Machine Learning with Applications 22 (2025) 100757

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Matthews correlation coefficient-based feature ranking in recursive ensemble feature selection for high-dimensional and low-sample size data

David Rojas-Velazquez [a,b], Aletta D. Kraneveld [c], Alberto Tonda [d], Alejandro Lopez-Rincon [a,*]

[a] Division of Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Faculty of Science, Universiteitsweg 99, Utrecht, 3508 TB, Utrecht, The Netherlands
[b] Department of Data Science, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Heidelberglaan 100, Utrecht, 3508 GA, The Netherlands
[c] Department of Neuroscience, Faculty of Science, Vrije Universiteit Amsterdam, De Boelelaan 1105, Amsterdam, 1081 HV, The Netherlands
[d] UMR 518 MIA - PS, INRAE, Université Paris-Saclay, Institut des Systèmes Complexes de Paris, Île - de - France (ISC-PIF) - UAR 3611 CNRS, 113 rue Nationale, Paris, 75013, Paris, France

ARTICLE INFO

Dataset link: https://github.com/steppenwolf0/REFS-MCC

Keywords:
Feature selection
Machine learning
Biomarker discovery
Deep learning

ABSTRACT

Identifying reliable biomarkers in omics data is challenging due to the high number of features and limited sample sizes, which often lead to overfitting, biased results, and poor reproducibility. These issues are further complicated by class imbalance, common in medical datasets. To address these challenges, we present MCC-REFS, an improved version of the Recursive Ensemble Feature Selection method. MCC-REFS uses the Matthews Correlation Coefficient (MCC) as a selection criterion, offering a more balanced evaluation of classification performance, especially in imbalanced datasets. Unlike traditional methods that require manual tuning or predefined feature counts, MCC-REFS automatically selects the most informative and compact feature sets using an ensemble of eight machine learning classifiers. We evaluated MCC-REFS on synthetic datasets and several real-world omics datasets, including mRNA expression profiles and multi-label breast cancer data. Compared to existing methods such as REFS, GRACES, DNP, and GCNN, MCC-REFS consistently achieved higher or comparable performance while selecting fewer features. Validation using independent classifiers confirmed the robustness of the selected features. Overall, MCC-REFS provides a scalable, flexible, and reliable approach for feature selection in biomedical research, with strong potential for diagnostic and prognostic applications.

1. Introduction

Advances in sequencing technologies for various types of *omics* data, such as genomics (genes), proteomics (proteins), and metabolomics (metabolites), have enabled deeper and more accurate sampling, thereby facilitating the identification of specific biomarkers for diverse medical applications (Li et al., 2022a). A biomarker can serve as a criterion for diagnosing or distinguishing a disease, monitoring its progression, evaluating its severity, and predicting the response to treatment, among other functions. Thus, there is a growing interest in finding new biomarkers (Ptolemy & Rifai, 2010). The main challenge in discovering new biomarkers is that *omics* data is characterized as high-dimensional and low-sample-size (HDLSS) (Visar et al., 2021). An additional challenge in the field of biomarker discovery is the lack of reproducibility, which can be attributed to several factors in HDLSS datasets, such as the *batch effect* (Goh et al., 2017; Loganathan et al., 2022; Papin et al., 2020).

Despite the availability of data, the high dimensionality of *omics* expression datasets presents several challenges. For example, mRNA expression data can include over 54,675 probesets (Robinson & Speed, 2007); DNA methylation profiling using Illumina arrays ranges from approximately 482,421 CpG sites with the HumanMethylation450 array (Pidsley et al., 2013) to around 930,000 CpG sites with the more advanced Infinium MethylationEPIC array (Carreras-Gallo et al., 2024); and miRNA datasets contain 1917 annotated hairpin precursors and 2654 mature sequences, according to miRBase v22 (Kozomara et al., 2019). In contrast, the number of samples in *omics* datasets are relatively small. In addition, studies show poor correlation between different measuring platforms (Rincon et al., 2020), specimen type (Donohue et al., 2019; Kuo et al., 2002) or patients used for biomarker selection (Ein-Dor et al., 2005).

As systems become more precise, the number of features (variables) to study increases, and consequently, the combinatorial joint value

* Corresponding author.
E-mail addresses: e.d.rojasvelazquez@uu.nl (D. Rojas-Velazquez), a.d.kraneveld@vu.nl (A.D. Kraneveld), alberto.tonda@inrae.fr (A. Tonda), a.lopezrincon@uu.nl (A. Lopez-Rincon).

https://doi.org/10.1016/j.mlwa.2025.100757
Received 23 July 2025; Received in revised form 8 October 2025; Accepted 12 October 2025
Available online 24 October 2025

they produce grows exponentially (Berisha et al., 2021). However, as the number of measurable variables increases, the sample size remains relatively small. These conditions impact model performance, computational requirements, data density, and the amount of training data needed. This is known as the *curse of dimensionality* in statistical learning theory (Theodoridis & Koutroumbas, 2006), which inevitably leads to a fundamental underdetermination of any prediction or classification in the model. Therefore, it is necessary to reduce the data to the critical variables to explain the underlying biological processes, known as feature selection (Bolón-Canedo et al., 2015).

Machine learning-based feature selection, commonly used for data reduction and biomarker discovery, is widely recognized as a powerful approach for analyzing high-dimensional, low-sample-size (HDLSS) data (Li et al., 2022b). Feature selection is a technique that selects the critical variables that best characterize a problem, removing those that are irrelevant or redundant (Bolón-Canedo et al., 2015). Selecting the best methodology for feature selection is not a trivial choice, since the performance of the method depends on the input data (Bolón-Canedo et al., 2016). A practical approach to this challenge is to use ensemble feature selection methods. The ensembles are based on the *Condorcet's jury theorem* which states that in situations of uncertainty, such as when dealing with model predictions, collective decisions of the majority often yield better results than individual choices (Austen-Smith & Banks, 1996; Boland, 1989). In practice, the ensemble classifiers method is an approach that takes a set or instances of classifiers, combines them to weigh a decision, and achieves results that outperform individual classifiers (Rokach, 2010).

The Recursive Ensemble Feature Selection (REFS) is an ensemble method for feature selection, designed to improve reproducibility in the biomarker discovery field as shown in Lopez-Rincon et al. (2019, 2020), composed of eight classifiers from the scikit-learn toolbox (Pedregosa et al., 2011) that rank the features at each iteration. REFS employs a nested approach within a 10-fold cross-validation scheme to produce robust, unbiased results and to prevent overfitting (Vabalas et al., 2019). REFS was designed to be accessible to researchers in the medical and pharmacological fields who may have no prior experience with machine learning, offering a simple and easy-to-use approach to feature selection without the need to manually adjust complex hyperparameters. However, depending on the study design, users can easily adjust hyperparameters such as the number of runs or the k value in k-fold cross-validation. In contrast to other feature selection methods where the evaluation of the effectiveness of the resulting feature set is provided by the algorithm itself (Ab Hamid et al., 2021; Ahsan et al., 2025; Chereda et al., 2024; Liu et al., 2017), we evaluate the efficiency of REFS results using different classifiers that are not part of the ensemble as well as external evaluations using datasets independent of the discovery dataset (when possible), to avoid overfitting and to provide unbiased and reliable results (Rojas-Velazquez et al., 2024).

We have used REFS successfully in previous works for feature selection in *omics* data, e.g. miRNA (Lopez-Rincon et al., 2019, 2020), mRNA (Kidwai et al., 2023; Metselaar et al., 2021; Rojas-Velazquez et al., 2023), and microbiome (16s rRNA sequences as input features) (Benner et al., 2021; Blankestijn et al., 2022; Kamphorst et al., 2023; Peralta-Marzal et al., 2024; Rojas-Velazquez et al., 2024). In comparison to other methods, REFS has been proven better in feature selection than using Univariate Feature Selection (UFS) (Lazo & Rathie, 1978), Recursive Feature Elimination (RFE) (Guyon et al., 2002), Elastic Net (EN) (Sokolov et al., 2016), Genetic Algorithms (GALGO) (Trevino & Falciani, 2006), Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) and Ensemble Feature Selection with Complete Linear Aggregation (EFS-CLA) (Abeel et al., 2010).

REFS uses accuracy as a metric to select the best set of features. Nevertheless, in practice, datasets tend to be imbalanced, e.g. if a dataset is 90% class *0* in a binary classification, classifying everything as class 0 would result in 90% accuracy, this means models tend to favor the majority class (Matharaarachchi et al., 2024). Therefore, it is necessary to use a more suitable metric, such as the Matthew's Correlation Coefficient (Chicco & Jurman, 2020).

The Matthew's Correlation Coefficient (MCC) is a robust statistical metric commonly used in machine learning to evaluate the performance of binary and multiclass classification models, especially under conditions of class imbalance. We selected MCC as our evaluation criterion because, in the context of imbalanced omics datasets, it offers a more reliable and informative assessment of classification performance compared to traditional metrics such as accuracy. Unlike accuracy, which can be misleading when one class dominates, the MCC accounts for all four confusion matrix categories (true positives, true negatives, false positives, and false negatives), providing a balanced measure even when class distributions are skewed (Chicco & Jurman, 2020). The MCC is essentially a coefficient value between −1 and +1, where +1 represents a perfect classification, 0 the same as random prediction, and −1 represents misclassification (Chicco & Jurman, 2020).

Changing metrics for feature selection may seem like a simple adjustment from a computational perspective, because the architecture of the algorithm is often subject to minor modifications or remains unchanged, but in the medical field it represents a meaningful methodological refinement that can significantly improve the reliability of the results and clinical relevance of new potential biomarkers to provide more accurate and robust diagnostic and prognostic tools for many complex diseases (Diaz-Uriarte et al., 2022).

In this study, we implemented an algorithm MCC-based, called MCC-REFS (Matheus Correlation Coefficient-Recursive Ensemble Feature Selection). First, we compared the performance between using accuracy-based feature selection criteria (REFS) versus using MCC for feature selection criteria (MCC-REFS) on two synthetic datasets and one real-world unbalanced dataset. Next, to validate our methodology, we made a comparison on binary real-world datasets against more recent feature selection techniques, such as GRAph Convolutional nEtwork feature Selector (GRACES) (Chen et al., 2023) and Deep Neural Pursuit (DNP) (Liu et al., 2017). Finally, we compared our results in a multilabel real-world dataset against Graph Convolutional Neural Network (GCNN) (Chereda et al., 2024).

2. Methods and materials

2.1. Feature selection

REFS uses eight different classifiers from the scikit-learn toolbox (Pedregosa et al., 2011) to rank the features in each iteration. The eight classifiers (Lopez-Rincon et al., 2019, 2018) were selected based on a study from The Cancer Genome Atlas (TCGA (Tomczak et al., 2015)) to separate the different types of cancer, the classifiers are:

- Stochastic Gradient Descent (SGD) (Zhang, 2004)
- Support Vector Classification (SVC) (Hearst et al., 1998)
- Gradient Boosting (Friedman, 2001)
- Random Forest (Breiman, 2001)
- Logistic Regression (Cox, 1958)
- Passive Aggressive (Crammer et al., 2006)
- Ridge Classifier (Tikhonov et al., 1943)
- Bagging (Breiman, 1999)

Each classifier ranks the features differently. For example, Fig. 1 graphically shows the ranking of the top 40 features from a synthetic dataset, where only the first 5 features are meaningful (Madelon dataset) (Guyon, 2003). It can be observed that all classifiers ranked the 5 meaningful features as important.

The classifiers rank the features in two ways, either by the weight of generated coefficients by each algorithm (Logistic Regression, Passive Aggressive, Ridge, SGD, SVC), or by how many times it was used

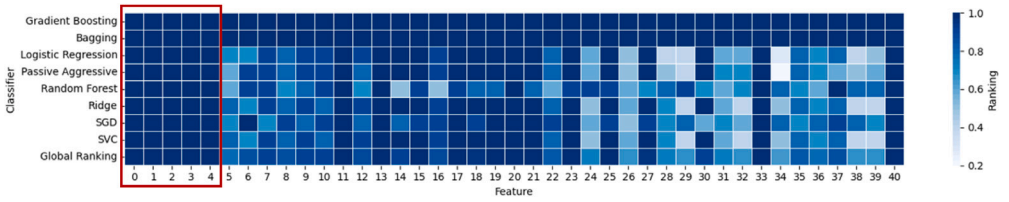

Fig. 1. Example of ranking of the first 40 features for one iteration in a *Madelon* dataset. The first five features are the informative features.

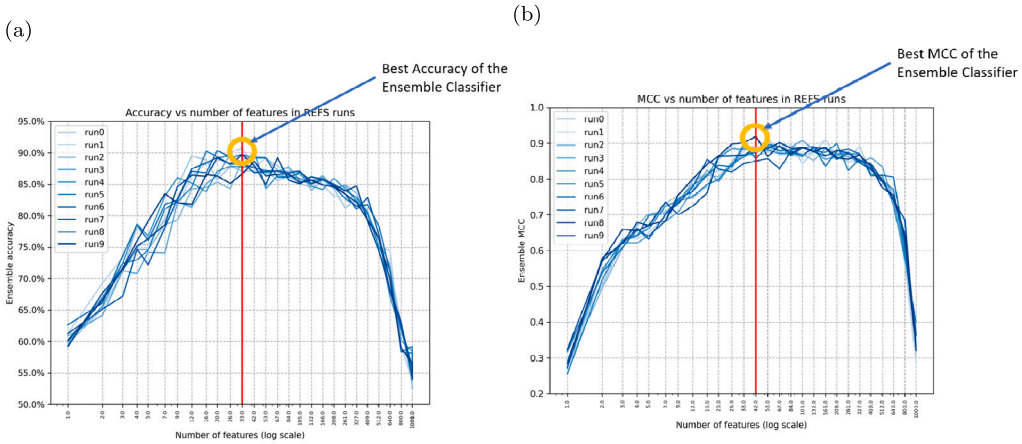

Fig. 2. (a) Selection of the best set of parameters in the best run ($n = 10$) by global accuracy in a 10-fold cross-validation scheme, (b) Selection of the best set of parameters in the best run ($n = 10$) by global MCC in a 10-fold cross-validation scheme.

in the decision trees (Bagging, Gradient Boosting and Random Forest). The first iteration reduces the dataset from the original size to 1000 and each subsequent iteration reduces the number of features by 20% (Lopez-Rincon et al., 2020). The features are evaluated using a nested approach within a 10-fold cross-validation scheme and reporting global accuracy (average accuracy of all classifiers). We repeat the overall procedure for n runs ($n = 10$). After the 10 runs, the final result is a reduced set of features, selected based on achieving the highest accuracy across the runs with the smallest number of features.

REFS uses global accuracy as discriminatory measure to choose the final set of features. However, class imbalance is a frequent issue encountered in *omics* data (Momeni et al., 2020), so as a solution, we propose to implement MCC as feature selection metric (Chicco & Jurman, 2020).

By implementing MCC as a metric for feature selection, REFS helps build robust and functional models, even with imbalanced data as an *out-of-the-box* algorithm for HDLSS data. This also avoids the need to tune complex hyperparameters while delivering reliable and repeatable results. Fig. 2 illustrates the reduction curve using accuracy and MCC, respectively.

After the feature selection process, the validation method consists of using the reduced set of features as input data for classifiers that are not part of the ensemble, e.g. Multilayer Perceptron (MLP) (Bishop, 1995). Evaluating the resulting set of features with independent algorithms, allows to avoid bias, overfitting and provide robust results.

2.2. Preprocessing and initialization

All datasets are divided into features, data, and labels, with missing values imputed using the mean for continuous variables and the mode for categorical ones. Ensemble classifiers are initialized using the default parameters provided by scikit-learn (Pedregosa et al., 2011), with the exception of Gradient Boosting Classifier and Random Forest Classifier where the initialization was *n_estimators = 300*. We ran MCC-REFS for 10 iterations using 10-fold cross-validation with stratification in a nested scheme (Zhong et al., 2023), selecting 10 folds based on the assumption that each class contains at least 10 samples; if fewer, the number of folds must be adjusted accordingly. Then, the data is normalized using z-score normalization as part of the algorithm.

2.3. Comparison between REFS and MCC-REFS

To evaluate feature selection criteria, we generated synthetic 'Madelon' datasets using the scikit-learn implementation adapted from Guyon (Guyon, 2003), comparing accuracy-based REFS with MCC-REFS. In addition to these synthetic comparisons, we assessed both methods on a imbalanced, binary-labeled real-world dataset obtained from Ciriello et al. (2015). For each dataset, we ran accuracy-based REFS and MCC-REFS for 10 runs, in a 10-fold cross-validation. Next, we tested the output feature set with a MLP classifier (not part of the ensemble) in a 10-fold cross-validation and reported the mean of the AUC (Area Under the Curve) (Šimundić, 2009)

2.3.1. Experiment 1

Data. Dataset 0 is a synthetic Madelon dataset consisting of 100 samples and 100 features, of which 5 are informative. It includes 51 samples labeled as class 0 and 49 as class 1.

2.3.2. Experiment 2

Data. Dataset 1 is a synthetic Madelon dataset consisting of 100 samples and 1000 features, of which 5 are informative. It includes 52 samples labeled as class 0 and 48 as class 1.

2.3.3. Experiment 3

Data. Dataset 2 obtained from Ciriello et al. (2015), contains multi-omics data designed to identify regulatory interactions across various omics layers, including copy number variations, somatic mutations, gene expression, and protein expression, that may predict survival outcomes in breast cancer. The dataset comprises 705 tumor samples from different patients, with 1936 features. The class distribution is highly imbalanced, with 611 samples labeled as *survival* and 94 as *fatal outcome*.

Table 1
Real-world datasets' characteristics from Li et al. (2018) used for comparing feature selection methods.

Dataset	Colon	SMK_CAN_187	ALLAML
Number of Samples	62	187	72
Number of Features	2000	19 993	7129

2.4. Comparison in binary classification

We compared MCC-REFS with recent feature selection techniques, including GRACES (Chen et al., 2023) and DNP (Liu et al., 2017), using three real-world, binary-labeled mRNA datasets from Li et al. (2018): **Colon**, **SMK_CAN_187**, and **ALLAML**. To ensure a fair comparison, we reproduced the methodology reported in Chen et al. (2023), splitting each dataset into 70% for training, and 30% for testing. In contrast with (Chen et al., 2023; Liu et al., 2017) where they divided the datasets in 70% for training, 20% validation and 10% testing. These datasets were selected because they were previously analyzed by the compared algorithms and are publicly available through the scikit-feature selection repository.[1] Table 1 presents an overview of the three real-world, binary-labeled selected datasets. We split each dataset randomly into 70% for training and 30% for testing, applied MCC-REFS on the training data in a 10-fold cross-validation, and evaluated the output feature set on the testing data using a multilayer perceptron (MLP) classifier (independent from the ensemble) 20 times, and compared the mean of the AUC with GRACES and DNP.

2.4.1. Experiment 4

Data. The **Colon** (Li et al., 2001) dataset contains gene expression profiles from colon tumor patients and healthy controls. It comprises 2000 mRNA features across 62 samples, with 40 samples labeled as patients and 22 as controls.

2.4.2. Experiment 5

Data. The **SMK_CAN_187** (Spira et al., 2007) dataset contains gene expression data from smokers diagnosed with lung cancer and smokers without lung cancer. It includes 19,993 mRNA features across 187 samples, with 90 samples labeled as lung cancer patients and 97 as non-cancer controls.

2.4.3. Experiment 6

Data. The ALLAML (Golub et al., 1999) dataset contains gene expression data from patients diagnosed with acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). It comprises 7129 mRNA features across 72 samples, with 47 samples labeled as ALL and 25 as AML.

2.5. Comparison in MultiLabel classification

To evaluate the performance of MCC-REFS in a multi-label classification setting, we conducted a comparative analysis using a real-world dataset on breast cancer subtypes from TCGA (the cancer genome atlas) (Weinstein et al., 2013), alongside the GCNN model proposed by Chereda et al. (2024).[2] For the dataset, we ran MCC-REFS for 10 runs, in a 10-fold cross-validation. Next, we tested the output feature set with a MLP classifier (not part of the ensemble) in a 10-fold cross-validation and reported the mean of the accuracy.

[1] https://jundongl.github.io/scikit-feature/datasets.html
[2] https://www.cbioportal.org/study/summary?id=brca_tcga_pan_can_atlas_2018

Table 2
Number of features in the resulting sets and AUC-ROC used to compare REFS and MCC-REFS performance in the synthetic datasets and the real-world dataset. Results are from using an MLP classifier for validation.

Method	Dataset 0	Dataset 1	Breast tumor
REFS	0.9000 (16 features)	0.9900 (105 features)	0.7100 (261 features)
MCC-REFS	0.9000 (16 features)	0.9900 (42 features)	0.7300 (132 features)

2.5.1. Experiment 7

Data. The TCGA dataset comprises gene expression profiles from breast cancer patient samples, totaling 20,531 features across 1082 samples. Following the sample selection criteria described by Chereda et al. (2024), we mapped sample IDs to corresponding clinical data containing subtype labels. The final dataset includes 981 samples categorized into five breast cancer subtypes: luminal A (499 samples), luminal B (197), basal-like (171), HER2-enriched (78), and normal-like (36).

3. Results

3.1. Comparison between REFS and MCC-REFS

3.1.1. Experiment 1

We ran REFS and MCC-REFS on **Dataset 0** to identify the most relevant features to separate class 0 from class 1. An in-depth analysis of both methodologies showed $k = 16$ as the top features in the solution, including the five meaningful features, see Fig. 3(a) for REFS and Fig. 3(b) for MCC-REFS. In both cases, in 10 different runs, the five informative features were selected as the best answer for each *run*, with the exception of feature 0 that appeared in both cases 9 out of 10 runs.

3.1.2. Experiment 2

We ran REFS and MCC-REFS on **Dataset 1**. The results of both methodologies showed in the case of accuracy, the solution was $k = 105$, whereas for MCC-REFS the solution was $k = 42$, see Fig. 4(a) for REFS and Fig. 4(b) for MCC-REFS. In both cases, the best solution had five informative features. In the case of MCC-REFS, in all 10 runs, the five meaningful features were selected in the best solution. For accuracy-based discrimination measure, the five informative features were selected in 10 out 10 runs, except for feature 3, which was selected in 9 out of 10 runs.

3.1.3. Experiment 3

We performed a compassion between REFS and MCC-REFS in **Dataset 2**, Fig. 5, by visual inspection, we can see that MCC-REFS performance is more stable (Fig. 5(b)) compared to REFS performance (Fig. 5(a)), demonstrating the effectiveness of the performance when using MCC as a feature selection metric when working with highly unbalanced datasets. The resulting sets of REFS and MCC-REFS are 261 and 132, respectively.

As mentioned, to validate and compare the performance of the resulting features set in each experiment, we calculated AUC-ROC with a classifier that is not part of the ensemble (MLP) in a 10-fold cross-validation scheme, with the output feature set. Table 2 presents the AUC-ROC results obtained from MLP validation in each experiment. The corresponding AUC-ROC plots can be found in *Supplementary file 1*.

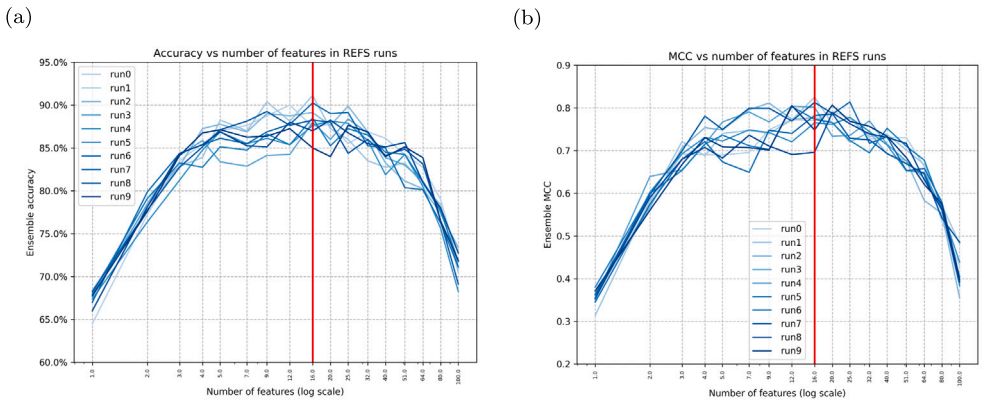

Fig. 3. Dataset 0 results: (a) Selection of the best set of features in the best run ($n = 10$) by global accuracy identify by a red line, (b) Selection of the best set of parameters in the best run ($n = 10$) by global MCC identify by a red line.

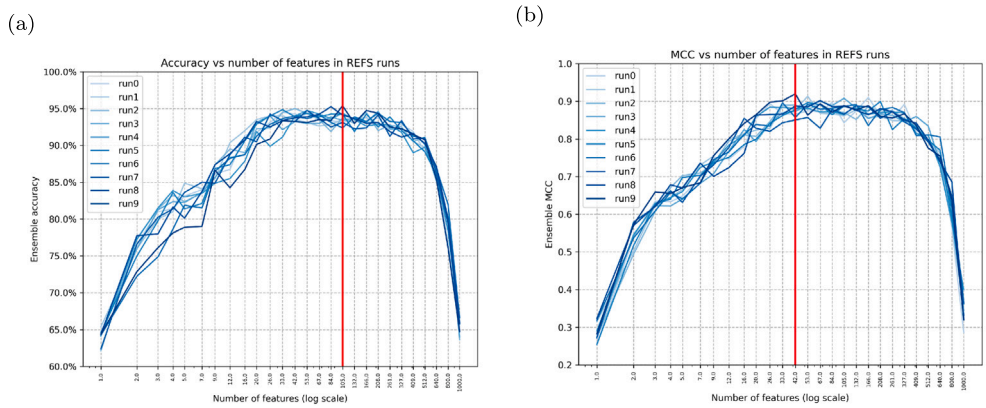

Fig. 4. Dataset 1 results: (a) Selection of the best set of features in the best run ($n = 10$) by global accuracy identify by a red line, (b) Selection of the best set of parameters in the best run ($n = 10$) by global MCC identify by a red line.

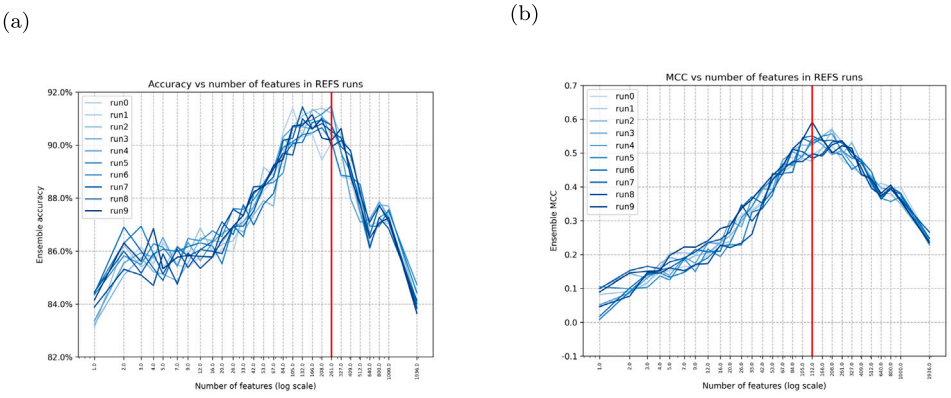

Fig. 5. Dataset 2 results: (a) Selection of the best set of features in the best run ($n = 10$) by global accuracy identify by a red line, (b) Selection of the best set of parameters in the best run ($n = 10$) by global MCC identify by a red line.

3.2. Comparison in binary classification

3.2.1. Experiment 4

We applied MCC-REFS to dataset Colon to identify the most relevant set of features. The optimal feature selection curve for the Colon dataset, which resulted in a subset of 12 features, is shown in Fig. 6(a). The ROC curve was generated by selecting features using MCC-REFS on 70% of the data designated for training, and evaluating performance on the remaining 30% using an MLP classifier, repeated over 20 independent runs (Fig. 6(b)).

3.2.2. Experiment 5

We applied MCC-REFS to dataset SMK_CAN_187 to identify the most relevant set of features. The optimal feature selection curve for the SMK _CAN_187 dataset, which resulted in a subset of 33 features, is shown in Fig. 7(a). The ROC curve was generated by selecting features using MCC-REFS on 70% of the data designated for training, and evaluating

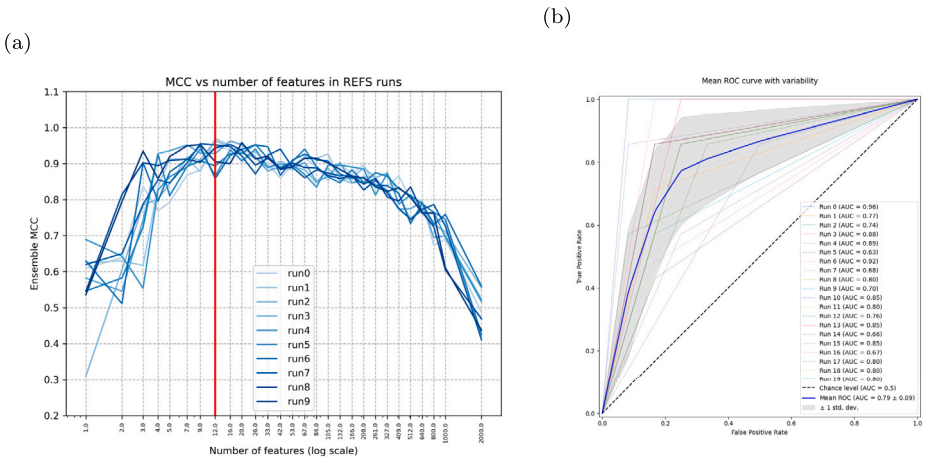

Fig. 6. (a) Selection of the best set of features in the best run (n = 12) red line, (b) ROC curve evaluating performance on the remaining 30% using an MLP classifier, repeated over 20 independent runs.

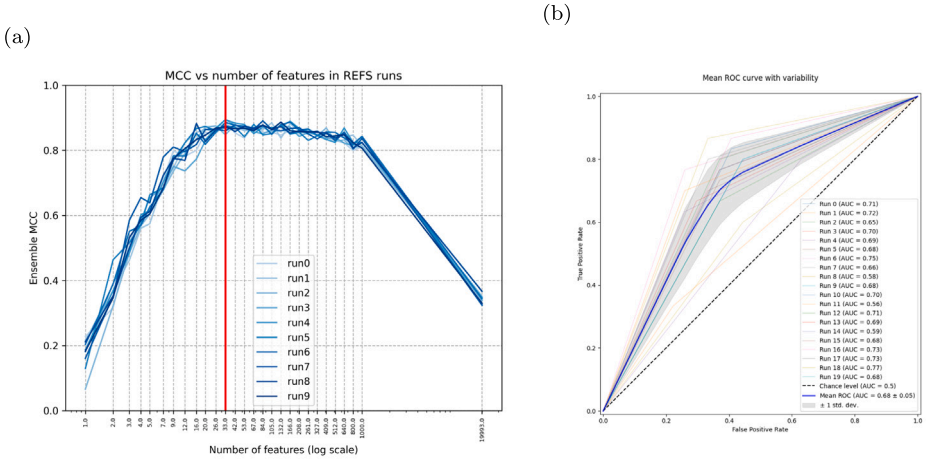

Fig. 7. (a) Selection of the best set of features in the best run (n = 33) red line, (b) ROC curve evaluating performance on the remaining 30% using an MLP classifier, repeated over 20 independent runs.

performance on the remaining 30% using an MLP classifier, repeated over 20 independent runs (Fig. 7(b)).

3.2.3. Experiment 6

We applied MCC-REFS to dataset ALLML to identify the most relevant set of features. The optimal feature selection curve for the ALLML dataset, which resulted in a subset of 12 features, is shown in Fig. 8(a). The ROC curve was generated by selecting features using MCC-REFS on 70% of the data designated for training, and evaluating performance on the remaining 30% using an MLP classifier, repeated over 20 independent runs (Fig. 8(b)).

In Table 3, we show the results reported for DNP (Liu et al., 2017) and GRACES (Chen et al., 2023), with the AUC-ROC obtained from the validation of the selected set using MCC-REFS with an MLP classifier for 20 evaluations for its comparison.

3.3. Comparison in MultiLabel classification

3.3.1. Experiment 7

To identify the most informative features for classification, MCC-REFS was applied to the TCGA dataset. From the original set of 20,531 genes, the method selected a subset of 327 genes, see Fig. 9(a). Fig. 9(b) illustrates the confusion matrix, highlighting the effectiveness of MCC-REFS in a multi-label classification setting. After validating the selected feature set using an MLP classifier, MCC-REFS achieved an accuracy of 0.9602, outperforming GCNN, which reached 0.9133 (Chereda et al., 2024).

Table 3
AUC-ROC resulting from the analysis of the three real-world datasets using DNP, GRACES and MCC-REFS. MLP classifier is not part of the ensemble.

Method	Dataset 3 (Colon)	Dataset 4 (SMK_CAN_187)	Dataset 5 (ALLAML)
GRACES	0.7591	0.6644	0.9025
DNP	0.7474	0.6454	0.8173
MCC-REFS μ All Classifiers	0.7578	0.6954	0.9270
MLP Classifier	**0.7900**	**0.6800**	**0.9500**

4. Discussion and conclusion

In this work, we implemented the Matthews Correlation Coefficient (MCC) as a feature selection criterion within the REFS framework, resulting in MCC-REFS. To validate its effectiveness, we tested MCC-REFS on two synthetic datasets and compared its performance with REFS, which uses global accuracy as its feature selection criterion. The results showed that REFS and **MCC-REFS** performed similarly on **Dataset 0** (100 samples and 100 features), with both methods identifying the informative features in the selected sets (k = 16). In

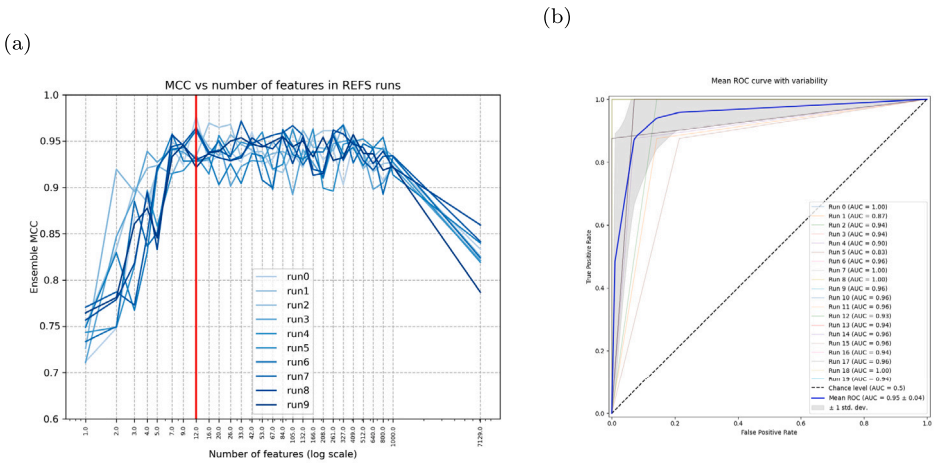

Fig. 8. (a) Selection of the best set of features in the best run (n = 12) red line, (b) ROC curve evaluating performance on the remaining 30% using an MLP classifier, repeated over 20 independent runs.

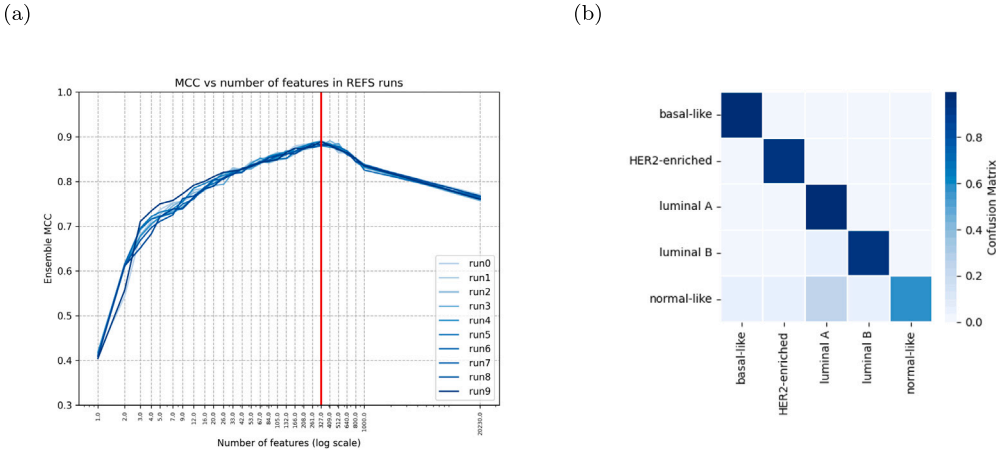

Fig. 9. (a) 10 runs in the Multi-Label breast cancer database, to find the best number of features (*k=327* genes) out of 20531 identify by a red line. (b) Normalized Confusion Matrix of the classification.

contrast, on **Dataset 1** (100 samples and 1000 features), MCC-REFS demonstrated better performance, identifying the informative features with a more compact set (k = 42) compared to REFS (k = 105). For the comparison using a highly imbalanced real-world dataset, MCC-REFS showed more stable performance than REFS. Although both methods achieved similar AUC-ROC scores, MCC-REFS produced a significantly smaller feature set (k = 132) compared to REFS (k = 261), highlighting its efficiency under class imbalance.

In real-world *omics* classification, the results showed that MCC-REFS had slightly better performance against GRACES and DNP (real-world binary-label datasets) and against GCNN (real-world multi-label datasets). It is important to highlight that **the number of features selected by MCC-REFS is not a fixed parameter** like in GRACES (k = 10) and DNP (k = 2, 5, 10, 25). This ensures that the number of features selected by MCC-REFS is not limited by user-defined constraints. Additionally, MCC-REFS does not require hyperparameter tuning and can be used out of the box.

During the comparisons, we noted that DNP (Liu et al., 2017) and GCNN (Chen et al., 2023) use the same feature selection algorithm to evaluate their effectiveness. This could limit the comparison between algorithms, because the evaluation is performed with the same selection model, which could suggest that the accuracy reported is a reflection of a model tailored to the resulting feature set due to the internal processes of the algorithms. To minimize this limitation, the databases used for this comparison do not present alterations, that is, they were used as reported in Li et al. (2018).

Another potential limitation we noted in the comparison is establishing the number of elements desired in the resulting set. An important question that arises from this is: *what is the argument for selecting the number of elements?* The automatic selection approach of MCC-REFS is a potential solution to this question, since the algorithm is given the freedom to select the set of features that obtains the best accuracy using the lowest number of elements in the set. Validation of the resulting set is done using algorithms that are not part of the ensemble, and the obtained AUC-ROC is validated using the diagnostic accuracy scale presented in Šimundić (2009).

Based on the results obtained from the experiments, we can conclude that MCC-REFS produces more robust and reliable results compared to REFS, GRACES, DNP, and GCNN. Although the results of MCC-REFS are promising and demonstrated better performance with real-world datasets (imbalanced classes), further experimentation is needed using this aproach. Therefore, the next step is to extend the analysis by exploring additional datasets available in public repositories such as NCBI,[3] GEO,[4] TCGA,[5] among others. Future work will investigate whether the increase in selected features under higher class imbalance reflects an adaptive behavior of MCC-REFS or a limitation

[3] https://www.ncbi.nlm.nih.gov/
[4] https://www.ncbi.nlm.nih.gov/geo/
[5] https://www.cancer.gov/ccg/research/genome-sequencing/tcga

in identifying truly informative features. To test this hypothesis, we plan to conduct experiments across a broader range of omics datasets, including mRNA, miRNA, methylation, and microbiome data, with varying imbalance levels.

CRediT authorship contribution statement

David Rojas-Velazquez: Methodology, Software, Formal analysis, Writing – original draft. **Aletta D. Kraneveld:** Supervision, Writing – review & editing. **Alberto Tonda:** Supervision, Validation, Writing – review & editing. **Alejandro Lopez-Rincon:** Conceptualization, Project administration, Funding acquisition, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary file 1

It contains all the AUC-ROC plots corresponding to the validation process for experiments using REFS and MCC-REFS.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.mlwa.2025.100757.

Data availability

The code is available in the following repository: https://github.com/steppenwolf0/REFS-MCC.

References

Ab Hamid, T. M. T., Sallehuddin, R., Yunos, Z. M., & Ali, A. (2021). Ensemble based filter feature selection with harmonize particle swarm optimization and support vector machine for optimal cancer classification. *Machine Learning with Applications*, *5*, Article 100054.

Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., & Saeys, Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, *26*(3), 392–398.

Ahsan, M. M., Raman, S., Liu, Y., & Siddique, Z. (2025). Hybrid oversampling technique for imbalanced pattern recognition: Enhancing performance with borderline synthetic minority oversampling and generative adversarial networks. *Machine Learning with Applications*, *20*, Article 100637.

Austen-Smith, D., & Banks, J. S. (1996). Information aggregation, rationality, and the condorcet jury theorem. *American Political Science Review*, *90*(1), 34–45.

Benner, M., Lopez-Rincon, A., Thijssen, S., Garssen, J., Ferwerda, G., Joosten, I., van der Molen, R. G., & Hogenkamp, A. (2021). Antibiotic intervention affects maternal immunity during gestation in mice. *Frontiers in Immunology*, *12*, Article 685742.

Berisha, V., Krantsevich, C., Hahn, P. R., Hahn, S., Dasarathy, G., Turaga, P., & Liss, J. (2021). Digital medicine and the curse of dimensionality. *NPJ Digital Medicine*, *4*(1), 153.

Bishop, C. (1995). *Neural networks for pattern recognition*: *vol. 2*, (pp. 223–228). Clarendon Press google schola.

Blankestijn, J., Lopez-Rincon, A., Neerincx, A., Vijverberg, S., Hashimoto, S., Gorenjak, M., Sardón-Prado, O., Corcuera, P., Korta-Murua, J., Pino-Yanes, M., et al. (2022). Classifying asthma control using salivary and fecal microbiome in children with moderate to severe asthma: results from the syspharmpedia study.

Boland, P. J. (1989). Majority systems and the condorcet jury theorem. *Journal of the Royal Statistical Society Series D: The Statistician*, *38*(3), 181–189.

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2015). Recent advances and emerging challenges of feature selection in the context of big data. *Knowledge-Based Systems*, *86*, 33–45.

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2016). Feature selection for high-dimensional data. *Progress in Artificial Intelligence*, *5*, 65–75.

Breiman, L. (1999). Pasting small votes for classification in large databases and on-line. *Machine Learning*, *36*, 85–103.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Carreras-Gallo, N., Dwaraka, V. B., Jima, D. D., Skaar, D. A., Mendez, T. L., Planchart, A., Zhou, W., Jirtle, R. L., Smith, R., & Hoyo, C. (2024). Creation and validation of the first infinium dna methylation array for the human imprintome. *Epigenetics Communications*, *4*(1), 5.

Chen, C., Weiss, S. T., & Liu, Y.-Y. (2023). Graph convolutional network-based feature selection for high-dimensional and low-sample size data. *Bioinformatics*, *39*(4), btad135.

Chereda, H., Leha, A., & Beißbarth, T. (2024). Stable feature selection utilizing graph convolutional neural network and layer-wise relevance propagation for biomarker discovery in breast cancer. *Artificial Intelligence in Medicine*, Article 102840.

Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 1–13.

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., et al. (2015). Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, *163*(2), 506–519.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *20*(2), 215–232.

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., & Singer, Y. (2006). Online passive aggressive algorithms. *Journal of Machine Learning Research*.

Diaz-Uriarte, R., Gómez de Lope, E., Giugno, R., Fröhlich, H., Nazarov, P. V., Nepomuceno-Chamorro, I. A., Rauschenberger, A., & Glaab, E. (2022). Ten quick tips for biomarker discovery and validation analyses using machine learning. *PLoS Computational Biology*, *18*(8), Article e1010357.

Donohue, D. E., Gautam, A., Miller, S.-A., Srinivasan, S., Abu-Amara, D., Campbell, R., Marmar, C. R., Hammamieh, R., & Jett, M. (2019). Gene expression profiling of whole blood: A comparative assessment of rna-stabilizing collection methods. *PloS One*, *14*(10), Article e0223065.

Ein-Dor, L., Kela, I., Getz, G., Givol, D., & Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, *21*(2), 171–178.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189–1232.

Goh, W. W. B., Wang, W., & Wong, L. (2017). Why batch effects matter in omics data, and how to avoid them. *Trends in Biotechnology*, *35*(6), 498–507.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, *286*(5439), 531–537.

Guyon, I. (2003). Design of experiments of the nips 2003 variable selection benchmark. In *NIPS 2003 workshop on feature extraction and feature selection*: *vol. 253*, (p. 40).

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*, 389–422.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their Applications*, *13*(4), 18–28.

Kamphorst, K., Lopez-Rincon, A., Vlieger, A. M., Garssen, J., van't Riet, E., & van Elburg, R. M. (2023). Predictive factors for allergy at 4–6 years of age based on machine learning: A pilot study. *PharmaNutrition*, *23*, Article 100326.

Kidwai, S., Barbiero, P., Meijerman, I., Tonda, A., Perez-Pardo, P., Lio, P., van der Maitland-Zee, A. H., Oberski, D. L., Kraneveld, A. D., & Lopez-Rincon, A. (2023). A robust mrna signature obtained via recursive ensemble feature selection predicts the responsiveness of omalizumab in moderate-to-severe asthma. *Clinical and Translational Allergy*, *13*(11), Article e12306.

Kozomara, A., Birgaoanu, M., & Griffiths-Jones, S. (2019). Mirbase: from microrna sequences to function. *Nucleic Acids Research*, *47*(D1), D155–D162.

Kuo, W. P., Jenssen, T.-K., Butte, A. J., Ohno-Machado, L., & Kohane, I. S. (2002). Analysis of matched mrna measurements from two different microarray technologies. *Bioinformatics*, *18*(3), 405–412.

Lazo, A. V., & Rathie, P. (1978). On the entropy of continuous probability distributions (corresp.). *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, *24*(1), 120–122.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018). Feature selection: A data perspective. *ACM Computing Surveys*, *50*(6), 94.

Li, C., Gao, Z., Su, B., Xu, G., & Lin, X. (2022a). Data analysis methods for defining biomarkers from omics data. *Analytical and Bioanalytical Chemistry*, *414*(1), 235–250.

Li, R., Li, L., Xu, Y., & Yang, J. (2022b). Machine learning meets omics: applications and perspectives. *Briefings in Bioinformatics*, *23*(1), bbab460.

Li, L., Weinberg, C. R., Darden, T. A., & Pedersen, L. G. (2001). Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the ga/knn method. *Bioinformatics*, *17*(12), 1131–1142.

Liu, B., Wei, Y., Zhang, Y., & Yang, Q. (2017). Deep neural networks for high dimension, low sample size data. In *IJCAI* (pp. 2287–2293).

Loganathan, T., et al. (2022). The influence of machine learning technologies in gut microbiome research and cancer studies-a review. *Life Sciences*, Article 121118.

Lopez-Rincon, A., Martinez-Archundia, M., Martinez-Ruiz, G. U., Schoenhuth, A., & Tonda, A. (2019). Automatic discovery of 100-mirna signature for cancer classification using ensemble feature selection. *BMC Bioinformatics*, *20*, 1–17.

Lopez-Rincon, A., Mendoza-Maldonado, L., Martinez-Archundia, M., Schönhuth, A., Kraneveld, A. D., Garssen, J., & Tonda, A. (2020). Machine learning-based ensemble recursive feature selection of circulating mirnas for cancer tumor classification. *Cancers*, *12*(7), 1785.

Lopez-Rincon, A., Tonda, A., Elati, M., Schwander, O., Piwowarski, B., & Gallinari, P. (2018). Evolutionary optimization of convolutional neural networks for cancer mirna biomarkers classification. *Applied Soft Computing*, *65*, 91–100.

Matharaarachchi, S., Domaratzki, M., & Muthukumarana, S. (2024). Enhancing smote for imbalanced data with abnormal minority instances. *Machine Learning with Applications*, *18*, Article 100597.

Metselaar, P. I., Mendoza-Maldonado, L., Li Yim, A. Y. F., Abarkan, I., Henneman, P., Te Velde, A. A., Schönhuth, A., Bosch, J. A., Kraneveld, A. D., & Lopez-Rincon, A. (2021). Recursive ensemble feature selection provides a robust mrna expression signature for myalgic encephalomyelitis/chronic fatigue syndrome. *Scientific Reports*, *11*(1), 4541.

Momeni, Z., Hassanzadeh, E., Abadeh, M. S., & Bellazzi, R. (2020). A survey on single and multi omics data mining methods in cancer data classification. *Journal of Biomedical Informatics*, *107*, Article 103466.

Papin, J. A., Mac Gabhann, F., Sauro, H. M., Nickerson, D., & Rampadarath, A. (2020). Improving reproducibility in computational biology research.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peralta-Marzal, L. N., Rojas-Velazquez, D., Rigters, D., Prince, N., Garssen, J., Kraneveld, A. D., Perez-Pardo, P., & Lopez-Rincon, A. (2024). A robust microbiome signature for autism spectrum disorder across different studies using machine learning. *Scientific Reports*, *14*(1), 814.

Pidsley, R., Y Wong, C. C., Volta, M., Lunnon, K., Mill, J., & Schalkwyk, L. C. (2013). A data-driven approach to preprocessing illumina 450k methylation array data. *BMC Genomics*, *14*(1), 1–10.

Ptolemy, A. S., & Rifai, N. (2010). What is a biomarker? research investments and lack of clinical integration necessitate a review of biomarker terminology and validation schema. *Scandinavian Journal of Clinical and Laboratory Investigation*, *70*(sup242), 6–14.

Rincon, A. L., Kraneveld, A. D., & Tonda, A. (2020). Batch correction of genomic data in chronic fatigue syndrome using cma-es. In *Proceedings of the 2020 genetic and evolutionary computation conference companion* (pp. 277–278).

Robinson, M. D., & Speed, T. P. (2007). A comparison of affymetrix gene expression arrays. *BMC Bioinformatics*, *8*, 1–16.

Rojas-Velazquez, D., Kidwai, S., Kraneveld, A. D., Tonda, A., Oberski, D., Garssen, J., & Lopez-Rincon, A. (2024). Methodology for biomarker discovery with reproducibility in microbiome data using machine learning. *BMC Bioinformatics*, *25*(1), 26.

Rojas-Velazquez, D., Tonda, A., Rodriguez-Guerra, I., Kraneveld, A. D., & Lopez-Rincon, A. (2023). Multi-objective evolutionary discretization of gene expression profiles: Application to covid-19 severity prediction. In *International conference on the applications of evolutionary computation (part of evoStar)* (pp. 703–717). Springer.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*, 1–39.

Šimundić, A.-M. (2009). Measures of diagnostic accuracy: basic definitions. *ejifcc*, *19*(4), 203.

Sokolov, A., Carlin, D. E., Paull, E. O., Baertsch, R., & Stuart, J. M. (2016). Pathway-based genomics prediction using generalized elastic net. *PLoS Computational Biology*, *12*(3), Article e1004790.

Spira, A., Beane, J. E., Shah, V., Steiling, K., Liu, G., Schembri, F., Gilman, S., Dumas, Y.-M., Calner, P., Sebastiani, P., et al. (2007). Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nature Medicine*, *13*(3), 361–366.

Theodoridis, S., & Koutroumbas, K. (2006). *Pattern recognition*. Elsevier.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *58*(1), 267–288.

Tikhonov, A. N., et al. (1943). On the stability of inverse problems. In *Dokl. akad. nauk sssr*: *vol. 39*, (pp. 195–198).

Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology*, (Współczesna Onkologia) *2015*(1), 68–77.

Trevino, V., & Falciani, F. (2006). Galgo: an r package for multivariate variable selection using genetic algorithms. *Bioinformatics*, *22*(9), 1154–1156.

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PloS One*, *14*(11), Article e0224365.

Visar, B., Chelsea, K., Richard, H. P., Hahn, S., Gautam, D., Pavan, T., & Liss, J. (2021). Digital medicine and the curse of dimensionality. *NPJ Digital Medicine*, *4*(1).

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., & Stuart, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, *45*(10), 1113–1120.

Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on machine learning* (p. 116).

Zhong, Y., Chalise, P., & He, J. (2023). Nested cross-validation with ensemble feature selection and classification model for high-dimensional biological data. *Communications in Statistics. Simulation and Computation*, *52*(1), 110–125.