Research Article

BIOMEDICAL Journal of Scientific & Technical Research

ISSN: 2574 -1241

DOI: 10.26717/BJSTR.2022.42.006722

Improving the Accuracy of Misclassified Breast Cancer Data Using Machine Learning

Rong-Ho Lin and Benjamin Kofi Kujabi*

Department of Industrial Engineering and Management, National Taipei University of Technology, Taiwan

***Corresponding author:** Benjamin Kofi Kujabi, Department of Industrial Engineering and Management, National Taipei University of Technology, 1, Sec. 3, Zhongxiao E. Rd., Taipei 10608 Taiwan, ROC, Taipei, Taiwan

ARTICLE INFO

Received: February 21, 2022

Published: March 01, 2022

Citation: Rong-Ho Lin, Benjamin Kofi Kujabi. Improving the Accuracy of Misclassified Breast Cancer Data Using Machine Learning. Biomed J Sci & Tech Res 42(2)-2022. BJSTR. MS.ID.006722.

Keywords: Misclassified Data; Classifiers; WEKA; myCBR; Protégé

Abbreviations: CBR: Case-Based Reasoning; MLP: Multi-Layer Perceptron; FN: False Negative; FP: False Positive; WBC: Wisconsin Breast Cancer

ABSTRACT

Background: Breast cancer is the most common cancer among women. Many studies have made significant gains in classifying cancer tumors, emphasizing the best algorithm and highest classification accuracy but with limited interest in correcting misclassified data (Type 1 and Type 2 errors).

Objective: This research proposes a novel hybrid integrated system of WEKA (Waikato Environment for Knowledge Analysis) and case-based reasoning (CBR) using myCBR plugin with protégé for the classification of breast cancer tumors and correction of misclassified data (Type 1 and Type 2 errors) of breast cancer tumors.

Methods: The Wisconsin breast cancer dataset retrieved from the Wisconsin university repository was used in this research. The dataset contained 699 instances, 2 classes (malignant and benign), and 9 integer-valued attributes. The breast cancer tumors were determined by applying the J48, IBK, LibSVM, JRip, and Multi-Layer Perceptron (MLP) classifiers to classify the breast cancer tumors. Later, the myCBR plugin with protégé was used as an advanced modeling technique to correct the misclassified data and enhance its accuracy.

Results: The proposed model performance evaluation was based on sensitivity, specificity, precision, and accuracy. Interestingly, based on the analyses, the IBK classifier had the highest misclassified data and the integrated system improved its classification accuracy from 95.61% to 98.53%.

Conclusion: The findings demonstrated that the integrating of WEKA and myCBR plugin with protégé had unprecedented results with misclassified data. Thus, providing accurate diagnostics procedures for distinguishing between benign and malignant.

Introduction

Globally, breast cancer is one of the most common cancers among women. According to the American Cancer Statistics Report 2020, an estimated 276,480 new breast cancer cases will be diagnosed in women and approximately 2,620 cases in men before the end of 2021 [1]. Early diagnosis of breast cancer can improve the prognosis and survival chances of patients with breast cancer. Over the last decade, researchers have proposed different algorithms and innovative diagnosis techniques to distinguish benign and malignant tumors. Clinically, the three prominent procedures used to detect breast cancer are fine-needle aspiration cytology, mammography, and physicians' clinical opinions [2-5]. Physicians might have different opinions on the interpretation of the examination results as the symptoms of breast cancer vary from patient to patient. This can lead to errors that might be detrimental

to the patients' health. For example, a malignant tumor might be interpreted as benign, resulting in a false negative (FN) (Type 1 error). Moreover, a benign tumor might be classified as malignant, resulting in a false positive (FP) (Type 2 error).

These misinterpretations of false-positive or false-negative cancer diagnoses can lead to unnecessary mastectomy. Furthermore, it might lead to life-threatening illnesses and patients taking the wrong drugs to cure the wrong illness [6]. To mitigate these common errors, researchers have applied numerous data mining techniques to assist clinicians in accurately diagnosing breast cancer. Data mining and machine learning constitute an integral part of breast cancer prediction and prognosis. These methods learn patterns that provide insight from historical data in order to enable new data prediction [7,8]. According to Ashutosh, et al. [9], data mining based on machine learning techniques can be used for classification, prediction, estimation, clustering, association rules, and visualization techniques. Of all these techniques, classification, prediction, and estimation are categorized as supervised learning techniques that entail model formulation based on the available data representation.

Additionally, classification is highly regarded among physicians in decision-making processes. Notably, the classification of breast cancer can be helpful to predict the outcome or discover the genetic behavior of tumors [10]. In most cases, the Wisconsin Breast Cancer (WBC) dataset and WEKA (Waikato Environment for Knowledge Analysis), which contains data mining algorithms, have been used to develop models for the classification of breast cancer. Researchers over the years have also adopted different rules to achieve the best classification accuracy. Abdar, et al. [11] proposed a nested ensemble approach that uses stacking and vote (voting) classification technique to distinguish benign breast tumors from malignant tumors using WBC. Aloraini [12] and Asri, et al. [13] have compared different learning algorithms, namely; Bayesian network, naïve Bayes, decision trees, J48, ADTree, multi-layer perceptron, and k-nearest neighbor, and reported that Bayesian network and SVM algorithms yield the highest accuracy levels.

Despite the focus on classification accuracy, data can be misclassified due to noise, and limited research on this problem is carried out. Smith and Martinez [14] proposed the PRISM (Preprocessing Instances that Should be Misclassified) method that identifies and removes instances to improve the data misclassification, achieving a 1.3% improvement in 53 datasets and a 1.9% increase in non-outlier instances. Although several machine learning algorithms and models have been established to help classify breast cancer, these algorithms and models have limitations and many imperfections. Thus it is crucial to develop a practical and effective model to perfect the classification of breast cancer tumors to avoid errors in clinical diagnosis of breast cancer and reduce the mortality rates of breast cancer patients. In this research, a novel hybrid system that comprises WEKA and case-based reasoning (CBR) using the myCBR plugin with protégé was employed to classify and of misclassified data (Type 1 and Type 2 errors) of breast cancer tumors. The CBR is useful in leveraging knowledge encapsulated in previously learned cases and resolves other cases to support new decisions [15]. Accordingly, the findings of this research can provide clinicians with diagnostics procedures for distinguishing between benign and malignant tumors.

Methodology

This research proposes a novel hybrid integrated system comprising WEKA and myCBR with protégé to classify breast cancer tumors and the correct misclassified data. The J48, IBK, LibSVM, JRip, and MLP classifiers were used to formulate this model to classify breast cancer tumors. Figure 1. presents the applied method.

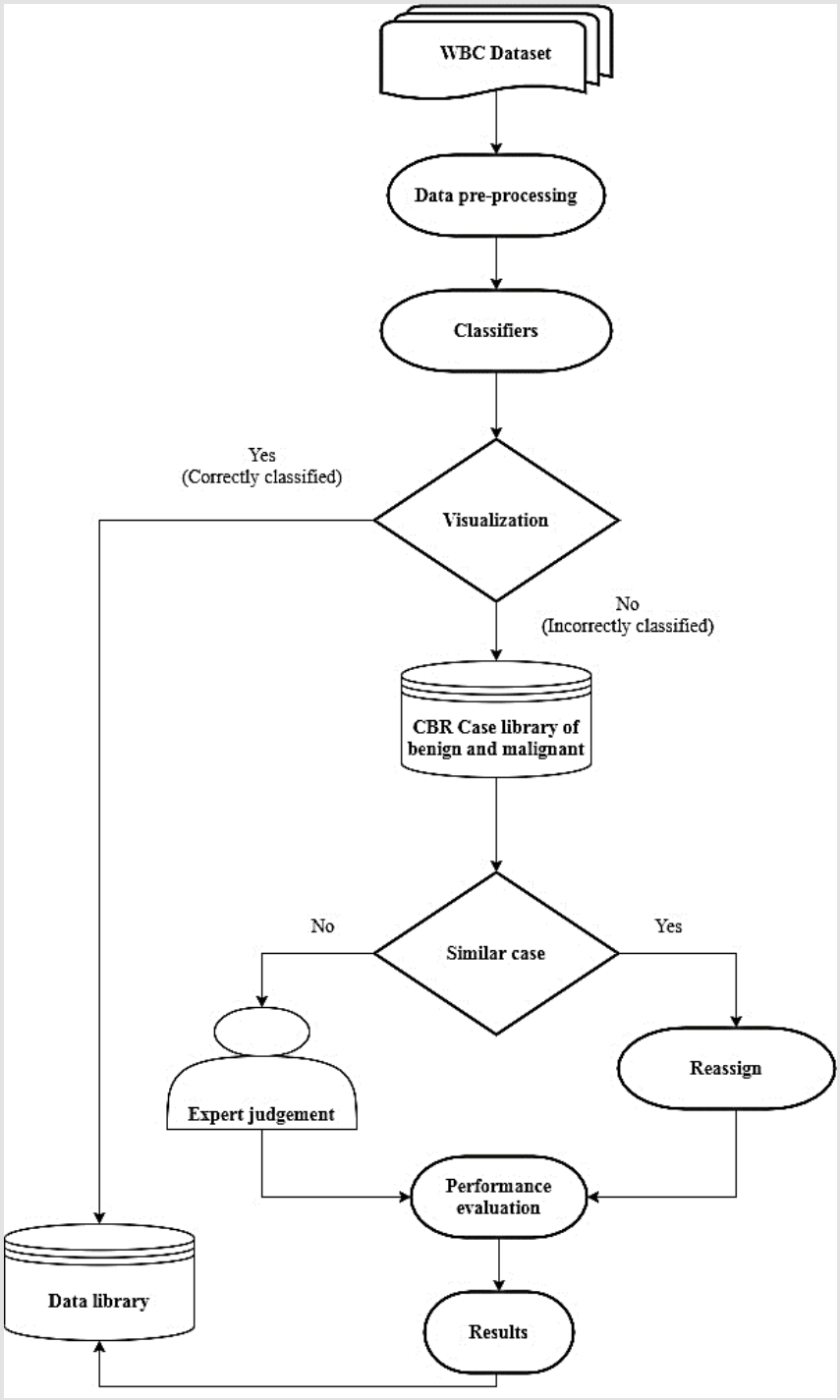

Figure 1: An integrated WEKA and myCBR with protégé model.

Data Description

Considering the emphasis on the classification of breast cancer tumors, several researchers have used the SEER (Surveillance, Epidemiology and End Results Program) [16] and WBC dataset to develop a prognosis and predictive models. The WBC dataset [17] collected by Dr. William H. Wolberg (1981 – 1991) at the University of Wisconsin Madison Hospital was used in this research. The dataset comprises 699 instances taken from fine-needle aspirates from the patient's breast. It contains nine attributes and one class; 458 (65.5%) were classified as benign and 241 (34.5%)

were malignant. The dataset has 16 missing instances, and in preprocessing the data, the 16 missing datasets were deleted. The remaining 683 instances were divided into training and testing sets using WEKA's resample unsupervised learning filter and sample the data without replacement. Table 1. presents the description of the nine attributes of the breast cancer dataset.

Table 1: Summary of attributes for WBC dataset.

Attribute	Values
Clump Thickness	1 ~ 10
Uniformity of Cell Size	1 ~ 10
Uniformity of Cell Shape	1 ~ 10
Marginal Adhesion	1 ~ 10
Single Epithelial Cell Size	1 ~ 10
Bare Nuclei	1 ~ 10
Bland Chromatin	1 ~ 10
Normal Nucleoli	1 ~ 10
Mitoses	1 ~ 10
Class	(2 for benign and 4 for malignant)

Experimental Tools

WEKA: The WEKA workbench [18] is a set of machine learning algorithms and data preprocessing tools for data mining tasks. The workbench comprises extensive algorithms that are applied directly to a dataset. It supports classification, data processing, regression, and visualization of results. In this research, WEKA 3.8.3 was used for the preprocessing, classification, and evaluation of the dataset.

myCBR (Final Protégé-Based Release): The myCBR is an open-source similarity-based retrieval plugin for ontology-based applications. It is a powerful tool that employed test highly sophisticated models, knowledge-intensive similarity measures, and it can be easily integrated with other applications [19]. Moreover, protégé is an open-source platform that provides a plug-and-play environment with a suite of tools to construct domain models and knowledge-based applications with ontologies [20,21]. Protégé and myCBR complement each other, because protégé defines classes and attributes in an objected-oriented way and manages instances of these classes and myCBR interprets them as cases. The integration of myCBR plugin and protégé facilitates the development of a more robust similarity model development and improves retrieval quality.

Experimental Procedure

Firstly, the data were preprocessed and then developed into a standard WEKA ARFF (Attributed Relation File Format) from the WBC dataset. Later the ARFF file was uploaded in the WEKA toolkit, where the feature selection method used ranker and IfoGain to attribute the evaluator to generate different feature subsets [22]. The data were later weighted down for classification and J48, IBK, LibSVM, JRip, and MLP classification algorithms were applied.

Testing the model

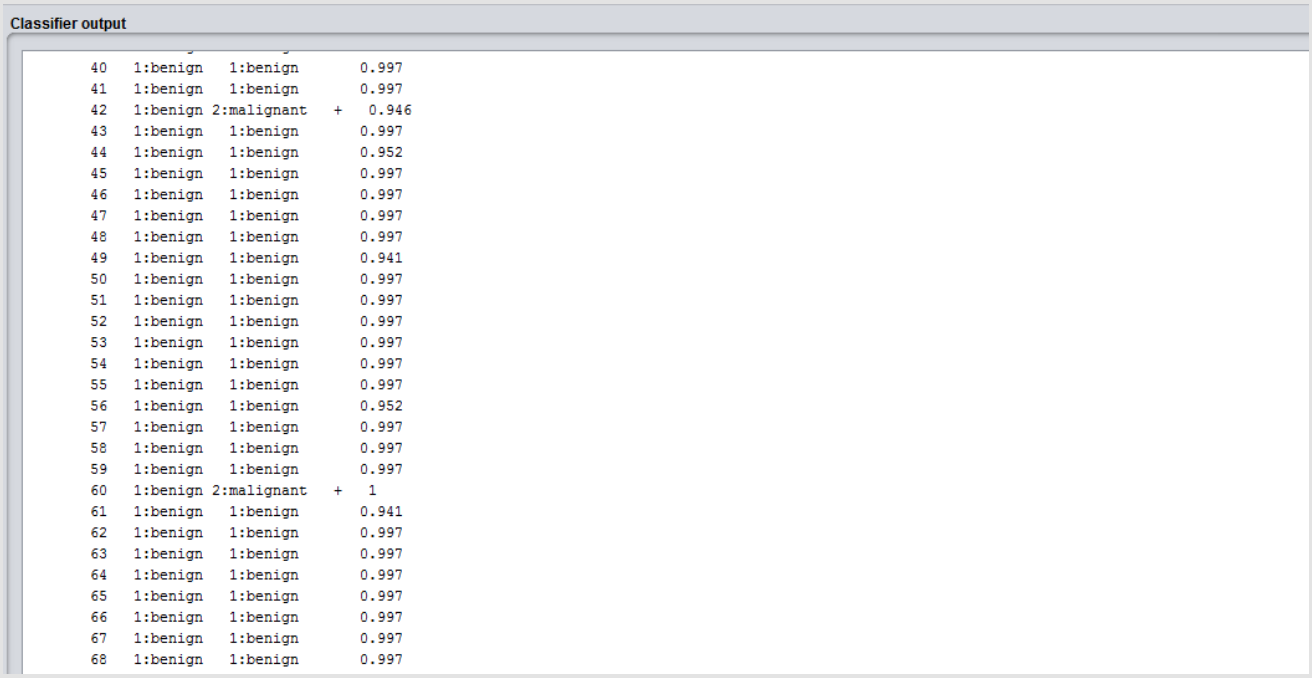

Figure 2: Correct and misclassified instances.

A k-fold cross-validation method was applied to test the generalization ability of the proposed method. Accordingly, the data were split into 10 subsets randomly. Subsequently, we used 10-1 of these subsets for training were used; that is, one-tenth of the dataset for testing the nine-tenth for training were used. This procedure was repeated 10 times until all these subsets were used for training and testing. The performance of the classifiers was averaged after the iteration of the 10 folds. In the second step, the effectiveness of case-based reasoning retrieval ability to assess the cases of all the misclassified data by the classifiers was utilized. This was done by visualizing the classification results with output prediction in WEKA. Figure 2. shows the instances that were classified correctly and those with "+" represent misclassified data. All the misclassified data were retrieved and processed into a CSV format and imported to myCBR. A class was created to be used as query case values during retrieval. The Euclidian distance was applied to measure the similarity and the values of weights were set to 1. The misclassified data were tested with respect to the 683 data.

Case-Based Reasoning

One of the most important tasks of CBR is to retrieve similar cases from the case base library. In this research, the similarity between cases was measured using the nearest neighbor algorithm, which computes the similarity between two cases using a global similarity measure. When a new case is loaded into the system, it will be compared with the cases in the CBR library system to determine whether a similar case with low-level or high-level characteristics can be found. The case with the highest similarity in the CBR library database would be retrieved. In this process, ten most similar cases were retrieved for analysis. Suppose no similarity exists in the CBR library database. In that case the system will refer to expert judgment where it will be evaluated by an expert before being validated and later stored in the data library. The CBR comprises three task or functions: Case library, Similarity measure, and Local similarity measure.

Case library: The case library stores historical solved data of known cases. New cases that is yet to be solved, the goal of the CBR is to retrieve cases from the case library that are most similar to the new case to support the prediction of the case value by the decision marker [23].

Similarity measure: When comparing two cases, their attribute values are compared using local similarity functions.

Local similarity: Local similarity can be used to sole with missing values and is cost-sensitive. It is widely used in medical applications [24].

$$Sim(a,b) = 1 - \frac{|a-b|}{range} \qquad (1)$$

Where:

A and B represent a new and previous feature from the local similarity respectively.

Range is the value of the difference between the upper and lower boundary of the set.

The global similarity function in myCBR is linked to compound attributes and to collect attributes in a unique similarity value by gathering their similarities.

$$Sim(A,B) = \frac{1}{\sum wi} . \sum_{i=1}^{p} wi.Simi(a,b) \qquad (2)$$

Where:

A and *B* represent new and previous cases respectively.

a is a new feature from the local similarity.

b is a previous feature from the local similarity.

p is the number of attributes.

i is the iteration.

wi is the weight of attributes $I\sum_{i=1} pWi$ = 1 and *sim i* is a local similarity to calculate for attribute *i*.

Distance Measure

The Euclidean distance of equation 3. has been the most widely used distance measure in various learning systems, and it measures the distance between two points. The formula proposed by Gu, et al. [23] proposed formula was applied to measure the distance during the retrieval process.

$$EU(t,r) = \sqrt{\sum_{i=1} n} w_i d_i^2(t,r) \qquad (3)$$

$$d_i(t,r) = diff(X_{t,i} - X_{r,i},$$

$$diff(X_{t,i} - X_{r,i} = X_{t,i} - X_{r,i}$$

The effects of measuring scales can be avoided by normalizing the input attributes and this could be achieved in CBR by weighting the attributes according to their importance. The weighted Euclidean distance for a stored case A in a case library and a target case B can be defined in equation 4 [25-26].

$$WEU(t,r) = \sqrt{\sum_{i=1} n} W_i d_i^2(t,r) \qquad (4)$$

A heterogeneous distance function handles both continuous and nominal attributes. The Euclidean distance is limited for continuous attributes and invents a discrete attribute if present.

Case reuse: Any retrieved case with the same features is either reused to solve the present case or modified using adaptation rules

to solve a new case. One of the approaches to case adaptation is mean values. For instance, if a parameter value v1 has to be updated to a value v2, the mean value method entails gathering cases that containing either v1 or v2. Therefore, the system sums the mean value of the two groups to produce the output [27].

Evaluation Method

The novel hybrid system performance is evaluated through a standard data classification system regarding the accuracy, sensitivity, specificity, geometric mean (G-mean), evaluation, and misclassified data. The widely used receiver operating characteristics curve (ROC) was applied to analyze the classifiers and the G-mean. Table 2 presents the confusion matrix used for evaluation. True Positive (TP) and true negative (TN) results represent correctly classified cases. For example, they represent a scenario in which a benign tumor is correctly classified as benign. A false positive (FP) or type 1 error represents a scenario in which the cases are misclassified by rejecting the null hypothesis (negative) when it is true. In another example, a malignant tumor (negative) is classified as benign (positive) when it is a malignant tumor (negative). An FN result (Type 2 error) represents a scenario in which cases are as well misclassified but the null hypothesis is not rejected when it is false. For example, a benign tumor (positive) is classified as malignant (negative) when it is classified a benign tumor (positive) [2].

Table 2: Confusion matrix.

	Predicted positive	Predicted negative
Actual positive	True positive (TP)	False negative (FN)
Actual negative	False positive (FP)	True negative (TN)

Accuracy: A test's accuracy is computed by estimating the fraction of true positive and negative instances in all cases computed as in equation 5.

$$\frac{TP+TN}{TP+FN+FP+TN}\times 100 \quad (5)$$

Sensitivity, correctly generate positive cases with either malignant or benign (also known as TP rate), as in equation 6.

$$\frac{TP}{TP+FN} \quad (6)$$

Specificity, correctly generate negative cases of those without benign or malignant (also known as the TN rate), as in equation 7.

$$\frac{TP}{TN+FP} \quad (7)$$

Equation 8, represent the G-mean, is used to evaluate the performance of the classifiers on incorrect data.

$$\sqrt{\frac{TP}{TP+FN}X\frac{TN}{TN+FP}} \quad (8)$$

Results

The J48, IBK, LibSVM, JRip, and MLP classifiers were first applied to classify benign and malignant tumors. As mentioned, the breast cancer dataset was obtained from the Wisconsin repository consisting of 699 cases. After preprocessing 683 cases were used for the classification process. Table 3. presents a confusion matrix used for evaluation. The accuracy, sensitivity, specificity, and G-mean were evaluated. From the analysis, the LibSVM had the highest accuracy (96.93%) among the classifiers. Thus, in this research, the improvement of misclassified data was highlighted. Accordingly, we embedded the myCBR plugin were embedded with protégé to build a flexible hybrid model. Table 4. shows the results for all misclassified data by the five classifiers. Notably, the accuracy of IBK improved considerably (2.92%), followed by J48 (2.83), MLP (2.5), LibSVM (1.8), and JRip (1.77).

Table 3: Classification results based on confusion matrix.

Classifiers	MLP	J48	IBK	Jrip	LibSVM
Accuracy	96	95.31	95.61	95.9	96.93
Sensitivity	96.85	96.82	96	97.71	97.74
Specificity	94.54	92.59	94.87	92.71	95.42
G-mean	95.69	94.68	95.43	95.18	96.57

Table 4: CBR corrected misclassified data.

Classifiers	Single	Reassigned	Improved
MLP	96	98.5	2.5
J48	95.31	98.14	2.83
IBK	95.61	98.53	2.92
Jrip	95.9	97.67	1.77
LibSVM	96.93	98.73	1.8

Discussion

The novel hybrid integrated system of WEKA (Waikato Environment for Knowledge Analysis) and case-based reasoning (CBR) using myCBR plugin with protégé for the classification of breast cancer tumors and correction of misclassified data (Type 1 and Type 2 errors) of breast cancer tumors was achieved as the main objective of this research. In Table 3, the derived breast cancer dataset from the Wisconsin breast cancer repository consists of 10 attributes. A confusion matrix was used for the evaluation of the classifiers. Although, different parameter settings were applied, the LibSVM classifier outperformed all other classifiers and achieved a 96.93% accuracy level. The LibSVM classifier outperformance is due to better numerical attributes and good computing speed and

memory processing. The MPL shows a comparable performance to LibSVM. It appears that when compared to instance based learning system, MLP tends to be a better technique for classification problems. However, it is also known for its best adaptive learning but lacks the power to represent interactions among variables [28].

Furthermore, observed that since IBK is an instance-based learning algorithm, it is reasonable to understand why MLP performs better than IBK. When finetuned, the IBK can be a much more effective tool for high classification accuracy when finetuned [29]. The J48 and Jrip are both decision trees thus Jrip performs slightly better than J48. Further, it can be attributed to their pruning methods or dataset adaption. In the case of J48, it adapts a subtree replacement which replaces nodes in decision trees with leaf and subtree raising, which involves moving nodes upwards toward the tree's root while replacing other nodes. In some instances, when the J48 performs poorly, it can be due to the complexity and heterogeneity of attributes values. Additionally, the Jrip isolates some data to reduce error pruning and adapts simple rules to improve the accuracy [30]. However, considerable emphasis was placed on improving the misclassified data and when the hybrid model was established, an upward spike inaccuracy (ranging from 1.77 to 2.92%) for IBK showed considerable improvement. The need to manage and correct the misclassified data is an essential factor for prognosis and diagnosis. The results demonstrate that the system is one of the best in correcting misclassified data compared to other models. It minimizes the risk of physicians misinterpreting tumors. The system can provide accurate diagnostic procedures for distinguishing between benign and malignant tumors.

Conclusion

Improving errors in misclassification will not only help physicians to make the proper judgment save the lives of breast cancer patients. The American Breast Cancer Society reported that breast cancer had been the leading cause of death in women leading to significant research in this domain. In this research, we devised a novel hybrid integrated system comprising WEKA and CBR using myCBR plugin and protégé for the classification of breast cancer tumors and the correction of misclassified data (Type 1 and Type 2 errors). A K-fold cross-validation technique was applied to the WBC and myCBR was embedded with protégé to correct the misclassified data. The findings demonstrate that integrating WEKA and the myCBR plugin with protégé had provided unprecedented results in correcting misclassified data. Thus, an extension of this model to accurately predict the stages of cancer will be highly recommended. Optimizing classifiers parameters to minimize misclassification will be vital for future research. This research paper will go a long way to assist physicians in making a swift decision regarding the prediction of breast cancer stages.

References

1. Cancer Facts & Figures 2021 | American Cancer Society.
2. S Wang, Y Wang, D Wang, Y Yin, Y Wang, et al. (2020)An improved random forest-based rule extraction method for breast cancer diagnosis. Applied Soft Computing 86: 105941.
3. L Borges (2015) Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection.
4. Y Hayashi, S Nakano (2015) Use of a Recursive-Rule eXtraction algorithm with J48graft to achieve highly accurate and concise rule extraction from a large breast cancer dataset. Informatics in Medicine Unlocked 1: 9-16.
5. DH Devi, DMI Devi (2016) Outlier Detection Algorithm Combined with Decision Tree Classifier For Early Diagnosis of Breast Cancer.
6. N Liu, J Shen, M Xu, D Gan, ES Qi, et al. (2018) Improved Cost-Sensitive Support Vector Machine Classifier for Breast Cancer Diagnosis. Mathematical Problems in Engineering.
7. RJ Kate, R Nadig (2017) Stage-specific predictive models for breast cancer survivability. International Journal of Medical Informatics 97: 304-311.
8. Data Mining (4th Edn.).,.
9. AK Dubey, U Gupta, S Jain (2016) Analysis of k-means clustering approach on the breast cancer Wisconsin dataset. Int J Comput Assist Radiol Surg 11(11): 2033-2047.
10. G Ravi Kumar (2019) An efficient prediction of breast cancer data using data mining techniques.
11. M Abdar, Moghadam MZ, Zhou X, Gururajan R, Tao X, et al. (2020) A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recognition Letters 132: 123-131.
12. A Aloraini (2012) Different Machine Learning Algorithms for Breast Cancer Diagnosis. International Journal of Artificial Intelligence & Applications 3: 21-30.
13. H Asri, H Mousannif, HA Moatassime, T Noel (2016) Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia Computer Science 83: 1064–1069.
14. MR Smith, T Martinez (2011)Improving classification accuracy by identifying and removing instances that should be misclassified. in The 2011 International Joint Conference on Neural Networks, pp. 2690–2697.
15. ZY Zhuang, L Churilov, F Burstein, K Sikaris (2009) Combining data mining and case-based reasoning for intelligent decision support for pathology ordering by general practitioners. European Journal of Operational Research 195(3): 662-675.
16. SEER Incidence Data, 1975-2018.
17. UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic).
18. Weka 3 - Data Mining with Open Source Machine Learning Software in Java.
19. myCBR.
20. protégé.
21. Protege Wiki.
22. N Liu, J Shen, M Xu, D Gan, ES Qi, et al. (2018)Improved Cost-Sensitive Support Vector Machine Classifier for Breast Cancer Diagnosis. Mathematical Problems in Engineering, p. e3875082.
23. D Gu, C Liang, H Zhao (2017) A case-based reasoning system based on weighted heterogeneous value distance metric for breast cancer diagnosis. Artif Intell Med 77: 31-47.

24. B López, C Pous, P Gay, A Pla, J Sanz, et al. (2011) eXiT*CBR: A framework for case-based medical diagnosis development and experimentation. Artificial Intelligence in Medicine 51(2): 81-91.

25. LIBSVM (2021) A Library for Support Vector Machines.

26. DR Wilson, TR Martinez (2019) Improved Heterogeneous Distance Functions.

27. MJ Huang, MY Chen, SC Lee(2007) Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. Expert Systems with Applications 32(3): 856-867.

28. CR Nicandro, Efrén MM, Yaneli AAM, Enrique MDC, Gabriel AMH, et al. (2013) Evaluation of the Diagnostic Power of Thermography in Breast Cancer Using Bayesian Network Classifiers. Computational and Mathematical Methods in Medicine 2013: e264246.

29. MA Ayu, SA Ismail, AFA Matin, T Mantoro (2012) A Comparison Study of Classifier Algorithms for Mobile-phone's Accelerometer Based Activity Recognition. Procedia Engineering 41: 224-229.

30. W Nor, H Mohamed, M Salleh, AH Omar (2013) A Comparative Study of Reduced Error Pruning Method in Decision Tree Algorithms. May 2013.

ISSN: 2574-1241

DOI: 10.26717/BJSTR.2022.42.006722

Benjamin Kofi Kujabi. Biomed J Sci & Tech Res