Honcharov Denys, PhD, Lecturer of the Department of Automation and Computer-Integrated Technologies, Petro Mohyla Black Sea National University, ORCID: https://orcid.org/0009-0004-1200-6677

GLOBAL VERSUS CLASS-BASED FEATURE RANKING FOR NOISY INSTANCE IDENTIFICATION IN BREAST CANCER DATA

Abstract

Background. The reliability of machine learning models in medical decision support systems depends not only on the classification algorithm but also on the quality of the training dataset. Previous studies evaluated the Coimbra Breast Cancer Dataset using a global consensus ranking of biomarkers obtained from multiple feature evaluators. However, such an approach assumes that biomarker relevance is identical for healthy individuals and breast cancer patients. **Objective**. This study investigates whether constructing independent biomarker rankings for each diagnostic class provides additional information about the dataset structure and improves the identification of potentially noisy patient records. **Methods**. The Coimbra Breast Cancer Dataset was divided into Healthy and Cancer subsets. For each subset, biomarker importance was independently evaluated using eight feature evaluation methods implemented in WEKA: Correlation Attribute Evaluation, Gain Ratio, Information Gain, OneR, ReliefF, SVM Attribute Evaluation, Symmetrical Uncertainty, and Classifier Attribute Evaluation. Consensus rankings were obtained by averaging the ranks assigned by the individual evaluators. The resulting class-specific rankings were compared with the conventional global ranking. Potentially noisy records were further analyzed using the Interquartile Range filter. **Results**. Statistical analysis confirmed that several biomarkers exhibit different distribution characteristics in the two diagnostic classes. Although the Healthy and Cancer subsets produced

similar consensus rankings, both differed substantially from the global ranking. In particular, MCP-1 changed from the lowest-ranked biomarker in the global analysis to the highest-ranked biomarker in both class-specific rankings, whereas Age became the least informative feature after class separation. The proposed approach identified five potentially noisy records in each diagnostic class. Most of these records were not statistical outliers according to the interquartile range criterion, indicating that they represent atypical combinations of biomarker values rather than isolated extreme measurements. **Conclusions**. The obtained results demonstrate that class-specific feature ranking provides additional information about the internal structure of medical datasets and complements conventional global feature ranking. The proposed approach can be used as an exploratory preprocessing technique for identifying potentially atypical patient records before applying machine learning methods.

Keywords: breast cancer, Coimbra Breast Cancer Dataset, feature selection, feature ranking, noisy data, outlier detection, machine learning, WEKA.

Problem statement. Medical decision support systems increasingly rely on machine learning methods trained on clinical datasets [1-3]. The reliability of such systems depends not only on the selected classification algorithm but also on the quality of the data used for training [1-4]. Even a small number of noisy, inconsistent or atypical records may distort the relationships between biomarkers and reduce the predictive performance of diagnostic models [4-11]. Therefore, the identification of potentially noisy instances has become an important stage of data preprocessing in medical applications.

A number of approaches have been proposed to assess the quality of medical datasets before model construction. One of them is based on the combined use of several feature evaluation methods to estimate the diagnostic importance of biomarkers and to identify records whose characteristics differ from the general

structure of the dataset. However, such an approach evaluates all patient records simultaneously, regardless of their class membership. In heterogeneous medical datasets this may reduce the sensitivity of the analysis because the significance of individual biomarkers often differs for healthy subjects and patients with a confirmed disease.

Although feature ranking methods have been extensively studied, little attention has been paid to the influence of class-specific biomarker rankings on the identification of noisy medical records. In particular, no comparative analysis between global and class-specific ranking strategies has been reported for the Coimbra Breast Cancer Dataset. This research gap motivates the present study.

Analysis of recent research and publications.

The Coimbra Breast Cancer Dataset has attracted considerable attention from researchers because it contains anthropometric characteristics and routine blood biomarkers that can be used for the primary diagnosis of breast cancer [14, 15]. Since its publication by Patrício et al., the dataset has been widely employed as a benchmark for evaluating machine learning methods in medical diagnostics [14].

The first studies based on the Coimbra dataset mainly focused on comparing the predictive performance of machine learning algorithms. Austria et al. evaluated several supervised classifiers implemented in WEKA and demonstrated that no single classification algorithm consistently outperformed the others for this dataset [2]. Their work primarily concentrated on classification accuracy and did not investigate the influence of individual patient records or dataset quality.

Further research shifted attention toward feature selection and biomarker analysis. Yue et al. investigated the diagnostic significance of clinical and biochemical indicators for breast cancer prediction and confirmed that only a subset of the available biomarkers makes the greatest contribution to classification performance [3]. Similarly, Alfian et al. proposed combining Support Vector

Machines with an Extra-Trees-based feature selection method, demonstrating that preliminary feature selection improves classification accuracy while reducing model complexity [4].

Recent advances in machine learning for medical diagnosis have increasingly focused on deep learning methods, large-scale medical data analysis and intelligent clinical decision support systems [5,6]. Although these studies report substantial improvements in predictive performance, they generally assume that the available datasets are reliable and therefore do not explicitly investigate the influence of noisy or atypical patient records.

A different research direction was presented in our previous study [12]. Instead of comparing classifiers, the emphasis was placed on evaluating the diagnostic value of the dataset itself. Eight feature evaluators were used to construct a consensus ranking of the nine biomarkers, after which every patient record was assigned, an integrated score reflecting its contribution to the overall dataset structure. This approach enabled the identification of potentially noisy instances, the analysis of dataset consistency and the construction of an interpretable JRip classifier based on the most informative biomarkers.

Despite the considerable number of studies devoted to the Coimbra Breast Cancer Dataset, previous research has mainly focused on improving classification accuracy or selecting informative biomarkers [2–6]. To the best of our knowledge, no published work has investigated whether constructing separate biomarker rankings for healthy individuals and breast cancer patients influences the identification of potentially noisy records. This unresolved issue constitutes the basis of the present study, in which the conventional global ranking strategy is compared with a class-specific ranking approach.

Formulation of the article's objectives. The purpose of this study is to investigate the influence of class-specific biomarker ranking on the identification of potentially noisy records in the Coimbra Breast Cancer Dataset. Unlike the

previously proposed approach, which evaluates all patient records using a single global ranking of biomarkers, the proposed method constructs independent rankings for healthy individuals and breast cancer patients. The study aims to compare both ranking strategies, analyze the differences in the detected potentially noisy instances and assess whether incorporating class-specific feature importance provides a more informative evaluation of dataset quality. The obtained results are expected to contribute to the development of more reliable data preprocessing procedures for machine learning applications in breast cancer diagnosis.

Main part. Motivation for Class-Specific Biomarker Ranking.

In our previous study, the diagnostic significance of the nine biomarkers included in the Coimbra Breast Cancer Dataset (CBCD) was evaluated by combining the results obtained from eight feature evaluation methods [9]. Based on the resulting consensus ranking, the contribution of each patient record to the overall dataset structure was estimated, making it possible to identify potentially noisy records and assess the suitability of the dataset for subsequent machine learning applications.

The proposed approach is based on global feature ranking, where the relative importance of each biomarker is determined using the complete dataset regardless of the patient's diagnostic class. Such an approach is well justified when the statistical properties of the biomarkers are similar across all observations. However, medical datasets are typically heterogeneous because the physiological and biochemical characteristics of healthy individuals and patients with a confirmed disease may differ substantially.

Under these conditions, the statistical characteristics of individual biomarkers may depend on the diagnostic class. The same biomarker may exhibit different distributions, variability, and diagnostic relevance in the Healthy and Cancer groups. Since most feature evaluation methods estimate feature importance according to the statistical distribution of the observed data, constructing a single

global ranking inevitably averages these class-specific characteristics. Consequently, biomarkers that are highly informative within one diagnostic class may receive lower importance scores when evaluated using the combined dataset, which may subsequently influence the assessment of individual patient records.

To investigate this hypothesis, a preliminary descriptive statistical analysis of the Coimbra Breast Cancer Dataset was performed separately for the Healthy and Cancer groups. For each biomarker, the mean (Mean), standard deviation (SD), coefficient of variation (CV), and the statistical significance of differences between the two diagnostic classes were calculated. The results are presented in Table 1.

Table 1 – Descriptive statistical characteristics of biomarkers Coimbra Breast Cancer Dataset

Biomarker	**Healthy (Mean±SD)**	**Cancer (Mean±SD)**	**CV Healthy**	**CV Cancer**	**p-value**
Age	58.08 ± 18.96	56.67 ± 13.49	32.6%	23.8%	0.479
BMI	28.32 ± 5.43	26.99 ± 4.62	19.2%	17.1%	0.202
Glucose	88.23 ± 10.19	105.56 ± 26.56	11.6%	25.2%	<0.001
Insulin	6.93 ± 4.86	12.51 ± 12.32	70.1%	98.4%	0.027
HOMA	1.55 ± 1.22	3.62 ± 4.59	78.5%	126.6%	0.003
Leptin	26.64 ± 19.34	26.60 ± 19.21	72.6%	72.2%	0.949
Adiponectin	10.33 ± 7.63	10.06 ± 6.19	73.9%	61.5%	0.766
Resistin	11.62 ± 11.45	17.25 ± 12.64	98.6%	73.2%	0.002
MCP-1	499.73 ± 292.24	563.02 ± 384.00	58.5%	68.2%	0.504

Source: compiled by the author based on a preliminary descriptive statistical analysis of the Coimbra Breast Cancer Dataset (CBCD).

where Mean denotes the arithmetic mean, SD is the standard deviation, CV is the coefficient of variation, and p-value represents the statistical significance of differences between the diagnostic groups.

To facilitate a visual comparison of biomarker distributions, boxplots were constructed separately for the Healthy and Cancer classes for each of the nine biomarkers (Fig. 1).

The obtained results indicate that the statistical characteristics of individual biomarkers indeed depend on the diagnostic class. The most pronounced differences were observed for Glucose, Insulin, and HOMA, which exhibited substantially greater variability among breast cancer patients than among healthy individuals. In contrast, the Resistin biomarker demonstrated a higher coefficient of variation in the Healthy group. Furthermore, statistically significant differences between the two diagnostic classes were identified for Glucose, Insulin, HOMA, and Resistin ($p < 0.05$), whereas no significant differences were found for the remaining biomarkers.

Overall, the preliminary statistical analysis confirms that the biomarkers exhibit different distribution patterns within the Healthy and Cancer groups. This finding suggests that their diagnostic relevance may also be class-specific. Consequently, constructing a single global feature ranking may partially obscure these class-dependent characteristics, whereas generating independent rankings for the Healthy and Cancer classes has the potential to provide a more representative assessment of the contribution of individual patient records and improve the identification of potentially noisy observations. The validity of this hypothesis is investigated in the following experimental study.

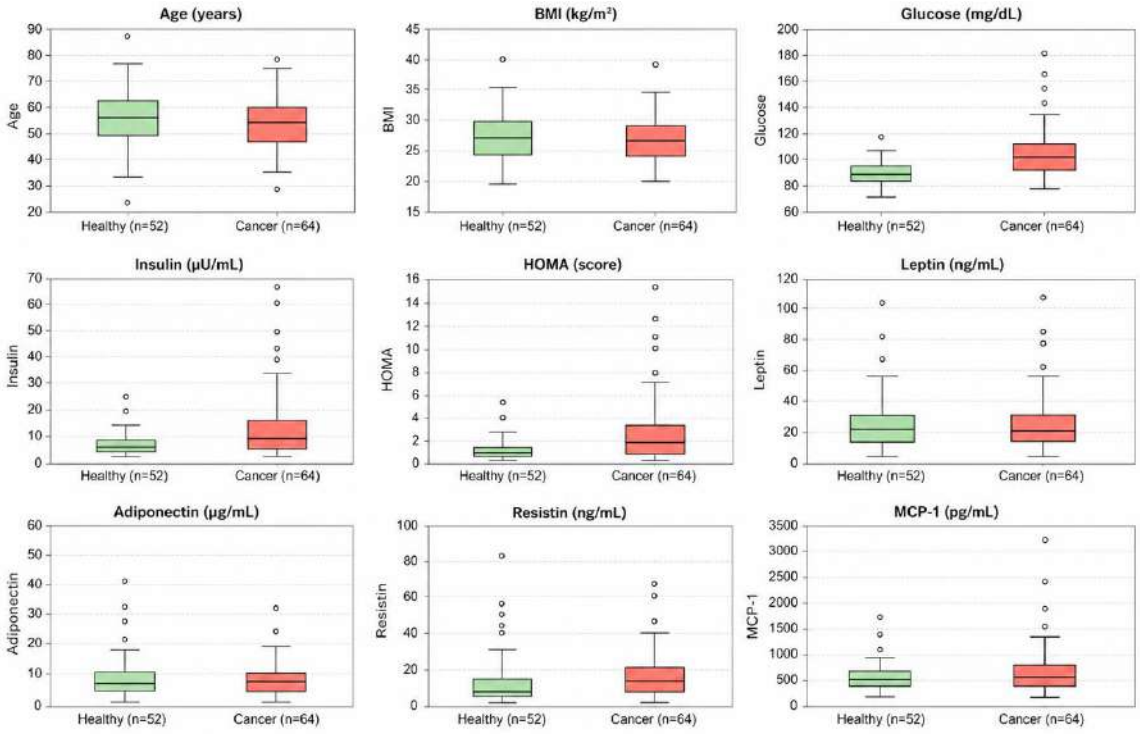


Fig. 1. Boxplot of the distribution of biomarker values for the classes healthy and patients

Source: author's development.

Experimental Dataset.

The experimental study was conducted using the Coimbra Breast Cancer Dataset (CBCD), which is publicly available in the UCI Machine Learning Repository [15]. The dataset was compiled from clinical examinations of patients and includes anthropometric measurements together with routine blood biochemical biomarkers that can be used for the primary diagnosis of breast cancer [14].

The dataset comprises 116 patient records, including 52 Healthy and 64 Cancer cases. Each record contains nine numerical biomarkers and one class attribute indicating the patient's diagnostic category. The structure of the dataset as displayed in the WEKA environment is presented in Fig. 2.

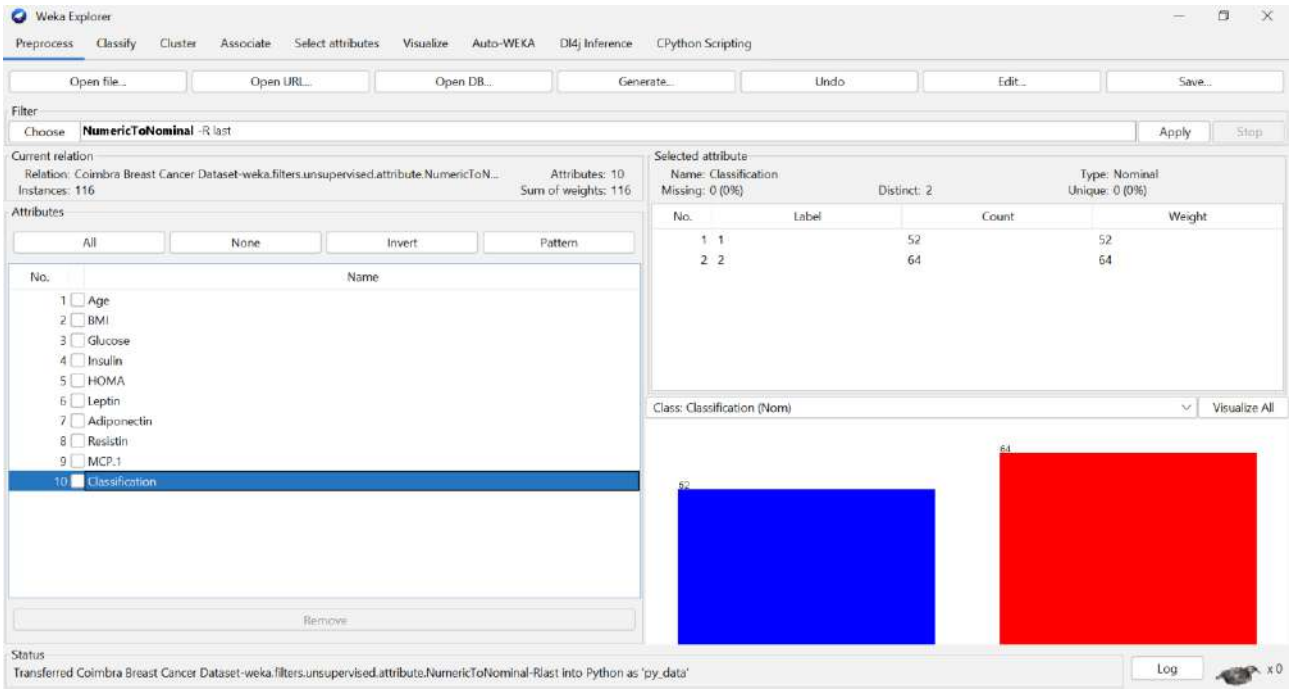


Fig. 2. The CBCD displayed in the WEKA environment.

The dataset includes the following attributes: Age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, and MCP-1. All attributes are numerical and represent either anthropometric characteristics or blood biochemical biomarkers. The target variable is the Classification attribute, which contains two possible values: Healthy and Cancer.

One of the main characteristics of the CBCD is the combination of a relatively small number of observations with heterogeneous clinical variables. This makes the dataset particularly suitable for investigating feature evaluation methods, assessing dataset quality, and identifying potentially noisy records. Furthermore, because all predictor attributes are numerical, most feature evaluation algorithms implemented in the WEKA environment can be applied directly without requiring additional data preprocessing or attribute type conversion.

Before conducting the experiments, the dataset was imported into the WEKA environment without modifying its original structure. All subsequent analyses

were performed both on the complete dataset and separately on the Healthy and Cancer subsets, enabling a comparison between the conventional global feature ranking and the proposed class-specific ranking approach.

Baseline: Global Biomarker Ranking.

In our previous study, biomarker relevance was evaluated using the complete Coimbra Breast Cancer Dataset without considering the diagnostic class of individual patient records. To accomplish this, eight independent feature evaluation methods were applied, each producing an individual ranking of the nine biomarkers. The resulting rankings were subsequently combined into a single consensus ranking (Table 2), which served as the basis for assessing the contribution of each patient record to the overall dataset structure [12].

Table 2. Consensus ranking of biomarkers obtained from eight feature evaluation methods [12].

Rank	Biomarker	Sum of Ranks	Score (Averege Rank)
1	Glucose	10	1.250
2	Age	24	3.000
3	HOMA	28	3.500
4	Resistin	31	3.875
5	BMI	43	5.375
6	Insulin	49	6.125
7	Leptin	52	6.500
8	Adiponectin	58	7.250
9	MCP.1	65	8.125

The Score represents the average rank of a biomarker calculated from the results of eight independent feature evaluation methods. A lower Score indicates a higher level of consensus regarding the diagnostic relevance of the corresponding biomarker.

Proposed Class-Specific Ranking.

Unlike the global approach, in which biomarker rankings are constructed using the complete dataset, the proposed method evaluates feature relevance

independently for each diagnostic class. To achieve this, the original dataset is divided into two subsets: Healthy and Cancer. A separate consensus biomarker ranking is then constructed for each subset using the same eight feature evaluation methods employed in the baseline approach.

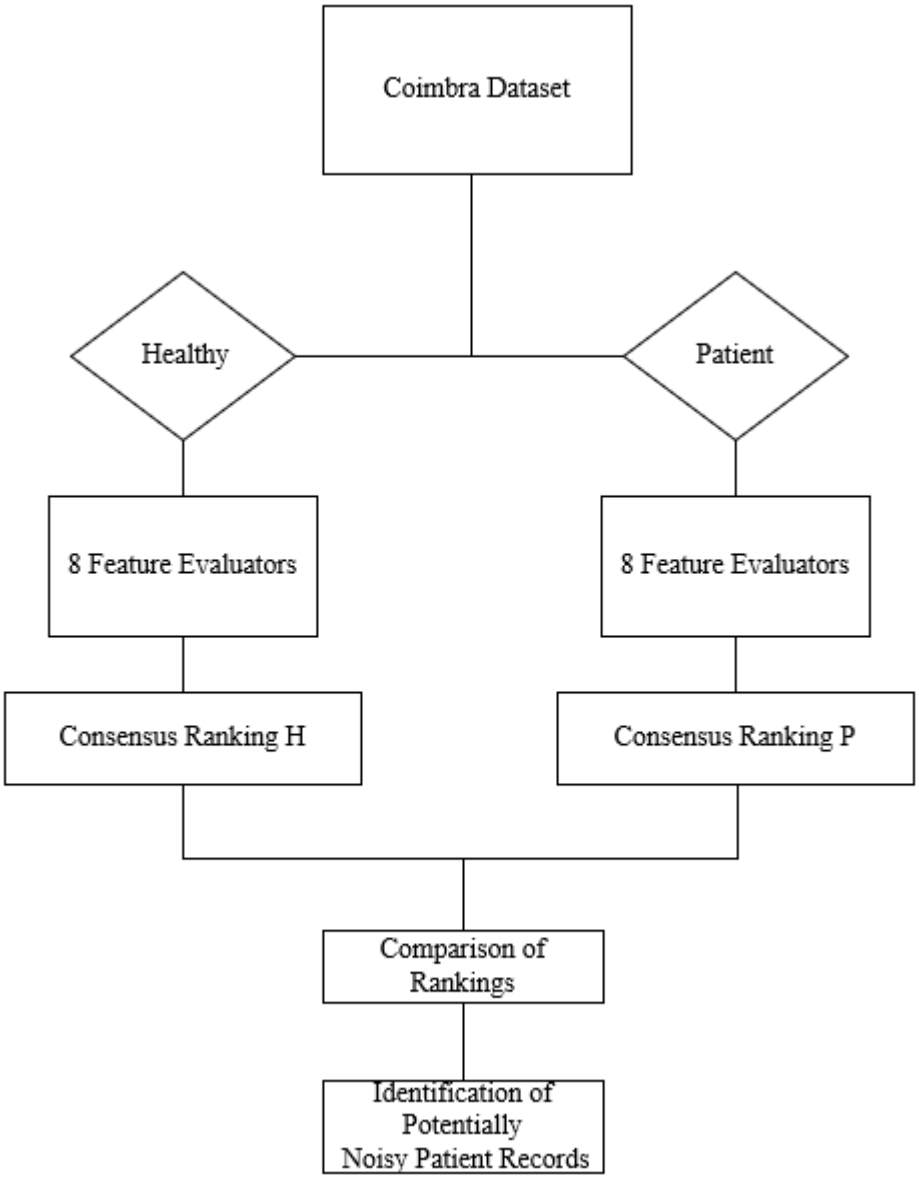


Fig. 3 Flowchart of the proposed class-specific biomarker ranking algorithm

Detailed Algorithm.

The proposed method extends the global biomarker ranking approach described in the previous section. The key difference is that feature relevance is evaluated independently for each diagnostic class rather than for the complete

dataset. As a result, separate consensus rankings are constructed for the Healthy and Cancer classes, providing a more representative characterization of the internal structure of each patient group.

In the first stage, the original dataset D is partitioned into two disjoint subsets according to the value of the class attribute:

$$D = DH \cup DC,\ DH \cap DC = \emptyset, \quad (1)$$

where DH contains the records of healthy individuals, and DC ontains the records of patients with a confirmed diagnosis of breast cancer.

For each subset, the same set of eight feature evaluators used in the baseline method is applied independently: Correlation Attribute Evaluation, Gain Ratio, Information Gain, OneR, ReliefF, SVM Attribute Evaluation, Symmetrical Uncertainty, and Classifier Attribute Evaluation [13]. Each evaluator generates its own ranking of the nine biomarkers based on their diagnostic informativeness within a specific class.

Let $R_{ij}^{(H)}$ denote the rank of the j-th biomarker assigned by the i-th feature evaluator for the Healthy subset, and let $R_{ij}^{(P)}$ – denote the corresponding rank for the Cancer subset. The consensus rank of each biomarker is then calculated as the arithmetic mean of the ranks assigned by all feature evaluators:

$$\mathrm{AR_j} = \frac{1}{m}\sum_{i=1}^{m} R_{(ij)}, \quad (2)$$

where m=8 is the number of feature evaluation methods.

As a result of this procedure, two independent consensus rankings are obtained:

$$\mathrm{AR^{(H)}} = \{\mathrm{AR_1^{(H)}}, \mathrm{AR_2^{(H)}}, \dots, \mathrm{AR_9^{(H)}}\}, \quad (3)$$

$$\mathrm{AR^{(P)}} = \{\mathrm{AR_1^{(P)}}, \mathrm{AR_2^{(P)}}, \dots, \mathrm{AR_9^{(P)}}\}. \quad (4)$$

The resulting rankings characterize the relative importance of each biomarker within the Healthy and Cancer classes independently. Unlike the conventional global ranking approach, the proposed method does not average the

statistical characteristics of the two diagnostic groups. Instead, it preserves their class-specific properties when estimating biomarker relevance.

The next stage involves evaluating individual patient records using the corresponding consensus rankings. For each record, an integrated score is calculated based on the consensus ranking associated with its diagnostic class. Consequently, records belonging to the Healthy class are evaluated using $AR^{(H)}$, whereas records belonging to the Cancer class are evaluated using $AR^{(P)}$. The resulting scores are subsequently used to rank patient records and identify potentially noisy observations.

The procedure for calculating the integrated record score is identical to that described in our previous study. The only difference is that each patient record is evaluated using the consensus ranking corresponding to its diagnostic class.

Experimental Procedure.

Experimental Procedure. To evaluate the proposed approach, the original Coimbra Breast Cancer Dataset was divided into two independent subsets according to the value of the class attribute. The first subset contained only records belonging to the Healthy class, whereas the second subset contained only records corresponding to patients with a confirmed diagnosis of breast cancer (Cancer). This partitioning made it possible to evaluate biomarker relevance independently within each diagnostic group.

The dataset partitioning was performed in the WEKA environment using the standard RemoveWithValues filter. To generate the Healthy subset, the following configuration was applied: *RemoveWithValues -S 0.0 -C last -L2.*

In this configuration, the parameter *-C last* specifies that filtering is performed according to the last attribute of the dataset, which contains the class labels, whereas

-L2 indicates the class value to be removed. After applying the filter and saving the result, a dataset containing only records from the Healthy class is obtained.

Similarly, the Cancer subset was generated using the filter: *RemoveWithValues -S 0.0 -C last -L1*. In this case, records belonging to the Healthy class are removed, resulting in a dataset containing only Cancer records.

After both subsets had been created, each of them was analyzed independently using the same eight feature evaluation methods employed in the baseline approach. For every evaluator, an individual ranking of the nine biomarkers was obtained. The consensus ranking for each diagnostic class was then calculated as the average rank across all evaluators. The resulting class-specific consensus rankings are presented in Table 3.

Table 3. Comparison of consensus biomarker rankings obtained using the global and class-specific approaches

Biomarker	**Global**	**Healthy**	**Cancer**
Age	3.000	9.000	9.000
BMI	5.375	4.125	4.000
Glucose	1.250	4.500	4.250
Insulin	6.125	2.625	2.500
HOMA	3.500	4.625	4.500
Leptin	6.500	6.500	6.125
Adiponectin	7.250	6.875	7.000
Resistin	3.875	4.875	5.750
MCP-1	8.125	1.875	1.750

Comparison of the global and class-specific rankings revealed that separating the dataset according to the diagnostic classes had only a minor effect on the relative ordering of biomarkers within the Healthy and Cancer subsets. The consensus rankings obtained for the two diagnostic classes were highly consistent with each other, indicating that the relative importance of the biomarkers remained stable within each class. At the same time, both class-specific rankings differed

noticeably from the global ranking, suggesting that the conventional approach partially averages the statistical characteristics of the two diagnostic groups.

At the same time, both class-specific rankings differed substantially from the global ranking. In particular, the biomarker MCP-1 moved from the last position in the global ranking to the first position in both class-specific rankings, whereas Age, which ranked second in the global analysis, became the least informative biomarker after partitioning the dataset. Similarly, Insulin improved from the sixth to the second position, while Glucose lost its status as the highest-ranked biomarker. These changes indicate that the importance of several biomarkers is strongly influenced by whether feature relevance is estimated using the complete dataset or within individual diagnostic classes. Consequently, the global ranking appears to average class-specific statistical characteristics, whereas the proposed approach preserves the relative importance of biomarkers within each diagnostic group.

To identify statistical outliers, the InterquartileRange filter with the configuration *–R first-last –O 3.0 –E 6.0* was applied independently to each dataset. The filter identified five outlier records in each dataset.

Table 4. Comparison of potentially noisy records identified in the Global, Healthy, and Cancer datasets.

Dataset	**Instances**	**Outliers**	**Extreme values**	**Non-extreme outliers**
Global	116	12	3	9
Healthy	52	5	1	4
Cancer	64	5	1	4

The analysis of the experimental results showed that applying the proposed class-specific ranking approach identified five potentially noisy records within each diagnostic group. However, only one of these records in each class was also classified as a statistical outlier according to the Interquartile Range (IQR)

criterion. The remaining potentially noisy records did not exhibit extreme values for any individual biomarker.

These findings suggest that the proposed approach is capable of identifying not only conventional statistical outliers but also records characterized by unusual combinations of biomarker values. Such records may not be detected when each biomarker is analyzed independently because their individual measurements remain within the expected range. Nevertheless, the combined biomarker profile differs noticeably from that of the majority of patients belonging to the same diagnostic class.

Conclusions

This paper investigated the influence of class-specific feature ranking on the preliminary analysis of the Coimbra Breast Cancer Dataset. Unlike the conventional approach, where feature relevance is estimated using the complete dataset, the proposed method constructs independent consensus rankings for the Healthy and Cancer classes by combining the results of eight feature evaluators.

The experimental results confirmed that the statistical properties of several biomarkers differ between diagnostic classes. Descriptive statistical analysis and boxplot visualization demonstrated differences in the variability and distribution of biomarker values, providing motivation for investigating class-specific feature ranking.

A comparison of the obtained rankings showed that separating the dataset by class changes the relative importance of several biomarkers. However, the Healthy and Cancer subsets produced similar consensus rankings, whereas both differed from the global ranking obtained from the complete dataset. This observation indicates that the global ranking partially reflects the averaging effect caused by combining heterogeneous patient groups.

The proposed ranking strategy was further applied to the identification of potentially noisy patient records. The experiment detected five potentially atypical

records in each diagnostic class. At the same time, only one detected record in each subset corresponded to an extreme statistical value according to the interquartile range criterion. This result suggests that most potentially noisy records are characterized not by individual extreme biomarker values but by unusual combinations of otherwise normal measurements.

Therefore, the conducted study indicates that class-specific consensus ranking should be regarded primarily as an additional exploratory data analysis tool rather than as a replacement for the conventional global ranking approach. It provides additional information about the internal structure of diagnostic classes and may support more informed preprocessing of medical datasets before applying machine learning methods.

Future research will focus on evaluating the proposed approach on larger medical datasets and investigating its influence on classification performance after removing or reweighting the identified potentially noisy records.

References

1) Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

2) Austria, Y. D., Goh, M. L., Sta. Maria Jr., L., Lalata, J.-A., Goh, J. E., & Vicente, H. (2019). Comparison of Machine Learning Algorithms in Breast Cancer Prediction Using the Coimbra Dataset. *International journal of simulation: systems, science & technology*. https://doi.org/10.5013/ijssst.a.20.s2.23

3) Yue, J., Zhao, N., & Liu, L. (2020). Prediction and Monitoring Method for Breast Cancer: A Case Study for Data from the University Hospital Centre of Coimbra. *Cancer Management and Research, Volume 12*, 1887–1893. https://doi.org/10.2147/cmar.s242027

4) Alfian, G., Syafrudin, M., Fahrurrozi, I., Fitriyani, N. L., Atmaji, F. T. D., Widodo, T., Bahiyah, N., Benes, F., & Rhee, J. (2022). Predicting Breast Cancer from Risk Factors Using SVM and Extra-Trees-Based Feature Selection Method. *Computers*, *11*(9), 136. https://doi.org/10.3390/computers11090136

5) Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

6) Beam, A. L., & Kohane, I. S. (2018). Big Data and Machine Learning in Health Care. *JAMA*, *319*(13), 1317. https://doi.org/10.1001/jama.2017.18391

7) García, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining*. Springer International Publishing. https://doi.org/10.1007/978-3-319-10247-4

8) Pei, J., Han, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier Science & Technology Books.

9) Hodge, V., & Austin, J. (2004). A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, *22*(2), 85–126. https://doi.org/10.1023/b:aire.0000045502.10941.a9

10) Zhu, X., & Wu, X. (2004). Class Noise vs. Attribute Noise: A Quantitative Study. *Artificial Intelligence Review*, *22*(3), 177–210. https://doi.org/10.1007/s10462-004-0751-8

11) Frenay, B., & Verleysen, M. (2014). Classification in the Presence of Label Noise: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(5), 845–869. https://doi.org/10.1109/tnnls.2013.2292894

12) Chuiko, G., & Honcharov, D. (2025). Breast Cancer Dataset from Coimbra: Pre-Ratings of Its Value to Machine Learning and Diagnosis. *Digital Technologies Research and Applications*, *4*(2), 182–193. https://doi.org/10.54963/dtra.v4i2.1348

13) Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517. https://doi.org/10.1093/bioinformatics/btm344

14) Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. (2018). Using Resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer*, *18*(1). https://doi.org/10.1186/s12885-017-3877-1

15) Patrcio, M.; Pereira, J.; Crisstomo, J.; et al., 2018. Breast Cancer Coimbra. UCI Machine Learning Repository. Available from: https://archive.ics.uci.edu/dataset/451/breast+cancer+coimbra