Advances in Industrial Engineering, Spring 2021, 55(2): 151-164
DOI: 10.22059/jieng.2021.326775.1789

RESEARCH PAPER

Hybrid Medical Data Mining Model for Identifying Tumor Severity in Breast Cancer Diagnosis

Faeze Araghi Niknam [a,b], Rouzbeh Ghousi [a,*], AmirHossein Masoumi [a], Alireza Atashi [e], Ahmad Makui [a]

a. *School of Industrial Engineering, Iran University of Science & Technology, Tehran, Iran.*
b. *Department Radiology, Tehran University of Medical Science, Tehran, Iran.*
c. *Department of Medical Informatics, Breast Cancer Research Center, Motamed Cancer Institute, ACECR, Tehran, Iran*

Received: 06 July 2021, Revised: 13 September 2021, Accepted: 16 September 2021

Abstract

Purpose: This study proposes a methodology for detecting tumor severity using data mining of databases relating to breast imaging modalities. In doing so, it proposes creating a software application that can serve as an efficient decision-making support system for medical practitioners, especially those in areas where there is a shortage of modern medical diagnostic devices or specialized practitioners, such as in developing countries.
Method: We investigated the data of approximately 3754 screened women by using "BI-RADS" categories as a quality assessment tool to screening, measure, and identify the size and location of lesions, determine the number of lymph nodes, collect biopsy samples, determine final diagnoses, prognoses, and age which were all available from the screening registry.
Result: The application of each algorithm on BI-RADS values 4 and 5 for Invasive Ductal Carcinoma lesions was assessed, and the following accuracy was acquired: CART: 84.71%. In order to get the best result, four optimum clusters based on tumor size were applied for constructing simple rules with significant confidence.
Conclusion: In this study, we present a hybrid approach - a combination of k-means with GRI and CART decision tree - to better assess breast cancer data sets.

Keywords:
Breast Cancer Prediction;
Mammography;
Ultrasonography;
Medical Data Mining;
Invasive Ductal Carcinoma

Introduction

Knowledge Discovery in Databases (KDD) and the use of other data mining techniques may be valuable tools in discovering hidden patterns to predict the outcomes of diseases at their early stages in an efficient manner [1]. Medical data is frequently produced and captured electronically in everyday clinical practice; however, the collected heterogeneous medical data lacks structural, functional, and semantic interoperability. Automated knowledge discovery techniques, which employ data mining techniques, can support clinicians and discover new relevant patterns in silos of electronic patient data [2,3].

The increasing survival rates and life expectancy of women exposed to breast cancer depend on two major factors. The first one is the diagnosis of the disease at an early stage through mammograms. The second one is advancing treatment such as chemotherapy, hormonal therapy, or surgery to prevent cancer progression. Achieving successful results requires early diagnosis and patient monitoring systems and better and more available health care systems

* Corresponding author: (R. Ghousi)
Email: ghousi@iust.ac.ir

[4,5]. Data mining is a reliable and practical process for image databases of breast cancer screenings, especially for classifying breast tissues, predicting outcomes, and discovering disease patterns to create a support system for each decision-maker involved in all process stages.

Numerous studies have been conducted about cancer diagnosis and recurrence using data mining approaches. Certain researchers have frequently applied two types of classification algorithms, namely "Decision Tree" and "Artificial Neural Network," to diagnose breast cancer with the overall prediction accuracy of 94% and 95.4%, respectively [6]. Another researcher demonstrates that the Decision Tree is the best predictor for the Wisconsin Dataset, which scored an accuracy rate of 94% [7]. Kharya applied the Artificial Neural Network algorithm and achieved an accuracy rate of 86.5%. Other classification algorithms, such as "Naïve Bayes" and "C4.5," produced 84.6% and 86.7% accuracy rates, demonstrating that these algorithms can significantly help clinicians diagnose breast cancer [8]. Researchers applied the Naïve Bayes and "RBF Network" and "J48" algorithms. The average rate of accuracy of reliance on breast cancer datasets indicate that the Naïve Bayes is the best predictor with an accuracy of 97.36% on the holdout sample, RBF Network algorithm came out second with an accuracy rate of 96.77%, and J48 came out third with an accuracy rate of 93.41 [9]. Gupta relied upon data mining techniques to discover hidden patterns of breast cancer diagnosis and revealed that ANN produced the highest accuracy than other classification techniques for early diagnosis and biopsy avoidance [10]. Sahu et al. suggested a hybrid feature selection method based on Principal Component Analysis and Artificial Neural Network for breast cancer classification and diagnosis [11]. Khamparia et al. proposed a hybrid transfer learning model based on MVGG and ImageNet on Mammography images to early detection of breast cancer [12]. Chaurasia et al. compared six different machine learning techniques on Wisconsin Breast Cancer data set before and after feature selection. Their results showed that a good feature selection could improve the performance of the algorithms [13]. Farid et al. proposed a synthetic model set of features to optimize the genetic algorithm to predict breast cancer. Their model significantly outperformed the single filter approaches and principal component analysis for optimum feature selection [14].

What distinguished the present study from its predecessors is its simultaneous reliance on two different data mining techniques to analyze breast cancer screening databases. The present study applies a combination of clustering algorithms (applying "K-means") and classification algorithms (applying "GRI" and "C&R Tree") to predict breast cancer and determine the extent of malignant tumors. In doing so, this study presents a new hybrid predictive model.

In order to conduct our research in an efficient manner and to come up with the most effective predictive model for the early detection of breast cancer, we were able to gain access to the raw medical data of 3754 breast screening data captured between 1997 and 2016 with mammogram and ultrasound interpretations. This dataset was obtained with the patients' consent from the Motamed Cancer Institute in Tehran, Iran. This dataset was categorized based on BI-RADS classification, according to patients' age (ranging from 20 to 80 years old) and size of the tumor (ranging from 1 to 5 cm). A unique identifier was generated automatically for each patient's record regarding the data's patient privacy and confidentiality. Thus the final dataset contains 3754 records and nine attributes (except ID). Our study used pathology (or cytology and altered kind of biopsy) where all patients were diagnosed with breast cancer.

Methodology

Clinicians use medical data mining as a tool to facilitate decision-making in respect of early diagnosis of disease by combining two or more elements of patient data to predict clinical outcomes. This method, which addresses medical data pre-processing, semantic

interoperability, and patient data privacy protection, was proposed and experimentally tested in healthcare management. It uses data analysis methods such as statistics, machine learning, and artificial intelligence to attain new non-trivial knowledge, e.g., prediction values, hidden patterns, and dependencies. CRISP-MED-DM's created data mining application methodology extends the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology [15].

In the following, we will briefly describe the activities carried out in each phase of the CRISP-MED-DM process using the raw data of breast cancer screenings in 6 steps to determine tumor severity in the breast tissue reduce unnecessary biopsies. Relying on the CRISP-MED-DM method and three Data Mining approaches, we developed our system as a hybrid model consisting of clustering and classification techniques.

Fig 1 illustrates the phases of data mining involved in our hybrid model and the algorithms used. The data set used consists of information relating to screenings of women aged between 20-80 years of age classified into two sets depending on the BI-RADS value. The data relating to BI-RADS value equaling or lower than the value of 3 was excluded. Contrary to them, the data relating to BI-RADS value equaling or higher than the value of 4, of which 3754 cases, was used in the data mining process.

The CRISP-MED-DM is a data mining process consisting of 6 stages. The first stage is "Defining the Problem," which is the problem of diagnosing breast cancer in its earlier stages through the volume of medical data accumulated through patient screenings. The second stage is "Data Understanding," which collects and classifies the raw data obtained from participating medical institutions. The data used in this study, its validity, integrity, and completeness were confirmed by several screening radiologists responsible for evaluating the mammographic examinations for all women sent to the hospital for diagnostic workup. A unique code was generated automatically for each patient's medical data to preserve patients' privacy and medical confidentiality. The mammography results, ultrasound, or combination of mammography with ultrasound examination were stated in detail by the Breast Imaging Reporting and Data System (BI-RADS) categories. This database also includes radiology interpretation data, biopsy technique data, tumor size, location, and lymph node number, which had all been identified and reported in mammography and ultrasound images. Besides, information regarding conventional risk factors, such as menopause family history, fertility, or other clinical data, such as the patient's chief complaint underlying the diagnostic work-up, are included.

The third stage is "Data Preparation," which harmonizes the data through data gap filling within a dataset. For example, supplementing missing data from mammography screenings with those obtained from ultrasounds or selecting the most relevant features of the radiological data within a dataset, especially through the BI-RADS lexicon.

The fourth stage, "Modeling," is the most complicated and essential stage of this medical data mining process. As explained above, for the classification of breast cancer data, a practical hybrid approach is proposed using data mining techniques consisting of clustering, association rule, and classification applied in three separate steps. In the first step, "clustering," the clustering method is applied to the dataset using the "K-means" algorithm. This algorithm can be applied in the primary stage of data pruning [16,17]. Most of the samples were clustered using the K-means algorithm (implemented by CLUTO). Because of its simplicity of description, it is susceptible to outliers. K-means is a semi-parametric method and is an informal technique that can classify data sets by assuming K clusters. K-means chief advantage is the fast computation for the large flexible data set if the number of clusters is small [18]. In the first step of this stage, 4 clusters of data are created based on the principle of most and slightest similarity within the groups, the accuracy of which is tested by the "Davies Bouldin Index," "Dunn's index," "SSE" and the "Silhouette Index" [19,20]

In the second step of the Modeling stage, "Association Rule," the Clustering step results are analyzed using the "GRI" algorithm. Association aims to form interactions between existing articles with organized order in a given record. Association rules are concerned with detecting existing association relationships, which are hidden in databases [21]. The most common tools for association modeling are statistics and *a priori* algorithm [22]. The application of the association rule consists of a two-pronged examination:

- Support (s): It indicates how frequently an item occurs or finds the frequency of an item within a dataset.

For example, Consider the rule A=>B; it is supported if it includes A and B together.

(A U B) $Support\ A => B = Support = P\ AUB$

- Confidence (c): Number of times the statement is found to be true.

For example, Consider the rule A=>B Confidence if it includes the above A together with B. $Confidence\ A => B = Support\ AUB / Support\ A$ = p (B/A) $Support\ A$

Both support and confidence represent the positive correlation and take the value between 0 and 1. It is used to find patterns of association among the attributes or variables and observations. The result of this algorithm's application produces rules defined to identify the different types of breast mass, which was subsequently tested against and confirmed by the Support, Confidence, and Lift indexes. The rules of law provide high-level rules of information based on synchronization (support) and precision (index) indices. The statement's rules are "if the front part of the rule is then the tale of the rule" [23]. Generalized Rules Induction (GRI) algorithm extracts several applicable rules from the source concerning the relationships among the data [24].

In the third and final step of the Modeling stage, the "Decision Tree" or "C&R Tree" algorithm uses recursive data partitioning to extract models that describe the trends of correlated predicting patterns whose performance may have been influenced by data segmentation. "Decision Tree Algorithm " Is a model that includes both predictive and descriptive approaches. It uses recursive data partitioning to extract models that describe the correlated trends predicting future data whose performance may be affected by data segmentation. It classifies various units into specific classes based upon the attributes of the objects. Internal nodes or leaves follow a root. Each leave is labeled with a question and includes a few instances to obtain reliable predictions. An arc associated with each node covers all possible responses in which decision rules in the form of 'IF condition-based-on attribute- values THEN outcome value may be constructed [25]. Classification and Regression Tree's main characteristic is generating regression trees and classifications; it can also handle numerical and categorical variables [26]. In the fifth stage of the CRISP-MED-DM data mining process, "Evaluation," the data resulting from the Modeling stage is randomly divided into "training" and "test" sets of data. The test set confirmed the results of the train set. Ultimately, the results obtained in the data mining process were subsequently tested against the results obtained through cytology.

In the sixth stage, "Deployment," the data resulting from the Evaluation stage are subsequently used and confirmed by radiologists in order to diagnose breast cancer within a unified knowledge-based system.

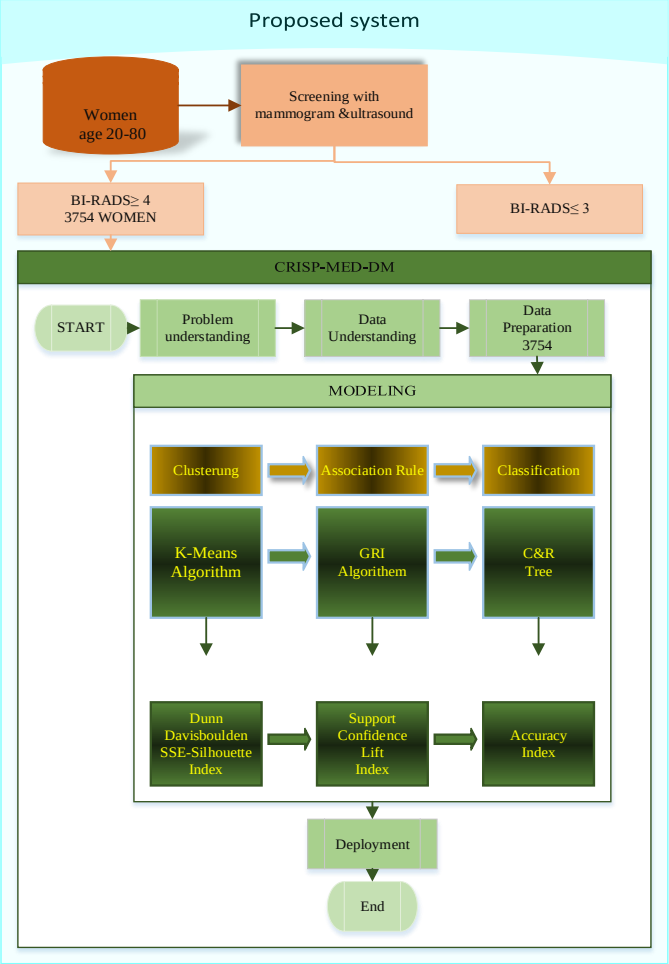


Fig 1. The framework of the proposed system

Results

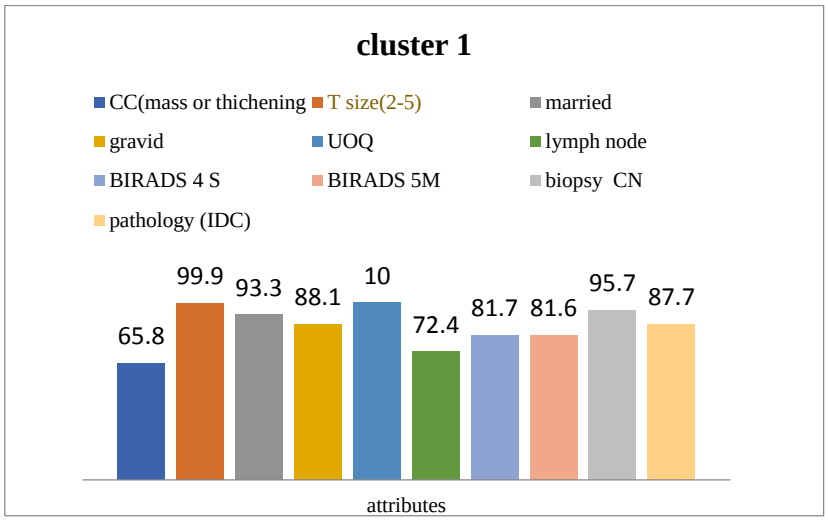
We propose an adaptive K-means algorithm to detect malignant tumors using datasets resulting from mammograms combined with ultrasound images in breast cancer screening. Our dataset consisted of medical data obtained from 3754 patients distributed in the four clusters. The clusters were distinguished based on tumor size, the number of lymph nodes, and their maturity. The chief complaint of these patients was sensitivity to the touch of the mass and thickening.

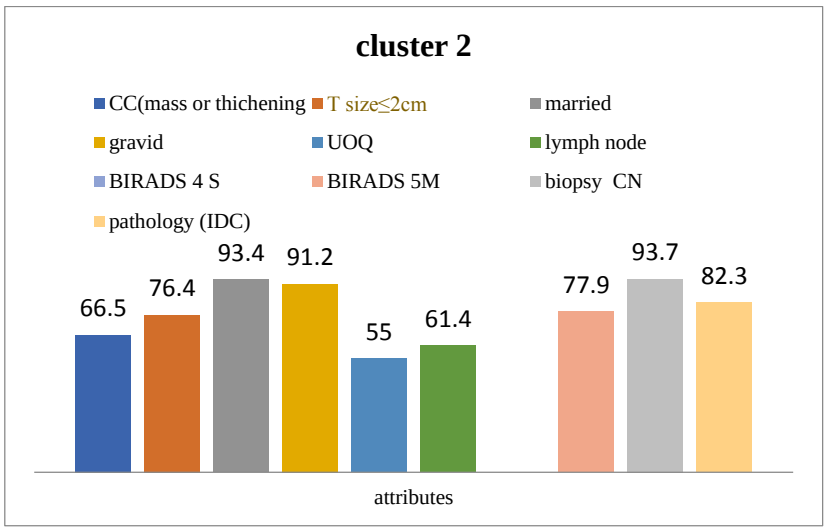
Mammographic lesions were categorized based on the BI-RADS Lexicon (mammography five and ultrasound 4) in the upper outer quadrant (UOQ) of the breast tissue, which was biopsied by core needle and their pathology-proven (Invasive Ductal Carcinoma). In cluster one, the most significant cluster, the datasets of 1794 women aged approximately 47 years old were diagnosed with a mass in 2-5 cm in their breast tissue. In the second cluster, the datasets of 594 patients aged around 48 were gathered based on tumors size ≤2cm. The data of 406 women around 46 years and the size of their tumors were ≥5cm and collected in the third cluster. Moreover, the data of 959 women around the age of 48 with tumor sizes of "2-5 cm" are gathered in the fourth cluster.

By studying the importance of the selected attributes, our analysis was that almost all of the selected attributes are at a high level of importance equal to 1. Analyzed and recorded Results highlight different tumor size measurements (≤2cm; ≥5cm; 2-5 cm) and have features in each of the four clusters, provided in Fig. 2. The data within each cluster indicates a high degree of dependency of the attributes to their clusters.

Association rules were extracted from the data contained in the four clusters through GRI algorithms. The association rules', consisting of antecedent and consequent elements, contain the clinical and preclinical information of patients and convert this information into rules that specify the types of tumors present in a given patient. In the context of our research, the association rule can be used to find the linking between a patient's characteristics and imaging techniques to diagnose breast cancer disease. We determined the attribute "biopsy pathology" as a consequent and other attributes as antecedents. The results obtained through cytology are used to test the results of the rules. In order to select interesting rules from the set of all possible rules, we used three different indexes: Support, Confidence, and lift. Support is an indication of how frequently the itemset appears in the dataset. Confidence is an indication of how often the rule has been found to be true. The lift value of an association rule is the ratio of the rule's confidence and the expected confidence of the rule. We chose the rules with lift more than one because making those rules is potentially helpful in predicting the consequent in future data sets. After that, we filtered the rules with higher support and confidence value.

Table 1 illustrates that approximately 100 association rules result from the 4 clusters. The difference between clusters 2 and 4 is in the number of people in each cluster at different ages. In cluster 2, the mean age of patients at the time of diagnosis was 47.3 years and 1794 patients, and in cluster 4, the mean age of patients at the time of diagnosis was 46.3 and 594 patients. So we displayed them in separate clusters.





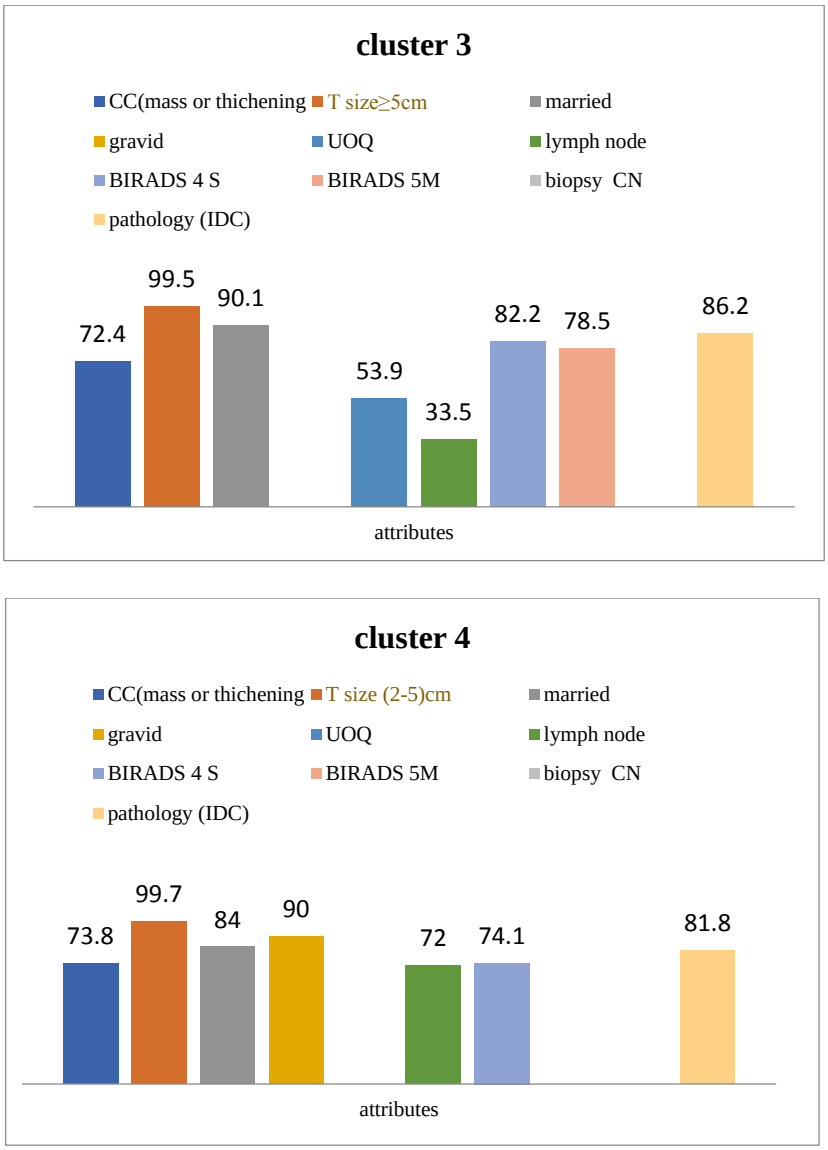


Fig. 2. Information of attributes in each cluster

Table 2 illustrates the most prominent rules extracted in each cluster, which are explained in the mathematical method. These are selected regarding the degree of certainty based on the opinion of the radiologists. The characteristics of the lesions were found in breast tissue are determined regardless of their frequency among patients. We used input factors such as lesion location, lesion size, personal information and made the feature selection by the Generalized Rules Induction algorithm.

Table 1. The outcome of the rules in each cluster

Clusters	Tumors size	GRI rules number	Consequents
Cluster 1	2-5 cm	8	Sarcoma
		25	Mucinous carcinoma
		2	Lobular carcinoma in situ
		28	Invasive lobular carcinoma
		12	Invasive ductal carcinoma
		33	Ductal carcinoma in situ
Cluster 2	≤ 2 cm	3	Paget's disease
		2	Micro invasion
		3	Medullary carcinoma
		6	Invasive ductal carcinoma
		8	Inflammatory carcinoma
Cluster 3	≥ 5 cm	23	Paget's disease
		21	Micro invasion
		10	Lobular carcinoma in situ
		15	Invasive tubular carcinoma
		7	Invasive ductal carcinoma
		19	Inflammatory carcinoma
Cluster 4	2-5 cm	13	Sarcoma
		6	Paget's disease
		40	Mucinous carcinoma
		4	Phyllodes tumor
		9	Invasive ductal carcinoma
		2	Ductal carcinoma in situ

Finally, we developed an application based on the C&R Tree algorithm, which was utilized as a tumor classification tool according to the criteria recommended by ACR (BI-RADS lexicon). The results of the application of the GRI algorithm are incorporated into the C&R Tree algorithm. This algorithm facilitates recognizing the dispersion of tumor severity to reduce the need for invasive surgeries such as mastectomy due to early detections of cancerous tumors. Fig. 3 is a descriptive image detailing the types of tumor severity and their distribution percent in the dataset. The accuracy rate acquired for the applications of the C&R Tree algorithm to the dataset is 84.97%.

Table 2. The most crucial extract association rules in four clusters

No.	Consequent	Antecedent	Support Index	Confidence Index
1	Mucinous carcinoma	CC :follow-up ∩_location _ uoq,∩ uiq ∩BI-RADS so: highly suggestive of malignancy	0.06%	100%
2	mucinous carcinoma	Family first degree (1 person) ∩LN = 4_9 ∩T_2-5cm ∩ BI-RADS so: suspicious finding	0.01%	100%
3	sarcoma	CC: mass or thickening ∩ BI-RADS ma: highly suspicious of malignancy ∩ location: lower half ∩ BI-RADS so: Highly suggestive of malignancy	0.1%	100%
4	invasive lobular carcinoma	BI-RADS ma: highly suspicious of malignancy∩ location: central(nipple areole) ∩ location lower half ∩ Biopsy = core needle	0.17%	100%
5	invasive lobular carcinoma	LN = 4_9 ∩mass or thickening ∩BI-RADS ma: highly suspicious of malignancy	2.496%	100%
6	Invasive ductal carcinoma	menopauses: no ∩ BI-RADS ma: highly suspicious of malignancy∩ tumors size: ≤2cm∩ BI-RADS so: malignancy	3.65%	100%
7	micro invasion	marital ∩ BI-RADS ma: highly suspicious of malignancy ∩ location: upper half ∩ sign Column < 1.500	0.34%	100%
8	ductal carcinoma in situ	CC: screening ∩ BI-RADS ma: suspicious finding ∩location: uoq ∩sign Column < 12.500	0.17%	100%
9	lobular carcinoma in situ	Family first degree (1 person) ∩CC: mass or thickening ∩ location: upper half ∩sign Column < 21.208	0.11%	100%
10	inflammatory carcinoma	CC: pain ∩ BI-RADS ma: highly suspicious of malignancy ∩ location: uoq ∩T _inflammatory	1.01%	100%
11	Paget's disease	CC: skin symptom ∩ location: central(nipple areole) ∩ sign Column < 7.500	0.34%	100.0%
12	invasive tubular carcinoma	Location: central (nipple areole) ∩ LN = 1_3 ∩ BI-RADS so : suspicious finding∩ T≥5cm	0.25%	100.0%
13	Phylloid tumor	CC: mass or thickening and skin symptom ∩ LN, ∩ T=2-5cm	0.31%	100.%

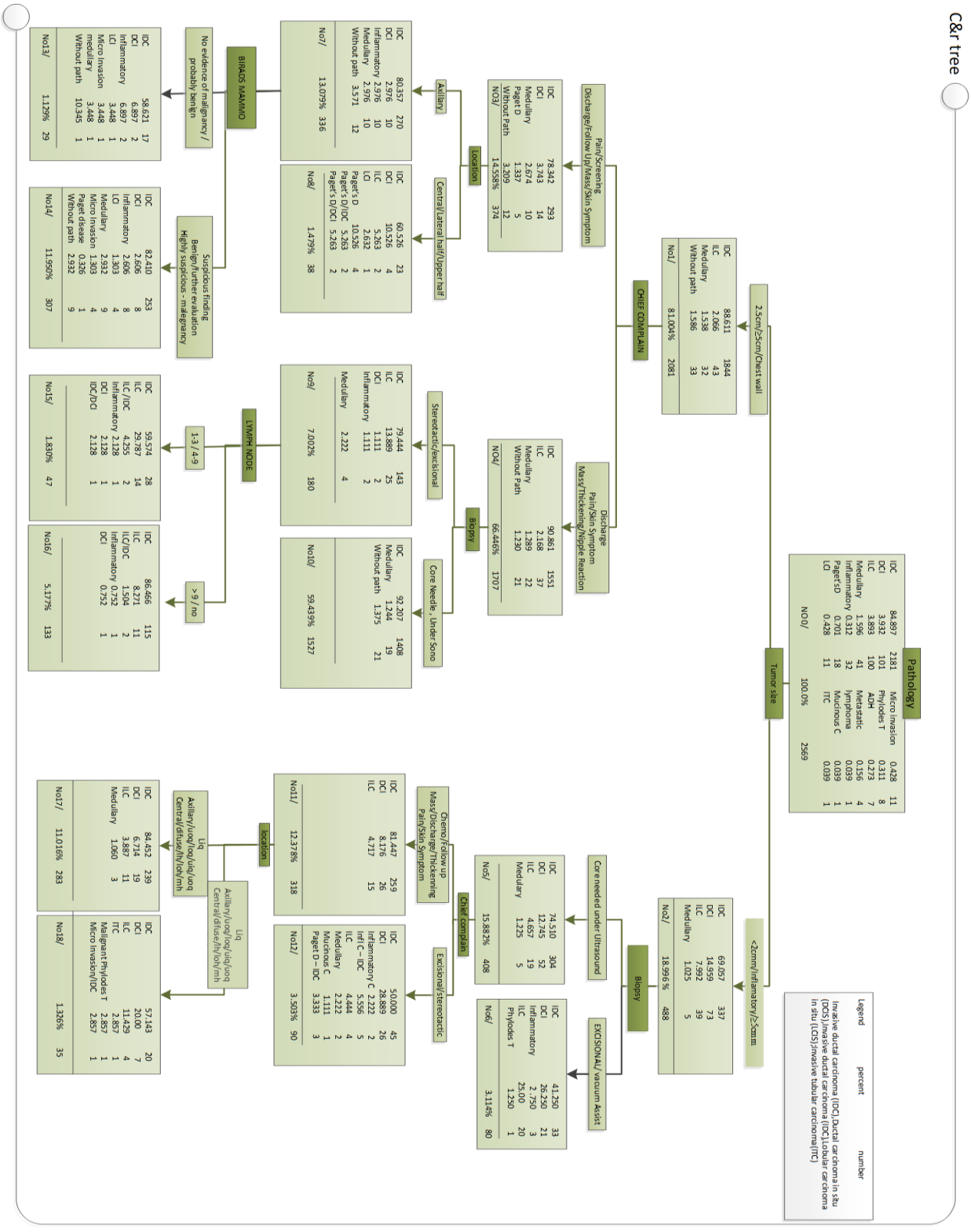


Fig. 3. Tree of C&R Tree for classifying the tumor severity

Discussion

The present study aimed to correctly identify tumor severity for the early detection of breast cancer among women. Medical data of 3754 patients between 20 and 80 years old who had

been screened from 1997 to 2016 was investigated. The data was analyzed in "WEKA," using attributes such as BI-RADS 4 lexicon and data relating to tumor size measurements, number of lymph nodes, and location of tumors in breast tissue, patient's age, and the patient's chief complaint. These attributes were used in the hybrid modeling process, and the results obtained were tested against and confirmed by the results of the pathologies taken from tumor biopsies.

Our results are similar to those found in other studies using the BI-RADS lexicon and were comparable with those from another study from the Netherlands that reported BI-RADS categories 4 and 5 in a screening population [27,28]. While our study is similar to some previous research conducted on this same issue (i.e., the use of medical data mining for the early detection of cancerous tumors), for example, through its reliance on similar parameters for data classification used. However, what sets this study apart from its predecessors is our proposal to use its results practically. We proposed for the results of our modeling process to be incorporated into an accessible software program, ideally through application technology, to design a practical decision-making support system application for medical practitioners involved in every stage of the cancer detection process. When medical practitioners discover tissues that appear to have tumor-like qualities, they can input the information concerning the attributes of the tumor into this proposed software ("Application"), and the Application will specify or predict whether the designated parameters coincide with the characteristics of the tumors that have been identified within the Application.

On the one hand, women at high-risk ages do not have access to skilled radiologists and modern technologies in many areas. However, on the other hand, it is not cost-beneficial to screen all of the patients with an expert radiologist who can diagnose breast cancer in its early stages. Therefore, by incorporating the rules obtained from a data mining process proposed in this study, problem-solving could help less experienced physicians to differentiate between benign and malignant tumors by a highly reliable decision support system, such as the hybrid model-based Application proposed in this study. One potential weakness of this study is that we could not obtain data relating to breast density. However, if we can obtain such information and incorporate it within the other parameters that have been applied in the hybrid modeling process proposed in this study, the results would indeed be more comprehensive [29,30]. Thus, our study results do not represent a final finding, and further imaging data of breast density must be performed.

Last but not least, the most important contributions of this study are as follows. The topic of evaluating breast cancer screening technology has been done in different ways in Iran. The present study is based on the Agency for Healthcare Research and Quality proposal that suggested using meta-analysis on breast cancer screening schemes. The development of the CRISP-DM method could create a hybrid model using the minimum parameters summarized from mammography and ultrasound reports. So, this study identified the different types of breast cancer tumors based on their characteristics, age indicators, and family records of patients to classify the malignant breast tumors.

The hybrid model set out in this study will drastically reduce costs and support the healthcare system through early diagnosis. It will reduce the need for unnecessary biopsies and possible side effects arising from invasive interventions. Therefore, this study's results can be used to establish a decision-making support system for physicians, especially surgeons in underprivileged communities and specialized clinical and para-clinical services providers.

Conclusion

Policymakers should be conscious of BI-RADS possibilities as a stratification tool and consider reviewing the current policies to reduce waiting times, costs, and unnecessary anxiety [31]. Since managers need precise decision-making, data mining methods may be used as effective

decision-making support systems [32]. This paper uses datasets of breast cancer diagnoses to demonstrate that medical data mining techniques are beneficial for discovering hidden patterns in the data that clinicians can utilize for faster and efficient clinical decision-making.

Based on the experiences and the results of implementing the proposed approach, some points have worth to be considered in future research. It seems that generating a tool allows a mathematical equation to be solved by aligning tumor size, location, lymph node numbers, and breast density could be studied as future research. This model could be matched with breast BI-RADS categories by practitioners like FRAX for clinical support in metabolic bone diseases and sarcoma nomogram diagram.

Acknowledgment

The authors sincerely thank Dr.Davood Aghakhani and Dr.Sayed Mahdi Tayebi Tafreshi. We also thank the radiologist Dr.Sayed Peyman Mosavi and Mr. Sayed Saeed Hashemiyan from the Imaging department of Milad Hospital. The present paper is derived from Miss. Faezeh Araghi Niknam's master thesis titled "Health technology assessment of breast cancer screening using medical data mining in imaging techniques to prevent and reduce cost."

References

[1] Keleş, M.K., *Breast cancer prediction and detection using data mining classification algorithms: a comparative study.* Tehnički vjesnik, 2019. 26(1): p. 149-155.

[2] Ghousi, R., *Applying a decision support system for accident analysis by using data mining approach: A case study on one of the Iranian manufactures.* Journal of Industrial and Systems Engineering, 2015. 8(3): p. 59-76.

[3] Sohrabei, S. and A. Atashi, *Performance Analysis of Data Mining Techniques for the Prediction Breast Cancer Risk on Big Data.* Frontiers in Health Informatics, 2021. 10(1): p. 83.

[4] Diz, J., G. Marreiros, and A. Freitas, *Applying data mining techniques to improve breast cancer diagnosis.* Journal of medical systems, 2016. 40(9): p. 1-7.

[5] Masoumi, A., et al., *A quantitative scoring system to compare the degree of COVID-19 infection in patients' lungs during the three peaks of the pandemic in Iran.* Journal of Industrial and Systems Engineering, 2021. 13(3): p. 61-69.

[6] Higa, A., *Diagnosis of breast cancer using decision tree and artificial neural network algorithms.* cell, 2018. 1: p. 10.

[7] Ghorbani, R. and R. Ghousi, *Predictive data mining approaches in medical diagnosis: A review of some diseases prediction.* International Journal of Data and Network Science, 2019. 3(2): p. 47-70.

[8] Kharya, S., *Using data mining techniques for diagnosis and prognosis of cancer disease.* arXiv preprint arXiv:1205.1923, 2012.

[9] Chaurasia, V., S. Pal, and B. Tiwari, *Prediction of benign and malignant breast cancer using data mining techniques.* Journal of Algorithms & Computational Technology, 2018. 12(2): p. 119-126.

[10] Gupta, S., D. Kumar, and A. Sharma, *Data mining classification techniques applied for breast cancer diagnosis and prognosis.* Indian Journal of Computer Science and Engineering (IJCSE), 2011. 2(2): p. 188-195.

[11] Sahu, B., S. Mohanty, and S. Rout, *A hybrid approach for breast cancer classification and diagnosis.* EAI Endorsed Transactions on Scalable Information Systems, 2019. 6(20).

[12] Khamparia, A., et al., *Diagnosis of breast cancer based on modern mammography using hybrid transfer learning.* Multidimensional systems and signal processing, 2021. 32(2): p. 747-765.

[13] Chaurasia, V. and S. Pal, *Applications of machine learning techniques to predict diagnostic breast cancer.* SN Computer Science, 2020. 1(5): p. 1-11.

[14] Farid, A.A., G. Selim, and H. Khater, *A Composite Hybrid Feature Selection Learning-Based Optimization of Genetic Algorithm For Breast Cancer Detection.* 2020.

[15] Niaksu, O., *CRISP data mining methodology extension for medical domain.* Baltic Journal of Modern Computing, 2015. 3(2): p. 92.
[16] Dubey, A.K., U. Gupta, and S. Jain, *Analysis of k-means clustering approach on the breast cancer Wisconsin dataset.* International journal of computer assisted radiology and surgery, 2016. 11(11): p. 2033-2047.
[17] Liu, Y., et al. *Understanding of internal clustering validation measures*. in *2010 IEEE international conference on data mining*. 2010. IEEE.
[18] Mahmud, M.S., M.M. Rahman, and M.N. Akhtar. *Improvement of K-means clustering algorithm with better initial centroids based on weighted average.* in *2012 7th International Conference on Electrical and Computer Engineering*. 2012. IEEE.
[19] Mughnyanti, M., S. Efendi, and M. Zarlis. *Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation*. in *IOP Conference Series: Materials Science and Engineering*. 2020. IOP Publishing.
[20] Sinaga, K.P. and M.-S. Yang, *Unsupervised K-means clustering algorithm.* IEEE Access, 2020. 8: p. 80716-80727.
[21] Brijs, T., et al., *Building an association rules framework to improve product assortment decisions.* Data Mining and Knowledge Discovery, 2004. 8(1): p. 7-23.
[22] Erpolat, S., *Comparison of Apriori and FP-Growth Algorithms on Determination of Association Rules in Authorized Automobile Service Centres.* Anadolu University Journal of Social Sciences, 2012. 12(2): p. 137-146.
[23] Özseyhan, C., B. Badur and O.N. Darcan, *An association rule-based recommendation engine for an online dating site.* Communications of the IBIMA, 2012. 2012: p. 1.
[24] Hu, R., *Medical data mining based on association rules.* Computer and information science, 2010. 3(4): p. 104.
[25] Vougas, K., et al., *Machine learning and data mining frameworks for predicting drug response in cancer: An overview and a novel in silico screening process based on association rule mining.* Pharmacology & therapeutics, 2019. 203: p. 107395.
[26] Breiman, L., et al., *Classification and regression trees*. 2017: Routledge.
[27] Yan, S., L. Zhang, and C. Song, *Applying a new maximum local asymmetry feature analysis method to improve near-term breast cancer risk prediction.* Physics in Medicine & Biology, 2018. 63(20): p. 205010.
[28] Mohapatra, S.K., et al., *The Positive Predictive Values of the Breast Imaging Reporting and Data System (BI-RADS) 4 Lesions and its Mammographic Morphological Features.* Indian Journal of Surgical Oncology, 2021. 12(1): p. 182-189.
[29] Trieu, P.D., et al., *Reader characteristics and mammogram features associated with breast imaging reporting scores.* The British Journal of Radiology, 2020. 93(1114): p. 20200363.
[30] Bihrmann, K., et al., *Performance of systematic and non-systematic ('opportunistic') screening mammography: a comparative study from Denmark.* Journal of Medical Screening, 2008. 15(1): p. 23-26.
[31] Elmore, J.G., et al., *International variation in screening mammography interpretations in community-based programs.* Journal of the National Cancer Institute, 2003. 95(18): p. 1384-1393.
[32] Van der Steeg, A., et al., *Effect of abnormal screening mammogram on quality of life.* Journal of British Surgery, 2011. 98(4): p. 537-542.

Appendix

Table A.1. the definition and analysis of clinical data models and clinical protocols used in data source systems

No.	Abbreviation	Definition Terminology
1	BC	Breast Cancer
2	family	Cancer history in the family
3	Age diag	Diagnosis breast cancer age
4	CC	Chief complain
5	BI-RADS	Breast Imaging and Reporting Data System
6	location	UOQ: upper outer quadrant - LIQ: Lower inner quadrant -Diffuse LOQ: Lower outer quadrant -Central / Medial- UIQ: Upper inner quadrant
7	mammograms	Using x-ray for breast imaging
8	ultrasound	Using ultrasound to detect a lesion in breast tissue
9	CT. Scan	Computerized Tomography Scan
10	MRI	Magnetic Resonance Imaging
11	T	Tumor size
12	LN	Lymph nodes Involvement
13	Bio	Biopsy
14	Bio. Path	Cytology of sample
15	Tumors severity	Carcinoma - Invasive ductal carcinoma- Sarcoma - Inflammatory breast cancer- Lobular carcinoma in situ - Paget s Disease - phyllodes
16	CRISP-MED-DM	Cross-Industry Standard Process for the Medical Data Mining