Computing and Informatics, Vol. 44, 2025, 124–150, doi: 10.31577/cai_2025_1_124

DECISION SUPPORT SYSTEM FOR IMPROVING BREAST CANCER DIAGNOSIS USING ENSEMBLE LEARNING

Mohammad Zahaby

Department of Computer Engineering
Borujerd Branch, Islamic Azad University
Borujerd, Iran

Mohammad Ebrahim Shiri[*]

Department of Computer Engineering
Borujerd Branch, Islamic Azad University
Borujerd, Iran
&
Department of Mathematics and Computer Science
Amirkabir University of Technology
Tehran, Iran
e-mail: shiri@aut.ac.ir

Hamid Haj Seyyed Javadi

Department of Mathematics and Computer Science
Shahed University, Tehran, Iran

Mostafa Boroumandzadeh

Department of Computer Engineering and Information Technology
Payame Noor University, Tehran, Iran

[*] Corresponding author

Abstract. Breast cancer (BC) is one of the leading causes of death in women worldwide. Early diagnosis of this disease can save many women's lives. The Breast Imaging Reporting and Data System (BIRADS) is a standard method developed by the American College of Radiology (ACR). However, physicians have had a lot of contradictions in determining the value of BIRADS, and all aspects of patients have not been considered in diagnosing this disease using the methods that have been used so far. In this article, a novel decision support system (DSS) has been presented. In the proposed DSS, firstly, c-mean clustering was used to determine the molecular subtype for patients who did not have this value by combining the mammography reports processing along with hospital information systems (HIS) obtained from their electronic files. Then several classifiers such as convolutional neural networks (CNN), decision tree (DT), multi-level fuzzy min-max neural network (MLF), multi-class support vector machine (SVM), and XGboost were trained to determine the BIRADS. Finally, the values obtained by these classifiers were combined using ensemble learning with the majority voting algorithm to obtain the appropriate value of BIRADS. This helps physicians in the early diagnosis of BC. Finally, the results were evaluated in terms of accuracy, specificity, sensitivity, positive predicted value (PPV), negative predicted value (NPV), f1-measure, and balanced accuracy by the confusion matrix. The obtained values were 87.77 %, 61.81 %, 92.74 %, 56.82 %, 92.75 %, 69.94 %, and 77.28 %, respectively.

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

1 INTRODUCTION

Nowadays, one of the main causes of death in the world is cancer. After cardiovascular diseases, cancer is the second most common cause of death in developed countries and the third most common cause of death in less developed countries, and causes more deaths than tuberculosis, AIDS and malaria [1]. So that if preventive measures are not taken, in the next 10 years we will witness the death of more than 85 million people in the world [1]. Currently, cancer is responsible for 12 % of deaths worldwide [2]. One of the most common types of cancer in women is breast cancer. According to the presented statistics, 19.9 % of deaths caused by cancer in women are related to breast cancer [3]. According to the statistics published by the world health organization (WHO), one out of every 8 to 10 women is diagnosed with breast cancer [4]. On the other hand, early diagnosis in the early stages is one of the important and fundamental factors in the treatment of this disease because when breast cancer is diagnosed early, the probability of treatment and survival is very high [5, 6]. Medical decision support systems (MDSS) are the result of the cooperation of physicians and engineers and are made to help and support health care staff in medical decisions [7, 8, 9]. Today, medical centers have realized the benefits of using MDSS in medical care for breast cancer [8]. The results of the research indicate that through decision support systems using patient data visualization, physicians have the ability to quickly access the necessary information to determine the appropriate treatment [10]. One of the things that can be used as an input in a MDSS system to help diagnose and treat breast cancer is mammography reports [11].

Based on the factors observed in mammography, ultrasonography, and MRI, radiologists use a classification system called BIRADS (created by the American College of Radiology) to describe imaging results in medical reports [12]. One of the reliable methods in evaluating and estimating the risk of breast lesions is BIRADS classification using mammography [13]. BIRADS is a label that is defined in mammography reports in 7 levels, between 0 and 6, and each of these numbers has a specific interpretation [14]. On the other hand, according to the preliminary studies conducted in this field, it was found that although various ideas based on medical decision support systems have been presented to diagnose cancer patients based on the information available in the electronic records of the patients [15, 16, 17, 18, 19, 20, 21], but so far no medical decision support system has been proposed to classify breast cancer patients based on the combination of information from mammography reports, electronic patient records (here HIS) and molecular subtypes.

Percha et al. in 2012 [19] processed the reports and assigned them to a BIRADS class; but the focus was solely on breast tissue. Nassif et al. in 2012 [18] extracted BIRADS features from clinical texts and compared them with manual reporting; but BIRADS was not graded. In 2013 [20], Sippo et al. automated BIRADS extraction from radiology reports by BIRADS Observation Kit and natural language processing (NLP). Gao et al. in 2015 [16] used NLP to extract information from unstructured mammography texts. Their method was limited to the diagnosis of four types of breast complications and only medical reports were used. In 2016, Bozkurt et al. presented an NLP-based decision support system for diagnosing malignancy from BIRADS reports and radiology text [22]. Castro et al. in 2017 [15] presented a rule-based NLP method for classifying radiology reports. Only one type of textual data was used. Gupta et al. in 2017 [17] presented a method based on parse tree and semantics for generating structured information from mammography reports. Only medical reports were used. In 2020, Esmaeili et al. presented a decision support system to help physicians interpret mammography text reports while creating a model capable of predicting a patient's need for a biopsy [11]. Achilonu et al. in 2022 [23] developed a rule-based NLP algorithm that retrieved important breast cancer parameters using pathology reports to exploring molecular subtypes. They used only molecular subtypes text reports. Higa in 2018 [24] used artificial neural network and decision tree classification algorithms to predict breast cancer using clinical information. Zhang et al. in 2019 [21] used deep learning to extract clinical information of breast cancer; but the complexity was high. Spaeth et al. in 2023 [25] used a model for breast cancer diagnosis that integrates influential factors associated with breast cancer from patient's clinical information, including long family history and polygenic risk, which allows to removes moderate factors to improve outcomes.

According to the study of the past research that was stated, mammography reports, HIS data, and molecular subtypes are used separately for the diagnosis of BIRADS. Therefore, in this research, the information of the electronic health record patients and molecular subtypes were placed next to the mammography reports to determine the significant difference in BIRADS diagnosis by adding this additional information. This work seeks to create a decision support system based on the prediction of BIRADS values and molecular subtypes. For this purpose, mammography reports were first processed using NLP and converted into vectors using word2vec [26]. 15 features of HIS were extracted from electronic files of patients. These variables include 2 numerical variables and 13 nominal variables, which were placed next to the vector extracted from the mammography report. Also, using the unsupervised c-mean method, the class of molecular subtypes of the samples was clustered and the molecular subgroup value was assigned to the data of each cluster. Convolutional neural networks (CNN), decision tree (DT), multi-level fuzzy min-max neural network (MLF), multi-class support vector machine (SVM), and XGboost was used for classification and determination of BIRADS. Next, the predicted values of BIRADS of each classifires are used as base learners to combine them using ensemble learning with majority voting algorithm to get better prediction.

This article is organized into four sections. In Section 1, the article context is introduced, its advances and limitations are specified in the field of study, also briefly enumerating its purposes. In Section 2, a decision support system is proposed and the description of its different stages and parts are discussed separately. After that, all the details needed to understand the operation of the system are described. In Section 3, the results obtained from the proposed system are analyzed and evaluated. Finally, in Section 4, the proposed system is discussed and conclusions are drawn regarding its feasibility and usefulness.

2 THE PROPOSED METHOD

In this work, a novel BIRADS diagnosis prediction model was presented as the proposed decision support system (DSS). The data set includes two resources of mammography reports and electronic patient records (extracted from HIS). This data set include of 250 mammography images along with their reports and electronic file records of Shahidzadeh Hospital Medical Training Center in Behbahan City in the period of 2020 to 2022. Mammography text reports have 210 features and other electronic records have 15 features. These 15 features can be seen in Table 1 includes 2 variables related to numerical features and Table 2 includes 13 variables related to nominal features, which together with 210 features related to mammography text reports, a total of 225 features were extracted for each patient.

Also, Table 3 contains information about the distribution of 250 patients who are placed in each of the BIRADS classes.

Figure 1 depicts the different stages of the proposed method in five phases. In the first phase, a data set is obtained, which includes mammography reports and

	Variable Name	Variable Description	Healthy People ($n = 17$) mean ± standard deviation	Patients ($n = 233$) mean ± standard deviation
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

Table 1. Numerical features extracted from HIS

HIS information of each person. Since mammography reports are free texts, they were processed and converted into vectors using NLP methods. In the second phase, important features in the HIS were selected with the physician's consultation. In the third phase, since only BIRADS is specified in the data set, the class of molecular subtypes must be determined first. Therefore, according to Table 4, all data were clustered into four clusters using c-mean algorithm, which is an unsupervised clustering method. Then, each cluster was assigned values related to the appropriate molecular subtype. In the fourth phase, a trained model for predicting BIRADS values by convolutional neural networks (CNN), decision tree (DT), multi-level fuzzy min-max neural network (MLF), multi-class support vector machine (SVM), and XGboost were presented, and in the fifth phase, ensemble learning with majority voting were used to combine the estimated BIRADS values of trained models from fourth phase, then the results were validated by evaluation parameters.

2.1 The First Phase: Dataset

Our dataset includes two main sources: mammography reports and electronic patient records (HIS subset). These data were obtained from the information available at Shahidzadeh Hospital Medical Training Center in Behbahan City, related to the years 2020 to 2022. This dataset includes mammography reports and electronic file records of 400 patients. Since the information of some patients was not complete, finally only the information of 250 patients who had complete information was used.

2.2 The Second Phase: Text Processing and Determining the Important Features of HIS

2.2.1 Preprocessing and Text Processing

Figure 2 illustrates the blocks of the proposed method for classifying medical reports and how to extract a vector from a mammography report. It should be noted that only the text processing flow is shown here.

In the preprocessing stage, mammography reports were stemmed by the NLTK library [27], and prepositions and punctuation marks were removed except for negative words to ensure the accuracy and relevance of the text analysis. NLTK provides a comprehensive suite of tools for natural language processing, including tokeniza-

	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	Breast-feeding	Presence/absence of a history of the Breastfeeding	No = 9
			Yes = 8
				No = 72
				Yes = 161

7	Shape	Mass 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

Table 2. Nominal features extracted from HIS

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

Table 3. Patients distribution according to BIRADS class

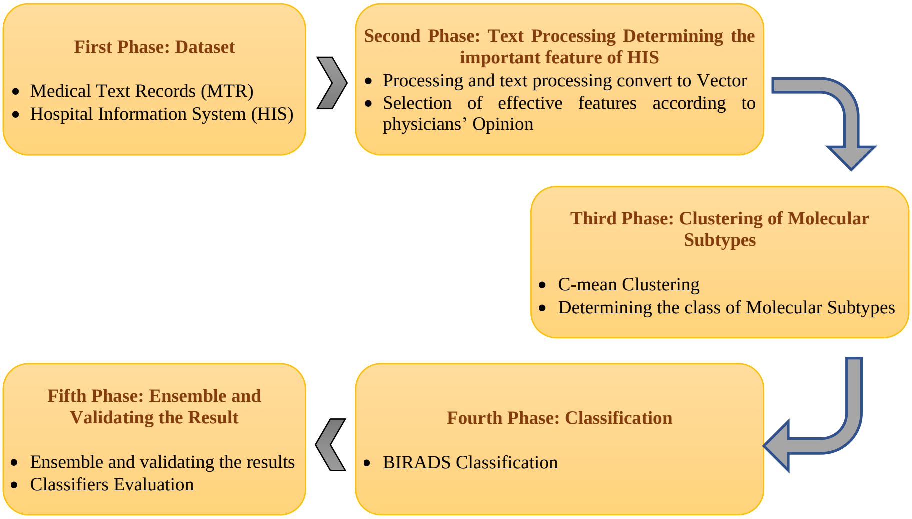
First Phase: Dataset
Medical Text Records (MTR)
Hospital Information System (HIS)
Second Phase: Text Processing Determining the important feature of HIS
Processing and text processing convert to Vector
Selection of effective features according to physicians' Opinion
Third Phase: Clustering of Molecular Subtypes
C-mean Clustering
Determining the class of Molecular Subtypes
Fourth Phase: Classification
BIRADS Classification
Fifth Phase: Ensemble and Validating the Result
Ensemble and validating the results
Classifiers Evaluation

Figure 1. Phasing of the proposed method

tion, stemming, and lemmatization, which are essential for accurate text preprocessing. NLTK was chosen over other libraries such as SpaCy and TextBlob for several reasons. Firstly, NLTK offers a broader range of NLP tools and resources, making it ideal for exploring different NLP techniques and customizing them for specific research requirements [28]. Secondly, NLTK's integration with other Python libraries such as NumPy, SciPy, and scikit-learn facilitates seamless data analysis and enhances the overall efficiency of the preprocessing stage [29]. Thirdly, NLTK's flexibility allows for fine-tuning preprocessing steps to meet the specific needs of our study, which is crucial for ensuring the accuracy and relevance of the text analysis [30]. For example, "No tangible mass in the breast or axillary seen" is a negative sentence and the negative sign is not removed. Integer and decimal numbers were converted to the corresponding string. To preserve local dependencies, bigram collection of possible pairs of words was calculated based on mutual information. To

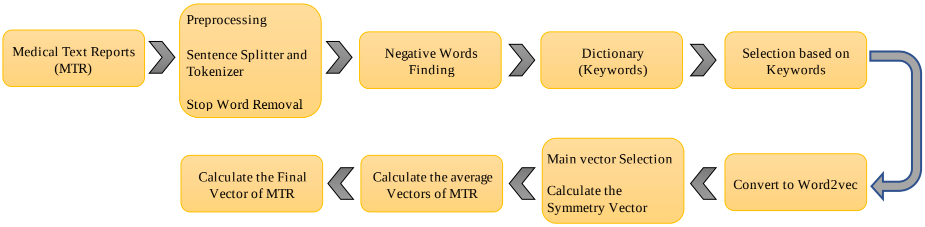
Medical Text Reports (MTR)
Preprocessing
Sentence Splitter and Tokenizer
Stop Word Removal
Negative Words Finding
Dictionary (Keywords)
Selection based on Keywords
Convert to Word2vec
Main vector Selection
Calculate the Symmetry Vector
Calculate the average Vectors of MTR
Calculate the Final Vector of MTR

Figure 2. Converting a report to a vector

improve the accuracy of word embedding, bigrams with less than 50 occurrences were removed and those with more than 1 000 occurrences were considered as single words.

Next, the key words in the dictionary were selected from the text. If there is a negative word in the sentence, it is contrasted, or its vector is reversed. For example, in the above sentence, it is possible that the person does not have breast cancer. Therefore, since it exists in the dictionary, "tangible mass" becomes the opposite, and if no opposite word is found for it, the result of the Word2vec is reversed. In order to reduce ambiguities and improve the semantic accuracy of reports, domain ontology was used in the text processing section. This was done by a lexical crawler [31] whose task is to identify the derived terms that have a common root with predefined terms that we mapped to controlled terms (key terms). In addition to the dictionary, we used commonly available terms (CLEVER) [31] are used in identifying clinical contexts and mapping.

After combining key terms and terms afforded from CLEVER, a total of 260 keys were obtained, which are mainly used for two purposes:

1. Reduce reports through mapping.
2. It helps to generate text-aware vectors.

Unsupervised method has been used to create word embedding using Word2vec model [26]. 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. Then the average of all the obtained vectors represents the vector of that text. Each report vector was calculated based on Equation (1):

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

In this equation, 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.

2.2.2 Selection of Effective Features of Electronic Health Record in HIS

Breast cancer specialists were consulted to select the most effective features in determining breast cancer from information obtained from HIS. HIS is extracted from picture archiving and communication system (PACS) and electronic files of patients in Shahidzadeh Hospital Medical Training Center in Behbahan City between 2020 and 2022. Electronic records including medical documents, images, and reports are stored in PACS. HIS is an integrated information system that includes aspects of hospital performance such as financial, patient health, legal and, administration services, etc. The database uses information related to the PACS system in medical training centers.

2.3 The Third Phase: Clustering of Molecular Subtypes

Breast cancer has various molecular subtypes that are recognized based on receptor and immunochemical status. Some receptors include estrogen receptor (ER), progesterone receptor (PR), epidermal growth factor receptor 2-neu (HER2), proliferation marker Ki67, and epidermal growth factor receptor (EGFR). There are four main sets of molecular subtypes in breast cancer: luminal a, luminal b, human epidermal growth factor (HER2), and breast cancer with basal-like molecular class (BLBC). Each of these molecular subtypes shows the rate of recurrence and survival, which is the most important factor in choosing different treatment techniques [32].

2.3.1 C-Mean Clustering

In order to obtain molecular subtypes, patients must perform an invasive procedure of biopsy from breast tissue. In this paper, only 156 samples out of 250 samples have molecular subtype features. Considering that this feature is very important and effective in determining the stages of breast cancer progression, c-mean clustering was used to assign molecular subtypes to samples that do not have this feature. This allows the system to be trained to accurately determine the molecular subtype for patients who are in the early stages of the disease and have not yet undergone biopsy. Thus, at first, all patients were placed in four clusters using c-mean (and the characteristics obtained from the second phase) and after the completion of the clustering steps, based on the values of the cluster centers, molecular subtypes were determined for each cluster. In the c-mean method, the samples are divided into c clusters, where c (the number of molecular subtypes) is specified in advance. The objective function is in the form of Equation (2):

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

In Equation (2), m is a real number greater than 1, which is chosen for m in most cases is 2, n is the number of samples, c is the cluster centre, u is the degree of membership, and x is the sample. To minimize the value of j, the membership degree and cluster centres are updated in each iteration with Equations (3) and (4), respectively [33, 34].

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}, \tag{3}$$

$$c_j = \frac{\sum_{i=1}^{n} u_{ij}^{m} . x_i}{\sum_{i=1}^{n} u_{ij}^{m}}. \tag{4}$$

The main criteria were as follows: 1. Only those with breast cancer were clustered. 2. Based on immunohistochemical results after surgery or biopsy according to the 13th St. Gallen International Breast Cancer Conference 2013, each cluster was identified by one of four different molecular subtypes [32]. Molecular subtypes along with immunophenotype are shown in Table 4.

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 %

Table 4. Molecular subtypes and immunophenotype [32]

Therefore, we obtained a logical relationship between BIRADS classification and molecular subtypes. Now, using a classifier, we can probabilistically identify molecular subtypes based on BIRADS information.

2.4 The Fourth Phase: Classification

2.4.1 Convolutional Neural Network (CNN)

Machine learning algorithms demonstrate effective performance within reasonable computational timeframes, enabling significant knowledge extraction from data [35]. Among these algorithms, Convolutional Neural Networks (CNNs) [36] are widely employed in image, speech, and text analysis within the machine learning domain. In this study, CNN is employed as a classifier for BIRADS detection, capable of recognizing intricate relationships between dependent and independent variables and handling noisy data. The input undergoes convolution operations, followed by pooling layers to reduce dimensionality and prevent overfitting [37]. During backpropagation, the parameter Θ is updated using error minimization. ReLU is typically used as the activation function in the first and second convolution layers, while the output layer employs the softmax process, and the loss function is the mean squared error. Additionally, the optimization algorithm employed here is Adam [38], known for its adaptive learning rate.

2.4.2 Decision Tree (DT)

Decision tree learning is a supervised machine learning algorithm that is widely used for classification and regression tasks. In this tree structures, the leaves represent class labels, and the branches represent combinations of features that lead to those class labels [39, 40, 41]. Decision trees are constructed based on minimizing a "quantity" called entropy [39, 41]. Early versions of decision trees could only use discrete variables, but newer algorithms can handle both discrete and continuous variables in learning [41, 42]. A decision tree's goal is to predict a variable's value based on the measures of input variables. One of the significant advantages of the decision tree algorithm is its interpretability and ease of understanding, which has made it popular [41, 42, 43]. However, its drawbacks include a lack of robustness and insufficient accuracy [42]. In this context, a decision tree has also been used for classification and BIRADS detection.

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

MLF is an advanced version of the Fuzzy Min-Max Neural Network [34], where the latter employs "hyper-boxes" for sample classification. A hyper-box is essentially an n-dimensional box characterized by a minimum point, a maximum point, and an associated membership function, with each hyper-box corresponding to a specific class. During network training hyper-boxes are generated and adjusted based on the arrival of training samples. Equation (5) provides the definition for the hyper-box:

$$B_j = \{X, V_j, W_j, f\left(X, V_j, W_j\right) \forall X \in I^n\}. \tag{5}$$

V_j and W_j denote the upper and lower bounds of a hyper-box. X represents an individual sample, and n represents the number of dimensions in the feature vectors. The dimensions of these hyper-boxes are regulated by Equation (6):

$$\forall_{i=1\ldots D}\left(max\left(w_b^i, x^i\right) - min\left(v_b^i, x^i\right)\right) \leq \Theta. \tag{6}$$

Equation (6) introduces the coefficient of expansion denoted by Θ. The algorithm comprises three layers. The first layer deals with the inputs, the second layer handles the hyper-boxes, and the third layer is responsible for the output or classes [34]. Here also the system was trained by MLF to recognize the BIRADS feature.

2.4.4 Multi-Class Support Vector Machine (SVM) Algorithm

Support Vector Machine (SVM) is an algorithm that aims to find the optimal separating hyperplane, which maximizes the margin between two classes. This hyperplane can be represented by the Equation (7) [44]:

$$W^T x + b = 0. \tag{7}$$

Here, x represents the input vector that contains the input features, b denotes the bias, W is the weight that determines the distance between the hyperplane and the data points, and W^T refers to the transpose of the matrix W. Selecting the best hyperplane involves identifying different hyperplanes that can classify the labels effectively. The algorithm then chooses the hyperplane farthest from the data points or the one with the maximum margin, as illustrated in Figure 3.

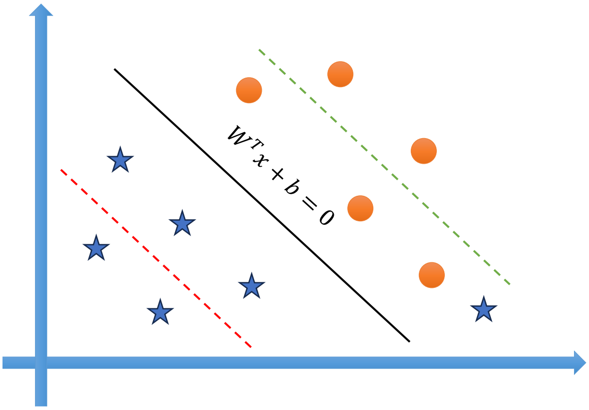

Figure 3. SVM hyperplane [44]

In this paper, the RBF kernel function was used and after extracting the model [45], the possible values for each class were obtained from BIRADS. Here, normalization by standard deviation method [46] was used. Seven support vector machines were used to detect BIRADS, which has seven classes. According to Table 5, seven support vector machines have made a decision for one sample, and considering that in this example, the fourth support vector machine shows the highest probability, so the sample belongs to the fourth class or "Probably benign" [12].

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

Table 5. SVM values

2.4.5 XGboost

XGBoost is a scalable and distributed machine learning library that utilizes Gradient Boosted Decision Trees (GBDT) for machine learning tasks. It is designed for improved speed and efficiency. Gradient Boosting is a machine learning method used for regression and classification problems. The Gradient Boosting model is a linear combination of a series of weak models created iteratively to form a robust final model [41]. This approach is a part of ensemble learning algorithms, and its performance is consistently better than fundamental or weak algorithms such as decision trees or bagging-based methods like Random Forest. However, the accuracy of this claim is somewhat influenced by the characteristics of the input data [47]. In this context, XGBoost has also been used for classification and BIRADS detection.

2.5 The Fifth Phase: Ensemble and Validating the Results

In the fifth phase, based on Figure 4, assuming that the person refers to the treatment system, initially based on the medical text reports (MTR) which in this work is mammography reports and also the patient's electronic file records from HIS, and after the data fusion, text processing, and clustering, the BIRADS values are predicted using the explained base learners (CNN, DT, MLF, SVM, XGboost), then the output was combined using ensemble learning.

2.5.1 Ensemble Learning

Ensemble learning is a machine learning approach where multiple models, often referred to as "base learners", are trained to solve the same problem and combined to achieve better results in classification or regression tasks. The combination is achieved by aggregating the outputs from each model, with two main objectives: reducing model error and maintaining generalization. Compared to a single model, ensembles increase final predictions' robustness and accuracy [48, 49]. A simple and intuitive ensemble technique is majority voting [50, 51]. Essentially, the ensemble selects the class for an object based on the majority choice from the individual classification results. Let us define the decision of the t^{th} classifier for class j as $d_{t,j} \in 0, 1$, where $(t = 1, 2, 3, \ldots, T;\ j = 1, 2, 3, \ldots, C)$. Here, T represents the number of results from base classifiers, and C represents the number of classes. If t^{th} classifier result chooses class j, then $d_{t,j}$ equals 1; otherwise, it is 0. The ensemble decision for class k, calculated using Equation (8), is determined by majority voting.

$$\sum_{t=1}^{T} d_{t,k} = \max_{j} \sum_{t=1}^{T} d_{t,j}. \tag{8}$$

Here, ensemble learning is used to predict the value of BIRADS using majority voting method. The performance of this method is compared to base classification algorithms.

2.5.2 Validation

The BIRADS results obtained from the base learners along with the final result obtained from the ensemble learning using the majority voting method, are validated by the evaluation parameters obtained from the confusion matrix such as accuracy, specificity, sensitivity, positive predicted value (PPV), negative predicted value (NPV), f1-measure, and balanced accuracy. Equations (13), (14), (15), (16), (17), (18) and (19) have been used to calculate these evaluation metrics, respectively.

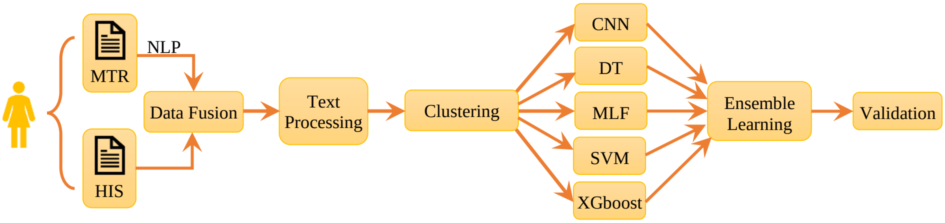

Figure 4. Road map of the proposed method

3 ANALYSIS AND EVALUATION OF RESULTS

To implement this plan, a computer with the following specifications was used:

Processor: Intel(R) Core(TM) i7-4790 CPU @ 3.60 GHz,

Installed memory (RAM): $2 * 8$ GB DDR RAM,

VGA: GT 730 2 GB,

HDD: 256 GB SSD + 1 TB SATA.

The operating system used in this research was Microsoft Windows 10 64 bit, and Python 3.8.7 was used in the Visual Studio Code environment to model the program.

3.1 Evaluation Parameters

According to Table 6, the confusion matrix is one of the evaluation criteria of classifiers and is an N by N square matrix; where N represents the number of classes, which is here there are 7 classes for BIRADS. The main diameter represents the number of correct detections, and the rest of arrays of matrix express the incorrect detections.

In binary classification models where only the positive or negative of the disease is diagnosed, in the confusion matrix there is a concept of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). But here the value of BIRADS is diagnosed for patients, which has 7 classes. Here, TP_i is the value of true positive of i^{th} class, which refers to cases where the actual class is i, and the detected class is i too. TP_i is obtained using Equation (9). There are two other concepts, false positive (FP) and false negative (FN), where FP_i is the value of false positive of i^{th} class, which refers to cases where the actual class is i but the detected class is other than i. FP_i is calculated using Equation (10). Also, FN_i, which is the

		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

Table 6. Confusion matrix [46]

value of false negative of i^{th} class, indicates the diagnosis of the class is i, but the actual class is other than i. FN_i is calculated using Equation (11). TN_i is the true negative value of class i, which refers to cases where the actual class is not i and the detected class also is not i. TN_i is obtained using Equation (12):

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

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

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

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

The parameters such as accuracy, specificity, sensitivity, positive predicted value (PPV), negative predicted value (NPV), f1-measure, and balanced accuracy are calculated using Equations (13), (14), (15), (16), (17), (18) and (19), respectively [52, 53].

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

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

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

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

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

$$\mathit{f1\text{-}measure} = \frac{2 \times PPV \times Sensitivity}{PPV + Sensitivity}, \tag{18}$$

$$Balanced\ Accuracy = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right), \tag{19}$$

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

3.2 Results

Figures 5, 6, 7, 8, 9 and 10 depict the level of accuracy, specificity, PPV, NPV, sensitivity, and f1-measure for Convolutional Neural Network (CNN), Decision Tree, Multi-Level Fuzzy Min-Max Neural Network (MLF), Support Vector Machine (SVM), XGboost, and proposed decision support system (DSS) for BIRADS detection and only using text mining. We can see that with the increase of the dimensions in the resulting vector of the text, the accuracy of the classification has increased, and this value has a downward trend in dimensions higher than 160. It has been found in many studies such as [54], by increasing the dimensions, quality of word2vector and subsequently accuracy were decreases. This issue was investigated by reducing and increasing the dimensions. Finally, 160 dimensions were used for further processing, since the best results was obtained in 160 dimensions.

Figure 5 shows the variation of accuracy for all classifiers used in this research in dimensions from 110 to 200. The best accuracy of the proposed decision support system occurred in dimension 160 with 87.77 %. In the same dimension, the accuracy for CNN, DT, MLF, SVM, and XGboost is 84.34 %, 80.46 %, 84.00 %, 81.37 %, and 83.66 %, respectively.

Figure 6 also shows the variation of specificity of all classifiers in mentioned dimensions. The best specificity of the proposed decision support system occurred in dimension 160 with 92.74 %. In the same dimension, the accuracy for CNN, DT, MLF, SVM, and XGboost is 91.11 %, 88.51 %, 91.40 %, 90.19 %, and 90.50 %, respectively.

Figure 7 also shows the variation of sensitivity of all classifiers in mentioned dimensions. The best specificity of the proposed decision support system occurred in dimension 160 with 61.81 %. In the same dimension, the accuracy for CNN, DT, MLF, SVM, and XGboost is 57.66 %, 34.18 %, 52.80 %, 56.84 %, and 44.63 %, respectively.

Figure 8 also shows the variation of positive predicted value (PPV) of all classifiers in mentioned dimensions. The best specificity of the proposed decision support

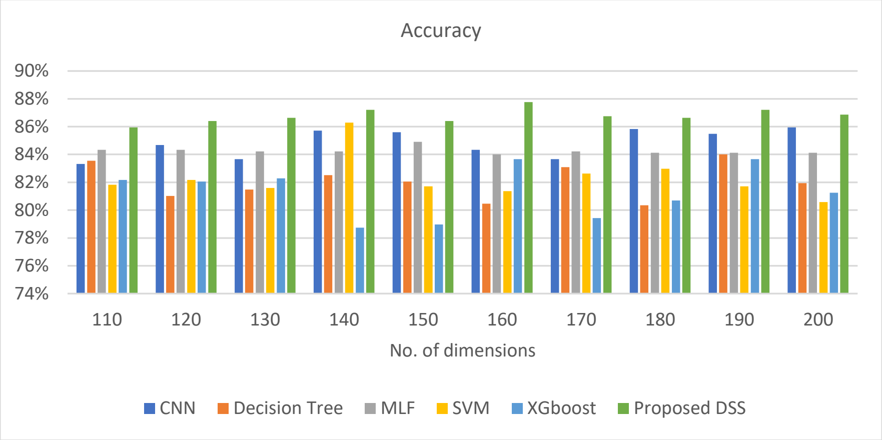

Figure 5. Variations of accuracy with the change of dimensions in the vector resulting from word2vec

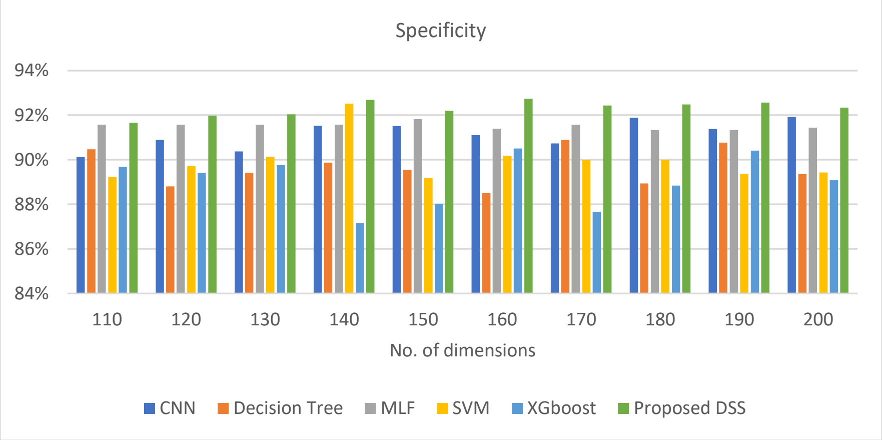

Figure 6. Variations of specificity with the change of dimensions in the vector resulting from word2vec

system occurred in dimension 160 with 56.82 %. In the same dimension, the accuracy for CNN, DT, MLF, SVM, and XGboost is 45.03 %, 33.24 %, 47.82 %, 30.97 %, and 42.62 %, respectively.

Figure 9 also shows the variation of negative predicted value (NPV) of all classifiers in mentioned dimensions. The best specificity of the proposed decision support system occurred in dimension 160 with 92.75 %. In the same dimension, the accuracy for CNN, DT, MLF, SVM, and XGboost is 90.65 %, 88.51 %, 91.00 %, 88.96 %, and 90.59 %, respectively.

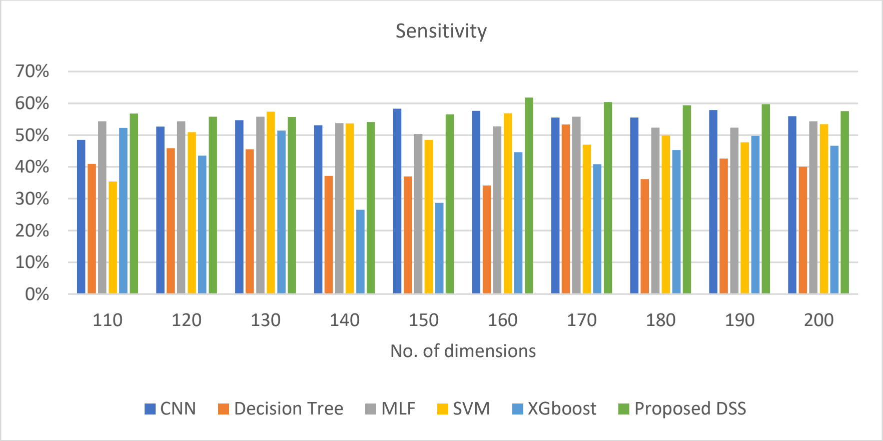

Figure 7. Variations of sensitivity with the change of dimensions in the vector resulting from word2vec

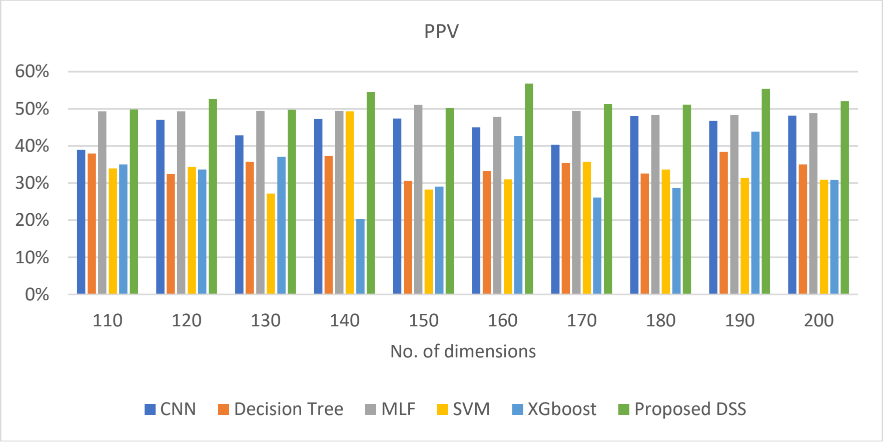

Figure 8. Variations of PPV with the change of dimensions in the vector resulting from word2vec

Figure 10 also shows the variation of f1-measure of all classifiers in mentioned dimensions. The best specificity of the proposed decision support system occurred in dimension 160 with 69.94 %. In the same dimension, the accuracy for CNN, DT, MLF, SVM, and XGboost is 57.83 %, 46.92 %, 57.24 %, 39.00 %, and 56.93 %, respectively.

Figure 11 also shows the variation of balanced accuracy of all classifiers in mentioned dimensions. The best specificity of the proposed decision support system occurred in dimension 160 with 77.28 %. In the same dimension, the balanced accu-

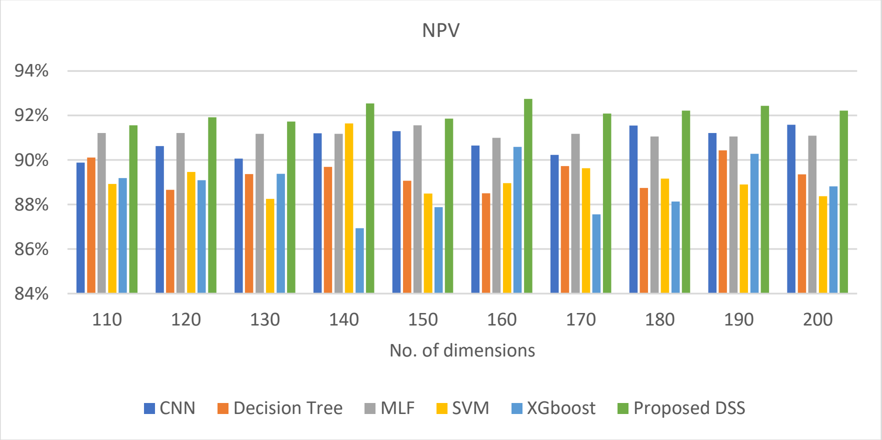

Figure 9. Variations of NPV with the change of dimensions in the vector resulting from word2vec

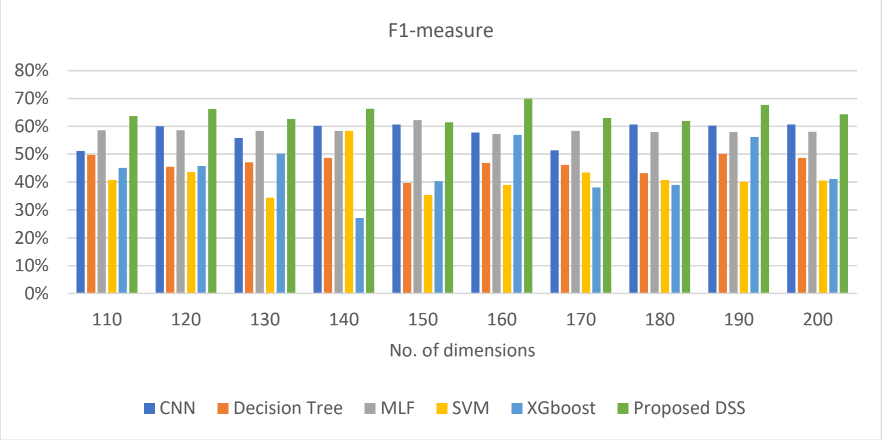

Figure 10. Variations of f1-measure with the change of dimensions in the vector resulting from word2vec

racy for CNN, DT, MLF, SVM, and XGboost is 74.38 %, 61.34 %, 72.10 %, 73.51 %, and 67.56 %, respectively.

Table 7 depicts the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), f1-measure, balanced accuracy, and accuracy for BI-RADS classification of proposed DSS. Classes one to seven indicate the corresponding values in BIRADS zero to six. Most disease classes were diagnosed with an accuracy of over 85 %. Follow-up of the disease corresponding to BIRADS = 6, which is seventh class, have the highest sensitivity of 90.00 %. The specificity value

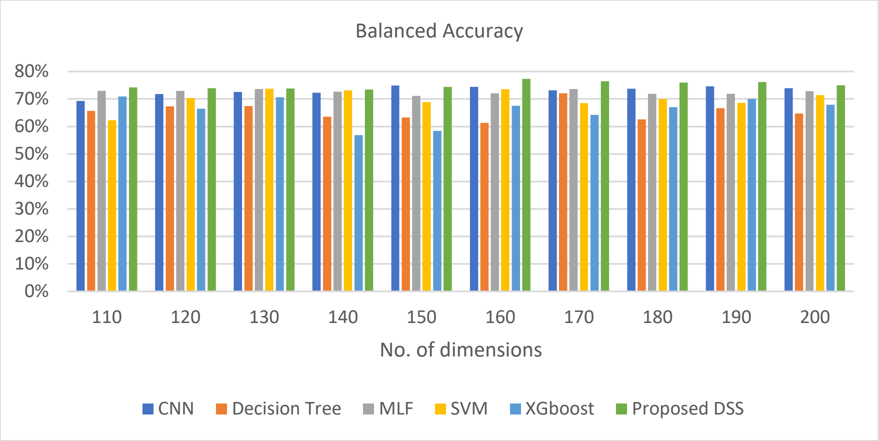

Figure 11. Variations of balanced accuracy with the change of dimensions in the vector resulting from word2vec

Confusion Matrix							ID Class	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1 Measure (%)	Balanced Accuracy (%)	Accuracy (%)
15	1	0	4	3	1	1	Class1	28.85	94.95	60.00	83.56	73.53	61.90	81.20
5	10	0	0	5	2	0	Class2	50.00	94.78	45.45	95.61	61.44	72.39	91.20
5	0	26	0	1	4	1	Class3	86.67	95.00	70.27	98.12	80.78	90.83	94.00
6	0	1	24	6	1	0	Class4	68.57	93.49	63.16	94.81	75.39	81.03	90.00
10	7	3	2	35	2	0	Class5	53.03	86.96	59.32	83.77	70.53	69.99	78.00
8	1	0	5	11	15	0	Class6	55.56	88.79	37.50	94.29	52.73	72.17	85.20
3	1	0	0	5	2	18	Class7	90.00	95.22	62.07	99.10	75.15	92.61	94.80

Table 7. Confusion matrix of proposed DSS

for healthy people is equal to 94.78 %, which illustrates the high performance of healthy people. The weighted average value of specificity is 90.66 %, the minimum value is for the fifth class, and the maximum value is related to the seventh class (95.22 %). The values show that the performance of the proposed method is suitable in terms of specificity values. The weighted average value of PPV is equal to 54.11 %, which is the maximum value is 70.27 % (third class). The value of NPV for healthy people is equal to 95.61 %, which shows that the proposed method has appropriate performance, the maximum value is 99.10 % (seventh class), and its minimum value is 83.56 % (first class). The weighted average value of f1-measure is 67.09 %, the maximum value is 80.78 % (third class) and its minimum is 52.73 % (sixth class), which shows that the proposed method has appropriate detection rate. The weighted average value of balanced accuracy is 76.29 %, the maximum value is 92.61 % (sixth class) and its minimum is 61.90 % (first class).The values show that the performance of the proposed method is suitable in terms of balanced accuracy values. The accuracy or ability of the test, in correctly differentiating sick and healthy cases in average is 87.77 %. The minimum and maximum value of the accuracy are 78.00 % and 94.80 % respectively.

As a result, by analyzing these parameters, it was found that the proposed method preformed well in the detection of BIRADS classes, which gratefully helps to diagnose the disease and determine the appropriate treatment method. Since here HIS values are used along with the results of text processing, therefore, the performance of detection of BIRADS has improved using the proposed method.

4 DISCUSSION AND CONCLUSION

The American College of Radiology (ACR) presented a standard called BIRADS to standardize mammography reports. This system led to the homogenization of reports and played a major role in advancing standard treatment planning, as it can be used to accurately prioritize the treatment progress. But this approach had disadvantages such as the difference of opinion among physicians to conclude the value of BIRADS. Therefore, in this work, it was suggested to use the information of the records of electronic files of people. Therefore, a hybrid approach of unstructured data (mammography reports) and structured data (electronic records file from HIS) has been used.

In this way, after preprocessing and processing the texts, the keywords were converted into vectors using Word2vec, and in each text, the average vectors of the keywords represented that text. A 210-dimensional vector was obtained for each text. After that, 15 features have been selected from the patients' electronic records. These variables include 2 numerical variables and 13 nominal variables, which were placed next to the vector extracted from the mammography report, and including 210 features related to mammography reports, a total of 225 features were used for classification. Also, CNN, DT, MLF, SVM, and XGboost was used to determine BIRADS classes, and then estimated BIRADS are combined using ensemble learning with majority voting algorithm. The results were evaluated in the form of different evaluation parameters such as sensitivity, specificity, PPV, NPV, f1-measure, and accuracy. The results show that the maximum evaluation parameters for BIRADS estimation are 90.00 %, 95.22 %, 70.27 %, 99.10 %, 80.78 % and 94.80 %, respectively. The accuracy of detecting BIRADS values for proposed method is 87.77 %. The proposed DSS helps the physician to make better decisions for the diagnosis of BIRADS by data fusion in HIS and medical text reports. This approach compared to similar works has improved the detection of the disease or the patient's health, as well as the determination of the level of the disease; Therefore, the physician can determine the individual's treatment routine more accurately.

In this work, data fusion was used to improve accuracy. It is suggested to use the weight for base learners for combination to increase the efficiency of the system in the next research. Also, since mammography images provide useful information to the physician, another suggestion is to use decision fusion, images and deep learning integration techniques to more accurately estimate the level of disease in order to help physicians in making more accurate decisions about treatment procedures. Another idea is the use of Random Forest as an alternative to Decision Tree. Random Forest is known for its ability to improve classification accuracy by reducing overfitting through the aggregation of multiple decision trees. This ensemble method can provide more robust and reliable results, which could enhance the performance of classification models. Additionally, other advanced machine learning algorithms were suggested to further improve the diagnostic accuracy and reliability of the system. These efforts will help in identifying the most effective models for breast cancer diagnosis and contribute to the development of more accurate and efficient clinical decision support systems.

5 COMPLIANCE WITH ETHICAL STANDARDS

In this study, there was no conflict of interest and none of the authors had contact with the patients, and their surnames were unknown to the authors.

Acknowledgments

We would like to express our sincere gratitude to the editor and the anonymous reviewers for their invaluable feedback and constructive comments, which have significantly improved the quality of this paper. Their expertise and insights have been instrumental in refining the research and enhancing its contribution to the field.

REFERENCES

[1] Balakumar, P.—Maung-U, K.—Jagadeesh, G.: Prevalence and Prevention of Cardiovascular Disease and Diabetes Mellitus. Pharmacological Research, Vol. 113, 2016, pp. 600–609, doi: 10.1016/j.phrs.2016.09.040.

[2] Bray, F.—Ferlay, J.—Soerjomataram, I.—Siegel, R. L.—Torre, L. A.—Jemal, A.: Global Cancer Statistics 2018: Globocan Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians, Vol. 68, 2018, No. 6, pp. 394–424, doi: 10.3322/caac.21492.

[3] U.S. Cancer Statistics Data Visualizations Tool. 2017, `https://www.cdc.gov/cancer/dataviz` [accessed 2023-01-10].

[4] Isfahani, P.—Hossieni Zare, S. M.—Shamsaii, M.: The Prevalence of Depression in Iranian Women with Breast Cancer: A Meta-Analysis. Internal Medicine Today, Vol. 26, 2020, No. 2, pp. 170–181, doi: 10.32598/hms.26.2.3207.1.

[5] Dehghan, P.—Mogharabi, M.—Zabbah, I.—Layeghi, K.—Maroosi, A.: Modeling Breast Cancer Using Data Mining Methods. Journal of Health and

Biomedical Informatics, Vol. 4, 2018, No. 4, pp. 266–278, http://jhbmi.ir/article-1-208-en.html.

[6] Ginsburg, O.—Yip, C. H.—Brooks, A.—Cabanes, A.—Caleffi, M.—Yataco, D. et al.: Breast Cancer Early Detection: A Phased Approach to Implementation. Cancer, Vol. 126, 2020, No. S10, pp. 2379–2393, doi: 10.1002/cncr.32887.

[7] Alaa, A. M.—Moon, K. H.—Hsu, W.—van der Schaar, M.: ConfidentCare: A Clinical Decision Support System for Personalized Breast Cancer Screening. IEEE Transactions on Multimedia, Vol. 18, 2016, No. 10, pp. 1942–1955, doi: 10.1109/TMM.2016.2589160.

[8] Mazo, C.—Kearns, C.—Mooney, C.—Gallagher, W. M.: Clinical Decision Support Systems in Breast Cancer: A Systematic Review. Cancers, Vol. 12, 2020, No. 2, Art. No. 369, doi: 10.3390/cancers12020369.

[9] Sim, L. L. W.—Ban, K. H. K.—Tan, T. W.—Sethi, S. K.—Loh, T. P.: Development of a Clinical Decision Support System for Diabetes Care: A Pilot Study. PLoS ONE, Vol. 12, 2017, No. 2, Art. No. e0173021, doi: 10.1371/journal.pone.0173021.

[10] Park, J.—Rho, M. J.—Moon, H. W.—Park, Y. H.—Kim, C. S.—Jeon, S. S.—Kang, M.—Lee, J. Y. J.: Prostate Cancer Trajectory-Map: Clinical Decision Support System for Prognosis Management of Radical Prostatectomy. Prostate International, Vol. 9, 2021, No. 1, pp. 25–30, doi: 10.1016/j.prnil.2020.06.003.

[11] Esmaeili, M.—Ayyoubzadeh, S. M.—Ahmadinejad, N.—Ghazisaeedi, M.—Nahvijou, A.—Maghooli, K.: A Decision Support System for Mammography Reports Interpretation. Health Information Science and Systems, Vol. 8, 2020, No. 1, Art. No. 17, doi: 10.1007/s13755-020-00109-5.

[12] Magny, S. J.—Shikhman, R.—Keppke, A. L.: Breast Imaging Reporting and Data System. StatPearls [internet], StatPearls Publishing, 2023.

[13] Farrokh, D.—Alamdaran, S. A.—Feizy, A.—Soleimany, H.: 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, Vol. 22, 2019, No. 6, pp. 1–6, doi: 10.22038/ijogi.2019.13738 (in Persian).

[14] Vanderheyden, R.—Xie, Y.: Mammography Image BI-RADS Classification Using OHPLall. 2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService), 2020, pp. 120–127, doi: 10.1109/BigDataService49289.2020.00026.

[15] Castro, S. M.—Tseytlin, E.—Medvedeva, O.—Mitchell, K.—Visweswaran, S.—Tanja, B.—Jacobson, R. S.: Automated Annotation and Classification of BI-RADS Assessment from Radiology Reports. Journal of Biomedical Informatics, Vol. 69, 2017, pp. 177–187, doi: 10.1016/j.jbi.2017.04.011.

[16] Gao, H.—Aiello Bowles, E. J.—Carrell, D.—M., B. D. S.: Using Natural Language Processing to Extract Mammographic Findings. Journal of Biomedical Informatics, Vol. 54, 2015, pp. 77–84, doi: 10.1016/j.jbi.2015.01.010.

[17] Gupta, A.—Banerjee, I.—Rubin, D. L.: Automatic Information Extraction from Unstructured Mammography Reports Using Distributed Semantics. Journal of

Biomedical Informatics, Vol. 78, 2018, pp. 78–86, doi: 10.1016/j.jbi.2017.12.016.

[18] Nassif, H.—Cunha, F.—Moreira, I. C.—Cruz-Correia, R.—Sousa, E.—Page, D.—Burnside, E.—Dutra, I.: Extracting BI-RADS Features from Portuguese Clinical Texts. 2012 IEEE International Conference on Bioinformatics and Biomedicine, 2012, pp. 1–4, doi: 10.1109/BIBM.2012.6392613.

[19] Percha, B.—Nassif, H.—Lipson, J.—Burnside, E.—Rubin, D.: Automatic Classification of Mammography Reports by BI-RADS Breast Tissue Composition Class. Journal of the American Medical Informatics Association, Vol. 19, 2012, No. 5, pp. 913–916, doi: 10.1136/amiajnl-2011-000607.

[20] Sippo, D. A.—Warden, G. I.—Andriole, K. P.—Lacson, R.—Ikuta, I.—Birdwell, R. L.—Khorasani, R.: Automated Extraction of BI-RADS Final Assessment Categories from Radiology Reports with Natural Language Processing. Journal of Digital Imaging, Vol. 26, 2013, No. 5, pp. 989–994, doi: 10.1007/s10278-013-9616-5.

[21] Zhang, X.—Zhang, Y.—Zhang, Q.—Ren, Y.—Qiu, T.—Ma, J.—Sun, Q.: Extracting Comprehensive Clinical Information for Breast Cancer Using Deep Learning Methods. International Journal of Medical Informatics, Vol. 132, 2019, Art. No. 103985, doi: 10.1016/j.ijmedinf.2019.103985.

[22] Bozkurt, S.—Gimenez, F.—Burnside, E. S.—Gulkesen, K. H.—Rubin, D. L.: Using Automatically Extracted Information from Mammography Reports for Decision-Support. Journal of Biomedical Informatics, Vol. 62, 2016, pp. 224–231, doi: 10.1016/j.jbi.2016.07.001.

[23] Achilonu, O. J.—Singh, E.—Nimako, G.—Eijkemans, R. M. J. C.—Musenge, E.: 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, Vol. 2022, 2022, No. 1, Art. No. 6157861, doi: 10.1155/2022/6157861.

[24] Higa, A.: Diagnosis of Breast Cancer Using Decision Tree and Artificial Neural Network Algorithms. International Journal of Computer Applications Technology and Research (IJCATR), Vol. 7, 2018, No. 1, pp. 23–27, `http://ijcatr.com/archieve/volume7/issue1/ijcatr07011004.pdf`.

[25] Spaeth, E. L.—Dite, G. S.—Hopper, J. L.—Allman, R.: Validation of an Abridged Breast Cancer Risk Prediction Model for the General Population. Cancer Prevention Research, Vol. 16, 2023, No. 5, pp. 281–291, doi: 10.1158/1940-6207.CAPR-22-0460.

[26] Guo, D.—Wang, Q.—Liang, M.—Liu, W.—Nie, J.: Molecular Cavity Topological Representation for Pattern Analysis: A NLP Analogy-Based Word2Vec Method. International Journal of Molecular Sciences, Vol. 20, 2019, No. 23, Art. No. 6019, doi: 10.3390/ijms20236019.

[27] Loper, E.—Bird, S.: NLTK: The Natural Language Toolkit. CoRR, 2002, doi: 10.48550/arXiv.cs/0205028.

[28] Wang, M.—Hu, F.: The Application of NLTK Library for Python Natural Language Processing in Corpus Research. Theory and Practice in Language Studies, Vol. 11, 2021, No. 9, pp. 1041–1049, doi: 10.17507/tpls.1109.09.

[29] Drozd, A.—Gladkova, A.—Matsuoka, S.: Python, Performance, and Natural Language Processing. Proceedings of the 5th Workshop on Python for High-Performance and Scientific Computing (PyHPC '15), ACM, 2015, doi: 10.1145/2835857.2835858.

[30] Chai, C. P.: Comparison of Text Preprocessing Methods. Natural Language Engineering, Vol. 29, 2023, No. 3, pp. 509–553, doi: 10.1017/S1351324922000213.

[31] Banerjee, I.—Bozkurt, S.—Alkim, E.—Sagreiya, H.—Kurian, A. W.—Rubin, D. L.: Automatic Inference of BI-RADS Final Assessment Categories from Narrative Mammography Report Findings. Journal of Biomedical Informatics, Vol. 92, 2019, Art. No. 103137, doi: 10.1016/j.jbi.2019.103137.

[32] Kao, K. J.—Chang, K. M.—Hsu, H. C.—Huang, A. T.: Correlation of Microarray-Based Breast Cancer Molecular Subtypes and Clinical Outcomes: Implications for Treatment Optimization. BMC Cancer, Vol. 11, 2011, No. 1, Art. No. 143, doi: 10.1186/1471-2407-11-143.

[33] Bezdek, J. C.—Ehrlich, R.—Full, W.: FCM: The Fuzzy c-Means Clustering Algorithm. Computers and Geosciences, Vol. 10, 1984, No. 2-3, pp. 191–203, doi: 10.1016/0098-3004(84)90020-7.

[34] Davtalab, R.—Dezfoulian, M. H.—Mansoorizadeh, M.: Multi-Level Fuzzy Min-Max Neural Network Classifier. IEEE Transactions on Neural Networks and Learning Systems, Vol. 25, 2014, No. 3, pp. 470–482, doi: 10.1109/TNNLS.2013.2275937.

[35] Alesheykh, R.: Comparative Analysis of Machine Learning Algorithms with Optimization Purposes. Control and Optimization in Applied Mathematics (COAM), Vol. 1, 2016, No. 2, pp. 63–75, `https://mathco.journals.pnu.ac.ir/article_3399_374.html`.

[36] Kalchbrenner, N.—Grefenstette, E.—Blunsom, P.: A Convolutional Neural Network for Modelling Sentences. CoRR, 2014, doi: 10.48550/arXiv.1404.2188.

[37] Xu, Q.—Zhang, M.—Gu, Z.—Pan, G.: Overfitting Remedy by Sparsifying Regularization on Fully-Connected Layers of CNNs. Neurocomputing, Vol. 328, 2019, pp. 69–74, doi: 10.1016/j.neucom.2018.03.080.

[38] Jais, I. K. M.—Ismail, A. R.—Nisa, S. Q.: Adam Optimization Algorithm for Wide and Deep Neural Network. Knowledge Engineering and Data Science (KEDS), Vol. 2, 2019, No. 1, pp. 41–46, doi: 10.17977/um018v2i12019p41-46.

[39] Hastie, T.—Tibshirani, R.—Jerome, F.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2001, doi: 10.1007/978-0-387-21606-5.

[40] Provost, F.—Fawcett, T.: Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media, Inc., 2013.

[41] Piryonesi, S. M.—El-Diraby, T. E.: Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index. Journal of Infrastructure Systems, Vol. 26, 2020, No. 1, Art. No. 04019036, doi: 10.1061/(ASCE)IS.1943-555X.0000512.

[42] Wu, X.—Kumar, V.: The Top Ten Algorithms in Data Mining. CRC Press, 2009.

[43] Piryonesi, S. M.—El-Diraby, T.: Using Data Analytics for Cost-Effective Prediction of Road Conditions: Case of the Pavement Condition Index. FHWA Publication No.: FHWA-HRT-18-065. Technical Report. United States. Federal Highway Administration. Office of Research, Development, and Technology, 2018, `https://rosap.ntl.bts.gov/view/dot/37578`.

[44] Vishwanathan, S. V. M.—Narasimha Murty, M.: SSVM: A Simple SVM Algorithm. Proceedings of the 2002 International Joint Conference on Neural Networks (IJCNN '02), Vol. 3, 2002, pp. 2393–2398, doi: 10.1109/IJCNN.2002.1007516.

[45] Chang, C. C.—Lin, C. J.: LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology (TIST), Vol. 2, 2011, No. 3, Art. No. 27, doi: 10.1145/1961189.1961199.

[46] Hafemeister, C.—Satija, R.: Normalization and Variance Stabilization of Single-Cell RNA-Seq Data Using Regularized Negative Binomial Regression. Genome Biology, Vol. 20, 2019, No. 1, Art. No. 296, doi: 10.1186/s13059-019-1874-1.

[47] Piryonesi, S. M.—El-Diraby, T. E.: Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems. Journal of Transportation Engineering, Part B: Pavements, Vol. 146, 2020, No. 2, Art. No. 04020022, doi: 10.1061/JPEODX.0000175.

[48] Garcia-Pedrajas, N.—Hervas-Martinez, C.—Ortiz-Boyer, D.: Cooperative Coevolution of Artificial Neural Network Ensembles for Pattern Classification. IEEE Transactions on Evolutionary Computation, Vol. 9, 2005, No. 3, pp. 271–302, doi: 10.1109/TEVC.2005.844158.

[49] Chen, S.—Luc, N. M.: RRMSE Voting Regressor: A Weighting Function Based Improvement to Ensemble Regression. CoRR, 2022, doi: 10.48550/arXiv.2207.04837.

[50] Dimitriadou, E.—Weingessel, A.—Hornik, K.: Voting-Merging: An Ensemble Method for Clustering. In: Dorffner, G., Bischof, H., Hornik, K. (Eds.): Artificial Neural Networks – ICANN 2001. Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 2130, 2001, pp. 217–224, doi: 10.1007/3-540-44668-0_31.

[51] Wang, H.—Yang, Y.—Wang, H.—Chen, D.: Soft-Voting Clustering Ensemble. In: Zhou, Z. H., Roli, F., Kittler, J. (Eds.): Multiple Classifier Systems (MCS 2013). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 7872, 2013, pp. 307–318, doi: 10.1007/978-3-642-38067-9_27.

[52] Shahabi, M.—Hassanpour, H.: Using the Artificial Intelligence Techniques for Diagnosing of Intensity of Non-Alcoholic Fatty Liver Disease by Clinical Parameters. Knowledge and Health in Basic Medical Sciences, Vol. 11, 2016, No. 3, pp. 69–75, doi: 10.22100/jkh.v11i3.1369 (in Persian).

[53] Tharwat, A.: Classification Assessment Methods. Applied Computing and Informatics, Vol. 17, 2021, No. 1, pp. 168–192, doi: 10.1016/j.aci.2018.08.003.

[54] Li, B.—Drozd, A.—Guo, Y.—Liu, T.—Matsuoka, S.—Du, X.: Scaling Word2Vec on Big Corpus. Data Science and Engineering, Vol. 4, 2019, No. 2, pp. 157–175, doi: 10.1007/s41019-019-0096-6.

Mohammad Zahaby received his B.Sc. degree in software engineering from the Islamic Azad University, Mobarakeh, Iran, in 2005 and his M.Eng. degree in computer science and engineering (information technology), from the Pune University, Pune, India in 2009. He is a full-time Ph.D. Candidate in computer engineering – software systems at the Borujerd Branch, Islamic Azad University, Borujerd, Iran. His research interests include medical decision support system, deep learning, image processing and machine learning.

Mohammad Ebrahim Shiri is Assistant Professor in the Department of Computer Sciences at Amirkabir University of Technology of Tehran, Iran. He received his Ph.D. from the Department of Computer Sciences at the University of Montreal, Canada, in 1999. His current research interests include artificial intelligence, multi-agent systems, intelligent tutoring systems, machine learning, image processing, and distributed systems.

Hamid Haj Seyyed Javadi received his M.Sc. and his Ph.D. degrees in the Amirkabir University of Technology, Tehran, Iran in 1996 and 2003, respectively. He has been working as a full-time faculty member and Associated Professor in the Department of Mathematics and Computer Science of Shahed University, Tehran, Iran. His research interests are wireless sensor networks, IoT, computer algebra, cryptography and security.

Mostafa Boroumandzadeh is Associated Professor in the Department of Computer Engineering and Information Technology, Payam Noor University, Tehran, Iran. He received his Ph.D. from the Department of Computer Engineering, Islamic Azad University branch of Shiraz, Iran, in 2021. His current research interests include decision support systems, data mining, machine learning, image processing, and distributed systems.