mathematics

MDPI

Article

Homogeneous Ensemble Feature Selection for Mass Spectrometry Data Prediction in Cancer Studies

Yulan Liang [1,*], Amin Gharipour [2], Erik Kelemen [3] and Arpad Kelemen [4]

[1] Department of Family and Community Health, University of Maryland Baltimore, Baltimore, MD 21201, USA
[2] School of Information and Communication Technology, Griffith University, Gold Coast Campus, Brisbane, QLD 4222, Australia; amin.gharipour@griffithuni.edu.au
[3] Department of Computer Science, University of Maryland College Park, College Park, MD 20742, USA; kelemen.erik@gmail.com
[4] Department of Organizational Systems and Adult Health, University of Maryland Baltimore, Baltimore, MD 21201, USA; kelemen@umaryland.edu
* Correspondence: liang@umaryland.edu

Abstract: The identification of important proteins is critical for the medical diagnosis and prognosis of common diseases. Diverse sets of computational tools have been developed for omics data reduction and protein selection. However, standard statistical models with single-feature selection involve the multi-testing burden of low power with limited available samples. Furthermore, high correlations among proteins with high redundancy and moderate effects often lead to unstable selections and cause reproducibility issues. Ensemble feature selection in machine learning (ML) may identify a stable set of disease biomarkers that could improve the prediction performance of subsequent classification models and thereby simplify their interpretability. In this study, we developed a three-stage homogeneous ensemble feature selection (HEFS) approach for both identifying proteins and improving prediction accuracy. This approach was implemented and applied to ovarian cancer proteogenomics datasets comprising (1) binary putative homologous recombination deficiency (HRD)-positive or -negative samples; (2) multiple mRNA classes (differentiated, proliferative, immunoreactive, mesenchymal, and unknown samples). We conducted and compared various ML methods with HEFS including random forest (RF), support vector machine (SVM), and neural network (NN) for predicting both binary and multiple-class outcomes. The results indicated that the prediction accuracies varied for both binary and multiple-class classifications using various ML approaches with the proposed HEFS method. RF and NN provided better prediction accuracies than simple Naive Bayes or logistic models. For binary outcomes, with a sample size of 122 and nine selected prediction proteins using our proposed three-stage HEFS approach, the best ensemble ML (Treebag) achieved 83% accuracy, 85% sensitivity, and 81% specificity. For multiple (five)-class outcomes, the proposed HEFS-selected proteins combined with Principal Component Analysis (PCA) in NN resulted in prediction accuracies for multiple-class classifications ranging from 75% to 96% for each of the five classes. Despite the different prediction accuracies of the various models, HEFS identified consistent sets of proteins linked to the binary and multiple-class outcomes.

Keywords: ensemble feature selection; machine learning; mass spectrometry proteomics; ovarian cancer; prediction

MSC: 62P10; 92C50

check for updates

Citation: Liang, Y.; Gharipour, A.; Kelemen, E.; Kelemen, A. Homogeneous Ensemble Feature Selection for Mass Spectrometry Data Prediction in Cancer Studies. *Mathematics* **2024**, *12*, 2085. https://doi.org/10.3390/math12132085

Academic Editors: Miguel Ángel Montero-Alonso and Juan De Dios Luna del Castillo

Received: 14 May 2024
Revised: 28 June 2024
Accepted: 30 June 2024
Published: 3 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Ovarian cancer is the deadliest gynecologic malignancy, with most patients diagnosed in late stages. Early detection and antineoplastic therapeutics are vital to treating ovarian cancer patients who may have heterogeneous responses [1–3]. Proteogenomics is an emerging approach integrating proteomics with genomics and transcriptomics, gaining

new insights for a more complete understanding of complex diseases and treatments to advance basic, translational, and clinical research [4–6]. Mass spectrometry (MS)-based proteomic technologies have enabled the profiling of thousands of global proteins and have made proteogenomic data available to examine the linkages among DNA, mRNA, proteins, and disease status and to determine which proteins are associated with gene mutation and disease status (such as cancer subtypes, disease stages, and patient treatment heterogeneity) [7–9].

Identifying protein and gene signatures from thousands of omics data generated from high-throughput technologies has been challenging from both computational and biomedical perspectives [10,11]. Standard statistical marker selection methods with association analysis, such as correlation coefficients, mutual information, *t*-tests, and chi-square tests, rely on *p*-values from statistical models. These methods have the advantages of being scalable, fast, and independent of any specific learning algorithm. However, these single-feature selection methods are sample-dependent and may have specific biases. These approaches also face the multi-testing burden of low power associated with a limited patient sample size. Moreover, high correlations among features such as protein markers, along with small to moderate effects on diseases, often lead to unstable selections that may cause reproducibility issues [12–15]. In addition, these methods may select redundant features and ignore feature correlations and dependencies. Another inherent challenge for genetic linkage analysis is that some recent studies have indicated modest/moderate correlations between genes, mRNAs, and proteins across different organisms (i.e., correlation coefficients from 0.09 to 0.46 in multicellular organisms) [16].

Machine learning (ML) with wrapper methods and ensemble feature selection has the advantage of alleviating and compensating such drawbacks [17,18]. Wrapper methods such as forward selection, backward elimination, stepwise selection, and recursive feature elimination consider feature dependencies and correlations, generally providing better performance. However, they are usually computationally expensive, especially with large feature sets, and prone to overfitting.

Embedded methods, including group selection techniques like Lasso or Ridge regression, Bayesian shrinkage or regularization models, or decision tree-based methods, overcome the drawbacks of filter and wrapper methods [19,20]. They combine the merits of filters and wrappers with improved computational efficiency. For example, Lasso, Ridge regression, and Elastic Net remain foundational techniques that are particularly useful for their simplicity and effectiveness in regularization [21]. Moreover, other more advanced methods like Smoothly Clipped Absolute Deviation and Minimax Concave Penalty provide significant improvements in high-dimensional feature selection and bias reduction, making them valuable tools in modern statistical and ML applications [22,23]. The drawbacks of these methods include requiring tuning additional parameters and being more computationally intensive than Lasso and Ridge regression. In addition, they may also be limited to specific algorithms and may not generalize well.

Ensemble methods have become increasingly popular in feature selection due to their ability to improve model performance, effectiveness, and robustness. They have the flexibility to combine different types of models (e.g., decision trees, linear models, neural networks) to leverage their unique strengths. This flexibility allows for tailoring the ensemble approach to the specific characteristics of the data and the problem at hand, providing a more reliable and comprehensive approach compared to single-model techniques. By combining multiple models, ensemble methods reduce the risk of overfitting, stabilize feature selection, and improve predictive performance. Additionally, ensemble methods are highly scalable and can incorporate different algorithms such as gradient boosting and model averaging techniques to handle large datasets efficiently, which makes them suitable for high-dimensional feature selection tasks in biomedical omics fields.

Moreover, ensemble models integrate predictions from multiple independent predictors to generate the strongest signals across predictors that rise to the top. Ensemble predictors consistently perform among the best across challenges and are the most robust

to noise in various datasets [24,25]. These methods provide different balances between computational efficiency and predictive accuracy, which makes them suitable for various scenarios in feature selection tasks for big data applications.

At the same time, different sets of ranked features from various ML methods could provide the same classification performance. Therefore, one key question is whether the ensemble ML feature selection approach could result in consistent and reproducible biomarkers in omics fields. This is important to reduce the burden of clinical validation. Furthermore, tuning various algorithms' parameters could vary the prediction accuracy and selection results. How best to measure a predictor's relative importance in the model and enhance interpretability of ML is a challenge despite the power and flexibility of ML ensemble modeling [26].

From the biomedical perspective, one important question is how well protein markers can be used to predict gene mutation status or mRNA status and which biomarkers are associated with those [27]. Recent studies on ensemble feature selection for mass spectrometry data prediction in cancer research have shown advancements [28,29]. For example, a study published in *Nature* adopted an ensemble systems biology approach to improve prognosis prediction by integrating multiple-feature selection techniques. Another study demonstrated the robustness of ensemble feature selection methods in identifying stable biomarkers for cancer diagnosis [25]. These studies underscore the importance of combining various feature selection methods to handle the high dimensionality and complexity of mass spectrometry data, leading to more accurate and reliable predictions in cancer studies.

In our earlier studies, we examined various statistical filter approaches for the reproducibility and stability of feature selection methods [15] and compared their performance using ovarian cancer proteogenomic data. In this paper, we evaluate and compare various ensemble feature selection approaches for their consistency in selecting biomarkers for predicting binary and multiple disease outcomes using mass spectrometry data in cancer research. Furthermore, we develop a robust and reproducible method, the "Three-stage Homogeneous Ensemble Feature Selection (HEFS)", for biomarker identification to enhance clinical validation. Our primary interest is addressing the key following questions: (1) Does this homogeneous ensemble feature selection approach find consistent and reproducible biomarkers? (2) Are the protein markers predictive of the HRD/gene mutation status or mRNA status?

This paper is organized as follows. In Section 2, various ensemble machine learning approaches, including HEFS, are presented and implemented on two ovarian cancer proteogenomic datasets for binary HRD and multiple-mRNA-class predictions, as well as protein importance ranking. Results are presented in Section 3. Discussion and conclusions are provided in Section 4.

2. Materials and Methods

2.1. Machine Learning Algorithms

Ensemble machine learning algorithms include error-correcting output coding, bagging, boosting, and stacking [24–27]. These approaches could combine multiple independent ML and statistical approaches to construct a set of classifiers into a single predictive model and then classify new data points by taking a (weighted) vote of their predictions. The reasoning behind such approaches is that all machine-learning methods are biased towards identifying method-specific patterns and features. Thus, combining multiple learners can produce better and more robust predictions for boosting in accuracy compared to an individual learner or model.

In this study, the following feature selection approaches were compared and tested prior to ensemble [24,30–33]: (1) Median: it utilizes the non-parametric Mann–Whitney U test (p values); (2) Spearman and Pearson r: it selects features that are highly correlated with the outcome/dependent variable, but have a low correlation with other features to avoid multicollinearity; (3) LogReg: it uses standardized β-coefficients of logistic regression (LR)

representing the importance measure and comparability between the different ranges of features; (4) Naive Bayes; (5) Random Forest (RF): ensemble of multiple decision trees based on the classification and regression tree (CART) algorithm; cforest is one type of RFs that uses conditional trees for classification and regression; (6) Neural network (NN). Moreover, since protein features may highly correlate in their expressions, Principal Component Analysis (PCA) was applied to all proteins, and then the constructed PCs were used in LR and RFs to see if there were classification accuracy differences either with PCA or without PCA.

Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values were used as performance evaluation criteria [27]. In LR models, the selected features with a leave-one-out cross validation (LOOCV) scheme were applied, followed by training an LR model with all available features to compare the two LR models based on their ROC curves and AUC values. In RFs, error rate-based and AUC-based feature selection and Gini index were used as performance evaluation criteria; the error rate measures the difference before and after permuting the class variable depending on the underlying trees. AUC is computed for each tree before and after permuting a feature for variable- importance measure. The Gini index measures the node impurity in the trees. Overall, Error rate RF, Gini RF, Error rate cforest, and AUC cforest were compared for their protein identification ability and prediction accuracy [25,27].

2.2. Ensemble Approaches

To improve each feature selection method discussed above, including Median, Spearman, Pearson, LogReg, Error-rate Random Forest, Gini Random Forest, Error-rate cforest, AUC cforest, ensemble learning was applied with the following steps to build the engine of the caret from each model. The flowchart in Figure 1 provides a structured outline of the ensemble process, where each step is described in sequence, and the arrows indicate the flow and direction of the process.

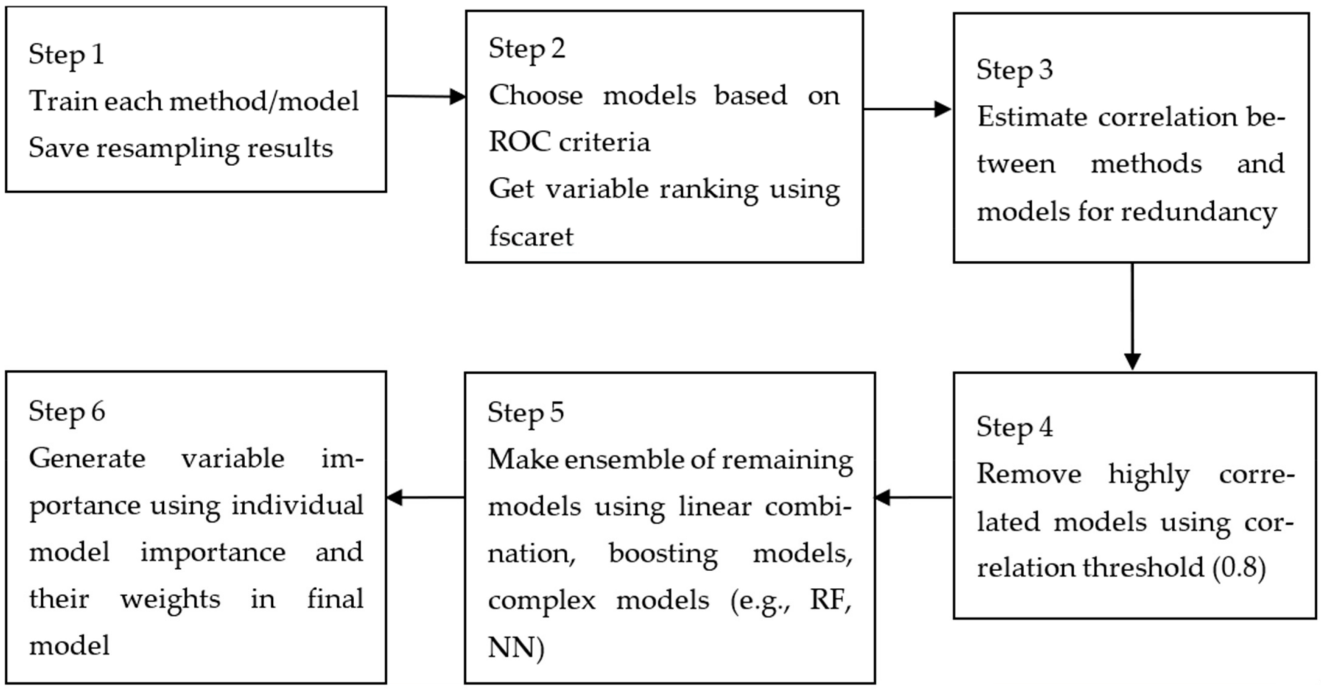

Figure 1. Flowchart of the ensemble process: training the models, selecting the best ones, handling redundancy, forming an ensemble, and finally generating variable-importance scores.

2.3. Three-Stage Homogeneous Ensemble Feature Selection

To overcome the reproducibility issues of the selection and further improve prediction accuracy [34–37], we refined the above ensemble process and generalized three-stage HEFS for both identifying consistent biomarkers and reducing errors. The flowchart in Figure 2 provides the steps involved in the biomarker selection and ML model comparison process.

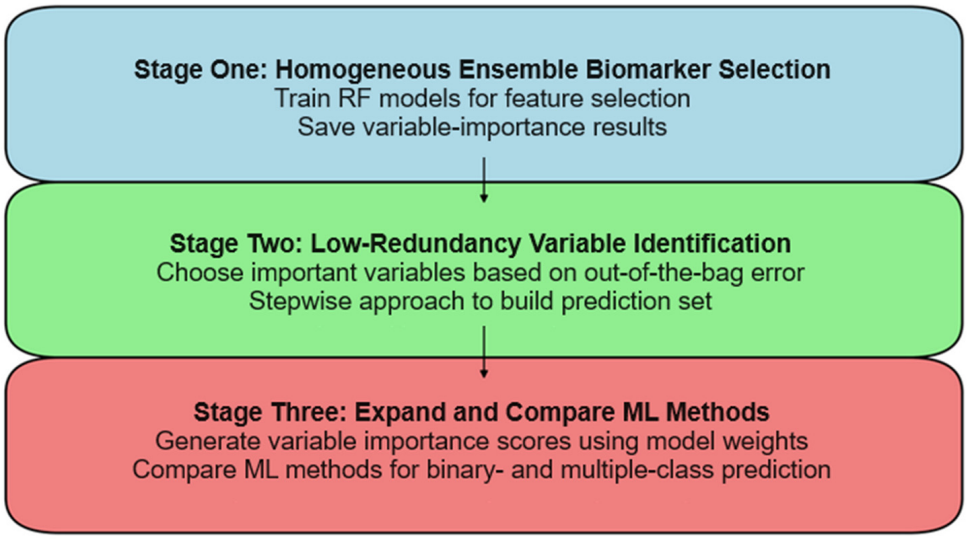

Figure 2. The three stages of HEFS include training the models, selecting important features, handling redundancy, forming an ensemble, and generating variable-importance scores.

The flowchart begins with Stage One, where a homogeneous ensemble biomarker selection based on a random forest approach was utilized to identify important biomarkers, even with some redundancy. At this stage, RF models were trained for feature selection, and variable-importance results were saved.

In Stage Two we refined this list to a small number of very low redundancy variables sufficient for better prediction. At this stage, two steps were taken, as follows: (1) We chose a set of important variables based on the out-of-bag error (interpretation set). (2) We utilized the stepwise approach for the interpretation set to build the prediction set.

Stage Three involved using the refined variable set from Stage One and Stage Two to expand and compare various ML methods, emphasizing the importance of ML comparison and validation. Variable-importance scores were generated using individual model importance, and their weights were saved in the final model. The various ML methods to be compared at this stage included the following: (1) gausprLinear (Gaussian processes for regression and classification with a linear kernel function); (2) gausprRadial (Gaussian processes for regression and classification with a radial-basis kernel function); (3) LogitBoost (Boosted Logistic Regression); (4) MLP (Multi-Layer Perceptron); (5) RF, parRF (Parallel Random Forest), wsrf (Weighted Subspace Random Forest); (6) SVM (Support Vector Machine); (7) treebag (Bagged Classification and Regression Tree).

To validate the stability of HEFS for reliable results, and to provide unbiased prediction accuracy, resampling with k-fold cross validation, leave-one-out cross validation, bootstrapping, and permutations was tested and employed [24,31–33,38]. We used a 10-fold cross-validation method in which the data were randomly divided into 10 equal subsets or "folds". The model was trained and validated ten times, each time using a different fold as the validation set and the remaining nine folds as the training set. During each iteration, the model was trained on the training set and evaluated on the validation set. This produced a performance metric (such as accuracy, precision, recall, etc.) for each of the 10 iterations. The performance metrics from the 10 iterations were averaged to obtain a final estimate of the model's performance. This average provided a robust estimate of how the model is expected to perform on unseen data. Ten-fold cross validation was

chosen due to its common use in ML and its suitability for our sample size, balancing computational feasibility and the trade-off between bias and variance in estimating the model performance.

We also evaluated the training data for the effect of the model tuning parameters on performance to guide the choice of the tuning parameter values. The training performance of all methods with automatic parameter tuning was also considered and tested since the process produced a profile of performance measures and the tuning parameters associated with the best measure value; then, the "optimal" model was chosen across these parameters. Permutation tests were further conducted to ensure the robustness of the resulting model. For instance, RF methods were run 100 times and averaged over the number of runs. An evaluation of the stability of feature importance was conducted by a bootstrapping algorithm.

The evaluation metrics for the predictions were as follows:

True Positive (TP): number of patients correctly identified as having the disease.
False Negatives (FN): number of patients who had the disease but were incorrectly identified as not having it.
True Negative (TN): number of patients correctly identified as not having the disease.
False Positives (FP): number of patients who did not have the disease but were incorrectly identified as having it.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}},\ \text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}},\ \text{PredictionAccuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$$

$$\text{Kappa Statistic}(\kappa) = \frac{\text{P}_\text{o}-\text{P}_\text{e}}{1-\text{P}_\text{e}} \quad \text{where } \text{P}_\text{o} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{TN}+\text{FP}+\text{FN})}, \text{ and } \text{P}_\text{e} = + \frac{[(\text{TP}+\text{FN})\times(\text{TP}+\text{FP})+(\text{FP}+\text{TN})\times(\text{FN}+\text{TN})]}{(\text{TP}+\text{TN}+\text{FP}+\text{FN})^2}$$

P_o is the observed agreement (accuracy), and P_e is the expected agreement by chance. Both are calculated based on TP, TN, FP, FN.

These metrics are included in the caret package in the Comprehensive R Archive Network (CRAN) (https://topepo.github.io/caret) (accessed on 15 June 2021) [26]. The results are reported based on the selected features/proteins. Additional analyses were conducted in R for ML with ensemble approaches and in SAS for data pre-processing.

3. Results

3.1. Datasets

MS proteomic ovarian cancer data were obtained from the Clinical Proteomic Tumor Analysis Consortium (NIH/NCI) and The Cancer Genome Atlas (TCGA), which include the genomic and transcriptomic characterizations of ovarian high-grade serous carcinomas (HGSCs) and 9606 global proteins measured from MS [5,8,39]. Two proteogenomics datasets were studied that were generated under similar experimental settings, with the binary disease status defined based on gene mutation and multiple mRNA classes. Primary outcomes:

(1) Binary outcomes were putative homologous recombination deficiency (HRD)-positive or -negative. HRD-positivity was defined by the presence of germline or somatic BRCA1 or BRCA2 mutations, BRCA1 promoter methylation, or homozygous deletion of PTEN. One hundred twenty-two serious ovarian carcinoma patient samples (67 HRD-positive, 55 HRD/-negative) with 9606 proteins were included.
(2) Five mRNA classes with 396 samples, as follows: differentiated (75), proliferative (72), immunoreactive (84), mesenchymal (75), and unknown (90). Figure 3 provides a general framework and a workflow of the datasets and proposed analytical procedures.

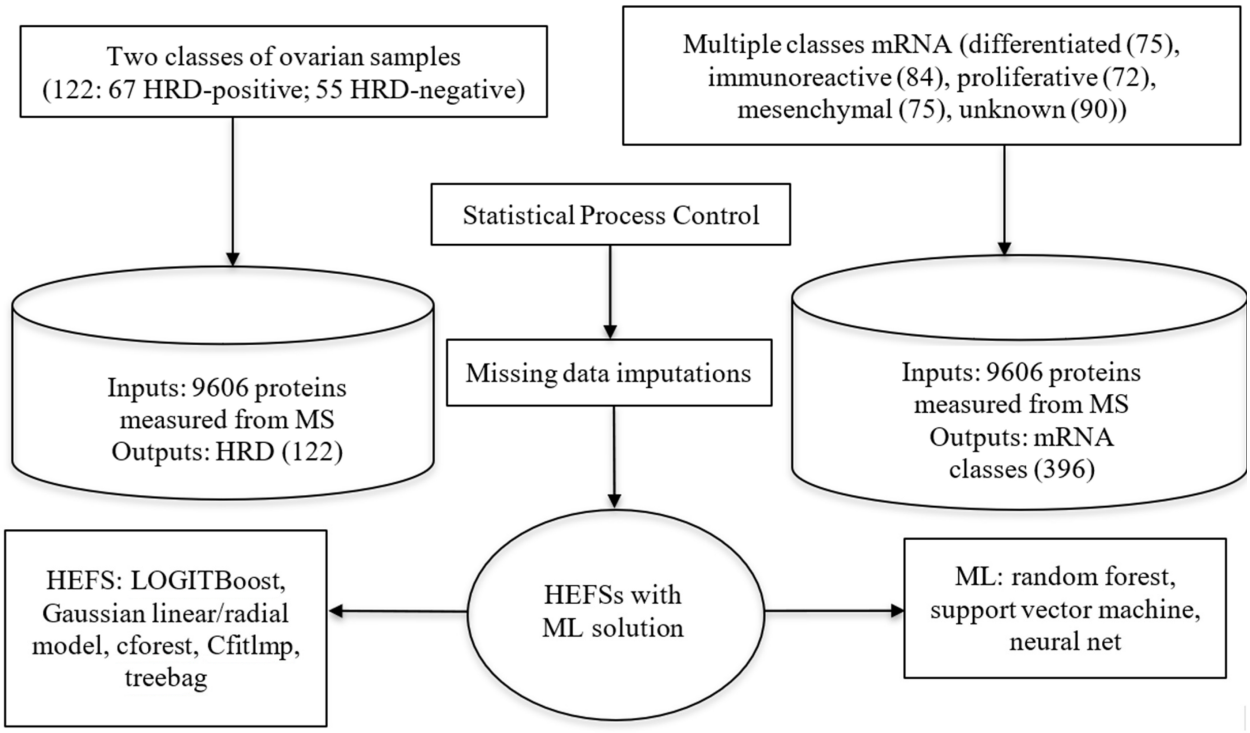

Figure 3. Machine learning solutions with ensemble feature selections for mass spectrometry data prediction.

3.2. Data Pre-Processing

Statistical process control for data quality examination (i.e., correcting technical variation, examining heterogeneity, high percentage of missing data, strong positive skewness, large proportions of zero) through measurement system analysis and process screening were conducted prior to our proposed approaches (see Figure 4). Variations and measure shifts were compared for HRD-positive (67) and HRD-negative (55) samples and glycosite versus nonglycosite samples. For non-glycosite samples, the largest upshift was found for the HRD-positive sample TCGA-29-1698-01A-01. Protein distributions and variations (CV: coefficient of variations) were also examined for outliers and irregularly distributed variables. Data transformation (transpose) was conducted due to high dimensionality with few samples (9606 proteins, 122 patients, with HRD-positive and -negative samples; 396 samples with five mRNA known classes).

Missing data evaluation and imputations: missing value patterns were examined. There was from 15% to 33% of missing values in the protein HRD status data. Several multiple imputations algorithms to impute the missing values (either biological or technical) were tested and compared [24,26] as follows: (1) Multivariate Normal Imputation (MNI), least squares prediction from non-missing variables; (2) Multivariate Singular Value Decomposition (SVD) (e.g., SVD took 29.08 s; MNI 6.32 s); (3) mixture model with clustering-based imputing; (4) Neural Network imputing. SVD provided better imputation accuracy (RMSE), classification error, and execution time. The multivariate SVD results were further used for the following ML approaches.

3.3. HEFS for Protein Ranking and Prediction Accuracy

Tables 1–3 provide prediction accuracies according to various metrics (kappa statistics for the agreement, sensitivity, specificity) resulting from different ML methods and HEFS approaches. The results revealed marked differences in prediction accuracies among different statistical or ML models for both binary and multi-class classifications, as discussed in Sections 2.1 and 2.2.

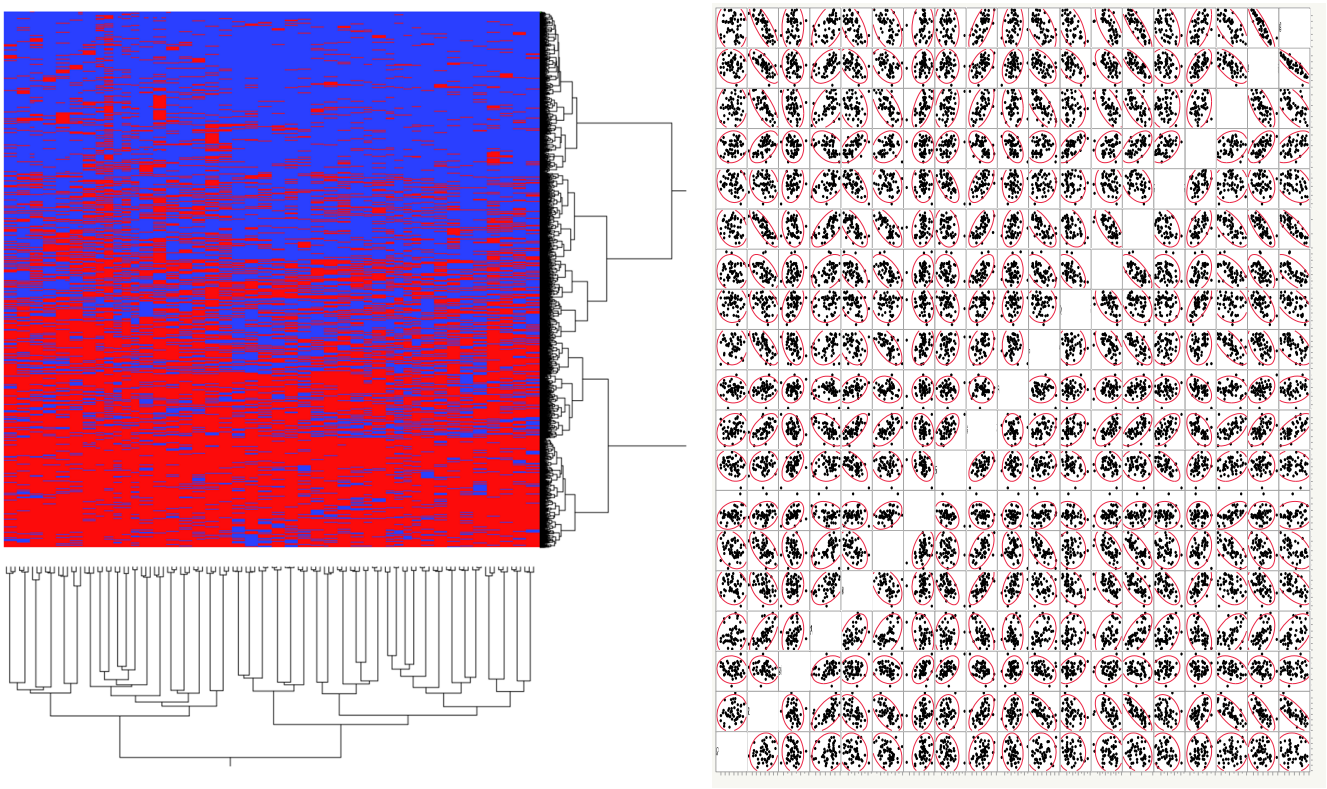

Figure 4. Left: Missing pattern examinations indicate that between 15% and 33% of expression missing occurred. **Right**: Multivariate scatter analysis shows the high correlations among the measured proteins.

For the binary HRD outcomes shown in Table 1, using the ensemble approaches proposed in Section 2.2 and with 34 selected interpretation proteins, LogitBoost and RF performed best, with 72% and 71% prediction accuracy, respectively. However, with our proposed three-stage HEFS approach described in Section 2.3 and using nine refined important prediction proteins with a sample size of 122, the prediction accuracy improved. The best ensemble ML method (Treebag) achieved 83% accuracy, 85% sensitivity, and 81% specificity. The sensitivity of the compared ensembling ML approaches was higher than the specificity for binary HRD class predictions (Tables 1 and 2). Regarding the performance of the various ML models with the proposed HEFS, RFs and NNs provided better prediction accuracies than simple Naive Bayes or logistic models for both binary HRD status and the five mRNA classes (see Table 3).

Table 1. Performance comparison of different ML methods with HEFS for binary HRD classes (34 interpretation proteins).

Model	Accuracy	95% CI	Kappa	Sensitivity	Specificity
gaussprLinear	0.61	(0.4, 0.8)	0.21	0.65	0.56
gaussprRadial	0.69	(0.5, 0.8)	0.36	0.85	0.50
LogitBoost	0.72	(0.5, 0.8)	0.44	0.70	0.75
Mlp	0.61	(0.4, 0.8)	0.20	0.70	0.50
mlpML	0.64	(0.5, 0.8)	0.26	0.70	0.56
parRF	0.69	(0.5, 0.8)	0.36	0.85	0.50
pcaNNet	0.61	(0.4, 0.8)	0.22	0.60	0.63
RF	0.71	(0.5, 0.8)	0.37	0.80	0.56
svmRadial	0.61	(0.4, 0.8)	0.18	0.80	0.38
Treebag	0.69	(0.5, 0.8)	0.39	0.70	0.69
Wsrf	0.58	(0.4, 0.7)	0.16	0.60	0.56

Table 2. Performance comparison of different ML methods with HEFS for binary HRD classes (9 prediction proteins).

Model	Accuracy	95% CI	Kappa	Sensitivity	Specificity
gaussprLinear	0.69	(0.5, 0.8)	0.38	0.75	0.63
gaussprRadial	0.75	(0.6, 0.9)	0.47	0.95	0.50
LogitBoost	0.58	(0.4, 0.7)	0.14	0.70	0.44
Mlp	0.69	(0.5, 0.8)	0.38	0.75	0.63
mlpML	0.69	(0.5, 0.8)	0.37	0.80	0.56
parRF	0.75	(0.6, 0.9)	0.48	0.85	0.63
pcaNNet	0.67	(0.5, 0.8)	0.33	0.70	0.63
RF	0.78	(0.6, 0.9)	0.54	0.90	0.63
svmRadial	0.69	(0.5, 0.8)	0.34	0.95	0.38
Treebag	0.83	(0.7, 0.9)	0.66	0.85	0.81
Wsrf	0.78	(0.6, 0.9)	0.55	0.80	0.75

Table 3. Performance comparison of different ML methods with HEFS for five mRNA classes.

	Interpretation Set (197 Biomarkers)			Prediction Set (20 Biomarkers)		
Model	**Accuracy**	**95% CI**	**Kappa**	**Accuracy**	**95% CI**	**Kappa**
Extra trees	0.63	(0.5, 0.7)	0.54	0.65	(0.5, 0.7)	0.56
mlpML	0.61	(0.5, 0.7)	0.51	0.49	(0.3, 0.5)	0.36
mlWeightDecay	0.66	(0.6, 0.7)	0.57	0.54	(0.4, 0.6)	0.41
NB	0.41	(0.3, 0.5)	0.26	0.55	(0.4, 0.6)	0.43
Pam	0.56	(0.4, 0.6)	0.44	0.46	(0.4, 0.5)	0.32
parRF	0.63	(0.5, 0.7)	0.54	0.64	(0.5, 0.7)	0.55
pcaNNet	0.54	(0.4, 0.6)	0.42	0.56	(0.5, 0.6)	0.45
protoclass	0.58	(0.5, 0.7)	0.48	0.48	(0.4, 0.6)	0.35
RF	0.65	(0.5, 0.7)	0.56	0.64	(0.5, 0.7)	0.55
RRF	0.65	(0.5, 0.7)	0.56	0.65	(0.5, 0.7)	0.56
Wsrf	0.67	(0.6, 0.7)	0.58	0.63	(0.5, 0.7)	0.54

Figure 5 shows NN with MLP (three layers, 10 hidden nodes) and the HEFS-selected proteins for HRD two-class prediction, with blue and red representing HRD-positive and -negative samples, respectively. The corresponding prediction accuracies with AUC from training (top left) and testing (bottom left) were included, with a testing accuracy of 74%. We further ran PCA with composite protein predictors in the same NN. Figure 6 shows the top selected important proteins from HEFS for HRD class prediction; blue and red represent HRD-positive and -negative ovarian cancer, respectively. PCA with composite protein predictors provided improved prediction accuracy (74% versus 93.8%).

Figure 7 presents NN with MLP (three layers, 10 hidden nodes) for mRNA class prediction (396 mRNA samples, with a two-thirds training and one-third testing split) using PCA with the top selected proteins from our proposed three-stage HEFS. The overall prediction accuracy for multi-class classifications with 396 samples and five principal components (PCs) ranged from 75% to 96% for each of the five classes (see Figure 7). The differences in the accuracy of each class may be due to the varied sample size of each class, ranging from 72 to 90. The ensemble workflow identified a consistent set of top selected important markers out of 9606 proteins linked either to the binary HRD status or to the multiple mRNA classes (see Tables 4 and 5). For example, the top 19 important proteins overlapped, but the 20th protein varied slightly. These markers were associated with either the mRNA classes or the binary HRD status and could be further examined through functional and pathway analysis [40–46].

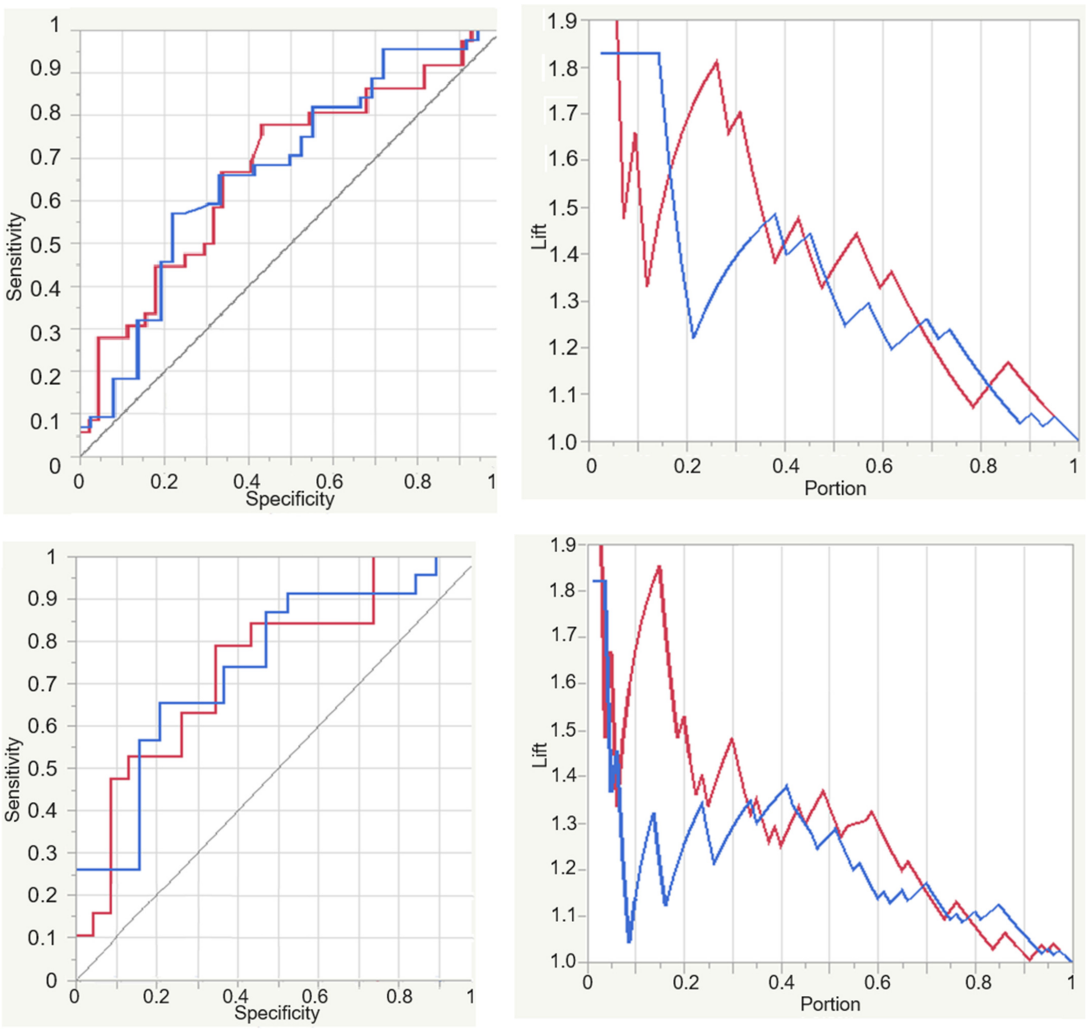

Figure 5. Neural Network with MLP (three layers, 10 hidden nodes) using 20 selected proteins for HRD two-class prediction: blue and red represent HRD-positive and -negative samples, respectively. **Left**: Prediction accuracy with ROC curves from training (**top**) and testing (**bottom**) sets; **right**: lift plot from training (**top**) and testing (**bottom**) sets.

Table 4. Top proteins (out of 9606) selected by the superior ML method (Treebag) with HEFS using the prediction set for the HRD classes.

1	calcium.uptake.protein.1..mitochondrial.isoform.1..calcium.uptake.protein.1..mitochondrial.isoform.2
2	acyl.coenzyme.A.thioesterase.2..mitochondrial..acyl.coenzyme.A.thioesterase.1
3	target.of.rapamycin.complex.2.subunit.MAPKAP1.isoform.2..target.of.rapamycin.complex.2.subunit.MAPKAP1.isoform.3..target.of.rapamycin.complex.2.subunit.MAPKAP1.isoform.1
4	peroxidasin.homolog.precursor
5	chromosome.19.open.reading.frame.29..chromosome.19.open.reading.frame.29
6	RING1.and.YY1.binding.protein
7	transmembrane.protein.9.precursor
8	arginyl.tRNA.synthetase..cytoplasmic

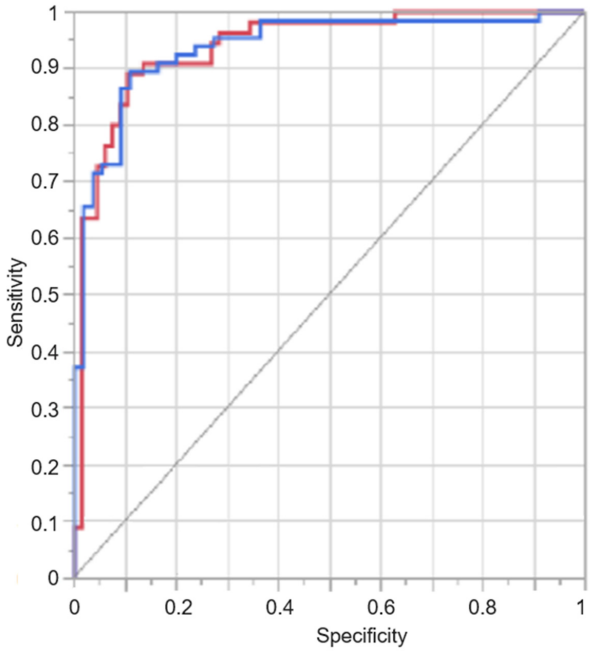

Figure 6. ROC curve with the selected important proteins from HEFS for binary HRD class predictions with PCA: blue represents HRD-positive ovarian cancer, and red represents HRD-negative ovarian cancer.

Table 5. Top proteins (out of 9606) selected by the superior ML method (wsrf) with HEFS using the prediction set for the multiple mRNA classes.

KH domain-containing, RNA-binding, signal transduction-associated protein.2
Protocadherin beta-8 precursor
Serine/threonine-protein kinase N3
Connector enhancer of kinase suppressor of ras 2 isoform 2; Connector enhancer of kinase suppressor of ras 2 isoform 3; Connector enhancer of kinase suppressor of ras 2 isoform 1
Protocadherin-9 isoform 1 precursor; Protocadherin-9 isoform 2 precursor
Ephrin type-A receptor 10 isoform 3
Tumor necrosis factor receptor superfamily member 11B precursor
Zinc finger protein 260
Family with sequence similarity 186 member A
Sushi domain-containing protein 1 precursor
Adenomatous polyposis coli protein isoform a; Adenomatous polyposis coli protein isoform b
Leucine-rich repeat transmembrane protein FLRT3 precursor
Glutamate receptor, ionotropic kainate 3 precursor
Pantothenate kinase 1 isoform alpha; Pantothenate kinase 1 isoform beta
AT-rich interactive domain-containing protein 3A
Dual-specificity protein phosphatase 22
Soluble scavenger receptor cysteine-rich domain-containing protein SSC5D isoform 1
Probable RNA polymerase II nuclear localization protein SLC7A6OS
Zinc finger protein 468 isoform 2
G kinase-anchoring protein 1 isoform b

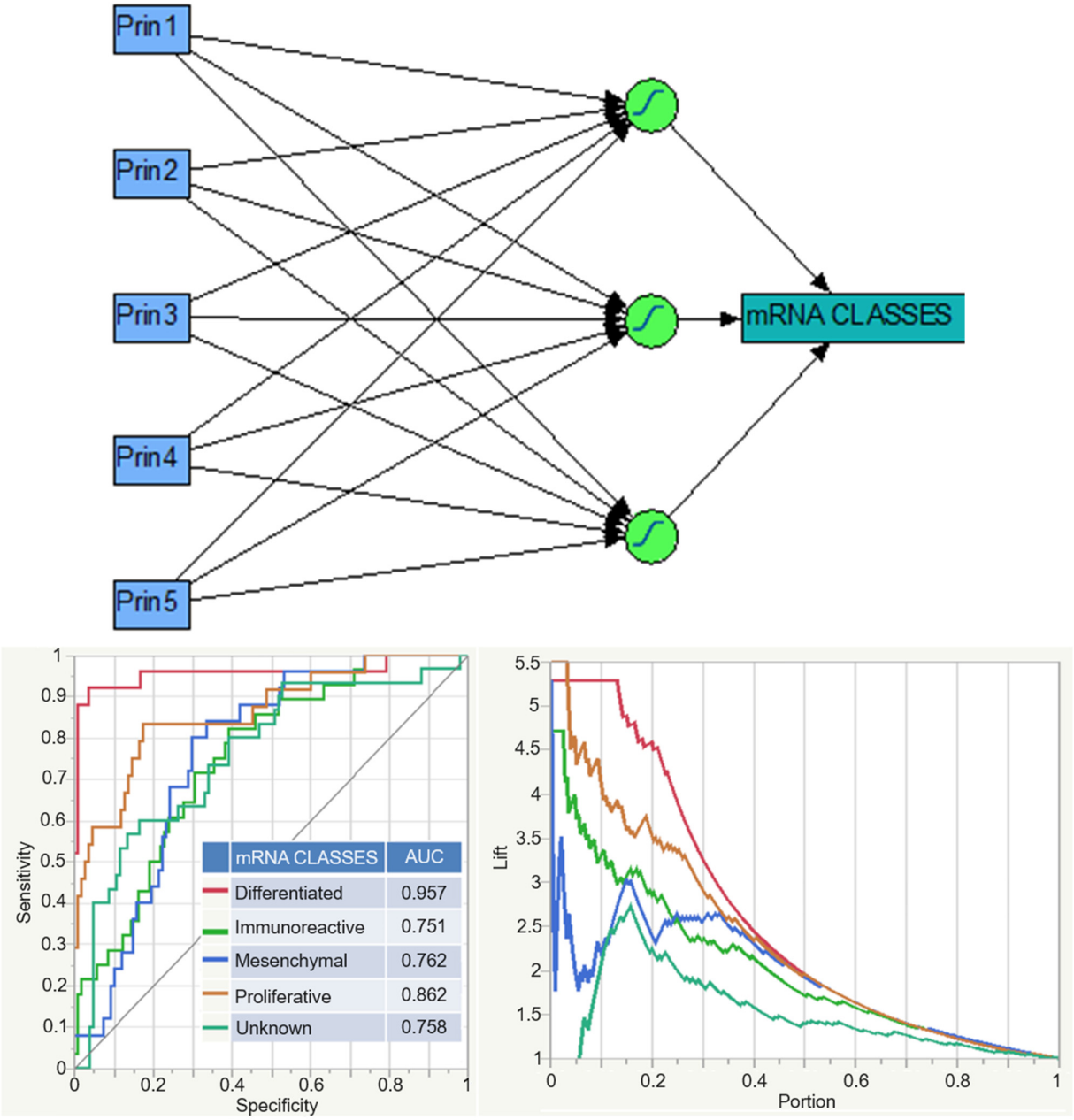

Figure 7. **Top**: Neural Network with MLP (three layers, 10 hidden nodes) and HEFS-selected important proteins with combined principal components for mRNA multiple-class prediction. **Bottom left**: prediction accuracies with ROC curves and AUC for five classes from testing set; **bottom right**: lift plot from testing set.

4. Discussion and Conclusions

This paper sought to evaluate ML with ensemble feature selection for stability, reproducibility of protein biomarker identifications, and prediction accuracy. We proposed and conducted a three-stage HEFS for biomarker identification, applied to ovarian protein data with binary and multiple-class outcomes. Our HEFS approach incorporates stability considerations into the algorithm design and has the advantage of alleviating and compensating for redundancy and reproducibility issues. The results showed that the ensemble approaches provided the stable selection of important biomarkers linked to ovarian cancer stages. Furthermore, prediction accuracy varied for both binary and multiple-class classifications using various ML approaches with the proposed HEFS method. Despite

the different prediction accuracies of the various models, HEFS identified consistent sets of proteins linked to both binary and multiple-class outcomes. Overall, the proposed ensemble approaches may hold promise with their better prediction power in addressing reproducibility issues. One potential drawback of HEFS is that it is computationally expensive (i.e., in the ensemble of RFs with a homogeneous feature selection algorithm using 10,000 trees), therefore requiring constant tuning of the model parameters for improved performance. More sophisticated ensemble models are much more susceptible to over-fitting than linear models, which generally require large sets of samples for training.

From biological and medical perspectives, the moderate correlations between protein and gene expression may be one reason why individual ML models did not produce high classification accuracies for HRD, in addition to sample size limitations. However, with the combined or principal components, the prediction accuracies were significantly improved. Additional subtypes of proteomic profiles with functional differences representing distinct subpopulations or hidden HRD stages within HRD-positive or -negative samples may explain how ensemble ML approaches performed better for multiple mRNA classes than for binary HRD status. The predictable proteins for HRD status identified from our proposed approaches did not include some well-known drug target proteins, such as those in the RAS family (i.e., HRAS, KRAS, NRAS). This may indicate some unique features of the proteomic data beyond the HRD-positive and -negative status, given the HRD status is defined by the presence of germline or somatic BRCA1 or BRCA2 mutations, BRCA1 promoter methylation, or homozygous deletion of PTEN [47–49].

The identification of these important markers is crucial for understanding underlying biological mechanisms and potential therapeutic targets. The identified high-ranked important protein markers/features from MS could be further examined to better understand the important biological processes and biomarkers influencing disease progression, and heterogeneity and the efficacy of platinum therapeutics [50–52]. The consistent set of markers can provide insights into the molecular pathways involved in ovarian cancer and help in the development of personalized treatment strategies. The functional analysis of these markers may reveal their roles in cell signaling, gene regulation, and other critical processes. Pathway analysis can further elucidate how these markers interact within the broader biological network, potentially identifying key nodes and interactions that are vital for disease progression. Further studies could focus on validating these markers in larger cohorts and different populations to ensure their reliability and applicability in clinical settings. Additionally, integrating these markers with other omics data, such as genomics and metabolomics data, could provide a more comprehensive understanding of ovarian cancer biology and improve the accuracy of predictive models.

Author Contributions: Conceptualization, Y.L. and A.G.; methodology, Y.L., A.G. and A.K.; formal analysis, Y.L., A.G. and E.K; writing original draft preparation, Y.L. and A.G.; writing review and editing, Y.L., A.K. and E.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We thank Johns Hopkins University for providing the data.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Walsh, C.S. Two Decades Beyond BRCA1/2: Homologous Recombination, Hereditary Cancer Risk and a Target for Ovarian Cancer Therapy. *Gynecol. Oncol.* **2015**, *137*, 343–350. [CrossRef] [PubMed]
2. Choi, J.; Ye, S.; Eng, K.H.; Korthauer, K.; Bradley, W.H.; Rader, J.S.; Kendziorski, C. IPI59: An Actionable Biomarker to Improve Treatment Response in Serous Ovarian Carcinoma Patients. *Stat. Biosci.* **2016**, *9*, 1–12. [CrossRef]
3. Tucker, S.L.; Gharpure, K.; Herbrich, S.M.; Unruh, A.K.; Nick, A.M.; Crane, E.K.; Coleman, R.L.; Guenthoer, J.; Dalton, H.J.; Wu, S.Y.; et al. Molecular Biomarkers of Residual Disease after Surgical Debulking of High-grade Serous Ovarian Cancer. *Clin. Cancer Res.* **2014**, *20*, 3280–3288. [CrossRef]

4. Ruggles, K.V.; Krug, K.; Wang, X.; Clauser, K.R.; Wang, J.; Payne, S.H.; Fenyo, D.; Zhang, B.; Mani, D.R. Methods, Tools and Current Perspectives in Proteogenomics. *Mol. Cell Proteom.* **2017**, *16*, 959–981. [CrossRef] [PubMed]
5. Zhang, H.; Liu, T.; Zhang, Z.; Payne, S.H.; Zhang, B.; McDermott, J.E.; Zhou, J.-Y.; Petyuk, V.A.; Chen, L.; Ray, D.; et al. Integrated Proteogenomic Characterization of Human High-Grade Ovarian Cancer. *Cell* **2016**, *166*, 755–765. [CrossRef]
6. Boja, E.S.; Rodriguez, H. Proteogenomic Convergence for Understanding Cancer Pathways and Networks. *Clin. Proteom.* **2014**, *11*, 22. [CrossRef]
7. Crutchfield, C.A.; Thomas, S.N.; Sokoll, L.J.; Chan, D.W. Advances in Mass Spectrometry-based Clinical Biomarker Discovery. *Clin. Proteom.* **2016**, *13*, 1. [CrossRef] [PubMed]
8. Wang, J.; Ma, Z.; Carr, S.A.; Mertins, P.; Zhang, H.; Zhang, Z.; Chan, D.W.; Ellis, M.J.C.; Townsend, R.R.; Smith, R.D.; et al. Proteome Profiling Outperforms Transcriptome Profiling for Coexpression Based Gene Function Prediction. *Mol. Cell Proteom.* **2017**, *16*, 121–134. [CrossRef] [PubMed]
9. Walsh, T.; Casadei, S.; Lee, M.K.; Pennil, C.C.; Nord, A.S.; Thornton, A.M.; Roeb, W.; Agnew, K.J.; Stray, S.M.; Wickramanayake, A.; et al. Mutations in 12 Genes for Inherited Ovarian, Fallopian Tube, and Peritoneal Carcinoma Identified by Massively Parallel Sequencing. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 18032–18037. [CrossRef]
10. Baggerly, K.A.; Morris, J.S.; Edmonson, S.R.; Coombes, K.R. Signal in Noise: Evaluating Reported Reproducibility of Serum Proteomic Tests for Ovarian Cancer. *J. Natl. Cancer Inst.* **2005**, *97*, 307–309. [CrossRef]
11. Baggerly, K.A.; Coombes, K.R. Deriving Chemosensitivity from Cell Lines: Forensic Bioinformatics and Reproducible Research in High-throughput Biology. *Ann. Appl. Stat.* **2009**, *3*, 1309–1334. [CrossRef]
12. Liang, Y.; Kelemen, A. Dynamic Modeling and Network Approaches for Omics Time Course Data: Overview of Computational Approaches and Applications. *Brief. Bioinform.* **2018**, *19*, 1051–1068. [CrossRef]
13. Wong, W.; Sue, A.C.-H.; Goh, W.W.B. Feature Selection in Clinical Proteomics: With Great Power Comes Great Reproducibility. *Drug Discov. Today* **2017**, *22*, 912–918. [CrossRef]
14. Goh, W.W.B.; Wong, L. Evaluating Feature-selection Stability in Next-generation Proteomics. *J. Bioinform. Comput. Biol.* **2016**, *14*, 1650029. [CrossRef]
15. Liang, Y.; Kelemen, A.; Kelemen, A. Reproducibility of Biomarker Identifications from Mass Spectrometry Proteomic Data in Cancer Studies. *Stat. Appl. Genet. Mol. Biol.* **2019**, *18*, 20180039. [CrossRef]
16. Koussounadis, A.; Langdon, S.P.; Um, I.H.; Haarrison, D.J.; Smith, V.A. Relationship Between Differentially Expressed mRNA and mRNA-protein Correlations in a Xenograft Model System. *Sci. Rep.* **2015**, *5*, 10775. [CrossRef]
17. Bannach-Brown, A.; Przybyla, P.; Thomas, J.; Rice, A.S.C.; Ananiadou, S.; Liao, J.; Macleod, M.R. Machine Learning Algorithms for Systematic Review: Reducing Workload in a Preclinical Review of Animal Studies and Reducing Human Screening Error. *Syst. Rev.* **2019**, *8*, 23. [CrossRef]
18. Capriotti, E.; Altman, R.B. A New Disease-specific Machine Learning Approach for the Prediction of Cancer-causing Missense Variants. *Genomics* **2011**, *98*, 310–317. [CrossRef]
19. Liang, Y.; Kelemen, A. Bayesian Models and Meta Analysis for Multiple Tissue Gene Expression Data Following Corticosteriod Administration. *BMC Bioinform.* **2008**, *9*, 354. [CrossRef]
20. Liang, Y.; Kelemen, A. Temporal Gene Expression Classification with Regularised Neural Network. *Int. J. Bioinform. Res. Appl.* **2005**, *1*, 399–413. [CrossRef]
21. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: Berlin/Heidelberg, Germany, 2009; ISBN 978-0387848570.
22. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
23. Zhang, C.-H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **2010**, *38*, 894–942. [CrossRef]
24. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: Berlin/Heidelberg, Germany, 2019; ISBN 978-1493979363.
25. Abeel, T.; Helleputte, T.; Van de Peer, Y.; Dupont, P.; Saeys, Y. Robust Biomarker Identification for Cancer Diagnosis with Ensemble Feature Selection Methods. *Bioinformatics* **2010**, *26*, 392–398. [CrossRef]
26. Neumann, U.; Genze, N.; Heider, D. EFS: An Ensemble Feature Selection Tool Implemented as R-package and Web-application. *Biodata Min.* **2017**, *10*, 21. [CrossRef]
27. Neumann, U.; Riemenschneider, M.; Sowa, J.P.; Baars, T.; Kälsch, J.; Canbay, A.; Heider, D. Compensation of Feature Selection Biases Accompanied with Improved Predictive Performance for Binary Classification by Using a Novel Ensemble Feature Selection. *BioData Min.* **2016**, *9*, 36. [CrossRef]
28. Cheng, L.-H.; Hsu, T.-C.; Lin, C. Integrating Ensemble Systems Biology Feature Selection and Bimodal Deep Neural Network for Breast Cancer Prognosis Prediction. *Sci. Rep.* **2021**, *11*, 14914. [CrossRef] [PubMed]
29. Budhraja, S.; Doborjeh, M.; Singh, B.; Tan, S.; Doborjeh, Z.; Lai, E.; Merkin, A.; Lee, J.; Goh, W.; Kasabov, N. Filter and Wrapper Stacking Ensemble (FWSE): A Robust Approach for Reliable Biomarker Discovery in High-dimensional Omics Data. *Brief. Bioinform.* **2023**, *24*, bbad382. [CrossRef] [PubMed]
30. Boulesteix, A.L.; Janitza, S.; Kruppa, J.; König, I.R. Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2012**, *2*, 493–507. [CrossRef]
31. Collins, G.S.; Moons, K.G.M. Reporting of Artificial Intelligence Prediction Models. *Lancet* **2019**, *393*, 1577–1579. [CrossRef]

32. Liang, Y.; Kelemen, A.; Tayo, B.O. Model-Based or Algorithms Based? Statistical Evidence for Diabetes and Treatments Using Gene Expression. *J. Stat. Methods Med. Res.* **2007**, *16*, 139–153. [CrossRef]
33. Chang, C.; Lin, C. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27. [CrossRef]
34. McShane, L.M.; Cavenagh, M.M.; Lively, T.G.; Eberhard, D.A.; Bigbee, W.L.; Williams, P.M.; Mesirov, J.P.; Polley, M.-Y.; Kim, K.Y.; Tricoli, J.V.; et al. Criteria for the Use of Omics-based Predictors in Clinical Trials: Explanation and Elaboration. *BMC Med.* **2013**, *11*, 220. [CrossRef] [PubMed]
35. Goh, W.W.B.; Wong, L. Advancing Clinical Proteomics via Analysis Based on Biological Complexes: A Tale of Five Paradigms. *J. Proteome Res.* **2016**, *15*, 3167–3179. [CrossRef] [PubMed]
36. Goh, W.W.B.; Wong, L. Advanced Bioinformatics Methods for Practical Applications in Proteomics. *Brief. Bioinform.* **2017**, *20*, 347–355. [CrossRef]
37. Wen, H.; Wang, H.-Y.; He, X.; Wu, C.-I. On the Low Reproducibility of Cancer Studies. *Natl. Sci. Rev.* **2018**, *5*, 619–624. [CrossRef] [PubMed]
38. Simon, R. Sensitivity, Specificity, PPV, and NPV for Predictive Biomarkers. *J. Natl. Cancer Inst.* **2015**, *107*, djv153. [CrossRef]
39. The Cancer Genome Atlas Research Network. Integrated Genomic Analyses of Ovarian Carcinoma. *Nature* **2011**, *474*, 609–615. [CrossRef]
40. Cavalcante, M.; Torres-Romero, J.C.; Lobo, M.D.P.; Moreno, F.B.M.B.; Bezerra, L.P.; Lima, D.S.; Matos, J.C.; Moreira, R.; Monteiro-Moreira, A.C. A Panel of Glycoproteins as Candidate Biomarkers for Early Diagnosis and Treatment Evaluation of B-cell Acute Lymphoblastic Leukemia. *Biomark. Res.* **2016**, *4*, 1. [CrossRef]
41. Ihle, N.T.; Byers, L.A.; Kim, E.S.; Saintigny, P.; Lee, J.J.; Blumenschein, G.R.; Tsao, A.; Liu, S.; Larsen, J.E.; Wang, J.; et al. Effect of KRAS Oncogene Substitutions on Protein Behavior: Implications for Signaling and Clinical Outcome. *J. Natl. Cancer Inst.* **2012**, *104*, 228–239. [CrossRef]
42. Logan, C.V.; Szabadkai, G.; Sharpe, J.A.; Parry, D.A.; Torelli, S.; Childs, A.-M.; Kriek, M.; Phadke, R.; Johnson, C.A.; Roberts, N.Y.; et al. Loss-of-function Mutations in MICU1 Cause a Brain and Muscle Disorder Linked to Primary Alterations in Mitochondrial Calcium Signaling. *Nat. Genet.* **2014**, *46*, 188–193. [CrossRef]
43. Perocchi, F.; Gohil, V.M.; Girgis, H.S.; Bao, X.R.; McCombs, J.E.; Palmer, A.E.; Mootha, V.K. MICU1 Encodes a Mitochondrial EF Hand Protein Required for Ca(2+) Uptake. *Nature* **2010**, *467*, 291–296. [CrossRef] [PubMed]
44. Robbins, P.F.; Lu, Y.-C.; El-Gamil, M.; Li, Y.F.; Gross, C.; Gartner, J.; Lin, J.C.; Teer, J.K.; Cliften, P.; Tycksen, E.; et al. Mining Exomic Sequencing Data to Identify Mutated Antigens Recognized by Adoptively Transferred Tumor-reactive T cells. *Nat. Med.* **2013**, *19*, 747–752. [CrossRef] [PubMed]
45. Sancak, Y.; Markhard, A.L.; Kitami, T.; Kovacs-Bogdan, E.; Kamer, K.J.; Udeshi, N.D.; Carr, S.A.; Chaudhuri, D.; Clapham, D.E.; Li, A.A.; et al. EMRE is an Essential Component of the Mitochondrial Calcium Uniporter Complex. *Science* **2013**, *342*, 1379–1382. [CrossRef] [PubMed]
46. Tran, E.; Robbins, P.F.; Lu, Y.-C.; Prickett, T.D.; Gartner, J.J.; Jia, L.; Pasetto, A.; Zheng, Z.; Ray, S.; Groh, E.M.; et al. T-Cell Transfer Therapy Targeting Mutant KRAS in Cancer. *N. Engl. J. Med.* **2016**, *375*, 2255–2262. [CrossRef] [PubMed]
47. Hathout, Y. Proteomic Methods for Biomarker Discovery and Validation. Are We There Yet? *Expert Rev. Proteom.* **2015**, *12*, 329–331. [CrossRef] [PubMed]
48. Alizadeh, A.A.; Aranda, V.; Bardelli, A.; Blanpain, C.; Bock, C.; Borowski, C.; Caldas, C.; Califano, A.; Doherty, M.; Elsner, M.; et al. Toward Understanding and Exploiting Tumor Heterogeneity. *Nat. Med.* **2015**, *21*, 846–853. [CrossRef]
49. Brenner, D.E.; Normolle, D.P. Biomarkers for Cancer Risk, Early Detection, and Prognosis: The Validation Conundrum. *Cancer Epidemiol. Biomark. Prev.* **2007**, *16*, 1918–1920. [CrossRef]
50. Tran, E.; Robbins, P.F.; Rosenberg, S.A. 'Final Common Pathway' of Human Cancer Immunotherapy: Targeting Random Somatic Mutations. *Nat. Immunol.* **2017**, *18*, 255–262. [CrossRef]
51. Schwarz, R.F.; Ng, C.K.Y.; Cooke, S.L.; Newman, S.; Temple, J.; Piskorz, A.M.; Gale, D.; Sayal, K.; Murtaza, M.; Baldwin, P.J.; et al. Spatial and Temporal Heterogeneity in High-grade Serous Ovarian Cancer: A Phylogenetic Analysis. *PLoS Med.* **2015**, *12*, e1001789. [CrossRef]
52. Tewari, D.; Java, J.J.; Salani, R.; Armstrong, D.K.; Markman, M.; Herzog, T.; Monk, B.J.; Chan, J.K. Long-term Survival Advantage and Prognostic Factors Associated with Intraperitoneal Chemotherapy Treatment in Advanced Ovarian Cancer: A Gynecologic Oncology Group Study. *J. Clin. Oncol.* **2015**, *33*, 1460–1466. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.