A NOVEL INTERPRETABLE WEB-BASED TOOL ON THE ASSOCIATIVE CLASSIFICATION METHODS: AN APPLICATION ON BREAST CANCER DATASET

A. K. Arslan, Z. Tunc, I. Balikci Cicek and C. Colak

Abstract*— *Aim: The second-largest cause of cancer mortality for women is breast cancer. The main techniques for diagnosing breast cancer are mammography and tumor biopsy accompanied by histopathological studies. The mammograms are not detective of all subtypes of breast tumors, particularly those which arise and are more aggressive in young women or women with dense breast tissue. Circulating prognostic molecules and liquid biopsy approaches to detect breast cancer and the death risk are desperately essential. The purpose of this study is to develop a web-based tool for the use of the associative classification method that can classify breast cancer using the association rules method.

Materials and Methods: In this study, an open-access dataset named "Breast Cancer Wisconsin (Diagnostic) Data Set" was used for the classification. To create this web-based application, the Shiny library is used, which allows the design of interactive web-based applications based on the R programming language. Classification based on association rules (CBAR) and regularized class association rules (RCAR) are utilized to classify breast cancer (malignant/benign) based on the generated rules.

Results: Based on the classification results of breast cancer, accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1-score values obtained from the CBAR model are 0.954, 0.951, 0.939, 0.964, 0.939, 0.964, and 0.939 respectively.

Conclusion: In the analysis of the open-access dataset, the proposed model has a distinctive feature in classifying breast cancer based on the performance metrics. The associative classification software developed based on CBAR produces successful predictions in the classification of breast cancer. The hypothesis established within the scope of the purpose of this study has been confirmed as the similar estimates are achieved with the results of other papers in the classification of breast cancer.

Keywords— Artificial intelligence, association rules, associative classification, web-based software, breast cancer.

Ahmet Kadir ARSLAN, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (arslan.ahmet@inonu.edu.tr)

Zeynep TUNC, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (zeynep.tunc@inonu.edu.tr)

İpek BALIKCI CICEK, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (ipek.balikci@inonu.edu.tr)

Cemil COLAK, Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey, (cemil.colak@inonu.edu.tr)

Manuscript received May 14, 2020; accepted June, 11, 2020.
Digital Object Identifier:

1. INTRODUCTION

The second most common reason for death in adult women is breast cancer, which is diagnosed annually in more than 2 million new cases. While in particular, in developing countries where 5-year survival rates have been 90 percent or higher for invasive breast cancer, breast cancer survival has increased dramatically over the previous three decades as a result of improved early detection and improved treatment. There is a sharp increase in the global incidence of breast cancer, according to the World Health Organization as a result of improvements in lifestyles, reproductive factors, and life expectancy. Among middle and low-income countries, 58 percent of all breast cancer deaths occur. While breast cancer survival rates in developed countries are approximately 80%, the rate drops to 60% in the middle- and 40% in low-income nations due to lack of early-screening programs which lead to incurable diagnoses in late-stage 80% of these tumors. Mammography and other expensive and technically complex methods cannot be done in middle and low-income countries due to high costs and shortages of trained staff. Furthermore, ER-positive breast cancer is more likely to be identified by mammograms and not indicated for younger people. Therefore, earlier diagnosis using traditional methods for all race classes are not predicted; for example, a small-sized African-American woman with metastases is more probable than a Caucasian woman. Hence, there is a significant ethnic difference in the survival of breast cancer with higher breast cancer mortality rates [1, 2].

Today, the development of computer technologies thanks to technological possibilities has led to the collection of large amounts of data in databases and to access these data more easily. As the amount of data collected grows and the complexity in the data structure collected increases, the need for much better analysis techniques also increases simultaneously. At this point, the concept of Knowledge Discovery in Databases has emerged in the past decades. Knowledge discovery is the process of finding new, previously unknown, and useful information in databases. The knowledge discovery process includes data selection, data preprocessing, data conversion, data mining, and evaluation stages. One of the important stages of the information discovery process is called data mining. Data mining is an interdisciplinary field defined as revealing the relationships and patterns hidden in the data. Data mining is the search for the relations and rules that will allow us to make predictions of a large amount of data using computer programs. According to the definition of data mining, the main purpose is to keep a large amount of data in the data warehouse and obtain meaningful information from these data [3, 4].

There are many models used in data mining. These models are examined under four main categories: association rules (ARs) analysis, classification, clustering, and predictive models [5]. ARs, which are one of the data mining models, are under the name of "association rules analysis", which has a wide usage area in many fields such as economy, education, telecommunication, and medicine. ARs are widely utilized in data mining due to their usefulness and easy understanding and are the process of finding associations, relationships, and patterns among the data as rules. ARs express the occurrence of events together with certain possibilities [6, 7].

Associative classification is a branch of scientific work, known as data mining in artificial intelligence. Associative classification combines the association rules and classification, two known methods of data mining, to create a model for predictive purposes. Specifically, associative classification is a type of classification approach that is built with a set of rules obtained by the association rule mining to form classification models. While the classification and association rules are the prediction of the class labels of the main purpose of the classification, they have similar tasks in data mining, except that the association rule is a method used to find common patterns, correlations, associations, or causal structures from datasets. In the past few years, association rules methods have been successfully used to create correct classifiers in associative classification. [8].

One of the main advantages of using a classification based on association rules according to classical classification approaches is that the output of an associative classification algorithm is represented by simple if-then rules, making it easier for the user to understand and interpret it. [8]. For this reason, the current study aims to develop a new user-friendly web-based software to realize the use of the associative classification method that can classify breast cancer using the association rules method. For this purpose, the main hypotheses of this study are to determine whether classification-based association rules models are successful in predicting breast cancer on the open-access dataset and evaluate the classification performance.

2. MATERIAL AND METHODS

2.1. Dataset

The open-access dataset named “Breast Cancer Wisconsin (Diagnostic) Data Set” was obtained from the UCI machine learning repository. The dataset consists of 569 samples examined for breast cancer with the ten predictors/inputs and one response/output variables. Of the individuals, 357 (62.7%) were diagnosed as benign, and 212 (37.3%) were diagnosed as malignant. A digitized image of a fine nail aspirate (FNA) of a breast mass is used to measure the characteristics. The characteristics of the presented cell nuclei are described in the image [9]. The explanations about the variables in the data set and their properties are given in Table 1.

TABLE I
EXPLANATIONS ABOUT THE VARIABLES IN THE DATASET AND THEIR PROPERTIES

Variable	Variable Explanation	Variable type	Variable role
Diagnosis	The diagnosis of breast tissues (M = malignant, B = benign)	Qualitative	Output
Radius	Mean distances from the center to perimeter points	Quantitative	Predictor
Texture	The standard deviation of gray-scale values	Quantitative	Predictor
Perimeter	Mean size of the core tumor	Quantitative	Predictor
Area	-	Quantitative	Predictor
Smoothness	Mean of local variation in radius lengths	Quantitative	Predictor
Compactness	(mean of perimeter)^2 / (area - 1)	Quantitative	Predictor
Concavity	Mean of the severity of concave portions of the contour	Quantitative	Predictor
Concave points	mean for the number of concave portions of the contour	Quantitative	Predictor
Symmetry	-	Quantitative	Predictor
Fractal dimension	mean for "coastline approximation" - 1	Quantitative	Predictor

3. ASSOCIATION RULES AND ASSOCIATIVE CLASSIFICATION

Data mining is often the analysis of large datasets to find unexpected relationships with the principle of being both understandable and useful for the owner of the data and summarizing the data with new methods. The relationships and summaries obtained from a data mining application are often called models or patterns. Linear equations, rules, sets, graphs, tree structures, and repetitive patterns in the time series are some of these Association rules that are among the most popular representatives of regional patterns in data mining [10].

Association rules mining is one of the unsupervised data mining tasks that look for the relationship between records in a data set. Association rules are often expressed as if it happens, then this happens. Mostly used for finding interpretable trends and relationships among variables [11]. Association rules are rules with support and confidence measurements in the form of “IF- precursor expression-, IF-successor expression” [12]. The value of support and confidence can be evaluated as units of measure showing the strength of an association rule. Confidence and support values, which are the measures of interestingness for association rules, are shown as follows [13].

D: Data,

ti: Records in data, D = {t1, t2, …, ti, …., tn}

X, Y: Items in rules (precursor and successor)

$X \rightarrow Y$, where X is the precursor and Y is the successor ($X \cap Y = 0$);

Support value (SV):

$$SV(X) = \frac{|t \in D, X \subset t|}{|D|}$$

Confidence value (CV):

$$CV(X \rightarrow Y) = \frac{SV(X \rightarrow Y)}{SV(X)}$$

As can be understood from the formulas, the support value is the ratio of repeated records in the data to the whole data. The confidence value is known as the ratio of the support value of a rule to the support value of the predecessors. The fact that the established rules are strong requires high trust and support values. At the beginning of the procedures, the rules that will remain above the minimum support and minimum confidence values to be determined by the researcher should be taken into consideration, and other information should be eliminated.

Association rules can be considered as a two-step process:

1. Find all common product sets. The predetermined minimum support value defines the frequency of these determined product clusters.
2. Establish strong association rules from common product sets. These rules are defined as the rules that provide minimum support and trust value. [14].

Some algorithms used and developed for association rules are; AIS [13], SETM [15], Apriori [16], Partition [17], RARM (Rapid Association Rule Mining) [18], and CHARM [19]. Among these algorithms, the first one is AIS, and the best known is the Apriori algorithm.

3.1. Apriori Algorithm

The name of the Apriori algorithm is Apriori, meaning "prior" since it gets the information from the previous step [16]. This algorithm is essentially iterative and is used to discover sets of passing items. It is necessary to browse the database many times to find frequently passing sets of items. In the first scan, there are frequently passed item sets that provide the minimum support criteria with one element, and in the following scans, the frequent element sets found in the previous scan are used to produce new potential favorite item sets called candidate sets. The support values of the candidate sets are calculated during scanning, and the sets that provide the minimum support criteria from the candidate sets are the frequent sets of items produced in that transition. Frequent sets of items become candidate sets for the next pass. This process continues until there is no new set of frequent [14]. According to the essence of the Apriori Algorithm, if the k-item set (item set with k elements) provides the minimum support criterion, the subset of this set also provides the minimum support criterion. That is, the support value of a set of items is not greater than the support value of the subset [14].

Association rules share many common features with classification. Both use rules to characterize regularities in a dataset. However, these two methods differ greatly in their goals. While classification focuses on prediction, association rules focus on providing information to the user. In particular, it focuses on detecting and characterizing unexpected relationships between data items [20].

3.2. Associative Classification

Associative classification is a data mining method that combines classification and association rules methods to make predictions. Particularly, an associative classification is an approach that uses rules obtained with association rules to create classification models. Associative classification is a special association rule mining with the target/response/dependent/class variable to the right of the rule obtained. In a rule such as $X \rightarrow Y$, Y must be the target / response / dependent / class variable. One of the primary advantages of employing a classification based on association rules according to classical classification approaches is that simple if-then rules represent the output of an associative classification algorithm. This rule makes it easier for the user to understand and interpret the result [8].

Associative classification and association rules are different methods. Relational classification takes into account only the class attribute in the relevant rules. However, association rules allow multiple attribute values in related rules. In other words, there is no class feature in association rules, an example of unsupervised learning, and a class is given in associative classification, an example of supervised learning. The purpose of the association rules is to discover the relationship between the items in the transaction database, while in the associative classification; the aim is to create a classifier that can predict the classes of test data objects. While association rules can have more than one attribute as a result of a rule, in associative classification, there is an only class attribute as a result of a rule. In association rules, over-fitting is not a problem, but in associative classification, over-fitting is a problem. Overfitting occurs when it performs well in the training data set and poorly in the test data set. Overfitting may be due to a variety of reasons, such as a small amount of training data object or noise [8]. The relational classification consists of three steps [21].

1) Determine the smallest support and confidence values,
2) Create rules and pruning,
3) Classification is made in the light of the meta-rules.

There are many algorithms for associative classification. Some of the methods are; CBAR (Classification based on

association rules), wCBAR (Weighted classification based on association rules), CARGBA (Classification based on association rule generated in a bidirectional approach), HMAC (Hierarchical Multi-label Associative Classification), GARC (Gain based association rule classification), and RCAR (Regularized class association rules).

3.3. Developed web-based software

To create this web-based application, the Shiny library was used to allow the design of interactive web-based applications based on the R programming language [22]. Also, in the development of the interface, shinythemes [23], shinyBS [24], shinyLP [25], shinyalert [26], shinyjs [27] were used. Boruta [28], arules [29], arulesCBAR [30], caret [31], visNetwork [32] packages were used to make the analysis. The main and submenus of the software are described below. The developed software includes three main sections: “Introduction”, “Data” and “Analysis”.

3.3.1. Introduction

This section includes an information section with general information about the software and information about the packages used during the software development phase. With the "Start" tab on the page, the "Data Transactions" menu is passed.

3.3.2. Data

There are three submenus under the "Data Transactions" main menu: "File upload", "Data viewing ", "Variable Types". In the "File upload" menu, the file containing the data set is loaded. This developed software supports data files with “.xls”, “.xlsx”, “.sav” and “.arff” extensions. After uploading the file, the "Data viewing" sub-tab becomes active, and we have the opportunity to see the data set. With the "Variable Types" tab, we can determine the type and role of the variables in the data set. If the response/output variable is not determined while determining the variable roles, the error screen “Missing variable definitions “appears in the developed software.

3.3.3. Analysis

The analysis will be made with the Response / Output variable and the predictive variable (s) with the "Analysis" tab. To carry out the analysis, it should be decided whether to select the variable with the section “Apply variable selection”. Then, "Support" and "Confidence" values should be determined. If no selection is made, the analysis is made by accepting the "Support" value as 0.2 and the "Confidence" value as 0.8. If there are numerical variables in the loaded data set, the "Discretization method" tab will be displayed in this tab. The conversion of numerical variables into categorical variables is performed by selecting one of the discretization methods included in this tab. If no selection is made, numerical variables are converted into categorical variables by using the “Ameva” method. Finally, with the "Classification algorithm", one of the CBAR (Classification based on association rules) and RCAR (Regularized class association rules) methods included in the software is selected, and analyzes are done with the "Analysis" button. The results can be printed with the "Print page".

3.3.4. Developed web-based software accessibility

The developed interactive web-based software can be accessed free of charge at http://biostatapps.inonu.edu.tr/ACS/.

4. RESULTS

Open access dataset named “Breast Cancer Wisconsin (Diagnostic) Data Set” was used to analyze how the web-based software developed in this study works and evaluates its outputs. First, the data set was loaded into the software. After this process, the "Analysis" step is started with the "Next" button. The classification performance metrics of the model are given in Table 2.

TABLE II
THE METRICS OF THE MODEL'S CLASSIFICATION PERFORMANCE

METRIC	MODELS	
	CBAR	RCAR
ACCURACY	0.954	0.951
BALANCED ACCURACY	0.951	0.951
SENSITIVITY	0.939	0.953
SPECIFICITY	0.964	0.95
POSITIVE PREDICTIVE VALUE	0.939	0.918
NEGATIVE PREDICTIVE VALUE	0.964	0.971
F1-SCORE	0.939	0.935

CBAR: Classification based on association rules; RCAR: Regularized class association rules

According to the findings of performance metrics, accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1-score values obtained from the CBAR model are 0.954, 0.951, 0.939, 0.964, 0.939, 0.964, and 0.939, respectively.

Association rules using the classification algorithm are given in Table 3. When radius=[15,28.1) and texture=[19.5,39.3) are considered, the probability of a woman getting breast cancer is about 100%. Similarly, as texture=[19.5,39.3) and area=[696,2.5e+03) are taken into account, the probability of a female having breast cancer is nearly 100%, and when texture=[19.5,39.3) and perimeter=[98.8,188) are regarded, the probability of a woman with breast cancer is almost 100%. In contrast to the above rules, as texture=[9.71,19.5), area=[144,696) and compactness=[0.0194,0.102) are reckoned, the probability of a female not having breast cancer is 99.5%. The other rules generated from the CBAR model can be interpreted as the rules described earlier (Table 3).

TABLE III
ASSOCIATION RULES USED TO CONSTRUCT THE BEST PERFORMING MODEL (CBAR)

Left-hand side rules	Right-hand side rules	Support	Conf.	Freq.
{radius=[15,28.1), texture=[19.5,39.3)}	{diagnosis=Malignant}	0.206	1	117
{texture=[19.5,39.3), area=[696,2.5e+03)}	{diagnosis=Malignant}	0.206	1	117
{texture=[19.5,39.3), perimeter=[98.8,188)}	{diagnosis=Malignant}	0.204	1	116
{texture=[9.71,19.5), area=[144,696),compactness=[0.0194,0.102)}	{diagnosis=Benign}	0.348	0.995	198
{area=[144,696), smoothness=[0.0526,0.0895),concavepoints=[0,0.0514)}	{diagnosis=Benign}	0.279	0.994	159
{area=[696,2.5e+03), concavity=[0.0933,0.427)}	{diagnosis=Malignant}	0.243	0.993	138
{perimeter=[43.8,98.8), smoothness=[0.0526,0.0895), fractal_dimension=[0.0553,0.0665)}	{diagnosis=Benign}	0.232	0.992	132
{texture=[9.71,19.5), area=[144,696), concavepoints=[0,0.0514)}	{diagnosis=Benign}	0.406	0.991	231
{texture=[9.71,19.5), area=[144,696),c oncavity=[0,0.0933)}	{diagnosis=Benign}	0.404	0.991	230
{texture=[9.71,19.5), perimeter=[43.8,98.8),compactness=[0.0194,0.102)}	{diagnosis=Benign}	0.351	0.99	200
{area=[144,696), smoothness=[0.0526,0.0895)}	{diagnosis=Benign}	0.285	0.988	162
{area=[144,696), concavity=[0,0.0933),concavepoints=[0,0.0514),symmetry=[0.172,0.304)}	{diagnosis=Benign}	0.246	0.986	140
{perimeter=[98.8,188), concavity=[0.0933,0.427)}	{diagnosis=Malignant}	0.244	0.986	139
{area=[696,2.5e+03), compactness=[0.102,0.345)}	{diagnosis=Malignant}	0.232	0.985	132
{texture=[19.5,39.3), smoothness=[0.0895,0.163),concavepoints=[0.0514,0.201)}	{diagnosis=Malignant}	0.223	0.984	127
{texture=[9.71,19.5), compactness=[0.0194,0.102),concavepoints=[0,0.0514),fractal_dimension=[0.0553,0.0665)}	{diagnosis=Benign}	0.281	0.982	160
{texture=[9.71,19.5), compactness=[0.0194,0.102),concavepoints=[0,0.0514)}	{diagnosis=Benign}	0.353	0.98	201
{area=[696,2.5e+03),concavepoints=[0.0514,0.201)}	{diagnosis=Malignant}	0.262	0.98	149
{texture=[9.71,19.5), area=[144,696),fractal_dimension=[0.0553,0.0665)}	{diagnosis=Benign}	0.327	0.979	186
{area=[144,696), compactness=[0.0194,0.102),concavity=[0,0.0933),concavepoints=[0,0.0514)}	{diagnosis=Benign}	0.466	0.978	265
{area=[144,696), concavity=[0,0.0933), concavepoints=[0,0.0514)}	{diagnosis=Benign}	0.534	0.977	304
{texture=[19.5,39.3), smoothness=[0.0895,0.163),compactness=[0.102,0.345),concavity=[0.0933,0.427)}	{diagnosis=Malignant}	0.209	0.975	119
{area=[144,696), compactness=[0.0194,0.102),concavepoints=[0,0.0514),symmetry=[0.106,0.172)}	{diagnosis=Benign}	0.278	0.975	158

 ISSN: 2548-0650 http://dergipark.gov.tr/jcs

5. DISCUSSION

The most common type of cancer among women is breast cancer (BC). Each year around the world, over half a million people die because of BC. BC is more prevalent in women and middle-aged people. If early diagnosis is not made promptly, cancer cells start spreading across the body. Operational intervention and intensive chemotherapy processes can become important for the treatment of patients with BC in the next step. Early diagnosis is so critical for those reasons. Advances in artificial intelligence technology predict that the efficiency of automated systems will be more dominant than the human factor in this field. In other words, the experts' decision-making processes should be converted through technical means. Nowadays, automated systems based on artificial intelligence models are commonly used to diagnose various diseases [33].

Studies of developing interpretable/explainable machine learning models and making black-box models interpretable/explainable have gained importance recently. In particular, the classification of traditional medical datasets with satisfactory accuracy and interpretation of model outputs may be the reason for these models to be preferred over classical statistical hypothesis tests that require many assumptions. CBAR and RCAR rule-based interpretable models used in this study are determined to create rules with a high ability to interpret with negligible classification performance losses compared to models (such as support vector machine, random forest, neural network-based model, etc.) employed in classifying breast cancer data in other studies.

In the current study, to predict breast cancer early, it is intended to develop a new user-friendly web-based software to realize the use of the associative classification method, which uses the association rules method. Association rules, one of the descriptive models of data mining, are methods that analyze the coexistence of events. Association rules use combinations such as statistical analysis, data mining, and database management to reveal existing hidden relationships. These relationships are based on the coexistence of data elements and express the co-occurrence of events together with certain possibilities [34].

Classification analysis is one of the basic methods of machine learning and is used by a large scientific community. Classification is an estimation process that assigns each observation in the dataset to the predetermined classes under certain rules [35]. Associative classification makes classification by combining two common data mining methods, association rules, and classification methods. In recent years, association rules methods have been successfully used to create correct classifiers in associative classification. The important advantage of using this method is that simple if-then rules represent its output by using a classification based on association rules according to classical classification approaches, and it facilitates to understand and interpret the rules [8].

In the current study, the proposed software based on the studied models generated promising predictions in classifying breast cancer for malignant and benign according to the metric values of the classification performance on the open-access "Breast Cancer Wisconsin (Diagnostic) Data Set". For this purpose, the main hypotheses of this study are to determine whether classification-based association rules models are successful in predicting breast cancer on the open-access dataset and evaluate the classification performance. According to the experimental results, the calculated accuracy metric was quite high (0.954), and the other metrics of balanced accuracy, sensitivity, specificity positive predictive value, negative predictive value, and F1-score were similarly so large (>0.930) from the proposed model and web-based software. As of 2020, several studies have been conducted to investigate the classification of breast cancer using machine learning and data mining techniques. A novel paper offers a comparative analysis by applying various machine learning algorithms such as Support Vector Machine, Naïve Bayes, Decision Tree, K-Nearest Neighbor, k-means clustering, and Artificial Neural Networks on Wisconsin Diagnostic Data Set to predict early breast cancer. The authors conclude, after analyzing all the implemented algorithms, that artificial neural network provides better prediction as 97.85% compared to all the other methods [36]. Another newly published work performed the experiments on the dataset Wisconsin Diagnostic Breast Cancer (WDBC), and the technique of k-fold cross-validation is used for model assessment. The proposed two-layer nested ensemble classifiers were compared with single classifiers (i.e., BayesNet and Naïve Bayes) in terms of classification precision, accuracy, recall, F1 score, the area under the ROC curve, computational durations of single and nested ensemble classifiers. The results show that the accuracy of SV-BayesNet-3-MetaClassifier and SV-Naïve Bayes-3-MetaClassifier was 98.07 percent, and the proposed two-layer nested ensemble models outperform the single classifiers and much of the preceding research [37]. Another up-to-date paper introduces a genetic algorithm for mutual information (MIGA), where MIGA is a combination of two algorithms: mutual information (MI) and genetic algorithm (GA) for detecting breast cancer using the Breast Cancer Wisconsin Diagnostic dataset. The results of the MIGA algorithm show that the highest accuracy (99 percent) with GA-based MI features was achieved [38]. The novel paper offers six classification algorithms of the medical diagnostic methods used in machine learning on the UCI three medical data sets, including the "Diagnostic Breast Cancer dataset for Wisconsin". Overall, three medical datasets, the experimental results indicate that the SVM classification algorithm has achieved the most promising prediction [39]. In another recent study, the dataset for Wisconsin Diagnostic Breast Cancer (WDBC) is analyzed with Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), Naïve Bayes (NB), Decision-Tree (DT) and Logistic Regression (LR) using 5-fold cross-validation method. Classification efficiency is calculated through the use of the confusion matrix through performance assessment parameters, i.e., precision, sensitivity, and specificity. The best result in that study by SVM is a 99.12 percent accuracy after the phase of normalization [40]. When the classification performances of the previous studies are outlined, the performance metrics values of the current study are so high (>0.930 for all the metrics evaluated) and similar to the reported other papers on the classification of breast

cancer. Besides, the present study develops free web-based software to classify the breast cancer, and defines the associated rules of any data sets (e.g., Breast Cancer Wisconsin (Diagnostic) Data Set) achieved from the associative classification methods. This research has important features compared to other studies in that it includes open access web-based software and association rules based on the classification of diseases (e.g., breast cancer).

As a result, in the analysis of the open-access dataset, the proposed model (CBAR) has a distinctive feature in classifying breast cancer based on the performance metrics. The associative classification software developed based on CBAR produces successful predictions in the classification of breast cancer. The hypothesis established within the scope of the purpose of this study has been confirmed as the similar estimates are achieved with the results of other papers in the classification of breast cancer.

REFERENCES

[1] K. Oktay et al., "A Computational Statistics Approach to Evaluate Blood Biomarkers for Breast Cancer Risk Stratification," Hormones and Cancer, vol. 11, no. 1, pp. 17-33, 2020.

[2] J. Ping et al., "Differences in gene-expression profiles in breast cancer between African and European-ancestry women," Carcinogenesis, 2020.

[3] H. Akpınar, "Veri tabanlarında bilgi keşfi ve veri madenciliği," İÜ İşletme Fakültesi Dergisi, vol. 29, no. 1, pp. 1-22, 2000.

[4] A. Koyuncugil and N. Özgülbaş, "Veri madenciliği: Tıp ve sağlık hizmetlerinde kullanımı ve uygulamaları," Bilişim Teknolojileri Dergisi, vol. 2, no. 2, 2009.

[5] L. T. Moss and S. Atre, Business intelligence roadmap: the complete project lifecycle for decision-support applications. Addison-Wesley Professional, 2003.

[6] Y.-L. Chen, J.-M. Chen, and C.-W. Tung, "A data mining approach for retail knowledge discovery with consideration of the effect of shelf-space adjacency on sales," Decision support systems, vol. 42, no. 3, pp. 1503-1520, 2006.

[7] A. K. Pujari, Data mining techniques. Universities press, 2001.

[8] F. Thabtah, "A review of associative classification mining," The Knowledge Engineering Review, vol. 22, no. 1, pp. 37-65, 2007.

[9] D. Dua and C. Graff, "UCI machine learning repository. School of Information and Computer Science, University of California, Irvine, CA," ed, 2019.

[10] A. S. Kumar and R. Wahidabanu, "Data Mining Association Rules for Making Knowledgeable Decisions," in Data Mining Applications for Empowering Knowledge Societies: IGI Global, 2009, pp. 43-53.

[11] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, "Advances in knowledge discovery and data mining," 1996: American Association for Artificial Intelligence.

[12] D. T. Larose and C. D. Larose, Discovering knowledge in data: an introduction to data mining. John Wiley & Sons, 2014.

[13] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in Proceedings of the 1993 ACM SIGMOD international conference on Management of data, 1993, pp. 207-216.

[14] J. Han and M. Kamber, "Data Mining: Concepts and Tecniques. ISBN 13: 978-1-55860-901-3," ed: Elsevier, USA, 2008.

[15] M. Houtsma and A. Swami, "Set-oriented mining for association rules in relational databases," in Proceedings of the eleventh international conference on data engineering, 1995, pp. 25-33: IEEE.

[16] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. 20th int. conf. very large data bases, VLDB, 1994, vol. 1215, pp. 487-499.

[17] A. Savasere, E. R. Omiecinski, and S. B. Navathe, "An efficient algorithm for mining association rules in large databases," Georgia Institute of Technology1995.

[18] A. Das, W.-K. Ng, and Y.-K. Woon, "Rapid association rule mining," in Proceedings of the tenth international conference on Information and knowledge management, 2001, pp. 474-481.

[19] M. J. Zaki and C.-J. Hsiao, "CHARM: An efficient algorithm for closed itemset mining," in Proceedings of the 2002 SIAM international conference on data mining, 2002, pp. 457-473: SIAM.

[20] N. Ye, The handbook of data mining. CRC Press, 2003.

[21] M. Azmi, G. C. Runger, and A. Berrado, "Interpretable regularized class association rules algorithm for classification in a categorical data space," Information Sciences, vol. 483, pp. 313-331, 2019.

[22] W. Chang, J. Cheng, J. Allaire, Y. Xie, and J. McPherson, "Shiny: web application framework for R," R package version, vol. 1, no. 5, 2017.

[23] W. Chang, T. Park, L. Dziedzic, N. Willis, and M. McInerney, "shinythemes: Themes for Shiny," R package version, vol. 1, no. 1, p. 144, 2015.

[24] E. Bailey, "shinyBS: Twitter bootstrap components for shiny," R package version 0.61, 2015.

[25] J. Dumas, "shinyLP: Bootstrap Landing Home Pages for Shiny Applications," R package version, vol. 1, p. 2, 2019.

[26] D. Attali and T. Edwards, "shinyalert: Easily Create Pretty Popup Messages (Modals) in'Shiny'," R package version 1.0, 2018.

[27] D. Attali, "Shinyjs: Easily improve the user experience of your shiny apps in seconds," R package version 0.9, 2016.

[28] M. B. Kursa and W. R. Rudnicki, "Feature selection with the Boruta package," J Stat Softw, vol. 36, no. 11, pp. 1-13, 2010.

[29] M. Hahsler et al., "Package 'arules'," ed, 2019.

[30] I. Johnson, "arulesCBA: Classification for Factor and Transactional Data Sets Using Association Rules."

[31] M. Kuhn, "The caret package," R Foundation for Statistical Computing, Vienna, Austria. URL https://cran. r-project. org/package= caret, 2012.

[32] B. Almende, B. Thieurmel, and T. Robert, "visNetwork: Network Visualization using "vis. js" Library," ed: CRAN, 2016.

[33] M. Toğaçar, B. Ergen, and Z. Cömert, "Application of breast cancer diagnosis based on a combination of convolutional neural networks, ridge regression and linear discriminant analysis using invasive breast cancer images processed with autoencoders," Medical Hypotheses, vol. 135, p. 109503, 2020/02/01/ 2020.

[34] İ. Perçın, F. H. Yağin, E. Güldoğan, and S. Yoloğlu, "ARM: An Interactive Web Software for Association Rules Mining and an Application in Medicine," in 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), 2019, pp. 1-5: IEEE.

[35] İ. PERÇİN, F. H. YAĞIN, A. K. ARSLAN, and C. ÇOLAK, "An Interactive Web Tool for Classification Problems Based on Machine Learning Algorithms Using Java Programming Language: Data Classification Software," in 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2019, pp. 1-7: IEEE.

[36] G. Rawal, R. Rawal, H. Shah, and K. Patel, "A Comparative Study Between Artificial Neural Networks and Conventional Classifiers for Predicting Diagnosis of Breast Cancer," in ICDSMLA 2019: Springer, 2020, pp. 261-271.

[37] M. Abdar et al., "A new nested ensemble technique for automated diagnosis of breast cancer," vol. 132, pp. 123-131, 2020.

[38] N. Vutakuri and A. U. J. I. J. o. A. I. P. Maheswari, "Breast cancer diagnosis using a Minkowski distance method based on mutual information and genetic algorithm," vol. 16, no. 3-4, pp. 414-433, 2020.

[39] P. S. Nishant, S. Mehrotra, B. G. K. Mohan, and G. Devaraju, "Identifying Classification Technique for Medical Diagnosis," in ICT Analysis and Applications: Springer, 2020, pp. 95-104.

[40] N. Panwar, D. Sharma, and N. J. A. a. S. Narang, "Breast Cancer Classification with Machine Learning Classifier Techniques," 2020.

BIOGRAPHIES

Ahmet Kadir ARSLAN received his BSc degree in Maths from Afyon Kocatepe University and MSc degree in Biostatistics and Medical Informatics from Inonu University. He is currently in his second year of his PhD in Biostatistics and Medical Informatics at Inonu University. His research interest are interpretable machine learning, decision support systems, neural networks, data preprocessing, image classification and dimension reduction.

Zeynep TUNÇ obtained her BSc. degree in mathematics from Çukurova University in 2010. She received MSc. degree in biostatistics and medical informatics from the Inonu University in 2018. She currently continues Ph.D. degrees in biostatistics and medical informatics from the Inonu University. In 2014, she joined the Department of Biostatistics and Medical Informatics at Inonu University as a researcher assistant. Her research interests are cognitive systems, data mining, machine learning, deep learning.

İpek BALIKÇI ÇİÇEK obtained her BSc. degree in mathematics from Çukurova University in 2010. She received MSc. degree in biostatistics and medical informatics from the Inonu University in 2018. She currently continues Ph.D. degrees in biostatistics and medical informatics from the Inonu University. In 2014, she joined the Department of Biostatistics and Medical Informatics at Inonu University as a researcher assistant. Her research interests are cognitive systems, data mining, machine learning, deep learning.

Cemil ÇOLAK obtained his BSc. degree in statisticsfrom Ondokuz Mayıs University in 1999. He received MSc. degree in Biostatistics from the Inonu University in 2001, and Ph.D. degree in the Graduate Department of Biostatistics and Medical Informatics of Ankara University in 2005. His research interests are cognitive systems, data mining, reliability, and biomedical system, genetics, and bioengineering. In 2016, he joined the Department of Biostatistics and Medical Informatics at Inonu University as a Professor, where he is presently a professor. He is active in teaching and research in the general image processing, artificial intelligence, data mining, analysis.