cdc.gov/cancer/dataviz*, released in June 2024.

Isfahani, P., S.M. Hossieni Zare, and M. Shamsaii (2020), The Prevalence of Depression in Iranian Women With Breast Cancer: A Meta-Analysis. *Quarterly of Horizon of Medical Sciences,* **26(2)**, 170-181.

Dehghan P, Mogharabi M, Zabbah I, Layeghi K, Maroosi A. (2018), Modeling Breast Cancer Using Data Mining Methods, *Journal of Health and Biomedical Informatics,* **4 (4)**266-278.

Ginsburg, O., Yip, C. H., Brooks, A., Cabanes, A., Caleffi, M., Dunstan Yataco, J. A., ... , and Anderson, B. O. (2020), Breast cancer early detection: a phased approach to implementation, *Cancer,* **126**, 2379-2393.

Alaa, A. M., Moon, K. H., Hsu, W., and Van Der Schaar, M. (2016), Confidentcare: A clinical decision support system for personalized breast cancer screening, *IEEE Transactions on Multimedia,* **18(10)**, 1942-1955.

Mazo, Claudia, Cathriona Kearns, Catherine Mooney, and William M. Gallagher (2020), Clinical Decision Support Systems in Breast Cancer: A Systematic Review, *Cancers,* **12(2)**, 369.

Sim, L.L.W., Ban K.H.K., Tan T.W., Sethi S.K., Loh T.P. (2017), Development of a clinical decision support system for diabetes care: A pilot study, *PLOS ONE,* **12(2)**, 1-15.

Park, J., Rho, M.J., Moon, H.W., Park, Y.H., Kim, C.S., Jeon, S.S., Kang, M. and Lee, J.Y. (2021), Prostate cancer trajectory-map: clinical decision support system for prognosis management of radical prostatectomy, *Prostate International,* **9(1)**, 25-30.

Esmaeili, M., Ayyoubzadeh, S. M., Ahmadinejad, N., Ghazisaeedi, M., Nahvijou, A., and Maghooli, K. (2020), A decision support system for mammography reports interpretation. *Health Information Science and Systems,* **8**, 1-8.

Magny SJ, Shikhman R, Keppke AL. (2023), Breast Imaging Reporting and Data System. [Updated 2023 Aug 28]. *StatPearls [Internet], Treasure Island (FL): StatPearls Publishing,* Available from: https://www.ncbi.nlm.nih.gov/books/NBK459169/.

Farrokh, D., Alamdaran, S. A., Feizy, A., and Soleimany, H. (2019), Diagnostic value of BIRADS method using sonography in evaluating the level of malignancy of breast masses compared with biopsy. *The Iranian Journal of Obstetrics, Gynecology and Infertility,* **22(6)**, 1-6.

Vanderheyden R. and Xie Y. (2020), Mammography Image BI-RADS Classification Using OHPLall. *IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService), Oxford, UK,* 120-127.

Gao, H., Bowles, E. J. A., Carrell, D., and Buist, D. S. (2015), Using natural language processing to extract mammographic findings, *Journal of biomedical informatics,* **54**, 77-84.

Zhang, X., Zhang, Y., Zhang, Q., Ren, Y., Qiu, T., Ma, J., and Sun, Q. (2019), Extracting comprehensive clinical information for breast cancer using deep learning methods, *International Journal of Medical Informatics,* **132**, 103985.

Achilonu, O. J., Singh, E., Nimako, G., Eijkemans, R. M., and Musenge, E. (2022), Rule-Based Information Extraction from Free-Text Pathology Reports Reveals Trends in South African Female Breast Cancer Molecular Subtypes and Ki67 Expression, *BioMed Research International,* **2022(1)** 6157861.

Higa, A. (2018), Diagnosis of Breast Cancer using Decision Tree and Artificial Neural Network Algorithms, *International Journal of Computer Applications Technology and Research,* **07(01)**, 023-027.

Spaeth, E. L., Dite, G. S., Hopper, J. L., and Allman, R. (2023), Validation of an abridged breast cancer risk prediction model for the general population, *Cancer Prevention Research,* **16(5)**, 281-291.

Castro, S. M., Tseytlin, E., Medvedeva, O., Mitchell, K., Visweswaran, S., Bekhuis, T., and Jacobson, R. S. (2017), Automated annotation and classification of BI-RADS assessment from radiology reports, *Journal of biomedical informatics,* **69**, 177-187.

Gupta, A., Banerjee, I., and Rubin, D. L. (2018), Automatic information extraction from unstructured mammography reports using distributed semantics. *Journal of biomedical informatics,* **78**, 78-86.

Dorothy A. Sippo, Graham I. Warden, Katherine P. Andriole, Ronilda Lacson, Ichiro Ikuta, Robyn L. Birdwell, Ramin Khorasani (2013), Automated extraction of BI-RADS final assessment categories from radiology reports with natural language processing, *Journal of Digital Imaging,* **26(5)**, 989-994.

Bethany Percha, Houssam Nassif, Jafi Lipson, Elizabeth Burnside, Daniel Rubin (2012), Automatic classification of mammography reports by BI-RADS breast tissue composition class. *Journal of the American Medical Informatics Association,* **19(5)**, 913–916.

Nassif, H., Cunha, F., Moreira, I.C., Cruz-Correia, R., Sousa, E., Page, D., Burnside, E. and Dutra, I. (2012), Extracting BI-RADS features from Portuguese clinical texts, *IEEE International Conference on Bioinformatics and Biomedicine,* 1-4.

Selen Bozkurt, Francisco Gimenez, Elizabeth S. Burnside, Kemal H. Gulkesen, Daniel L. Rubin (2016), Using automatically extracted information from mammography reports for decision-support, *Journal of Biomedical Informatics,* **62**, 224-231.

Guo, D., Wang, Q., Liang, M., Liu, W., and Nie, J. (2019), Molecular cavity topological representation for pattern analysis: A NLP analogy-based word2vec method, *International Journal of Molecular Sciences,* **20(23)**, 6019.

Loper, E. and S. Bird (2002), Nltk: The natural language toolkit, *arXiv preprint,* **cs/0205028**.

Imon Banerjee, Selen Bozkurt, Emel Alkim, Hersh Sagreiya, Allison W. Kurian, Daniel L. Rubin (2019), Automatic inference of BI-RADS final assessment categories from narrative mammography report findings, *Journal of Biomedical Informatics,* **92**, 103137.

Kao, K. J., Chang, K. M., Hsu, H. C., and Huang, A. T. (2011), Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization, *BMC cancer,* **11(1)**, 1-15.

Bezdek, J.C., R. Ehrlich, and W. Full (1984), FCM: The fuzzy c-means clustering algorithm, *Computers & Geosciences,* **10(2)**, 191-203.

Chang, C. C. and C. J. Lin (2011), LIBSVM: A library for support vector machines, *ACM transactions on intelligent systems and technology (TIST),* **2(3)**, 1-27.

Hafemeister, C. and R. Satija (2019), Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression, *Genome biology,* **20(1)**, 1-15.

Vishwanathan, S.V.M. and M.N. Murty (2002), SSVM: a simple SVM algorithm, *Proceedings of the 2002 International Joint Conference on Neural Networks, IJCNN'02, (Cat. No.02CH37290).*

Davtalab, R., M.H. Dezfoulian, and M. Mansoorizadeh (2013), Multi-level fuzzy min-max neural network classifier, *IEEE transactions on neural networks and learning systems,* **25(3)**, 470-482.

Shahabi, M. and H. Hassanpour, Using the Artificial Intelligence Techniques for Diagnosing of intensity of Non-Alcoholic Fatty Liver Disease by Clinical Parameters, *Journal of Knowledge & Health,* **11(3)**, 69-75.

Tharwat, A. (2021), Classification assessment methods, *Applied Computing and Informatics,* **17(1)**, 168-192.

Bofang Li, Aleksandr Drozd, Yuhe Guo, Tao Liu, Satoshi Matsuoka, Xiaoyong Du (2019), Scaling Word2Vec on Big Corpus, *Data Science and Engineering,* **4(2)**, 157-175.

Trevor Hastie, J.F., Robert Tibshirani (2001), The Elements of Statistical Learning. *Springer Series in Statistics, Springer New York, NY. 536.*

Provost, F. and T. Fawcett (2013), Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking, *O'Reilly Media.*

Piryonesi, S.M. and T.E. El-Diraby (2020), Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index, *Journal of Infrastructure Systems,* **26(1)**, 04019036.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S. and Zhou, Z.H. (2008), Top 10 algorithms in data mining, *Knowledge and Information Systems,* **14(1)**, 1-37.

Piryonesi, S.M. and T. El-Diraby (2018), Using Data Analytics for Cost-Effective Prediction of Road Conditions: Case of The Pavement Condition Index:[summary report], *United States, Federal Highway Administration, Office of Research*

Piryonesi, S.M. and T.E. El-Diraby (2020), Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems, *Journal of Transportation Engineering,* **146(2)**, 04020022.

Alesheykh, R. (2016), Comparative Analysis of Machine Learning Algorithms with Optimization Purposes, *Control and Optimization in Applied Mathematics,* **1(2)**, 63-75.

Nal Kalchbrenner, E.G., Phil Blunsom (2014), A Convolutional Neural Network for Modelling Sentences, *arXiv preprint arXiv:1404.2188.*

Xu, Q., Zhang, M., Gu, Z., and Pan, G. (2019), Overfitting remedy by sparsifying regularization on fully-connected layers of CNNs, *Neurocomputing,* **328**, 69-74.

Jais, I. K. M., Ismail, A. R., and Nisa, S. Q. (2019), Adam optimization algorithm for wide and deep neural network, *Knowl. Eng. Data Sci.,* **2(1)**, 41-46.

García-Pedrajas, N., C. Hervás-Martínez, and D. Ortiz-Boyer (2005), Cooperative coevolution of artificial neural network ensembles for pattern classification, *IEEE transactions on evolutionary computation,* **9(3)**, 271-302.

Chen, Shikun, Nguyen Manh Luc (2022), RRMSE Voting Regressor: A weighting function based improvement to ensemble regression, *arXiv preprint arXiv:2207.04837.*

Dimitriadou, E., A. Weingessel, and K. Hornik (2001), Voting-Merging: An Ensemble Method for Clustering, *Berlin, Heidelberg: Springer Berlin Heidelberg.*

Wang, H., Yang, Y., Wang, H., and Chen, D. (2013), Soft-Voting Clustering Ensemble, *Multiple Classifier Systems: 11th International Workshop, MCS 2013, Nanjing, China, Proceedings 11. Berlin, Heidelberg: Springer Berlin Heidelberg,* 307-318.