Enhancing Breast Cancer Classification Using Ensemble Techniques and Feature Selection Algorithms

1st Nadia Azri
dept. of Computer Science
Mohamed Kheider University
Biskra, Algeria
nadia.azri@univ-biskra.dz

2nd Sadek Labib Terrissa
dept. of Computer Science
Mohamed Kheider University
Biskra, Algeria
terrissa@univ-biskra.dz

3rd Fadila Madaci
Faculty of Medicine
University of Algiers 1
Algiers, Algeria
fadila.madaci@gmail.com

4th Nourddine Zerhouni
Institut FEMTO-ST, UMR CNRS 6174 - AS2M
UFC,ENSMM
Besancon, France
noureddine.zerhouni@ens2m.fr

***Abstract*—Over the past ten years, breast cancer (BC) has become a major health concern due to its high fatality rate. It remains the most common cancer in women to be diagnosed. The urgency to address this disease has prompted extensive research efforts toward developing effective classification models and diagnostic tools. With an emphasis on the study of classification outcomes utilizing the well-known WDBC dataset, we examine the efficacy of several feature selection (FS) methods and classifiers within the context of the majority voting ensemble technique in this work. We explore a diverse set of FS algorithms, including PCA, Relief, RF, LASSO, GA, and CFS, combined with various classifiers such as SVM, MLP, and RF. Our study reveals that the PCA FS algorithm consistently achieves high accuracy results across multiple classifiers, notably 99.12% accuracy when paired with RF and LR. This consistency highlights its effectiveness in capturing informative features and its compatibility with the majority voting ensemble technique, which also achieved an accuracy of 99.12%. These findings contribute to advancing our understanding of ensemble techniques and FS algorithms, particularly in the context of the WDBC dataset, providing valuable insights for future development of robust classification models for similar datasets.**

***Keywords*—Computer-Aided Diagnosis, Breast Cancer, Machine Learning, Deep Learning, Feature Selection, Majority Voting Ensemble, WDBC dataset.**

I. Introduction

Breast cancer (BC) is the most frequent cancer among people, accounting for about 2.3 million cases yearly. It is a widespread global health concern. Breast cancer is the first or second most common cause of cancer-related deaths in women in 95% of the world's countries. Nonetheless, there is a notable difference in survival rates between and within nations, with low- and middle-income nations accounting for about 80% of breast cancer fatalities. [1]. Since early detection raises the chance of a successful intervention, it is essential for enhancing treatment results. The diverse nature of breast cancer types further complicates the classification process, making it a challenging task. Human classification is not always accurate, thus the most effective therapy techniques can only be implemented when breast cancer kinds are accurately classified [2].

Classifier systems have a great deal of promise and utility in the diagnosis of medical conditions. By leveraging artificial intelligence techniques, these systems aid in mitigating potential errors resulting from inexperienced experts, while also facilitating a more efficient and detailed analysis of medical data. It is important to note that expert decision-making remains paramount in the diagnostic process, but the integration of artificial intelligence techniques for classification can significantly augment the capabilities of these experts.

The application of classifier systems in medical diagnosis has consistently demonstrated its value, offering numerous benefits. It not only helps mitigate errors arising from inexperienced experts but also provides detailed analyses of medical data within shorter timeframes. Extensive research and experimentation have been conducted on medical datasets, particularly those related to breast cancer, showcasing consistently high rates of classification accuracy.

Our main goal in this study work is to investigate the potential of various feature selection methods and classifier systems in breast cancer classification. We seek to evaluate their performance and assess their effectiveness in accurately classifying breast cancer using the WDBC datasets. Our goal is to add to the current body of knowledge in the creation of reliable and accurate breast cancer classification models by carrying out thorough experiments and analyses. To improve the overall accuracy and dependability of the classification findings, we also look into the applicability of the majority voting technique, which combines the predictions of many classifiers. Through our research, we aspire to advance the field of breast cancer classification and provide valuable in-

sights for improved diagnosis and treatment strategies.

This paper follows a structured format to present our research on breast

cancer classification. An summary of the relevant field research is given in Section 2. Section 3 provides a detailed explanation of our suggested strategy, including the methods and techniques applied. In Section 4, the acquired results are showcased and examined. Section 5 brings the investigation to a close with a summary of the results and a discussion of potential directions.

II. Related Work

The field of breast cancer diagnostics has witnessed notable advancements in the use of machine learning and data mining approaches to increase the accuracy of classification and facilitate early detection. This section discusses notable contributions from various studies that have advanced the field of breast cancer classification:

Nguyen et al. [3] suggested a machine learning method for breast cancer diagnosis that makes use of the feature selection strategy and random forest classifier. Using the Wisconsin Breast Cancer and Prognostic Dataset, they evaluated the system and obtained an amazing average classification accuracy of about 99.8%.

Liu et al. [4] conducted a comprehensive study on FS methods for breast cancer mass classification. Comparing techniques such as F-Score, Relief, and mRMR, they demonstrated the efficacy of these methods with fewer selected features, achieving an accuracy of up to 96.15%.

Aličković et al. [5] utilized a Genetic Algorithm (GA) to classify breast cancer data. They attained a remarkable accuracy of 99.48% by employing the Rotation Forest classifier.

Wang et al. [6] presented an enhanced approach for breast cancer classification known as Improved RF-based Rule Extraction (IRFRE). Their evaluation on benchmark datasets, including WOBC, WDBC, and SEER BC dataset, showcased promising results.

Abdar et al. [7] introduced a data mining technique for breast cancer prediction, utilizing Support Vector Machine (SVM) and Artificial Neural Network (ANN). Their study focused on the analysis of the Wisconsin Breast Cancer Dataset (WBCD).

Alharbi et al. [8] developed an automated Computer-Aided Diagnosis (CAD) system aimed at early breast cancer detection. They integrated a genetic-fuzzy algorithm into the Saudi breast cancer diagnosis database, offering valuable support to medical practitioners.

Jafari-Marand et al. [9] introduced the Life-Sensitive Self-Organizing Error Drive (LSSOED) Artificial Neural Network (ANN) for diagnosing breast cancer, employing the WBCD and WOBC datasets. This approach enhanced decision-making quality by minimizing misclassification costs.

Sahu et al. [10] introduced a hybrid approach that integrated Principal Component Analysis (PCA) with Artificial Neural Network (ANN) for the classification of breast cancer tumors. This method demonstrated superior performance in terms of accuracy, sensitivity, and F-measure compared to other algorithms.

Gopal et al. [11] integrated machine learning techniques with IoT technology for BC early diagnosis. Their approach, utilizing PCA for feature extraction and MLP, Logistic Regression, and Random Forest for tumor classification, showed promising results on the WBCD dataset.

Jabbar [12] endeavored to construct a decision support system utilizing an ensemble model incorporating Bayesian Networks (BN) and Radial Basis Function (RBF) for the classification of breast cancer data. The model demonstrated outstanding performance with an accuracy of 97.42% on the WBCD dataset.

Abdur Rasool et al. [13] suggested exploratory data techniques employing SVM, LR, KNN, and ensemble models for the detection of breast cancer. Notably, their approach achieved an accuracy of 99.3% with SVM, contributing to intelligent diagnosis.

Chaurasi et al. [14] conducted a study comparing six ML algorithms on the WDBC dataset for breast cancer prediction. They found that feature selection improves accuracy, and integrated models and a stacking classifier (Voting Classifier) enhance accuracy, with all algorithms performing best, exceeding 90% accuracy on a data subset.

Lahoura et al. [15] introduced a breast cancer diagnosis framework based on the cloud, employing Extreme Learning Machine (ELM) as a rapid and straightforward classifier. The cloud-based ELM model delivered reliable and efficient services, demonstrating elevated accuracy on the WBCD dataset and surpassing alternative techniques.

Haq et al. [16] introduced an approach that incorporated supervised (Relief algorithm) and unsupervised (Autoencoder, PCA algorithms) techniques for feature selection in the detection of breast cancer. Their method, employing a support vector machine classifier, attained a notable accuracy of 99.91%, surpassing baseline methods.

Badr et al. [17] presented three key contributions: 1) Improving the performance of Support Vector Machine (SVM) through the integration of Grey Wolf Optimizer (GWO) for breast cancer diagnosis, achieving an accuracy of 98.60% on the WDBC dataset. 2) Introducing three efficient scaling techniques, resulting in a high accuracy of 99.30% with rapid convergence. 3) Implementing a parallel technique that demonstrated a speedup of 3.9 on four CPU cores, achieving an accuracy of 93.26% on Electronic Health Records (EHR) dataset compared to 82.05% for SVM.

Ed-daoud et al. [18] introduced a two-stage methodology for breast cancer classification. In the initial stage, Association Rules (AR) are employed to eliminate irrelevant features, thereby reducing the dimension of the feature space. In the subsequent stage, diverse classifiers are applied to distinguish tumors. Utilizing Support Vector Machine (SVM) in conjunction with AR, the approach achieved a maximum classification accuracy of 98.00% for eight attributes and 96.14% for four attributes. This approach not only accelerates the training

process but also enhances accuracy, rendering it well-suited for swift and automated classification systems.

These related works have significantly contributed to the advancement of breast cancer classification and diagnosis. They have paved the way for further research and the development of more effective and accurate classification models. Expanding on the established knowledge base, our investigation delves into the effectiveness of ensemble techniques and feature selection algorithms in improving the precision and dependability of breast cancer classification. Drawing on insights gleaned from earlier research, our goal is to provide meaningful contributions to the domain of breast cancer diagnosis.

III. Methodology

A. Dataset

This study uses the Breast Cancer Wisconsin Diagnostic dataset from the UCI Machine Learning Repository, which was created by Wolberg, Street, and Mangasarian at the University of Wisconsin [19]. This dataset is widely used in the fields of machine learning and cancer research. Each of the 569 occurrences in this dataset has 30 numerical attributes that were obtained from digital images of breast mass fine needle aspirates. These characteristics, which included clump thickness, homogeneity of cell size and shape, and others, were calculated from the digital images of the FNA samples. With 212 samples categorized as malignant and 357 as benign, the dataset provides a balanced class distribution. Its broad application in classification tasks and the assessment of machine learning algorithms for breast cancer diagnosis accounts for its significance.

B. Data Pre-processing

The role of this step in enhancing result accuracy is crucial. In this sub-sections, we will outline the techniques employed to enhance the dataset's quality.

The dataset is pre-processed to exclude instances where an attribute's value is missing. Missing data cause a number of issues. First, the statistical power—the likelihood that the test would reject the null hypothesis when it is false—decreases when there are no data. Second, faulty parameter estimation may result from lost data. Third, it might make the samples less representative. Fourth, it can make the study's analysis more challenging. Each of these errors has the potential to undermine the reliability of the trials and produce false results [20].

In the diagnostic data set for breast cancer, the attributes "diagnosis" are substituted for B's 0 and M's 1. When the units of measurement used in the data gathering are different, we must normalize the data. The process of standardizing involves rescaling one or more attributes to have a mean of 0 and a standard deviation of 1 [14].

For this study, we employed the z-score standardization method. The following formula can be used to get the Z-score of a data point given the distribution's mean (μ) and standard deviation (σ):

$$Z = \frac{(X - \mu)}{\sigma} \tag{1}$$

Where:

- Z is the Z-score
- X is the value being standardized
- μ is the mean of the distribution
- σ is the standard deviation of the distribution

C. Features Selection Techniques

We often prioritize specific characteristics that hold the greatest significance for making accurate predictions or achieving desired outcomes. To accomplish this, we employ a feature selection technique, which is a procedure utilized to identify the most relevant output variable [21]. In our study, we employed six distinct feature selection techniques and subsequently compared their respective results with each other.

1) Principal Component Analysis (PCA): is an unsupervised method employed to decrease the feature count in high-dimensional data. Its objective is to convert the data into a lower-dimensional representation that retains the majority of the data's variance while minimizing reconstruction error. PCA accomplishes this by creating new variables called principal components, which are linear combinations of the initial features. These components are intentionally orthogonal, signifying the absence of redundant information. Collectively, the principal components establish an orthogonal basis for the data space, presenting a quantitatively rigorous method to streamline the data representation [22].

2) The Relief algorithm: pioneered by Kira and Rendell in 1992, is a feature selection method employing a filter-based approach. Renowned for its heightened sensitivity to feature interactions [23], it functions as an independent evaluation filter. The algorithm calculates a surrogate statistic for each feature, serving as a measure of the "quality" or "relevance" of the feature concerning the target concept, such as forecasting the value of an endpoint [24].

3) Random Forest (RF): is a widely employed algorithm for measuring feature importance. It serves the purpose of both feature selection and exhibits excellent accuracy and classification robustness [25]. With its rapid training speed, high precision, and the absence of complex parameter adjustments, the RF model offers clear advantages in feature selection [26].

4) Least Absolute Shrinkage and Selection Operator (LASSO): achieves feature selection and shrinkage by adjusting the absolute value of coefficient functions. This leads to certain coefficients being reduced to zero, effectively removing corresponding features from the subset [27]. Additionally, features with negative coefficients can also be eliminated. LASSO performs remarkably well when dealing with features that have small coefficients, retaining them in the selected feature subset. On the other hand, features with large coefficient values are more likely to be included in the chosen subset. LASSO proves valuable in identifying unnecessary features [28].

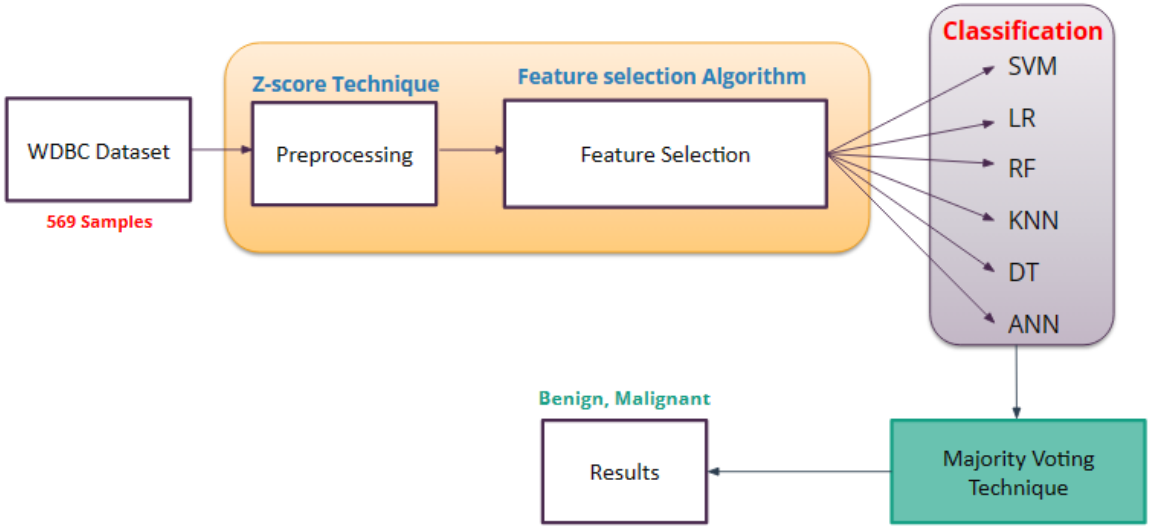

Fig. 1: The workflow of our approach.

5) Genetic Algorithms (GA): are heuristic search and optimization techniques that draw inspiration from natural evolution [29]. They mimic the principles of natural evolution in the realm of computers. GAs operate on the principle of "survival of the fittest," emulating genetic inheritance and the Darwinian struggle for survival. These algorithms are versatile optimization methods that incorporate a probabilistic element, enabling them to explore complex and poorly understood solution spaces [30].

6) Correlation-based Feature Selection (CFS): is a filter algorithm that selects features based on a heuristic correlation function [31]. The objective of this function is to identify feature subgroups that exhibit strong correlations with the class variable while being uncorrelated with each other. Inessential features are disregarded as they lack a relationship with the class, while redundant features are excluded since they exhibit high correlations with the remaining features. A feature's capacity to predict classes in areas of the instance space that other features haven't yet predicted determines whether or not it gets included [32].

D. Model Description

Our approach combines both traditional machine learning (ML) classifiers and deep learning (DL) classifiers to enhance the accuracy and robustness of our classification results. We employed a total of six classifiers, including ML classifiers such as SVM, RF, LR, KNN, DT, as well as DL classifiers like ANN. Each of these classifiers underwent training on the dataset, contributing its predictions to the final decision-making process (see Fig. 1).

The Artificial Neural Network (ANN) utilized in our study is a multi-layer model comprising three hidden layers. Each hidden layer is composed of 64 units and employs the 'relu' activation function, which introduces non-linearity into the network. To mitigate overfitting, dropout layers with a dropout rate of 0.5 are incorporated between the hidden layers. The output layer of the ANN consists of a single unit and utilizes the 'sigmoid' activation function, which allows for binary classification. This architecture and configuration of the ANN enable it to learn complex patterns and make accurate predictions on the classification task at hand.

Furthermore, our approach utilizes the majority voting ensemble technique as the final decision-making algorithm. Majority voting is a widely used ensemble technique that combines the predictions of multiple classifiers in a parallel fashion. The predictions from the ML and DL classifiers are aggregated through the majority voting process to generate the final solution. This ensemble approach leverages the collective wisdom of the diverse classifiers, leading to improved accuracy and robustness in the classification results.

In addition to classifier selection, we conducted an extensive evaluation of feature selection algorithms. We tested six different feature selection algorithms to determine the most effective one for the WDBC dataset. The objective was to identify the algorithm that produced the most optimal feature subset for our classification task. By considering both the diversity of classifiers and the effectiveness of feature selection, our approach aims to achieve superior performance and accuracy in the classification task.

IV. Experimental Results

We split the WDBC dataset using our suggested strategy to assess its efficacy. Eighty percent of the dataset was put aside for training, and the remaining twenty percent was set aside for testing. Because of this partitioning, we were able to train our model on a sizable chunk of the data while keeping a separate set aside for evaluation of its effectiveness and capacity for generalization.

We expanded our research to include the effect of several feature selection (FS) algorithms on each individual classifier,

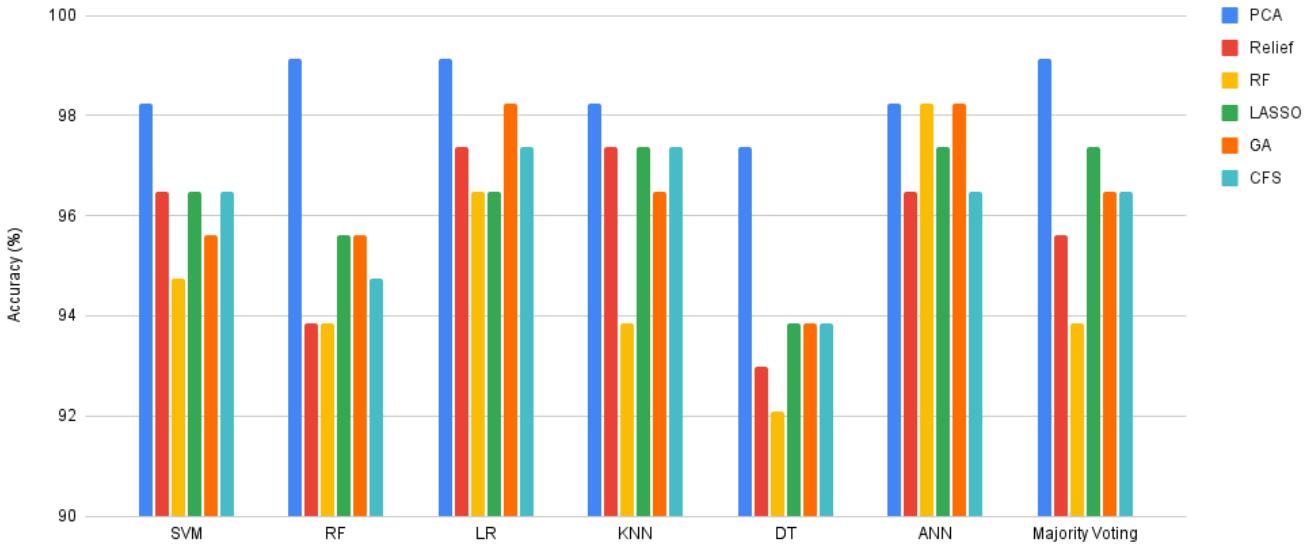

Fig. 2: Comparing the classification results' performance across several FS algorithms.

TABLE I: Comparative analysis of the classification outcomes using various classifiers and feature selection techniques.

	SVM	**RF**	**LR**	**KNN**	**DT**	**ANN**	**Majority Voting**
PCA	98.24%	99.12%	99.12%	98.24%	97.36%	98.24%	99.12%
Relief	96.49%	93.85%	97.36%	97.36%	92.98%	96.49%	95.61%
RF	94.73%	93.85%	96.49%	93.85%	92.10%	98.24%	93.85%
LASSO	96.49%	95.61%	96.49%	97.36%	93.85%	97.36%	97.36%
GA	95.61%	95.61%	98.24%	96.49%	93.85%	98.24%	96.49%
CFS	96.49%	94.73%	97.36%	97.36%	93.85%	96.49%	96.49%

specifically with regard to accuracy, in order to fully assess the efficacy of our suggested approach. We applied many FS algorithms to each classifier used in our method and evaluated the impact of each technique on the classification results' accuracy. Through this examination, we were able to establish which FS method would maximize accuracy in our classification task and evaluate the impact that various FS strategies had on the performance of each classifier. By carrying out this assessment, we wanted to make sure that each classifier and the chosen FS algorithm worked in perfect harmony, improving the accuracy of the final classification results.

As shown in table I, we obtained notable results that showcased the effectiveness of different feature selection (FS) algorithms on each classifier. The SVM classifier yielded the highest accuracy of 98.24% when coupled with the PCA FS algorithm. Similarly, the RF and LR classifiers achieved their best performance with the PCA FS algorithm, both achieving an accuracy of 99.12%. The KNN classifier also demonstrated exceptional accuracy of 98.24% when paired with the PCA FS algorithm. The DT classifier achieved a strong accuracy of 97.36% using the PCA FS algorithm. Lastly, the ANN classifier delivered impressive results with the PCA, RF, and GA FS algorithms, all achieving an accuracy of 98.24%.

These results highlight the potency of the PCA FS algorithm across various classifiers, consistently producing superior accuracy (as shown in Fig 2). Additionally, the RF and GA FS algorithms also showcased their effectiveness when paired with the ANN classifier. This comprehensive evaluation of FS algorithms on each classifier helps identify the optimal combinations that yield the highest accuracy, enabling us to make informed decisions about feature selection for improved classification outcomes.

Furthermore, we observed that the PCA FS algorithm achieved exceptional results when paired with the RF and LR classifiers, both attaining an impressive accuracy of 99.12%. Similarly, the Relief FS algorithm showcased its effectiveness when used with the LR and KNN classifiers, both achieving a notable accuracy of 97.36% (see Fig 2).

Moreover, the RF FS algorithm demonstrated its strength when combined with the ANN classifier, achieving a high accuracy of 98.24%. Likewise, the LASSO FS algorithm yielded remarkable results when paired with the ANN and KNN classifiers, both achieving an accuracy of 97.36%.

Additionally, the GA FS algorithm delivered excellent outcomes when used with the ANN and LR classifiers, both achieving an accuracy of 98.24%. Similarly, the CFS FS algorithm showcased its effectiveness with the LR and KNN classifiers, both attaining an accuracy of 97.36%.

These results highlight the compatibility and performance of specific FS algorithms with different classifiers. It emphasizes the importance of selecting the most suitable FS algorithm for

each classifier to optimize accuracy and enhance the overall performance of the classification task.

Additionally, we examined the results of the majority voting ensemble technique on the SVM, DT, LR, KNN, ANN, and RF classifiers. When utilizing the PCA FS algorithm, the majority voting approach achieved an accuracy of 99.12%, which is consistent with the accuracy achieved by the LR and RF classifiers individually.

Moreover, when employing the Relief FS algorithm, the majority voting ensemble technique yielded an accuracy of 95.61%, surpassing the DT and RF classifiers but falling short of the other classifiers.

Similarly, when utilizing the RF FS algorithm, the majority voting ensemble technique achieved an accuracy of 93.85%, which is lower than the accuracies achieved by the ANN and LR classifiers.

Furthermore, the majority voting technique ensemble utilizing the LASSO FS algorithm resulted in an accuracy of 97.36%, which is identical to the higher accuracies achieved by the KNN and ANN classifiers.

Additionally, the majority voting ensemble technique using the GA FS algorithm achieved an accuracy of 96.49%, which aligns with the accuracy achieved by the KNN classifier and falls short of the accuracies achieved by the ANN and LR classifiers.

Lastly, the majority voting ensemble technique utilizing the CFS algorithm obtained an accuracy identical to the GA FS algorithm. It achieved the same accuracy as the ANN and SVM classifiers, while falling short of the accuracies achieved by the KNN and LR classifiers.

These findings illustrate the effectiveness of the majority voting ensemble technique in combining the predictions of multiple classifiers. The accuracy of the majority voting ensemble results can be greatly impacted by the feature selection algorithm used, with different classifiers and algorithms exhibiting differing accuracies.

V. Conclusion

In conclusion, our study presented a comprehensive evaluation of feature selection (FS) algorithms and the majority voting ensemble technique in the context of classification tasks using different classifiers. The results of our analysis shed light on key findings.

Firstly, the PCA FS algorithm consistently demonstrated high accuracy results across various classifiers, indicating its effectiveness in capturing informative features and its compatibility with the majority voting ensemble technique.

Secondly, the Relief FS algorithm exhibited competitive accuracy, surpassing some classifiers while falling behind others. This highlights its value in feature selection for the majority voting ensemble.

Thirdly, the RF FS algorithm achieved relatively lower accuracy compared to the ANN and LR classifiers, suggesting that it may not be the most suitable choice for the majority voting ensemble.

On the other hand, the LASSO FS algorithm showed exceptional accuracy, matching the performance of the KNN and ANN classifiers, demonstrating its compatibility with the majority voting ensemble technique.

Regarding the GA and CFS FS algorithms, they yielded similar accuracies to each other. While achieving the same accuracy as the ANN and SVM classifiers, they fell short of the accuracies attained by the KNN and LR classifiers.

Overall, our study underscores the significance of selecting an appropriate FS algorithm that complements the majority voting ensemble. The results highlight the potential of combining diverse classifiers to enhance accuracy, while emphasizing the need for careful consideration of FS algorithms to optimize the ensemble's performance. These findings provide valuable insights for practitioners and researchers aiming to leverage ensemble techniques and FS algorithms for classification tasks.

References

[1] https://www.who.int/news/item/03-02-2023-who-launches-new-roadmap-on-breast-cancer, [Online; accessed July 19, 2023].

[2] Dora, L., Agrawal, S., Panda, R., & Abraham, A. (2017). Optimal breast cancer classification using Gauss–Newton representation based algorithm. Expert Systems with Applications, 85, 134-145, doi: 10.1016/j.eswa.2017.05.035.

[3] Nguyen, C., Wang, Y., & Nguyen, H. N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. J. Biomedical Science and Engineering, 6, 551-560.

[4] Liu, X., & Tang, J. (2013). Mass classification in mammograms using selected geometry and texture features, and a new SVM-based feature selection method. IEEE Systems Journal, 8(3), 910-920.

[5] Aličković, E., & Subasi, A. (2017). Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Computing and applications, 28, 753-763.

[6] Wang, S., Wang, Y., Wang, D., Yin, Y., Wang, Y., & Jin, Y. (2020). An improved random forest-based rule extraction method for breast cancer diagnosis. Applied Soft Computing, 86, 105941.

[7] Abdar, M., & Makarenkov, V. (2019). CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer. Measurement, 146, 557-570.

[8] Alharbi, A., & Tchier, F. (2017). Using a genetic-fuzzy algorithm as a computer-aided diagnosis tool on Saudi Arabian breast cancer database. Mathematical biosciences, 286, 39-48.

[9] Jafari-Marandi, R., Davarzani, S., Gharibdousti, M. S., & Smith, B. K. (2018). An optimum ANN-based breast cancer diagnosis: Bridging gaps between ANN learning and decision-making goals. Applied Soft Computing, 72, 108-120.

[10] Sahu, B., Mohanty, S., & Rout, S. (2019). A hybrid approach for breast cancer classification and diagnosis. EAI Endorsed Transactions on Scalable Information Systems, 6(20).

[11] Gopal, V. N., Al-Turjman, F., Kumar, R., Anand, L., & Rajesh, M. (2021). Feature selection and classification in breast cancer prediction using IoT and machine learning. Measurement, 178, 109442.

[12] Jabbar, M. A. (2021). Breast cancer data classification using ensemble machine learning. Engineering and Applied Science Research, 48(1), 65-72.

[13] Rasool, A., Bunterngchit, C., Tiejian, L., Islam, M. R., Qu, Q., & Jiang, Q. (2022). Improved machine learning-based predictive models for breast cancer diagnosis. International journal of environmental research and public health, 19(6), 3211.

[14] Chaurasia, V., & Pal, S. (2020). Applications of machine learning techniques to predict diagnostic breast cancer. SN Computer Science, 1(5), 270.

[15] Lahoura, V., Singh, H., Aggarwal, A., Sharma, B., Mohammed, M. A., Damaševičius, R., ... & Cengiz, K. (2021). Cloud computing-based framework for breast cancer diagnosis using extreme learning machine. Diagnostics, 11(2), 241.

[16] Haq, A. U., Li, J. P., Saboor, A., Khan, J., Wali, S., Ahmad, S., ... & Zhou, W. (2021). Detection of breast cancer through clinical data using supervised and unsupervised feature selection techniques. IEEE Access, 9, 22090-22105.

[17] Badr, E., Almotairi, S., Salam, M. A., & Ahmed, H. (2022). New sequential and parallel support vector machine with grey wolf optimizer for breast cancer diagnosis. Alexandria Engineering Journal, 61(3), 2520-2534.

[18] Ed-daoudy, A., & Maalmi, K. (2020). Breast cancer classification with reduced feature set using association rules and support vector machine. Network Modeling Analysis in Health Informatics and Bioinformatics, 9, 1-10.

[19] 'UCI Machine Learning Repository'. http://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic (accessed May. 24, 2023).

[20] Kang, H. (2013). The prevention and handling of the missing data. Korean journal of anesthesiology, 64(5), 402-406.

[21] Kavitha, R., & Kannan, E. (2016, February). An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining. In 2016 international conference on emerging trends in engineering, technology and science (icetets) (pp. 1-5). IEEE.

[22] Vidal, R., Ma, Y., Sastry, S. S., Vidal, R., Ma, Y., & Sastry, S. S. (2016). Principal component analysis. Generalized principal component analysis, 25-62.

[23] Kira, K., & Rendell, L. A. (1992, July). The feature selection problem: Traditional methods and a new algorithm. In Proceedings of the tenth national conference on Artificial intelligence (pp. 129-134).

[24] Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection: Introduction and review. Journal of biomedical informatics, 85, 189-203.

[25] Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. Pattern recognition letters, 31(14), 2225-2236.

[26] Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS journal of photogrammetry and remote sensing, 67, 93-104.

[27] Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. J. M., Ignatious, E., ... & De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. IEEE Access, 9, 19304-19326.

[28] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1), 267-288.

[29] McCall, J. (2005). Genetic algorithms for modelling and optimisation. Journal of computational and Applied Mathematics, 184(1), 205-222.

[30] Oreski, S., & Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection in credit risk assessment. Expert systems with applications, 41(4), 2052-2064.

[31] Hall, M. A. (1999). Correlation-based feature selection for machine learning (Doctoral dissertation, The University of Waikato).

[32] Singh, S., & Singh, A. K. (2018). Web-spam features selection using CFS-PSO. Procedia computer science, 125, 568-575.