



Article

Breast Cancer Dataset from Coimbra: Pre-Ratings of Its Value to Machine Learning and Diagnosis

Gennady Chuiko*  and Denis Honcharov 

Department of Computer Engineering, Petro Mohyla Black Sea National University, 54003 Mykolaiv, Ukraine

* Correspondence: henadiy.chuyko@chmnu.edu.ua**Received:** 25 June 2025; **Revised:** 22 July 2025; **Accepted:** 30 July 2025; **Published:** 19 August 2025

Abstract: This study aimed to evaluate a relatively new dataset developed to facilitate the primary diagnosis of breast cancer, collected by the University Hospital Centre of Coimbra in Portugal. Based on these assessments, the authors sought to develop a clear visual classifier to assist medical professionals in prediction and monitoring. This classifier utilizes routine blood test results along with physical data, offering a more straightforward and cost-effective alternative to traditional mammographic studies. The Coimbra Breast Cancer Dataset (CBCD) includes the following attributes: Age, Body Mass Index (BMI), Glucose, Insulin, Homeostatic Model Assessment for Insulin Resistance (HOMA-IR), Leptin, Adiponectin, Resistin, and Monocyte Chemoattractant Protein-1 (MCP1). The visual classifier was designed using Java-based machine learning algorithms within the Java-based WEKA software (version 3.9.6). Its well-designed interface enables clinicians, even those without expertise in machine learning, to use these algorithms effectively. The nine attributes of the CBCD were statistically categorized into three subsets based on their relevance to the overall model. This organization may help reduce the dimensionality of the diagnostic dataset while allowing specific classifiers to exhibit their unique preferences. A properly tuned JRip classifier demonstrated acceptable performance with the entire dataset and was effective in reducing it to six or even four attributes. The primary advantage of this classifier lies in its decision rules, which are easy for medical professionals to interpret and apply.

Keywords: Breast Cancer; Machine Learning; Biomarkers; Visual Classifying; Diagnostics

1. Introduction

Breast cancer (BC) is the most prevalent cancer among Ukrainian women [1] and one of the most widespread and fearsome types of cancer worldwide [2]. As Women are the bearers of the national genetic pool, the diagnosis and monitoring of breast cancer remain a significant challenge for healthcare providers. The authors advocate for the use of Machine Learning (ML) techniques and modern software to analyze unique datasets dedicated to BC monitoring. Simultaneously, the results of such data mining must be presented in a highly visualized form to be accessible to clinicians.

A few widely known datasets related to BC are recognized among Machine Learning experts. For example, many still use the breast cancer dataset from the Institute of Oncology, University Medical Center, Ljubljana (1988) [3,4], which includes ten attributes and 286 instances. The binary class is nominal, with “no-recurrence-events” and “recurrence-events” cases (201 and 85 instances, respectively). This dataset is quite noisy, and the best classifier achieves only a precision of 0.713 using 10-fold cross-validation.

The newer (1995) Wisconsin Breast Cancer Dataset contains 31 attributes and 569 instances [5]. It includes

two classes: malignant (212) and benign (357). The newer BreakHis database, versions [6,7], may be a source of extension for the previous dataset, providing additional information about biopsy, tumor class, tumor type, patient ID, and magnification factor. Both datasets and databases are based on mammography and share the same courses.

The three datasets introduced above focus on what may be termed the task of “secondary diagnostics”: breast tumors already exist, and the goal is to classify them as benign or malignant, or as able/unable to recur. In contrast, the Coimbra BC dataset [8] addresses the primary diagnostic task. This relatively new (2018) and compact dataset comprises 116 cases, with 52 instances classified as “healthy” and 64 instances classified as “patient”. Specifically, this dataset is the focus of our study.

The dataset comprises ten attributes, nine of which are numeric and one nominal. The numeric attributes represent anthropometric data and indicators obtained from routine blood analysis. These include Age, Body Mass Index (BMI), Glucose, Insulin, Homeostatic Model Assessment for Insulin Resistance (HOMA-IR), Leptin, Adiponectin, Resistin, and Monocyte chemoattractant protein-1(MCP1). Despite being a new dataset, CBCD has garnered the attention of ML experts, as evidenced by papers [2,9–14] that have explored this dataset.

The main aims of the papers cited studies dedicated to the CBCD were as follows: identifying high-performance classifiers [2,12,13]; conducting attribute relevance tests and developing selection methods [11,12,14]; comparing the efficacy of various ML algorithms’ efficacies [9]; performing deep statistical analyses of the dataset [10]; and studying of trends in BC prognosis using ML [13]. These goals are generally regarded as legitimate among ML experts. Still, they appear to be far beyond the competence of clinicians, who are the primary decision-makers in breast cancer cases.

The CBCD has the potential to be a valuable resource for clinical decision-making, focusing on “primeval” diagnosis and monitoring using a concise set of available biomarkers. However, there are still aspects that require further study. After careful analysis, we have identified some unclear elements that need to be addressed, including:

- Clinicians need evaluations of data diagnostic validity, as well as constructive criticisms.
- The absence of a grounded attribute selection path would order the dataset, allowing the reduction of the number of attributes and grouping them by rank, understanding how dataset reduction or classification options can impact its performance.
- Classifiers with acceptable performance and an easy-to-understand visual presentation are needed, as many clinicians may be unfamiliar with machine learning and its associated mathematics.
- The cluster structure of the CBCD is unclear, as is the extent to which these clusters correspond to preassigned classes.

The above list outlines the tasks and structure of the paper. First, we will describe the dataset and the techniques used for handling it (Section 2). Next, we will present the results of relevancy evaluations related to the dataset’s attributes (Section 3.1) and further discuss our findings on outlier detection and noise estimations (Section 3.2). Following that, we will showcase the results of classifying both the complete dataset and its reduced versions (Section 3.3), along with some straightforward rules in plain text for clinicians who work with this dataset. Finally, the discussion and conclusions are presented in Sections 4 and 5.

There is much potential for the CBCD to be a more effective tool for clinical decisions. However, it requires a deeper understanding of its peculiarities and potential enhancements. First of all, the relevance of its attributes is unclear. The possibilities of dataset reduction remain unexplored. The evaluation of noise and outliers has not been performed. These aspects highlight the research gap, a lacuna that should be addressed. Additionally, the availability of visual ML tools for clinical decision-making is greatly appreciated. The paper presented here is an effort to move in this direction.

2. Materials and Methods

Data of various forms serve as “material” within machine learning. **Table 1** provides a brief description of the attributes, units, and ranges of CBCD data collected by the University Hospital Centre of Coimbra (Portugal). The UCI Machine Learning Repository offers free access to this dataset [8]. It is customary in machine learning to normalize numerical features within such datasets. However, the authors chose not to normalize the data in order to make it easier for clinicians to deal with familiar and understandable units of measurement. Excellent and detailed

descriptions of the study participants (64 women with breast cancer and 52 healthy volunteers) and details of the blood tests from the direct authors of the CBCD can be found in [14]. It is hardly worth repeating the same thing here a second time.

Table 1. Coimbra Breast Cancer dataset description.

#	Attributes	Type	Units	Range
1	Age	numeric	years	24–89
2	BMI (Body Mass Index)	numeric	kg/m ²	18.37–38.58
3	Glucose	numeric	mg/dL	60–201
4	Insulin	numeric	μU/mL	2.43–58.46
5	HOMA (Homeostatic Model Assessment for Insulin Resistance)	numeric	-	0.467–25.05
6	Leptin	numeric	ng/mL	4.31–90.28
7	Adiponectin	numeric	μg/mL	1.656–38.04
8	Resistin	numeric	ng/mL	3.21–82.1
9	MCP1 (Monocyte chemoattractant protein-1)	numeric	ng/mL	45.84–1698.44
10	Classes	nominal	-	{healthy, patients}

Note: The ordering of attributes is arbitrary.

WEKA [15,16] has become an excellent choice over the past two decades for clinicians who want to utilize machine learning to analyze biomedical signals and datasets. This powerful ML tool is convenient for data visualization, making it easier for clinicians to access important information and helping them make more informed decisions. WEKA is Java-based software designed to be helpful for those with no expertise in Machine Learning. WEKA achieves this effect due to its well-designed and ubiquitous graphical user interface. An example of WEKA applied to the Wisconsin Breast Cancer dataset can be found in [17]. The latest version of WEKA (3-9-6) was used in our paper.

Figure 1 is an example of “visuality” in WEKA. An experienced clinician or data scientist may note that healthy individuals predominate only in the two bins with the lowest Glucose Levels. The relation reverses to the opposite opposite starting from the third bin. The histogram states that the higher the glucose level, the higher the risk of breast cancer.

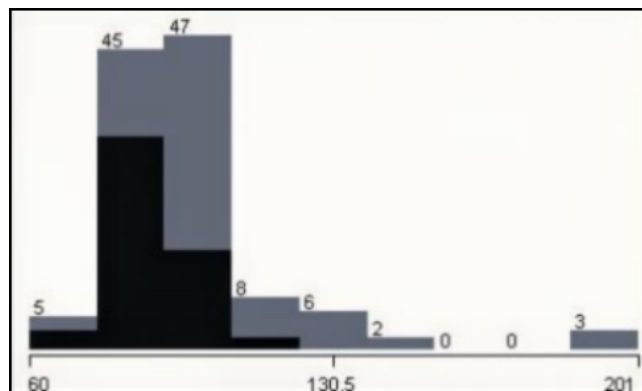


Figure 1. WEKA histogram for glucose concentration with nine bins (containers); red indicates patients with breast cancer, and blue indicates healthy people; the labels show the number of individuals covered by each bin.

Only the histogram of Glucose is presented here, as we will subsequently prove that Glucose is the most relevant (primary) attribute in the CBCD, ranking first. This conclusion contradicts the results of [9] and [12], where Glucose ranked second or third, but matches the latter outcome in [11]. However, similar histograms are generated by WEKA automatically for all numerical features. This is helpful for preliminary visual analysis of the dataset.

3. Results

3.1. Feature Selection and Ranking Concerning CBCD

Feature selection and ranking enhance performance in data mining [12,15]. One of the most significant advantages is the ability to eliminate irrelevant attributes. Fewer features in the dataset lead to a reduced workload and

quicker diagnostics. WEKA provides a variety of algorithms for attribute selection, ordering, and ranking based on relevance.

Table 2 presents eight of these algorithms applied to the CBCD. We can form rank vectors for each attribute, resulting in nine vectors, each with eight integer components ranging from 1 to 9. For instance, the vector for the Glucose attribute is (1, 2, 1, 1, 2, 1, 1, 1). Each vector contains eight components that correspond to the rows in **Table 2**. By determining the median rank of each attribute, we can arrange the set of units in ascending order: Glucose, Age, HOMA, Resistin, BMI, Insulin, Leptin, Adiponectin, and MCP1.

Table 2. CBCD attributes ranking (ordering) evaluations.

#	WEKA Evaluator	Attribute Ranking (Best First)
1	Correlation Ranking Filter	Glucose, HOMA, Insulin, Resistin, BMI, MCP1, Age, Adiponectin, Leptin
2	Gain Ratio feature evaluator	Age, Glucose, HOMA, Leptin, Resistin, BMI, Adiponectin, Insulin, MCP1
3	Information Gain Ranking Filter	Glucose, HOMA, Age, Leptin, Resistin, BMI, Adiponectin, Insulin, MCP1
4	OneR feature evaluator	Glucose, Age, Resistin, HOMA, Insulin, BMI, Adiponectin, Leptin, MCP1
5	RELIEF Ranking Filter	Age, Glucose, Resistin, Insulin, HOMA, BMI, Adiponectin, Leptin, MCP1
6	SVM feature evaluator	Glucose, BMI, Resistin, Insulin, HOMA, Age, Leptin, Adiponectin, MCP1
7	Symmetrical Uncertainty Ranking Filter	Glucose, Age, HOMA, Leptin, Resistin, BMI, Adiponectin, Insulin, MCP1
8	J48 feature evaluator	Glucose, Age, Resistin, HOMA, MCP1, BMI, Adiponectin, Leptin, Insulin

Next, we can calculate the Manhattan distance matrix for the rank vectors of the nine attributes in the specified order. The Manhattan metric (L1) is utilized because the vector components are integers. This distance matrix is symmetrical and has a size of 9×9 , with zeroes on the diagonal. **Figure 2** illustrates the heat maps of this matrix using a white-blue color scale, where white indicates a zero distance.

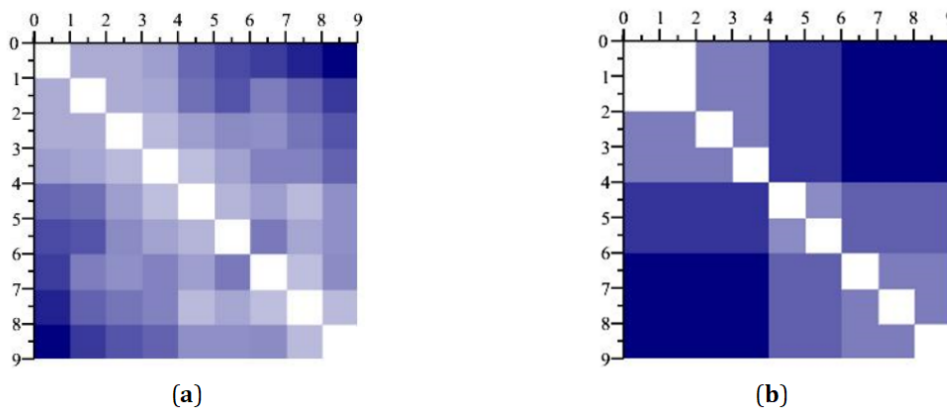


Figure 2. Heat maps of the distance matrix: “as is” (a) and with contrasting by distances averaging within matrix blocks (b); the white-blue scale implies that white represents zero distances.

The visual representation of the heat maps in **Figure 2** suggests the presence of three subsets within the CBCD attribute set. The first subset comprises four attributes, ranked from 1 to 4: Glucose, Age, HOMA, and Resistin. The second subset, sharing ranks 5 and 6, includes BMI and Insulin. The third subset, with ranks ranging from 7 to 9, consists of Leptin, Adiponectin, and MCP1. Someone may combine the second and third subsets, resulting in two subsets for such a reader.

We utilized the Two-Sample Paired T-test function to compute the paired T-test on the set of rank vector pairs [18]. This test allows for determining the significance of the difference between two means when the population standard deviation is unknown. Statistical analysis confirms the preliminary outcomes of visual insight. In summary, our results can be presented as follows:

- The differences in ranks among the pairs of attributes are generally insignificant within each subset, except for a few rare cases, at a 95% confidence level;
- However, the differences among the three subsets are statistically significant at the same confidence level;

- In the first subset, Glucose is likely to hold the highest rank (1), while MCP1 holds the lowest rank (9) in the third subset.

Figure 3 displays the statistical Box Plot (John Tukey chart) for the three CBCD subsets mentioned earlier. It shows some split-up among the subsets, although it would be desirable for it to be more conclusive.

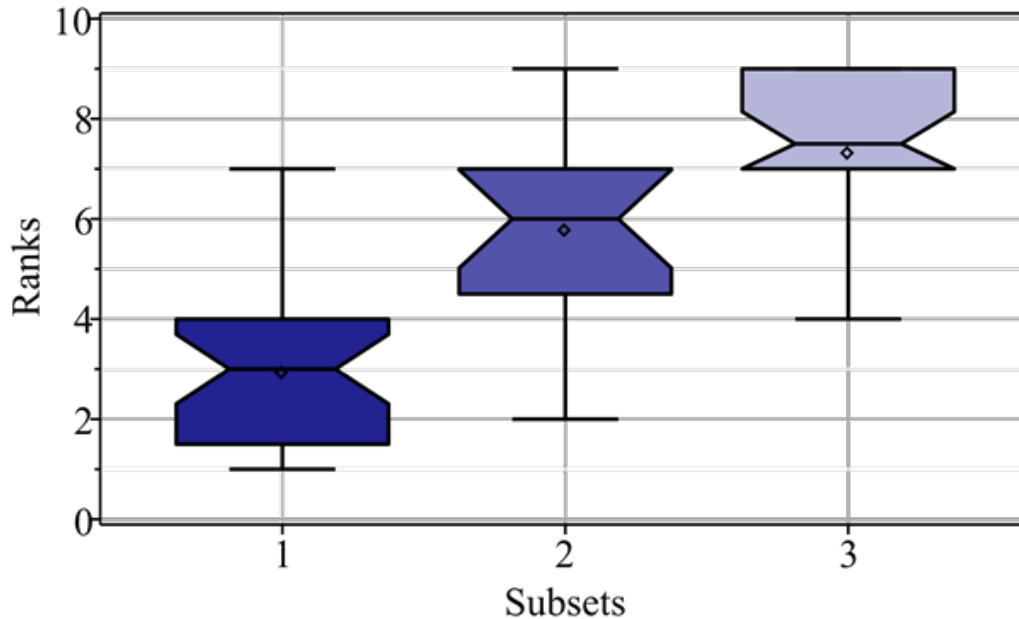


Figure 3. Box Plot (box-and-whisker plot) for CBCD subsets; ranges of subsets (whiskers), sample median (horizontal line on the “waist”), sample mean (the point), interquartile ranges (heights of boxes).

It is interesting to compare the subsets ranked by relevance that are proposed here with those from the cited articles (Table 3). As the statistical analysis mentioned above suggests, one could choose to disregard the specific order inside these subsets and instead focus on their content.

Table 3. Comparison of the contents and capacities of subsets with the highest rank.

Reference	Top-Ranked Subsets
[11]	{Resistin, Glucose, Age, BMI}
[12]	1. {Age, Glucose, Insulin, Leptin, Adiponectin} 2. {Age, BMI, Glucose, HOMA, Leptin, Adiponectin}
[14]	{Age, HOMA, Leptin, Adiponectin}
Our results	{Glucose, Age, HOMA, Resistin}

Although the results in Table 3 are not overwhelmingly conclusive, some common traits can be identified, and overall compliance seems to be moderate. The grouping errors that we aimed to avoid may become more pronounced when relying solely on intuitive selection [14], utilizing a single rater [11], or following the recommendations of casual evaluators [12]. The findings suggest that the Coimbra Breast Cancer Dataset (CBCD) could potentially be streamlined by focusing on a less relevant subset that includes three attributes: leptin, adiponectin, and MCP1. Any further reduction should be approached with caution and requires additional justification.

Now, one can transform the above-mentioned rank matrix, augmented with the entries indicating to which of the three classes (subsets) each of the nine rank vectors belongs. Glucose, Age, Resistin, and HOMA belong to the first (higher ranking) class. BMI and Insulin to the middle class, Leptin, Adiponectin, and MCP1 to the last (low) class. We can consider it as a small dataset with nine attributes, including the nominal class attribute, comprising nine instances.

The authors processed the dataset using the Kohonen self-organizing map (weka.clusterers.SelfOrganizingMap) with the following parameters: -L 1.0, -O 2000, -C 1000, -H 2, and -W 2. This WEKA algorithm was used for unsuper-

vised clustering with a neural network. We then analyzed the correspondence matrix that links the clusters to the three predefined classes. In this matrix, the classes are represented by the rows, while the columns represent the identified clusters. A diagonal matrix indicates a perfect match between the clusters and the classes. In our analysis, we achieved exactly that—a diagonal matrix with the values 4, 2, and 3 on the main diagonal. These numbers represent the number of attributes in each class (cluster, subset) and are consistent with the previously obtained results.

The Kohonen map, while useful, does not provide information about the ordering of attributes within each cluster (class or subset). However, it confirms the existence of three clusters and the appropriate quantities of characteristics within them. Interestingly, the algorithm indicated four clusters instead of three, but one of these clusters was empty and contained no attributes.

3.2. Outliers and Noisy Data in CBCD

Three WEKA filters were employed to identify outliers, or “noisy data,” within the CBCD dataset. The first filter uses interquartile ranges and established statistical methods to detect extreme values. This outlier detection filter identified a total of 12 instances—three from healthy individuals and nine from patients—out of 116 cases. As a result, the outliers account for approximately 10% of the dataset on average. The level of noise within the “patients” class is even higher, at about 14%.

The second filter used is CAIRAD (Co-appearance-based Analysis for Incorrect Records and Attribute-value Detection) [19]. This filter allows the labeling of “noisy data” as missing. Importantly, it identifies four instances (three from the healthy class and one from the patient’s class) that are misclassified. Misclassification is particularly problematic for small datasets, such as CBCD, as it directly contributes to off-diagonal elements of the confusion matrix.

Additionally, CAIRAD enables the definition of a “noisy values percentage” for each attribute. **Table 4** presents these results. The results in **Table 4** confirm the existence of significant noise in the dataset. While the percentages of “missing” data might seem concerning, they are less critical than the aforementioned misclassification errors.

Table 4. Percentage of “noisy values” for attributes in accord with the CAIRAD filter [19].

Attrib.	Cluc.	Age	HOMA	Resist.	BMI	Insul.	Leptin	Adip.	MCP1
Noise %	6	14	6	6	9	12	8	16	0

The Local Outlier Factor (LOF) is a well-known local anomaly detection algorithm introduced in 2000, and it is widely used in various applications, including implementations within the WEKA framework. This algorithm combines the concepts of nearest neighbors and local density to compute the Local Outlier Factor (LOF) score. An LOF score of approximately 1 indicates that an instance is part of the “core” of the dataset. Instances with LOF scores greater than one are typically considered outliers.

However, the boundary between what constitutes the “core” and what should be classified as outliers is uncertain. Consequently, it can be somewhat blurred because it relies on the “thumb rule” forecast, which is a recognized drawback of the method. Nonetheless, the LOF algorithm has a long history of success in anomaly detection and is deserving of application to CBCD.

An array of LOF estimates for all 116 instances served as the population for proper statistical analysis. The histogram (**Figure 4**) illustrates an asymmetrical distribution with a “long tail” of outliers that merges with the “core” at approximately LOF = 1.3. Therefore, this point could be used as the threshold to distinguish between the “core” and outliers for CBCD. Additionally, a cumulative distribution function (CDF) can be derived for the LOF array, assuming an empirical statistical distribution for this one-dimensional array, which behaves as a random variable.

The chosen threshold of LOF = 1.3, along with the derived CDF, enables us to predict that the “core” of the dataset comprises 95 instances, while the “tail” contains 21 outliers. One can assume that the “core” instances form the diagonal elements of the confusion matrix, while the outliers occupy the off-diagonal elements. Then, the accuracy of the prognosis must be approximately 82%. Such a level might be sufficient for preliminary screening, but it is hardly enough for serious decision-making.

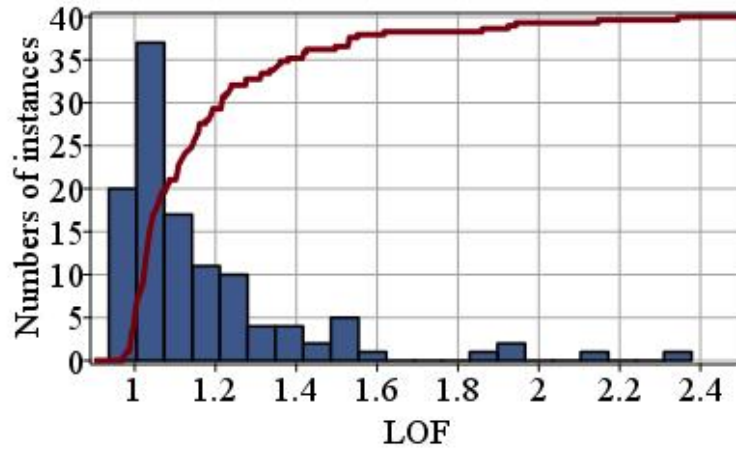


Figure 4. Histogram of Local Outlier Factors (LOF) for the Coimbra Breast Cancer Dataset. The curve represents the Cumulative Distribution Function ($0 \leq \text{CDF} \leq 1$) derived from the LOF array by numeric methods. The threshold was chosen at LOF 1.3.

3.3. Classifying and Reducing the CBCD Dataset

Some of the machine learning experts working with CBCD try to find the “best” classifier with the highest performance empirically. However, the issue is that there are many performance indicators for classifiers. Medics, for instance, traditionally prefer sensitivity (accuracy) and specificity (accuracy in remembering, recall). At the same time, ML practitioners often refer to the F1 measure, Matthew’s Correlation Coefficient (MCC), Cohen’s kappa statistic, or AUC-ROC when discussing a single index or evaluating a confusion matrix for overall performance. The “best” classifier problem is essentially a multi-criteria optimization problem but is often treated as a single-criteria one. While this approach is known and legitimate, supplementary criteria to the main one must be set at acceptable levels or incorporated into the problem’s constraints. These aspects are often omitted or unclear in works dedicated to CBCD.

The optimization of classifiers using a single-criterion approach can be automated [20]. Experience indicates that a classifier’s performance on the same dataset can vary significantly depending on the chosen optimization criteria and the tuning of its hyperparameters. In our study, we utilized the algorithm [20] with three different optimization criteria for the dataset: (1) the number of correctly classified instances, (2) the error rate, and (3) the Kappa Statistics (also known as Cohen’s Kappa). Among these, the Kappa statistic is widely recognized by machine learning experts. Kappa values over 0.75 are considered excellent, values between 0.40 and 0.75 are seen as fair to good, and values below 0.40 indicate poor agreement between the pre-assigned and classified classes. The highest achieved Kappa value achieved in our analysis was approximately 0.65.

Figure 5 illustrates the results of these beneficial exercises. **Figure 5a** indicates that metrics such as accuracy, the number of correctly classified instances, and the number of incorrectly classified ones are relatively the same across different criteria. Furthermore, these metrics align closely with the estimates provided in the previous section. In contrast, **Figure 5b** highlights the significant impact of criterion selection on the distribution of incorrectly classified instances among the classes, specifically regarding False Negatives and False Positives.

It is important to note that the confusion matrices, and consequently all performance metrics, are significantly influenced by the testing methods used. Let’s compare the confusion matrix obtained by using the dataset as the training set (Matrix **Cm0**) with the matrix resulting from the 6-fold cross-validation test method (Matrix **Cm6**). Both matrices were obtained for CBCD using the same classifier, with parameters that remained constant (weka.classifiers.rules.JRip -F 3 -N 4.435982504230809 -O 2 -S 1 -P). One can see a marked increase in False Positives (FP, from 0 to 12), which refers to the most critical errors, where a healthy individual is misclassified as sick. Cross-validation leads to a deterioration in all metrics related to both classes.

$$\mathbf{Cm0} = \begin{pmatrix} 40 & 12 \\ 0 & 64 \end{pmatrix}, \mathbf{Cm6} = \begin{pmatrix} 32 & 20 \\ 12 & 52 \end{pmatrix} \quad (1)$$

The confusion matrix **Cm0** likely indicates an overfitted model, a common issue that arises with small and noisy datasets when the entire dataset is used for training [21]. Aside from that, while cross-validation is generally effective for large datasets, its accuracy may drop when applied to smaller datasets. Thus, these matrices roughly define the upper and lower boundaries of the classification performance for CBCD in its current form (see **Table 5**).

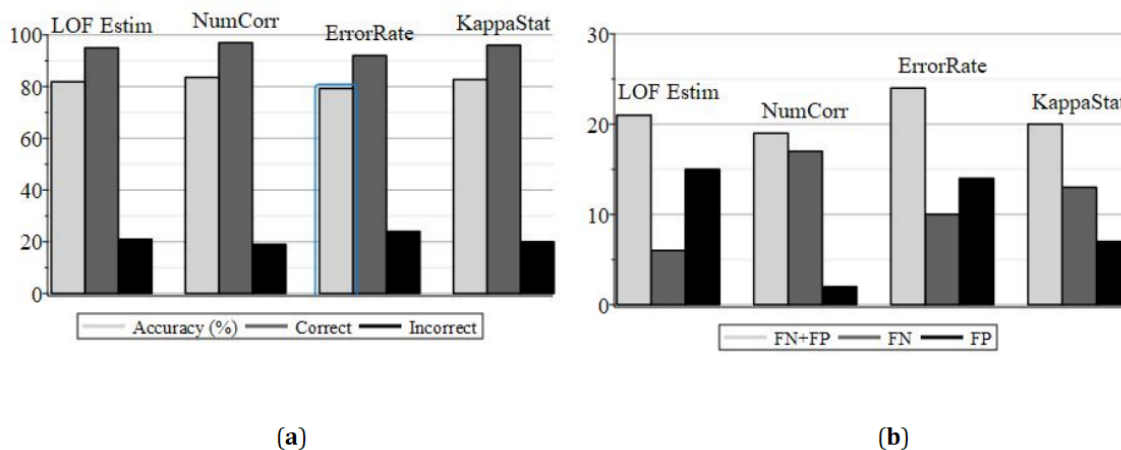


Figure 5. Some classification performance metrics: (a)—accuracy (in %), the numbers of correctly and incorrectly classified instances according to our prior estimation (Section 3.2) and three different optimization criteria; (b)—the numbers of incorrectly classified instances, including False Positives (FP) and False Negatives (FN). The typical standard deviations were (8.5–9.5) % at a confidence level of 0.95.

Table 5. Estimated upper and lower boundaries for some performance metrics of CBCD.

Metrics	Accuracy (%)	FP Ratio	Kappa	F1-Measure	ROC Area
Range	72–90	0.00–0.20	0.43–0.79	0.72–0.89	0.72–0.86

Filtering outliers or misclassified instances is advisable for noisy data. However, it’s important to proceed with caution to avoid removing essential components or entire minority classes, especially in smaller datasets. We utilized one of the WEKA filters (`weka.filters.unsupervised.instance.RemoveMisclassified`) with the following configuration: `-W "weka.classifiers.rules.PART -C 0.25 -M 2" -C-1 -F 0 -T 0.1 -I 0`. This process removed 16 instances out of a total of 116, primarily affecting the classes of healthy individuals (14 instances) and patients (2 instances). This result aligns with the estimates presented in Section 3.2, which indicated a range of 12 to 21 instances.

Let us define some notation:

“ds0”: This is the initial dataset containing nine numeric attributes (excluding class labels) and 116 instances.

“ds1”: This is the filtered dataset, which contains nine numeric attributes and 100 instances after removing outliers.

“ds2”: This is the reduced dataset, consisting of six attributes and 100 instances, which was achieved by eliminating the attributes Leptin, Adiponectin, and MCP1.

“ds3”: This is an additional reduced dataset with four attributes and 100 instances, created by excluding BMI and Insulin as attributes.

Note that both ways to attribute reduction are consistent with the conclusions of Section 3.1, while the filtering of outliers (denoising) is based on the results of Section 3.2.

We performed 1,000 tests for each dataset using the classifier mentioned above, employing 5-fold cross-validation mode. This approach allowed us not only to calculate the averaged values of performance metrics but also their standard deviations and the statistical significance of differences in each performance metric among datasets.

The results show an increase in all metrics from ds0 to ds1, then to ds2, and finally to ds3. Suppose one takes the accuracy (in percentages) as an example. Then the row of indexes displays this growth: 70.2 (8.7) < 79.7 (9.1)

$< 81.1(8.6) < 81.6(8.6)$. However, the standard deviations (values in brackets) indicate why such an apparent rise remains statistically insignificant. This experiment does not confirm the benefits of outlier cleaning and attribute reduction. However, it also does not rule out the possibility of improvements, as the results from these methods are at least comparable to those from the original dataset. Therefore, even a very compact dataset, such as ds3, which contains not only four attributes and 100 instances, can be effective when performance metrics are not overly demanding.

The JRip algorithm implements the propositional rule learner known as Repeated Incremental Pruning to Produce Error Reduction (RIPPER). William W. Cohen proposed this algorithm as an optimized version of IREP [22]. Based on the dataset, JRip generates simple “IF-THEN” rules, making it beneficial and transparent for clinicians.

1. IF Glucose \leq 90 AND Resistin \leq 12.9361 AND HOMA \geq 0.827271 THEN Classification is healthy (1) (leverages 17 cases, probability 100%)
2. IF Age \leq 36 AND Glucose \leq 90 THEN Classification is healthy (1) (leverage 7 cases, probability 100%)
3. IF Age \geq 66 AND Glucose \leq 102 AND Resistin \leq 12.766 THEN Classification= healthy 1 (leverages 9, probability 100%)
4. ELSE Classification is patient (2) (leverages 67, probability 93 %)

The listed rules were obtained from the shortest data set, ds3. This set of rules is easily programmable, allowing for the automatic processing of results from analyses of only four (and in practice, even three) indicators. If the aim is to conduct preliminary screening, such an automatic system would save time, labor, and costs.

4. Discussion

The Coimbra Breast Cancer Data Set, a relatively recent development, was the focus of this study. It includes various biomarkers and anthropometric indices as attributes and is divided into two classes: healthy individuals and patients. The primary purpose of this dataset is to aid in the preliminary diagnosis of breast cancer. Its attributes, which can be derived from routine blood analyses, make the Coimbra Breast Cancer Data Set promising for quick assessments. However, its clinical value remains under discussion and requires further exploration.

Firstly, both ML experts and clinicians should recognize the importance of attribute rankings (in terms of relevance and value) to identify opportunities for dataset reduction. By minimizing the number of features, diagnostic processes can be completed more quickly and with reduced effort.

Using statistical analysis and machine learning methods, the authors categorized CBCD attributes into three subgroups: highly relevant attributes (Glucose, Age, Resistin, HOMA), moderately relevant attributes (BMI, Insulin), and low-relevance attributes (Leptin, Adiponectin, MCP1). The differences between these subsets are statistically significant at a 95% confidence level; however, the specific ordering of attributes within each subset remains uncertain. Therefore, reducing the CBCD by removing the low-relevant subgroup (and perhaps the mid-relevant subgroup as well) is a reasonable approach. The reduction of attribute quantity depends on the used classifier and its tuning parameters, therefore requiring careful consideration. The experience from this study—particularly with classifiers like JRip and J48 — supports the cautious approach.

Clinical diagnosis is considered a classification problem in ML [15]. Any classification model maps an attribute set to a binary (nominal) class attribute. The JRip classifier [22] was chosen for its creation of easily interpretable “Decision Rules” (see Section 3.3) that can be directly used in clinical decision-making and its relatively high performance regarding CBCD. Some “tuning” of JRip, including the selection of its optimal hyperparameters, allows acceptable performance for both complete and reduced datasets.

The Coimbra Breast Cancer Dataset comprises 116 instances, approximately evenly distributed between the two classes: 52 healthy individuals and 64 patients. The dataset’s small size and the presence of a considerable amount of noise limit its overall utility. However, the attribute structure shows promise if the dataset is expanded and cleaned.

Secondly, clinicians and machine learning experts need to understand the reliability of the Coimbra dataset. Our study tested four filters for outliers, revealing that the dataset contains a significant amount of noise, with 12 to 21 instances identified as outliers. This represents a relatively high level of noise, especially given that the dataset contains only 116 cases. Therefore, high-performance metrics from any machine learning classifier cannot

be expected unless the noise is addressed.

Simply removing outliers is not the most effective approach to improvement, since the dataset is small. Instead, noise reduction should be accompanied by the necessary expansion of the dataset in terms of the number of instances. Our experience suggests that reducing outliers, as well as reducing attribute subsets, does not lead to significant improvements in performance metrics. Fortunately, these procedures do not degrade the metrics either.

5. Conclusions

The main conclusions of this study can be summarized as follows:

1. WEKA software is excellent for analyzing medical datasets such as CBCD. Its advanced visualization tools are particularly helpful for clinicians who rely primarily on visual aids.
2. Nine attributes of CBCD are statistically significantly divided into three subsets by relevance. This separation may guide diagnostic dataset reduction, although specific classifiers may exhibit their own “preference nuances”.
3. On average, approximately 15% ($\pm 5\%$) of the data in CBCD is questionable, indicating that it is a relatively “noisy” dataset. While denoising is desirable, it should be balanced with a significant expansion of the dataset in terms of the number of instances.
4. The duly “tuned” JRip classifier could be clinically helpful, offering acceptable performance and generating “Decision Rules” that are easy to interpret.
5. The Coimbra Breast Cancer Dataset is promising for preliminary diagnostics because it relies on a small set of relatively inexpensive and accessible markers. Its limitations include small dataset size and a relatively high level of noise from outliers.

Future research could be more effective if the dataset size were expanded and the class balance were improved [23]. This suggestion is particularly relevant to the authors of this promising dataset.

Author Contributions

Conceptualization, G.C.; methodology, G.C.; software, G.C., D.H.; validation, G.C. and D.H.; formal analysis, G.C., D.H.; investigation, G.C., D.H.; resources, G.C.; data curation, G.C., D.H.; writing—original draft preparation, G.C.; writing—review and editing, G.C., D.H.; visualization, G.C., D.H.; supervision, G.C.; project administration, G.C. All authors have read and agreed to the published version of the manuscript.

Funding

This work received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The Coimbra Breast Cancer Dataset (CBCD) is available in the UCI Machine Learning Repository as a CSV file. <https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic>

The datasets ds1, ds2, and ds3 mentioned in the text are derived from CBCD by the WEKA filter mentioned in the text (`weka.filters.unsupervised.instance.RemoveMisclassified` with the configuration: `-W "weka.classifiers.rules.PART -C 0.25 -M 2" -C -1 -F 0 -T 0.1 -I 0`), and the removal of one or two subsets of attributes. The reader can repeat these operations themselves, using WEKA.

Conflicts of Interest

The authors declare that they have no conflict of interest.

References

1. The World Bank. *Breast Cancer in Ukraine: The Continuum of Care and Implications for Action*; The World Bank: Washington, DC, USA. 2018. [CrossRef]
2. Abdulkareem, A.H.; Kasapbaşı, M.C. Enhancing Detection Method of Breast Cancer Using Coimbra Dataset. *İstanbul Ticaret Üniversitesi Teknoloji ve Uygulamalı Bilimler Dergisi* **2020**, *3*, 51–59. Available from: <https://dergipark.org.tr/tr/pub/icujtas/issue/57160/824195>
3. Zwitter, M.; Soklic, M.; 1988. Breast Cancer. UCI Machine Learning Repository. Available from: <https://archive.ics.uci.edu/dataset/14/breast+cancer>
4. Mohamed, T.S.; Khalifah, S.M. Breast Cancer Prediction: The Classification of Non-Recurrence-Events and Recurrence-Events Using Functions Classifiers. In Proceedings of the 3rd Information Technology to Enhance e-Learning and Other Applications (IT-ELA 2022); Baghdad, Iraq, 27–28 December 2022; pp. 55–60. [CrossRef]
5. Street, W.; Wolberg, W.; Mangasarian, O. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. 1995. Available from: <https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic>
6. BreakHis. Breast Cancer Histopathological Database (BreakHis). Available from: <https://www.kaggle.com/datasets/ambarish/breakhis>
7. Spanhol, F.A.; Oliveira, L.S.; Petitjean, C.; et al. A Dataset for Breast Cancer Histopathological Image Classification. *IEEE Trans. Biomed. Eng.* **2016**, *63*, 1455–1462. [CrossRef]
8. Patrcio, M.; Pereira, J.; Crisostomo, J.; et al., 2018. Breast Cancer Coimbra. UCI Machine Learning Repository. Available from: <https://archive.ics.uci.edu/dataset/451/breast+cancer+coimbra>
9. Austria, Y.D.; Goh, M.L.; Maria, L.B.S., Jr.; et al. Comparison of Machine Learning Algorithms in Breast Cancer Prediction Using the Coimbra Dataset. *Int. J. Simul. Syst. Sci. Technol.* **2019**, *20*, 233–240. [CrossRef]
10. Yue, J.; Zhao, N.; Liu, L. Prediction and Monitoring Method for Breast Cancer: A Case Study for Data from the University Hospital Centre of Coimbra. *Cancer Manag. Res.* **2020**, *12*, 1887–1893. [CrossRef]
11. Alfian, G.; Syafrudin, M.; Fahrurrozi, I.; et al. Predicting Breast Cancer from Risk Factors Using SVM and Extra-Trees-Based Feature Selection Method. *Computers* **2022**, *11*, 1367. [CrossRef]
12. Kayaalp, F.; Basarslan, M.S. Performance Analysis Of Filter Based Feature Selection Methods On Diagnosis Of Breast Cancer And Orthopedics. In Proceedings of the 6th International Congress on Fundamental and Applied Sciences (ICFAS 2019); Tirana, Albania, 18–20 June 2019; pp. 1–11. Available from: https://www.researchgate.net/publication/334401380_Performance_Analysis_Of_Filter_Based_Feature_Selection_Methods_On_Diagnosis_Of_Breast_Cancer_And_Orthopedics#fullTextFileContent
13. Salad, Z.; Singh, Y. A Five-Year (2015 to 2019) Analysis of Studies Focused on Breast Cancer Prediction Using Machine Learning: A Systematic Review and Bibliometric Analysis. *J. Public Health Res.* **2020**, *9*, 65–75. [CrossRef]
14. Patrício, M.; Pereira, J.; Crisóstomo, J.; et al. Using Resistin, Glucose, Age and BMI to Predict the Presence of Breast Cancer. *BMC Cancer* **2018**, *18*, 29. [CrossRef]
15. Bouckaert, R.R.; Frank, E.; Kirkby, R.; et al. WEKA Manual for Version 3-9-5. The University of Waikato; 2020. Available from: <https://sourceforge.net/projects/weka/>
16. Chuiko, G.P.; Darnapuk, Y.S.; Dvornik, O.V.; et al. Efficacy of Weka for Medical Data Mining: Ambulatory Blood Pressure Monitoring as a Case-Study. *Online J. Cardiol. Res. Rep.* **2023**, *7*, 7–9. [CrossRef]
17. Srikanth, K.; Zahoor, S.; Huq, U.L.; et al. Analysis, Implementation, and Comparison of Machine Learning Algorithms on Breast Cancer Dataset Using WEKA Tool. *Int. J. Recent Technol. Eng.* **2019**, *7*, 330–333. Available from: https://www.researchgate.net/publication/333115193_Analysis_implementation_and_comparison_of_machine_learning_algorithms_on_breast_cancer_dataset_using_WEKA_tool
18. Sun, J., Ed. Volume 171. *Progress in Molecular Biology and Translational Science. The Microbiome in Health and Disease*; Academic Press: Cambridge, MA, USA, 2020. p. 397.
19. Rahman, M.G.; Islam, M.Z.; Bossomaier, T.; et al. CAIRAD: A Co-Appearance Based Analysis for Incorrect Records and Attribute-Values Detection. In Proceedings of the 2012 International Joint Conference on Neural Networks (IJCNN); Brisbane, Australia, 10–15 June 2012; pp. 1–10. [CrossRef]
20. Thornton, C.; Hutter, F.; Hoos, H.H.; et al. Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge

- Discovery and Data Mining; Chicago, IL, USA, 11–14 August 2013; pp. 847–855. [CrossRef]
21. Ying, X. An Overview of Overfitting and Its Solutions. *J. Phys. Conf. Ser.* **2019**, *1168*, 022022. [CrossRef]
 22. Cohen, W.W. Fast Effective Rule Induction. In Proceedings of the 12th International Conference on Machine Learning; Tahoe City, CA, USA, 9–12 July 1995; pp. 115–123. Available from: <https://dl.acm.org/doi/abs/10.5555/3091622.3091637>
 23. Adebayo, O.J.; Omotayo, O.A.; Olaleye, I. Enhanced Breast Cancer Prediction Using ADASYN and Optimized LightGBM. *FECCUPIT Bull.* **2024**, *2*, 11–20. Available from: <https://bulletin.feccupit.ro/archive/pdf/20240202.pdf>



Copyright © 2025 by the author(s). Published by UK Scientific Publishing Limited. This is an open access article under the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Publisher's Note: The views, opinions, and information presented in all publications are the sole responsibility of the respective authors and contributors, and do not necessarily reflect the views of UK Scientific Publishing Limited and/or its editors. UK Scientific Publishing Limited and/or its editors hereby disclaim any liability for any harm or damage to individuals or property arising from the implementation of ideas, methods, instructions, or products mentioned in the content.