Eximia Journal

(ISSN 2784-0735)

Vol. 4

2022

Eximia Journal
Vol. 4, 19-32, May, 2022
ISSN: 2784-0735
www.eximiajournal.com

Improving the Accuracy of Misclassified Breast Cancer Data using Machine Learning

Rong-Ho Lin[1], Benjamin Kofi Kujabi[2], Chun-Ling Chuang[3], Yueh-Chung Chen[4], Chang-Ming Chen[5],

[1 2]National Taipei University of Technology, Department of Industrial Engineering and Management. 1, Sec. 3, Zhongxiao E. Rd., Taipei 10608 Taiwan, ROC. Taipei, Taiwan, [3]Kainan University, Department of Information Management, [4]Division of Cardiology, Department of Internal Medicine, Taipei City Hospital, Renai Branch, Taipei, [4]Department of Radiation Oncology, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan

rhlin@mail.ntut.edu.tw, kofi_ben3@hotmail.com, clchuang@mail.knu.edu.tw, chenyuehchung.tw@yahoo.com.tw, c5766.c5766@gmail.com

Abstract. **Background**: Breast cancer is the most common cancer among women. Many studies have made significant gains to classify breast cancer tumors with much emphasis on 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. To determine the breast cancer tumors, we applied the J48, IBK, LibSVM, JRip, and Multi-Layer Perceptron (MLP) classifiers to classify the breast cancer tumors. Next, 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 integration of WEKA and myCBR plugin with protégé had unprecedented results with misclassified data. Thus, providing accurate diagnostics procedures for distinguishing between benign and malignant.

Keywords. Misclassified data, Classifiers, WEKA, myCBR, protégé

1. 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 cases of breast cancer 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 detection 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 health of patients. For example, a malignant tumor might be interpreted as a benign tumor, resulting in a false negative (FN) (Type 1 error). Moreover, a benign tumor might be classified as a malignant tumor, resulting in a false positive (FP) (Type 2 error). These misinterpretations of false-positive or false-negative diagnoses of cancer 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 to accurately diagnose 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 prediction on new data [7], [8]. According to Ashutosh et al., (2016) [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 useful 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., (2018) [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 (2012) [12] and Asri et al., (2016) [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 research on this problem is limited. To alleviate the problem of data misclassification Smith and Martinez (2011) [14] proposed a PRISM method that identifies and removes instances that should be misclassified, 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 classification of breast cancer, but these algorithms and models have limitations and a lot of imperfections, thus it is crucial to develop a practical and effective model to perfect the classification of breast cancer tumors in order to avoid errors in clinical diagnosis of breast cancer and reduce the mortality rates of breast cancer patients. In this research, we develop a novel hybrid system that comprises WEKA and case-based reasoning (CBR) using the 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. The CBR is useful in leveraging knowledge encapsulated in previously learned cases and resolves other cases to support the creation of new decisions [15]. Accordingly, the findings of this research can provide clinicians with diagnostics procedures for distinguishing between benign and malignant tumors.

2. Literature Review

Breast cancer detection and classification have been reported in several studies [16]–[19]. These studies applied different approaches using data mining techniques. In this section, we present details of some previous work done in the area of breast cancer diagnosis by other researchers. In the work of Chaurasia and Tiwari (2018) [20], they applied Naïve Bayes, RBF networks, and J48 algorithms using the Wisconsin breast cancer dataset to predict the survivability of breast cancer patients. Their results focus on the performance of the classifiers. In another research similar to [21], the authors proposed to diagnose and analyze breast cancer disease by applying multilayer perceptron and SVM algorithms and thereafter assess their performance. Their results obtained 96.71% accuracy. Although, many researchers focus on the classification accuracy of the algorithms with little interest in improving misclassified data. Pruengkarn et al., (2015) [22] proposed a framework for misclassification analysis using fuzzy C-means and ensemble techniques, which obtained an improved performance of 14.36% after using majority voting. In the work of Liu et al., (2018) [6], they tackled the deficiency of unequal misclassification costs for breast cancer diagnosis and further improve the classification accuracy of the diagnosis of breast cancer. In this approach, the authors proposed an improved cost-sensitive support vector machine classifier (ICS-SVM) to obtain an optimal result that outperformed all the existing methods. Furthermore, Smith and Martinez (2011) [14] proposed the PRISM method that identifies and removes instances that should be misclassified. They applied nine learning algorithms to improve the accuracy with an average of 1.3 to 1.9%.

In an effort to merge applications and built a robust model, Abdrabou and Salem [20] presented a breast cancer classification model based on a combination of ontology and case-based reasoning to effectively classify breast cancer tumors. In their approach, they used JCOLIBRL and myCBR object-oriented frameworks.

In comparing classifiers for breast cancer tumors, in the work of Banu et al (2017) [24] , they compared J48, One R, Zero R and decision stump using the WBC dataset. The experimental results showed that J48 had the highest classification accuracy.

T. Sanli et al., (2020) [25] in their study to assess the performance of datamining on three datasets using the WEKA application, reported that the KNN, SMO, and J48 were the three most successful classification algorithms.

H.Z.M Shafri and F.S.H Ramle (2009) [26] investigates a new approach in image classification using SVM and Decision tree method, in which the results showed that the accuracy of SVM was 73% and the overall accuracy of DT method was 69%.

Balogun A.O et al., (2017) [27] develop an intrusion detection system using health care, breast cancer, diabetes dataset from the WEKA repertory and the KDDCup’99 dataset. They applied Naïve Bayes, Radial Basis Function and Ripper algorithm on the datasets. The results showed Ripper algorithm gave the best accuracy (99.76%) on the KDDCap’99 dataset.

Md Akizur Rahman and Ravie Chandren Muniyandi (2020) [28] applied a two-step feature selection (FS) technique with 15-neuron neural network to classify cancer using (WDBC) dataset, their classification accuracy results showed a significant improvement of 99.4%.

The aforementioned literature showed the lack of research in correcting misclassification. Therefore, this research was to build a novel hybrid integrated system for the classification of breast cancer tumors and correction of misclassified data. Most importantly, our hybrid system is better suited for correcting misclassified data with an optimal correction percentage when compared to Na Liu et al., (2018) [6], Smith and Martinez (2011) [14], and Chaurasia et al., [20].

3. Methodology:

This research proposes a novel hybrid integrated system comprising WEKA and myCBR with protégé for the classification of breast cancer tumors and the correction of 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.

3.1. Data Description

Considering the emphasis made on the classification of breast cancer tumors, several researchers have used the SEER (Surveillance, Epidemiology and End Results Program) [29] and WBC dataset to develop a prognosis and predictive models. In this research, we used the WBC dataset [30] collected by Dr. William H. Wolberg (1981 – 1991) at the University of Wisconsin Madison Hospital. 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 the process of 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)

3.2. Experimental Tools

3.2.1. WEKA

The WEKA workbench [31] 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.

3.2.2. myCBR (Final Protégé-Based Release)

MyCBR is an open-source similarity-based retrieval plugin for ontology-based applications. It is a powerful tool that can be used to test highly sophisticated models, knowledge-intensive similarity measures, and it can be easily integrated with other applications [32]. 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 [33-34].

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.

3.3. Experimental Procedure:

In this section, we detail the experimental procedures carried out in this research. First, the data were preprocessed and then developed a standard WEKA ARFF (Attributed Relation File Format) from the WBC dataset. The ARFF file was uploaded in the WEKA toolkit, where the feature selection method using the ranker and IfoGain attribute evaluator to generate different feature subsets [35]. The data were later weighted down for classification and J48, IBK, LibSVM, JRip, and MLP classification algorithms were applied.

Eximia Journal
Vol. 4, 19-32, May, 2022
ISSN: 2784-0735
www.eximiajournal.com

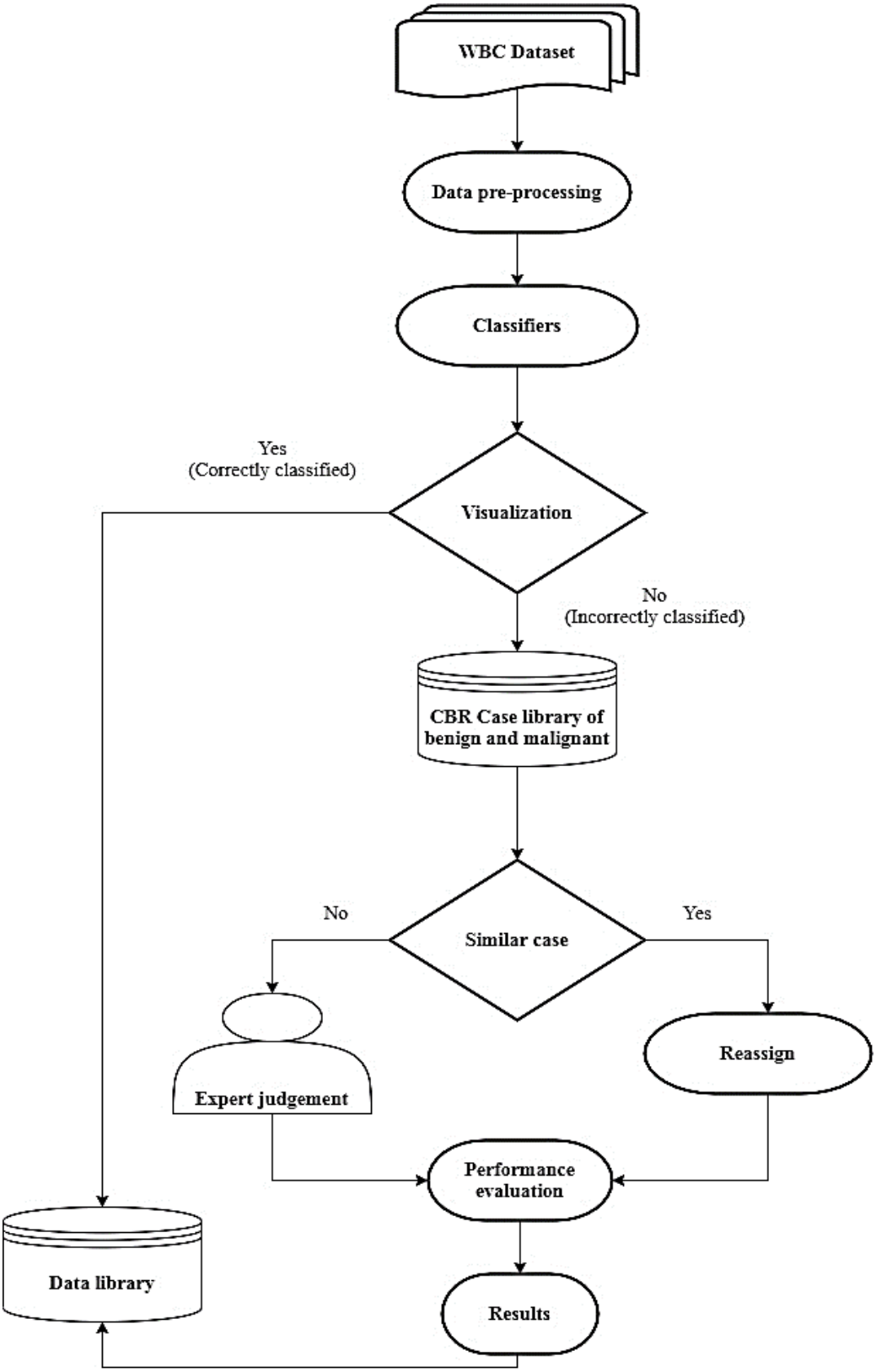

Figure 1. Proposed model of our integrated WEKA and myCBR with protégé

Eximia Journal
Vol. 4, 19-32, May, 2022
ISSN: 2784-0735
www.eximiajournal.com

3.4 Testing the model

A k-fold cross-validation method was applied to test the generalization ability of our proposed method. Accordingly, the data were split into 10 subsets randomly. Subsequently, we used 10-1 of these subsets for training; that is, we used one-tenth of the dataset for testing the nine-tenth for training. 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 the option of 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 data format and imported into 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.

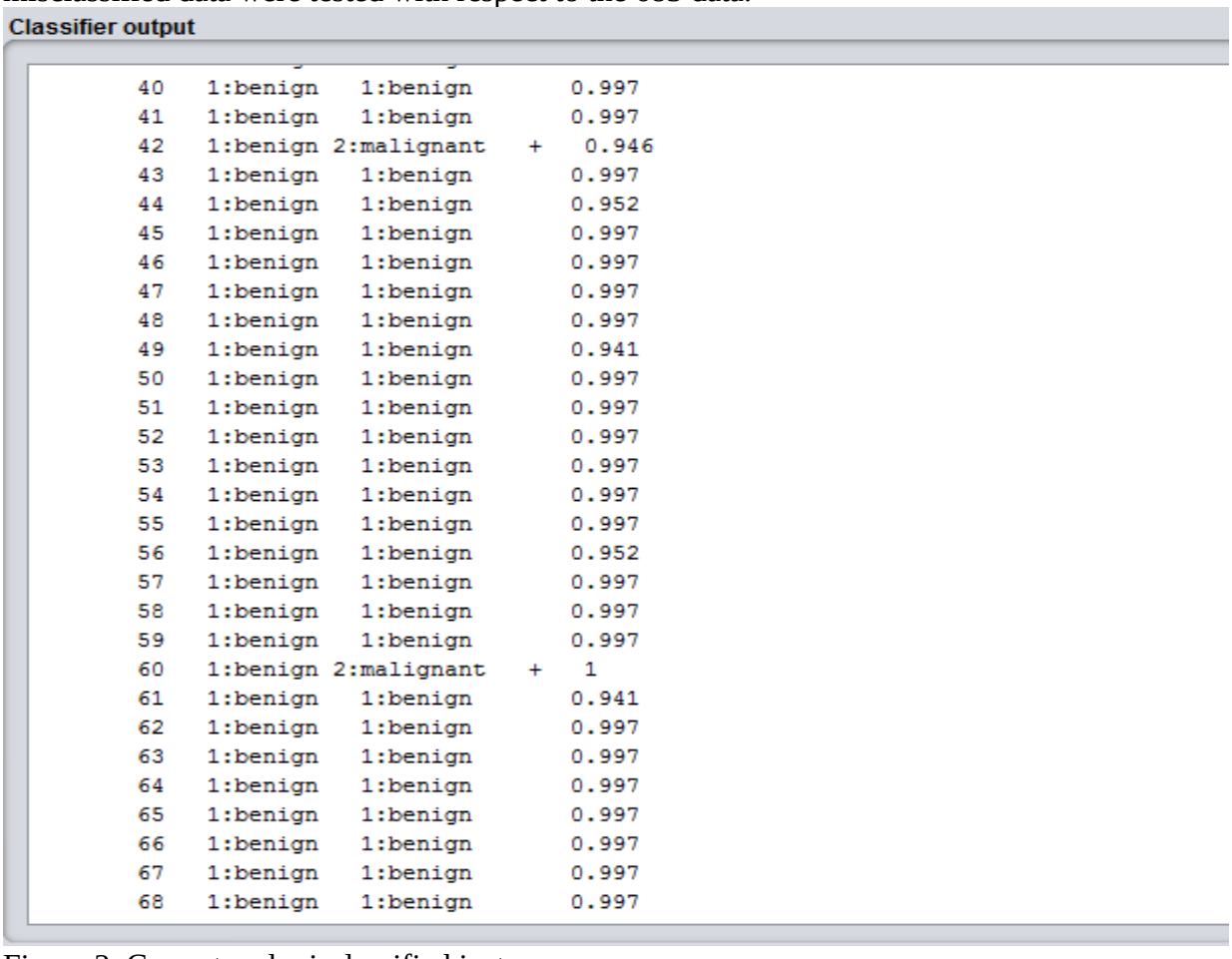

Figure 2. Correct and misclassified instances.

CBR

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 by applying the nearest neighbor algorithm, which computes the similarity between two cases using 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 characteristics 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. If no similarity exists in the CBR library database, then 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 of three task or functions such as Case library, Similarity measure, and Local similarity measure.

Case library: The case library stores historical solved data of known cases. For a new case 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 [36].

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

***Local similarity*:** Local similarity can be used to deal with missing values and is cost-sensitive. It is widely used in medical applications [37].

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

Where

A and B represents a new and previous feature.

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.

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

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 $i \sum_{i=1}^{p} wi = 1$ 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, it measures the distance between two points. The formula proposed by Gu et al., (2017) [36] was applied to measure the distance during the retrieval process.

$$\text{EU(t, r)} = \sqrt{\sum_{i=1}^{n} w_i d_i^2(t, r)}, \text{ where}$$
$$\text{d}_i\text{ (t, r)} = diff(x_{t,i}, x_{r,i}),$$
$$diff(x_{t,i}, x_{r,i}) = x_{t,i} - x_{r,i} \quad (3)$$

The effects of measuring scales can be avoided by normalizing the input attributes. 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 as follows in equation 4 [38], [39].

$$WEU\text{ (t, r)} = \sqrt{\sum_{i=1}^{n} w_i d_i^2(t, r).} \quad (4)$$

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.

Eximia Journal
Vol. 4, 19-32, May, 2022
ISSN: 2784-0735
www.eximiajournal.com

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 method of mean values entails gathering cases that contain either v1 or v2. Therefore, the system sums the mean value of the two groups to produce the output [40].

4. Evaluation method

To evaluate the performance of the novel hybrid system, a standard data classification system was used for accuracy, sensitivity, specificity, geometric mean (G-mean), evaluation, and determining 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. For example, a malignant tumor (negative) is classified as benign (positive) when it is actually 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 actually is a benign tumor (positive) [2].

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

$$\frac{TP+TN}{TP+FN+FP+TN} \text{ x } 100 \tag{5}$$

Where equation 6 represent Sensitivity, which is to correctly generate positive cases with either malignant or benign also known as TP rate.

$$\frac{TP}{TP+FN} \tag{6}$$

Where equation 7 represent the Specificity, this is to correctly generate negative cases of those without benign or malignant (also known as the TN rate).

$$\frac{TP}{TN+FP} \tag{7}$$

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

$$\sqrt{\frac{TP}{TP+FN} x \frac{TN}{TN+FP}} \tag{8}$$

Table 2. Confusion matrix.

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

5. Experimental Results

A thorough experiment was conducted and the results are discussed in this section. The J48, IBK, LibSVM, JRip, and MLP classifiers were first applied to classify benign and malignant tumors. As mentioned, we obtained a breast cancer dataset from the Wisconsin repository consisting of 699 cases. After preprocessing 683 cases were used for classification process. Table 3. presents a confusion matrix used for evaluation. The accuracy, sensitivity, specificity, and G-mean were evaluated. The LibSVM had the highest accuracy (96.93%) among the classifiers. In this research, emphasis was placed on improving misclassified data; accordingly, we embedded the myCBR plugin 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%). This was followed by J48 (2.83), MLP (2.5), LibSVM (1.8), and JRip (1.77).

Table 3. Classification results based on confusion matrix.

Classifier	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

6. Discussion

The objective of this research is to proposed 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. The breast cancer dataset used in this research was derived from the Wisconsin breast cancer repertory and consist of 10 attributes as shown in table. In Table 3. a confusion matrix was used for the evaluation of the classifiers. Although, different parameter settings were applied but nonetheless, the LibSVM classifier outperformed all other classifiers and achieved a 96.93% accuracy level. This is as a result that LibSVM performs better with numerical attributes and its good at computing speed and memory. 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 lack the power to represent interactions among variables [41]. Looking at the results of IBK, it can be observed that since IBK is an instance-based learning algorithm, it is reasonable to understand why MLP performs better than IBK. IBK can be much more an effective tool to produce a high classification accuracy when finetuned [42]. J48 and Jrip are both decision trees but it can be seen that Jrip performs slightly better than J48. This can be attributed to their pruning methods

or their adaption to dataset. In the case of J48, it adapts a subtree replacement which replaces nodes in decision trees with leaf and subtree raising, involves moving nodes upwards toward the tree's root while replacing other nodes. In certain cases, when the J48 performs poorly, it can be due to the complexity and heterogeneity of values of attributes. On the other hand, the Jrip isolates some data to reduce error pruning and adapts simple rules to improve the accuracy [43]. However, considerable emphasis was placed on improving the misclassified data and when the hybrid model was established, an upward spike in accuracy (ranging from 1.77 to 2.92%) for IBK showed considerable improvement. The need to manage and correct the misclassified data serves as an essential factor for prognosis and diagnosis. It minimizes the risk of physicians misinterpreting tumors. The system can provide accurate diagnostic procedures for distinguishing between benign and malignant tumors. The results demonstrate that the system is one of the best in terms of correcting misclassified data when compared to other models.

7. Conclusion

Improving errors in misclassification will not only help physicians to make the right judgment but also save the lives of breast cancer patients. The American Breast Cancer Society reported that breast cancer has been the leading cause of death in women and this has led 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 the integration of WEKA and the myCBR plugin with protégé had provided unprecedented results regarding the correction of misclassified data.

In future research, we will extend this model to accurately predict the stages of cancer. Therefore, we will optimize the parameters of the classifiers to minimize misclassification. We expect this to help physicians make swift decisions regarding the prediction of breast cancer stages.

References

1. Cancer Facts & Figures 2021 | American Cancer Society [Internet]. [cited 2021 Jul 8]. Available from: https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2021.html
2. Wang S, Wang Y, Wang D, Yin Y, Wang Y, Jin Y. An improved random forest-based rule extraction method for breast cancer diagnosis. Applied Soft Computing\019 Nov 18;105941. https://doi.org/10.1016/j.asoc.2019.105941
3. Borges L. Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection. In 2015. https://www.researchgate.net/publication/311950799_Analysis_of_the_Wisconsin_Breast_Cancer_Dataset_and_Machine_Learning_for_Breast_Cancer_Detection
4. Hayashi Y, Nakano S. 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. 2015 Jan 1;1:9–16. https://doi.org/10.1016/j.imu.2015.12.002
5. Devi DH, Devi DMI. OUTLIER DETECTION ALGORITHM COMBINED WITH DECISION TREE CLASSIFIER FOR EARLY DIAGNOSIS OF BREAST CANCER R.

In 2016. https://www.technicaljournalsonline.com/ijeat/VOL%20VII/IJAET%20VOL%20VII%20ISSUE%20II%20APRIL%20JUNE%202016/20167217.pdf
6. Liu N, Shen J, Xu M, Gan D, Qi E-S, Gao B. Improved Cost-Sensitive Support Vector Machine Classifier for Breast Cancer Diagnosis [Internet]. Mathematical Problems in Engineering. 2018 [cited 2020 Feb 20]. Available from: https://www.hindawi.com/journals/mpe/2018/3875082/
7. Kate RJ, Nadig R. Stage-specific predictive models for breast cancer survivability. International Journal of Medical Informatics. 2017 Jan 1;97:304–11. DOI:10.1016/j.ijmedinf.2016.11.001
8. Data Mining - 4th Edition [Internet]. [cited 2019 Dec 26]. https://doi.org/10.1016/C2015-0-02071-8
9. Dubey AK, Gupta U, Jain S. Analysis of k-means clustering approach on the breast cancer Wisconsin dataset. Int J Comput Assist Radiol Surg. 2016 Nov;11(11):2033–47. DOI: 10.1007/s11548-016-1437-9
10. Ravi Kumar G. An efficient prediction of breast cancer data using data mining techniques. In 2019. http://ijiet.com/wp-content/uploads/2013/09/18.pdf
11. Abdar M, Zomorodi-Moghadam M, Zhou X, Gururajan R, Tao X, Barua PD, et al. A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recognition Letters. 2020 Apr 1;132:123–31. https://doi.org/10.1016/j.patrec.2018.11.004
12. Aloraini A. Different Machine Learning Algorithms for Breast Cancer Diagnosis. International Journal of Artificial Intelligence & Applications. 2012 Nov 30;3:21–30. DOI: 10.5121/ijaia.2012.3603
13. Asri H, Mousannif H, Moatassime HA, Noel T. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia Computer Science. 2016 Jan 1;83:1064–9. https://doi.org/10.1016/j.procs.2016.04.224
14. Smith MR, Martinez T. Improving classification accuracy by identifying and removing instances that should be misclassified. In: The 2011 International Joint Conference on Neural Networks. 2011. p. 2690–7. **DOI:** 10.1109/IJCNN.2011.6033571
15. Zhuang ZY, Churilov L, Burstein F, Sikaris K. Combining data mining and case-based reasoning for intelligent decision support for pathology ordering by general practitioners. European Journal of Operational Research. 2009 Jun 16;195(3):662–75. https://doi.org/10.1016/j.ejor.2007.11.003
16. Aruna S, Rajagopalan D, Nandakishore L. Knowledge based analysis of various statistical tools in detecting breast cancer. Computer Science & Information Technology. 2011 Jul 1;2. DOI: 10.5121/csit.2011.1205
17. Mandal SK. Performance Analysis of Data Mining Algorithms for Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree. International Journal of Engineering and Computer Science [Internet]. 2017 Feb 26 [cited 2021 Jul 7];6(2). Available from: http://www.ijecs.in/index.php/ijecs/article/view/2616
18. Deshmukh B, Patil A, Pawar B. Comparison of Classification Algorithms using WEKA on Various Datasets [Internet]. 2012 [cited 2021 Jul 7]. Available from: https://www.semanticscholar.org/paper/Comparison-of-Classification-Algorithms-using-WEKA-Deshmukh-Patil/a832e0e87e5e2273ab7c601bf080693c0c1049e7
19. Sharma T, Jain M. WEKA Approach for Comparative Study of Classification Algorithm [Internet]. 2013 [cited 2021 Jul 7]. Available from:

https://www.ijarcce.com/upload/2013/april/60-trilok-WEKA%20approach%20for%20comparative.pdf
20. Chaurasia V, Pal S, Tiwari B. Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology. 2018 Jun;12(2):119–26. https://doi.org/10.1177/1748301818756225
21. Ghosh S, Mondal S, Ghosh B. A comparative study of breast cancer detection based on SVM and MLP BPN classifier. In: 2014 First International Conference on Automation, Control, Energy and Systems (ACES). 2014. p. 1–4. DOI: 10.1109/ACES.2014.6808002
22. Pruengkarn R, Fung CC, Wong KW. Using misclassification data to improve classification performance. In: 2015 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). 2015. p. 1–5. DOI: 10.1109/ECTICon.2015.7206950
23. Lotfy Abdrabou EAM, Salem A-BM. A breast cancer classifier based on a combination of case-based reasoning and ontology approach. In: Proceedings of the International Multiconference on Computer Science and Information Technology. 2010. p. 3–10. DOI: 10.1109/IMCSIT.2010.5680045
24. Banu G, Perumal P, Bashier I. Applications of Data Mining Classification Techniques on Predicting Breast Cancer Disease. IJLTET. 2017 Mar 28;8:321–5. DOI: 10.21172/1.82.043.
25. Sanlı T, Sıcakyüz Ç, Yüregir OH. Comparison of the accuracy of classification algorithms on three data-sets in data mining: Example of 20 classes. International Journal of Engineering, Science and Technology. 2020 Sep 15;12(3):81–9. **DOI:** 10.4314/ijest.v12i3.8
26. A Comparison of Support Vector Machine and Decision Tree Classifications Using Satellite Data of Langkawi Island [Internet]. Science Alert. [cited 2021 Jul 8]. Available from: https://scialert.net/fulltext/?doi=itj.2009.64.70 **DOI:** 10.3923/itj.2009.64.70
27. Balogun A, BALOGUN A, Sadiku P, Adeyemo V. Heterogeneous Ensemble Models for Generic Classification. Scientific Annals of Computer Science. 2017 May 5;VX:2017. https://www.researchgate.net/publication/320383367_Heterogeneous_Ensemble_Models_for_Generic_Classification
28. Rahman A, Muniyandi R. An Enhancement in Cancer Classification Accuracy Using a Two-Step Feature Selection Method Based on Artificial Neural Networks with 15 Neurons. Symmetry. 2020 Feb 10;12:271. https://doi.org/10.3390/sym12020271
29. SEER Incidence Data, 1975 - 2018 [Internet]. [cited 2021 Jul 7]. Available from: https://seer.cancer.gov/data/
30. UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set [Internet]. [cited 2021 Jul 7]. Available from: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
31. Weka 3 - Data Mining with Open Source Machine Learning Software in Java [Internet]. [cited 2021 Jul 7]. Available from: https://www.cs.waikato.ac.nz/ml/weka/
32. myCBR [Internet]. [cited 2021 Jul 7]. Available from: http://www.mycbr-project.org/
33. protégé [Internet]. [cited 2021 Jul 7]. Available from: https://protege.stanford.edu/
34. Protege Wiki [Internet]. [cited 2021 Jul 7]. Available from: https://protegewiki.stanford.edu/wiki/Main_Page
35. Liu N, Shen J, Xu M, Gan D, Qi E-S, Gao B. Improved Cost-Sensitive Support Vector Machine Classifier for Breast Cancer Diagnosis. Mathematical Problems in Engineering. 2018 Nov 28;2018:e3875082. https://doi.org/10.1155/2018/3875082

36. Gu D, Liang C, Zhao H. A case-based reasoning system based on weighted heterogeneous value distance metric for breast cancer diagnosis. Artif Intell Med. 2017 Mar;77:31–47. DOI: 10.1016/j.artmed.2017.02.003
37. López B, Pous C, Gay P, Pla A, Sanz J, Brunet J. eXiT*CBR: A framework for case-based medical diagnosis development and experimentation. Artificial Intelligence in Medicine. 2011 Feb 1;51(2):81–91. https://doi.org/10.1016/j.artmed.2010.09.002
38. LIBSVM -- A Library for Support Vector Machines [Internet]. [cited 2021 Jul 8]. Available from: https://www.csie.ntu.edu.tw/~cjlin/libsvm/
39. Wilson DR, Martinez TR. Improved Heterogeneous Distance Functions. arXiv:cs/9701101 [Internet]. 1996 Dec 31 [cited 2019 Dec 26]; Available from: http://arxiv.org/abs/cs/9701101. https://doi.org/10.48550/arXiv.cs/9701101
40. Huang M-J, Chen M-Y, Lee S-C. Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. Expert Systems with Applications. 2007 Apr 1;32(3):856–67. https://doi.org/10.1016/j.eswa.2006.01.038
41. Nicandro C-R, Efrén M-M, María Yaneli A-A, Enrique M-D-C-M, Héctor Gabriel A-M, Nancy P-C, et al. Evaluation of the Diagnostic Power of Thermography in Breast Cancer Using Bayesian Network Classifiers. Computational and Mathematical Methods in Medicine. 2013 May 22;2013:e264246. https://doi.org/10.1155/2013/264246
42. Ayu MA, Ismail SA, Matin AFA, Mantoro T. A Comparison Study of Classifier Algorithms for Mobile-phone's Accelerometer Based Activity Recognition. Procedia Engineering. 2012 Jan 1;41:224–9. https://doi.org/10.1016/j.proeng.2012.07.166
43. Nor W, Mohamed H, Salleh M, Omar AH. A Comparative Study of Reduced Error Pruning Method in Decision Tree Algorithms. 2013 May 16; **DOI:** 10.1109/ICCSCE.2012.6487177