Ensemble Machine Learning Approaches for Proteogenomic Cancer Studies

Yulan Liang (liang@son.umaryland.edu)
University of Maryland, Balitimore https://orcid.org/0000-0001-6792-488X
Amin Gharipour
Griffith University
Erik Kelemen
University of Maryland, College Park
Arpad Kelemen
University of Maryland Baltimore

Research

Keywords: Ensemble Machine Learning, Homogeneous Ensemble Feature Selection, Proteogenomics, Ovarian Cancer

Posted Date: November 5th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-101902/v1

Ensemble Machine Learning Approaches for Proteogenomic Cancer Studies

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

[1]University of Maryland Baltimore,

655 West Lombard Street, Baltimore, MD 21201

[2]Griffith University, Gold Coast Campus, Australia

Parklands Dr, Southport QLD 4215, Australia

[3]University of Maryland, College Park

College Park, MD 20742

*corresponding author: liang@umaryland.edu

Abstract

Background: The identification of important proteins is critical for medical diagnosis and prognosis in common diseases. Diverse sets of computational tools were developed for omics data reductions and protein selections. However, standard statistical models with single feature selection involve the multi-testing burden of low power with the available limited 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 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 approach for both identifying proteins and improving prediction accuracy. This approach was implemented and applied to ovarian cancer proteogenomics data sets: 1) binary putative homologous recombination deficiency positive or negative; and 2) multiple mRNA classes (differentiated, proliferative, immunoreactive, mesenchymal, and unknown). We conducted and compared various machine learning approaches with homogeneous ensemble feature selection including random forest, support vector machine, and neural network for predicting both binary and multiple class outcomes. Various performance criteria including sensitivity, specificity, kappa statistics were used to assess the prediction consistency and accuracy.

Results: With the proposed three-stage homogeneous ensemble feature selection approaches, prediction accuracy can be improved with the limited sample through continuously reducing errors and redundancy, i.e. Treebag provided 83% prediction accuracy (85% sensitivity and 81%

specificity) for binary ovarian outcomes. For mRNA multi-classes classification, our approach provided even better accuracy with increased sample size.

Conclusions: Despite the different prediction accuracies from various models, homogeneous ensemble feature selection proposed identified consistent sets of top ranked important markers out of 9606 proteins linked to the binary disease and multiple mRNA class outcomes.

Keywords: Ensemble Machine Learning, Homogeneous Ensemble Feature Selection, Proteogenomics, Ovarian Cancer

I. 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 in order 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, diseases stages, and patient treatment heterogeneity) [7-9].

Identifications of protein and gene signatures from thousands of omics data generated from high throughput technologies have been challenging from both computational and biomedical perspectives [10-11]. Standard statistical marker selection methods with association analysis such as forward, backward, stepwise selection methods are based on the p-values of statistical models;

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 the features such as protein markers, along with small to moderate effects on disease status (often accumulated over time), may lead to unstable selections that may cause reproducibility issues [12-14]. Group selection techniques such as lasso or ridge regressions or Bayesian shrinkage models may overcome the above drawbacks [15]. One inherited challenge for the proteogenomic data is that some recent studies have indicated modest/moderate correlations between genes, mRNAs and proteins across different organisms (i.e., correlation coefficients 0.09 to 0.46 in multi-cellular organisms) [16].

Machine learning (ML) with wrapper methods and ensemble feature selection has the advantage to alleviate and compensate for some of the demerits of statistical approaches [17-18]. Ensemble learning models integrate predictions from multiple independent predictors to generate the strongest signals across predictors that rise to the top. Ensemble predictors consistently performed among the best across challenges and tended to be the most robust to noise in the datasets [19-20].

But different sets of ranked features from various ML methods could provide the same classification performance. Therefore, one key question is whether we can ensemble the ML feature selection approach that could result in consistent and reproducible biomarkers. This is important to reduce the burden of clinical validation. From the biomedical perspective, one important question is how well the protein markers are predictable for gene mutation or mRNA or disease status, and which biomarkers are associated with those status [21]. Last but not least, despite the power and flexibility of ML ensemble modeling, tuning various algorithms' parameters

could vary the prediction accuracy and selection results. How to best measure a predictor's relative importance in the model and enhance interpretability of ML is a challenge [22].

II. METHODS

A. Ensemble Machine Learning

Ensemble machine learning algorithms include bagging, boosting, stacking, and error-correcting output coding [19-22]. These approaches 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 based on the fact that all machine-learning approaches are biased to identify 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, prior to ensemble the following individual feature selection approaches were compared and tested [19, 24-27]: 1) Median: utilize the non-parametric Mann-Whitney-U Test (p values); 2) Spearman and Pearson r: select features that are highly correlated with the outcome variable, but the low correlation with other features to avoid multicollinearity; 3) LogReg: standardized β-coefficients of the logistic regression (LR) to represent the importance and comparability between the different ranges of protein features; 4) Naïve Bayes; 5) Neural network (NN); 5) Random forests (RFs): ensembles of multiple decision trees based on the classification and regression tree (CART) algorithm, cforest is one type of RF that uses conditional trees for classification and regression. Moreover, since protein features may highly correlate in their expressions, principal component analysis (PCA) was applied to all proteins, and then the

constructed PCAs were used in LR and RFs to see if there are classification accuracy differences either with PCA or without PCA.

Receiver Operating Characteristics (ROC) curves and Area Under Curve (AUC) values were used as performance evaluation criteria [21]. In LR models, the selected features with a leave-one-out cross-validation (LOOCV) scheme was applied, followed by training an LR model with all available feature in order to compare the two LR models based on their ROC curves and AUC values. In RFs, error-rate based, AUC based feature selection, and Gini-index were used as performance evaluation criteria for feature selection and importance rank.

The error-rate-measures the difference before and after permuting the class variable depending on the underlying trees, AUC is then computed for each tree before and after permuting a feature for the importance measure. The Gini-index measures the node impurity in the trees of RF [20-21]. Overall, Error-rate RF, Gini RF, Error-rate cforest, AUC cforest could be considered as error-correcting output coding for correcting bias and variances to improve the learning.

To reduce the redundancy and generate the feature importance, the following ensemble steps were conducted to build the engine of the caret from each model:

1- train each method/model, save all resampling results.

2- choose a set of methods/models using ROC criteria, get the variable ranking from them by using the fscaret from R for the scaling process.

3-estimate/find the correlation between the methods/models for redundancy

4- Use the resampling results to remove highly correlated models. Pearson's Correlation Coefficient is calculated for all possible pairs of trained models and then we removed the highly correlated models using 0.8 as a utilized threshold.

5- make the ensemble of remained models using a linear combination, boosting models, and complex models like RF (VSURF) or NNs for variable/feature selection.

6- generate the variable importance using individual model variable importance and their associated weight in the final model. The results of each individual method are normalized to a common scale to quantitative ensemble importance.

B. Three stage homogeneous ensemble feature selection

To further stabilize the selected features for the reproducibility and to improve prediction accuracy, we continue to refine the above ensemble process, and generalize three-stage homogeneous ensemble feature selection (HEFS) for both identifying biomarkers and reducing errors as follows:

In stage one: a homogeneous ensemble biomarker selection based on random forest approach is utilized to identify important biomarkers even with some redundancy.

In stage two: a smaller number of variables, with very low redundancy and sufficient for a better prediction, are identified.

In stage three: utilizing the selection results of stage one and stage two, expand and compare various more advanced ML methods including the following 1) gaussprLinear (Gaussian Processes For Regression And Classification with Linear kernel function); 2) gaussprRadial (Gaussian Processes For Regression And Classification with Radial Basis kernel function); 3) LogitBoost (Boosted Logistic Regression); 4) MLP, MLPML (Multi-Layer Perceptron, multiple layers); 5) RF, parRF (Parallel Random Forest), wsrf (Weighted Subspace Random Forest; 6) SVM (Support Vector Machine); 7) treebag (Bagged Classification And Regression Tree).

The following specific steps are employed:

1. train RF models for feature selection and save variable importance results
2. Choose a set of important variables based on the out-of-bag error (interpretation set)
3. Utilize the stepwise approach for interpretation set to build the prediction set
4. Utilize the interpretation set and the prediction set, and compare various ML methods for predicting binary and multiple classes.

The key questions: 1) Does this homogeneous ensemble feature selection approach find consistent and reproducible biomarkers? 2) Are the protein markers predictable for HRD/gene mutation status or mRNA status?

To validate the stability of the HEFS for reliable results, and to provide the unbiased prediction accuracy, resampling with k-fold cross-validation, bootstrapping and permutations are tested and employed [19, 25-28]. We used a 10-fold cross-validation method for a stratified random sample of the data into the training set and hold out a test set for the very end, and only use the training and parameter tuning. We also evaluated the training data for the effect of model tuning parameters on performance in order to guide which tuning parameter values should be chosen. The training performance of all methods with automatic parameter tuning is also considered and tested since the process produces a profile of performance measures and the tuning parameters associated with the best measure value, then the "optimal" model is chosen across these parameters. Permutation tests are further conducted for the robustness of the resulting model. For instance, RF methods are conducted 100 times and averaged over the number of runs. An evaluation of the stability of feature importance is conducted by a bootstrapping algorithm. Sensitivity and specificity are reported based on the selected feature/proteins, which has been included in the caret package in Comprehensive R Archive Network (CRAN) https://topepo.github.io/caret) [22]. Additional analyses were conducted in R for ML with ensemble approaches and SAS for data pre-processing.

III. RESULTS

A. Proteogenomic Datasets

MS proteomic ovarian cancer data was 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 carcinoma (HGSC) and 9606 global proteins measured from MS [5, 8, 33]. Two proteogenomics datasets were studied that were generated under similar experimental settings: both binary disease status defined based on gene mutation and multiple mRNA classes. Primary outcomes:

1) binary outcomes are putative homologous recombination deficiency (HRD): positive or negative. HRD positive is 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 non-HRD/negative) from 9606 proteins were included.

2) five mRNA classes with 396 samples: differentiated (75), proliferative (72), immunoreactive (84), mesenchymal (75), and unknown (90). Figure 1 provides a general framework and workflow of the data sets and proposed analytical procedures.

B. Data pre-processing

Statistical process control for data quality examination (i.e., correcting technical variation, examining heterogeneity, high percentage missing, strong positive skewness, large proportions of zero) through Measurement system analysis, process screening were conducted prior to our proposed approaches (See Figure 2). Variations and measure shifts are compared for HRD positive (67) versus HRD negative (55); Glycosite versus nonglycosite. i.e., for non-glyco the largest

upshift was found for HRD positive sample for TCGA-29-1698-01A-01. Protein distributions and variations (CV: Coefficient of variations) are also examined to examine outliers and irregularly distributed variables. Data (transpose) transformation was conducted due to high dimensionality with the few samples (9606 proteins, 122 patients with HRD: positives versus negative; 396 samples with 5 mRNA known classes).

Missing data evaluation and imputations: missing value patterns are examined. There are 15% to 33% missing in the protein-HRD status data. Several multiple imputations algorithms to impute missing (either biological or technical) were tested and compared [19, 22]: 1) Multivariate Normal Imputation (MNI): Least squares prediction from non-missing variables; 2) Multivariate Singular value decomposition (SVD). (e.g., SVD took 29.08 seconds; MNI 6.32 seconds); 3) Mixture model with clustering-based imputing; 4) neural network imputing. SVD provides better imputation accuracy (RMSE), classification error and execution time. Multivariate SVD results were further used for the follow-up ML approaches.

C. HEFS for protein rankings and prediction accuracy

Tables 1–3 provide the prediction accuracies according to the discrimination (kappa statistics for the agreement, sensitivity, specificity) resulting from different ML methods and HEFS approaches. Results revealed marked differences for the prediction accuracies with different individual statistical or ML models as discussed in section 2. PCA with composite protein predictors didn't provide better prediction accuracies, either with PCA or without PCA ranged from 45% (LR) to 58% (RFs with ctree).

Sensitivities of presented ML approaches are higher than specificities for binary HRD class predictions. RF and NN provided better prediction accuracies than simple Naïve Bayes. The

overall prediction accuracies for multiple classes' classification are comparable to predict binary HRD classes (see Table 3).

The key questions: 1) Are the protein markers selected through multiple refining steps proposed in section 2 improved the prediction accuracy for both binary HRD/gene mutation status or multiple mRNA classes? 2) Does the refined homogeneous ensemble feature selection approach find consistent and stable biomarkers?

With proposed three-stage HEF's approaches, prediction accuracy can be improved with the limited sample through continuously reducing errors and redundancy, i.e. Treebag provided 83% prediction accuracy (85% sensitivity and 81% specificity); RFs provided 71% prediction accuracy. For mRNA multi-classes classification, HEFS provided even better accuracy with increased sample size.

Figure 3 shows the top selected proteins (list on the top right) from mlpML (three layers, 10 hidden nodes) for HRD two classes prediction, 10-fold cross-validation was utilized given 122 samples. Fig. 4 provides Neural network (three layers, 10 hidden nodes) for mRNA classes prediction (396 mRNA sample, 1/3 testing, 2/3 training) with the top selected important proteins. Fig. 5 shows the top selected important proteins from HEFs for HRD classes' prediction with discriminant analysis: Red and green dots represent OV HRD Positive versus negative and normal samples, respectively.

Ensemble workflow proposed from section 2 identified a consistent set of top ranked important markers out of 9606 proteins linked either to the binary HRD status or multiple mRNA classes (see Tables 4 & 5), e.g., top 19 important proteins are overlapped but number 20 is slightly varied. Some of them i.e., peroxidasin. homolog. precursor; transmembrane. protein.9. precursor; calcium. uptake.protein.1..mitochondrial.isoform.1..calcium.uptake.protein.1..mitochondrial.isoform.2

linked to the mRNA classes or binary HRD status, and could be further examined through functional and pathway analysis [34-40].

IV. DISCUSSIONS AND CONCLUSIONS

This paper seeks to evaluate ML with ensemble feature selection for the protein biomarker identifications' stability and reproducibility in addition to prediction accuracy. We proposed and conducted three-stage HEFS for biomarker identification and applied the ovarian proteogenomics data. Our HEFS approach incorporates stability considerations into the algorithm design stage and has the advantage to alleviate and compensate the reproducibility issues. Results showed that the ensemble approaches provided the stable selection of the important biomarkers linked to ovarian cancer stages. Overall, ensemble approaches may hold promise with better prediction power for reproducibility issues.

One potential drawback of the utilized HEFS is that it is computationally expensive (i.e. in the ensemble of RF with homogeneous feature selection algorithm using 10000 trees), therefore requiring constant tuning of the model parameters for improved performance. More sophisticated ensemble models are much more susceptible to over-fitting than simple linear weights, which generally require large sets of samples to train.

From biological and medical perspectives, the potential correlations between protein expression and genes are moderate, which may be one reason why individual ML model did not result in high classification accuracies for HRD in addition to the sample size limitations. Additional subtypes of proteomic profiles with functional differences representing distinct subpopulations or hidden HRD stages within HRD positive or negative may explain how ensemble ML approaches perform

better for multiple mRNA classes than 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 RAS (i.e., HRAS, KRAS, NRAS), which may indicate some unique values of proteomic data beyond HRD positive vs negative, given HRD status is defined by the presence of germline or somatic BRCA1 or BRCA2 mutations, BRCA1 promoter methylation, or homozygous deletion of PTEN [41-43]. The identified high ranked important protein markers/features from MS could be further examined to better understand the important biological processes and biomarkers influencing the disease progression, heterogeneity, and efficacy of platinum therapeutics [44-46].

FUNDING

There is no funding support for the current study.

COMPETING INTERESTS

The authors declare that they have no competing interests.

ACKNOWLEDGMENT, DECLARATIONS AND CONSENT

Not applicable.

DATA SHARING

not applicable to this article as no datasets were generated, the public data available from NIH/NCI:

https://www.ncbi.nlm.nih.gov/projects/gap/cgibin/study.cgi?study_id=phs000892.v1.p

AUTHORS' CONTRIBUTIONS

YL & AG carried out study and methods' design, computing and manuscript writing; EK and AK participated the computing, implementations, and manuscript writing. All authors read and approved the final manuscript.

REFERENCES

1. Walsh CS. (2015) Two decades beyond BRCA1/2: Homologous recombination, hereditary cancer risk and a target for ovarian cancer therapy. Gynecol Oncol;137:343–50.
2. Choi J, Ye S, Eng KH, et al. (2016) IPI59: An actionable biomarker to improve treatment response in serous ovarian carcinoma patients. Statistics in Biosciences;9(1):1-12.
3. Tucker SL, Gharpure K, Herbrich SM, et al. (2014) Molecular biomarkers of residual disease after surgical debulking of high-grade serous ovarian cancer. Clin Cancer Res; 20:3280–8.
4. Ruggles KV, Krug K, Wang X, et al. Methods, tools and current perspectives in proteogenomics. Mol Cell Proteomics 2017;6(6):959-981.
5. Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, et a; (2016), Rodland KD, and the CPTAC investigators. Deep proteogenomic characterization of human ovarian cancer. Cell. 2016; 166: 755-765.
6. Boja,E.S,, Rodriguez, H. (2014). Proteogenomic convergence for understanding cancer pathways and networks Clin Proteomics. 2014; 11(1): 22. doi: 10.1186/1559-0275-11-22
7. Crutchfield, C.A., Thomas, S.N., Sokoll, L.J., Chan, D.W., (2016) Advances in mass spectrometry-based clinical biomarker discovery. Clin Proteomics.7;13:1
8. Wang J, Ma Z, Carr SA, Mertins P, Zhang H, et al. (2017). Proteome profiling outperforms transcriptome profiling for co-expression based gene function prediction. Mol Cell Proteomics, 16(1):121-134
9. Walsh T, Casadei S, Lee MK, Pennil CC, Nord AS, Thornton AM, et al. (2011). Mutations in 12 genes for inherited ovarian, fallopian tube, and peritoneal carcinoma identified by massively parallel sequencing. Proc Natl Acad Sci,108:18032–7.

10. Baggerly KA, Morris JS, Edmonson SR, Coombes KR. (2005) Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. J Natl Cancer Inst. 97(4):307-9.
11. Baggerly, Keith A., Coombes, Kevin R. (2009) Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. Ann. Appl. Stat. 3 (4), 1309--1334.
12. Liang, Y., Kelemen, A., (2017) Computational Dynamic Approaches for Temporal Omics Data with Applications to System Medicine., BioDataMining 10:20.
13. Wang W, Sue ACH, Goh WW (2017). Feature selection in clinical proteomics: with great power comes great reproducibility. Drug Discovery Today; 22:6, 912-918.
14. Wilson Wen Bin Goh and Limsoon Wong, (2016) Evaluating feature-selection stability in next-generation proteomics. J. Bioinform. Comput. Biol. 14, 1650029
15. Liang, Y., Kelemen, A. (2008). Bayesian models and meta analysis for multiple tissue gene expression data following corticosteriod administration. BMC Bioinformatics. 9:354.
16. Koussounadis, A. et al. (2015) Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. Sci. Rep. 5, 10775
17. Alexandra Bannach-Brown, Piotr Przybyła, James Thomas, Andrew S. C. Rice, Sophia Ananiadou, Jing Liao and Malcolm Robert Macleod, (2019).Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error, Systematic Reviews, 10.1186/s13643-019-0942-7, 8, 1
18. Capriotti E, Altman RB (2011) A new disease-specific machine learning approach for the prediction of cancer-causing missense variants. Genomics, 98: 310–317.
19. Kuhn, M. and Johnson, K. (2018) Applied Predictive Modeling, Springer

20. 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. 2010; 26(3):392–8.
21. Neumann U, Riemenschneider M, Sowa JP, Baars T, Kälsch J, Canbay A, Heider D. (2016) Compensation of feature selection biases accompanied with improved predictive performance for binary classification by using a novel ensemble feature selection. BioData Min.; 9(1):36.
22. Neumann, U., Genze, N. & Heider, D. (2017) EFS: an ensemble feature selection tool implemented as R-package and web-application. Biodata Mining 10, 21.
23. Crutchfield CA1, Thomas SN2, Sokoll LJ2, Chan DW2. Goh WW, Wong L. (2016) Evaluating feature-selection stability in next-generation proteomics. J Bioinform Comput Biol;14(5):1650029
24. Boulesteix AL, Janitza S, Kruppa J, König IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip Rev Data Min Knowl Discov. 2012; 2(6):493–507.
25. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet. 2019;393(10181):1577-1579.
26. Liang, Y., Kelemen, A., Tayo, B. O., (2007) Model Based or Algorithms Based? Gene Expression Based Statistical Methods to Find Evidence of Diabetes. Journal of Statistical Methods for Medical Research, 16(2): 139-153
27. Chang, C. & Lin, C. LIBSVM: A library for support vector machines. Acm Transactions on Intelligent Systems & Technology 2, 389–96 (2011).

28. Simon R.(2015) Sensitivity, specificity, PPV, and NPV for predictive biomarkers. J Natl Cancer Inst; 107(8):djv153

29. McShane LM, Cavenagh MM, Lively TG, et al. (2013) Criteria for the use of omics-based predictors in clinical trials: explanation and elaboration. BMC Med;11-220.

30. Wilson Wen Bin Goh, Limsoon Wong. (2016) Advancing Clinical Proteomics via Analysis Based on Biological Complexes: A Tale of Five Paradigms. Journal of Proteome Research 15:9, 3167-3179.

31. Wilson Wen Bin Goh, Limsoon Wong. (2017) Advanced bioinformatics methods for practical applications in proteomics. Briefings in Bioinformatics.

32. Haijun Wen, Hurng-Yi Wang, Xionglei He, Chung-I Wu, (2018) On the reproducibility of cancer studies, National Science Review, 5 (5), 619–624.

33. The Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. Nature; 474:609–15.

34. Cavalcante MS, Torres-Romero JC, Lobo MDP, et al. A panel of glycoproteins as candidate biomarkers for early diagnosis and treatment evaluation of B-cell acute lymphoblastic leukemia Biomark Res 2016;4(1)

35. Ihle NT, Byers LA, Kim ES, et al. 2012 Effect of KRAS Oncogene Substitutions on Protein Behavior: Implications for Signaling and Clinical Outcome. Journal of the National Cancer Institute;104(3):228–39.

36. 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., Bonthron, D. T., Pysden, K. A., et al. (2014). Loss-of-function mutations in MICU1 cause a brain and muscle disorder linked to primary alterations in mitochondrial calcium signaling. Nature Genet. 46: 188-193.

37. Perocchi, F., Gohil, V. M., Girgis, H. S., Bao, X. R., McCombs, J. E., Palmer, A. E., Mootha, V. K. (2010) MICU1 encodes a mitochondrial EF hand protein required for Ca(2+) uptake. Nature 467: 291-296.

38. Robbins PF, Lu YC, El-Gamil M, et al. (2013) Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. Nat Med;19:747-752.

39. 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., Calvo, S. E., Goldberger, O., Mootha, V. K. (2013) EMRE is an essential component of the mitochondrial calcium uniporter complex. Science 34 2: 1379-1382.

40. Tran E, Robbins PF, Lu YC, et al (2016). T-Cell Transfer Therapy Targeting Mutant KRAS in Cancer. N Engl J Med;375(23):2255-62.

41. Hathout Y. (2015) Proteomic methods for biomarker discovery and validation. Are we there yet? Expert Rev Proteomics;12(4):329-31.

42. Alizadeh AA, Aranda V, Bardelli A, et al. (2015) Toward understanding and exploiting tumor heterogeneity. Nature Methods 21:846-53. doi:10.1038/nm.3915

43. Brenner DE, Normolle DP. (2007) Biomarkers for cancer risk, early detection, and prognosis: the validation conundrum. Cancer Epidemiol Biomarkers Prev;16:1918–1920.

44. Tran E, Robbins PF, Rosenberg SA. (2017) 'Final common pathway' of human cancer immunotherapy: targeting random somatic mutations. Nat Immunol:18(3):255-262.

45. Schwarz RF, Ng CK, Cooke SL, et al. (2015) Spatial and temporal heterogeneity in high-grade serous ovarian cancer: a phylogenetic analysis. PLoS Med;12: e1001789.

46. Tewari D, Java JJ, Salani R, et al (2015). Long-term survival advantage and prognostic factors associated with intraperitoneal chemotherapy treatment in advanced ovarian cancer: a gynecologic oncology group study. J Clin Oncol;33:1460–66.

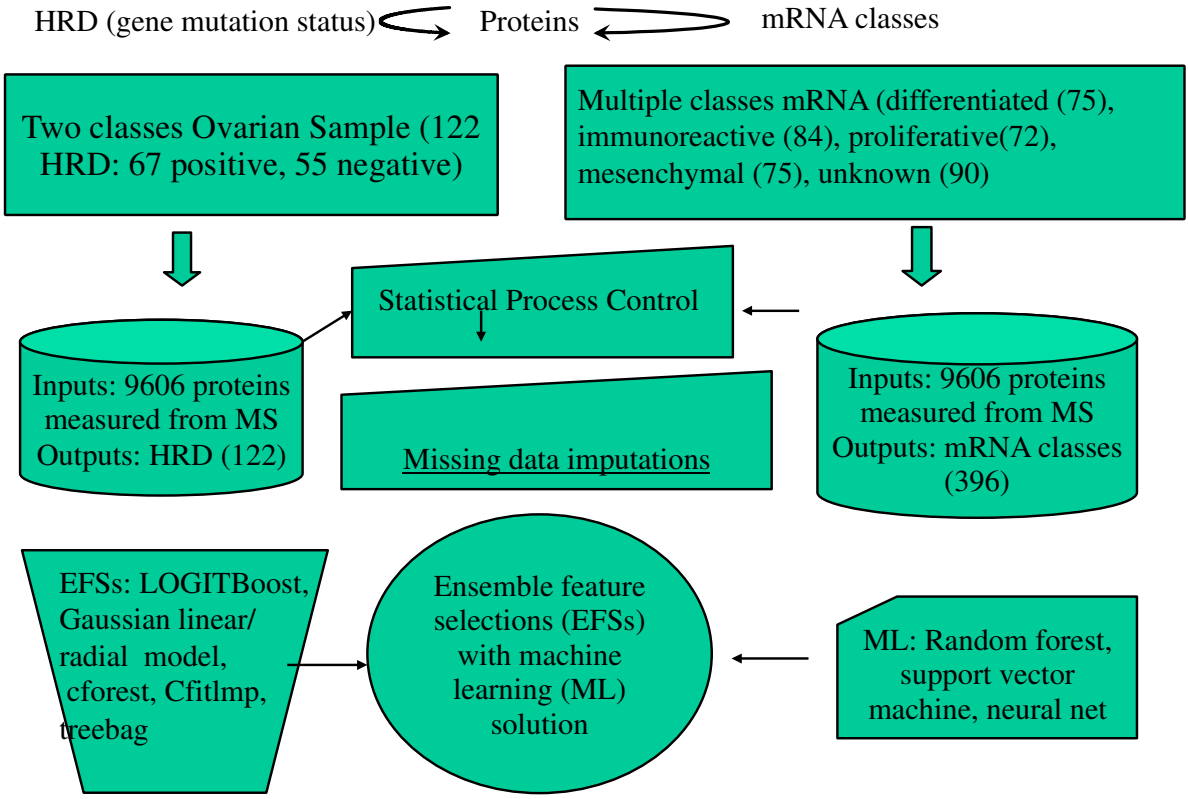

Figure 1. Machine learning Solutions with Ensemble Feature Selections for Proteogenomics Data Prediction

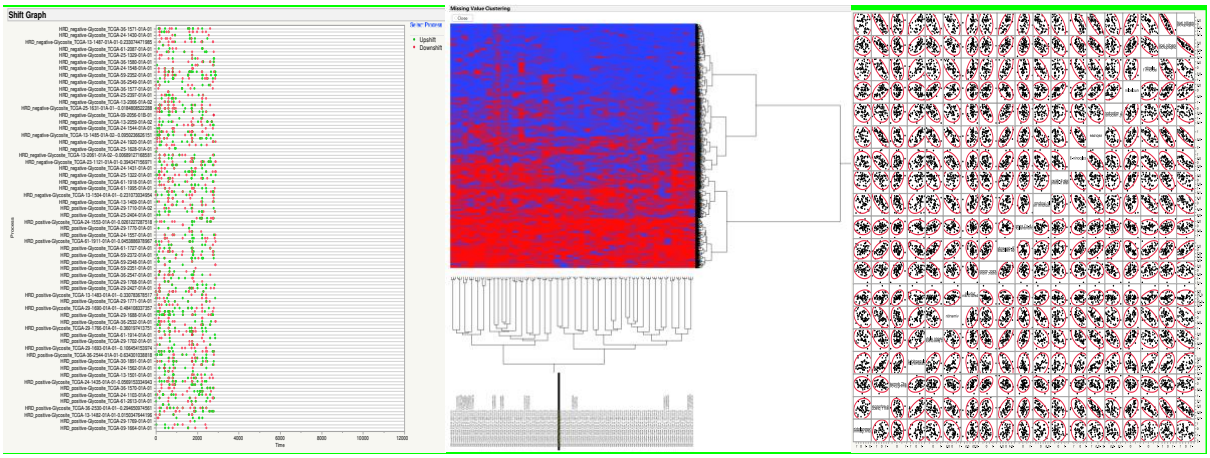

Figure 2. Left: Statistical process control for quality examination; **Middle:** Missing patterns examinations indicates between. 15% - 33% expression missing occurred; **Right:** Multivariate scatter analysis shows the high correlations among the measured proteins

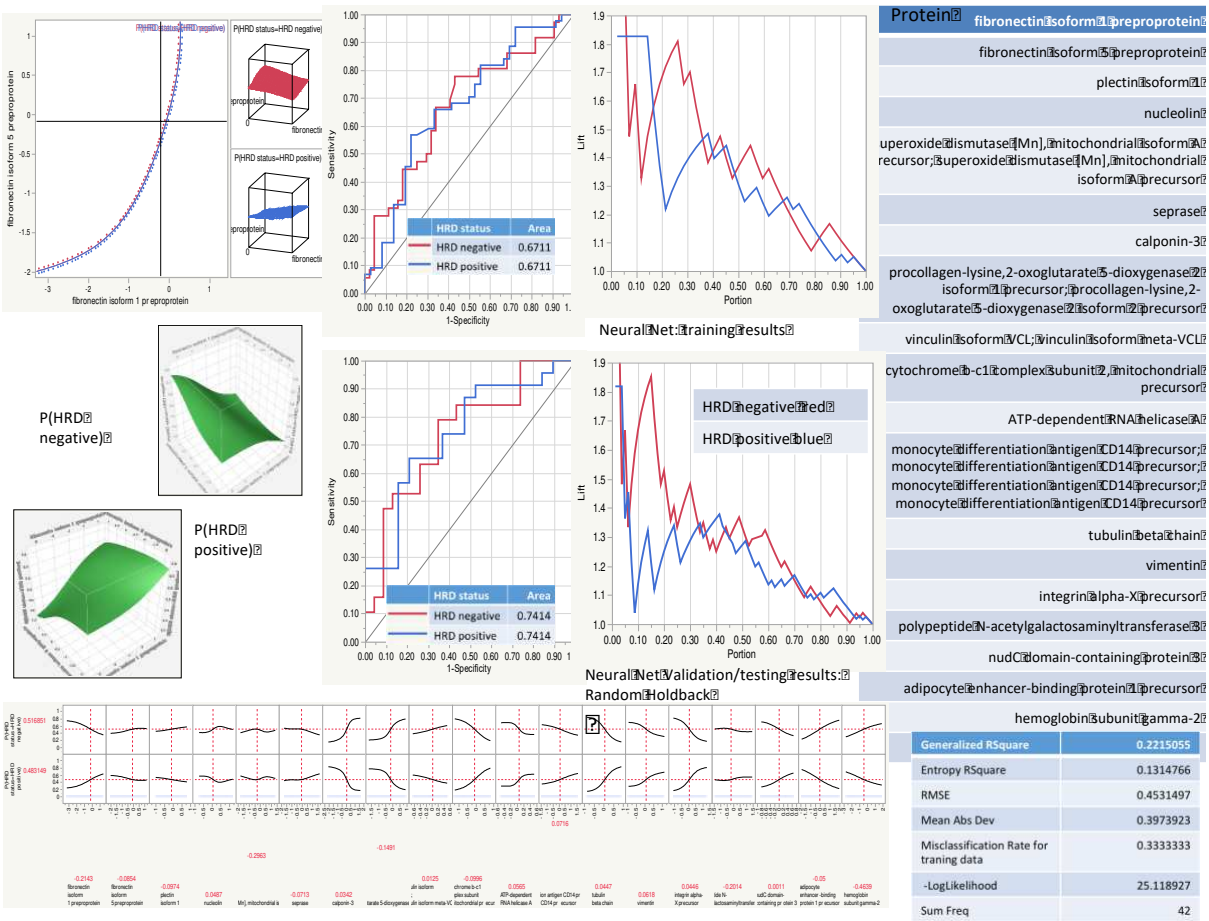

Figure 3. Top selected proteins using mlpML (NN: three layers, 10 hidden nodes) for HRD two classes prediction: Red and blue represent HRD Positive versus negative, respectively.

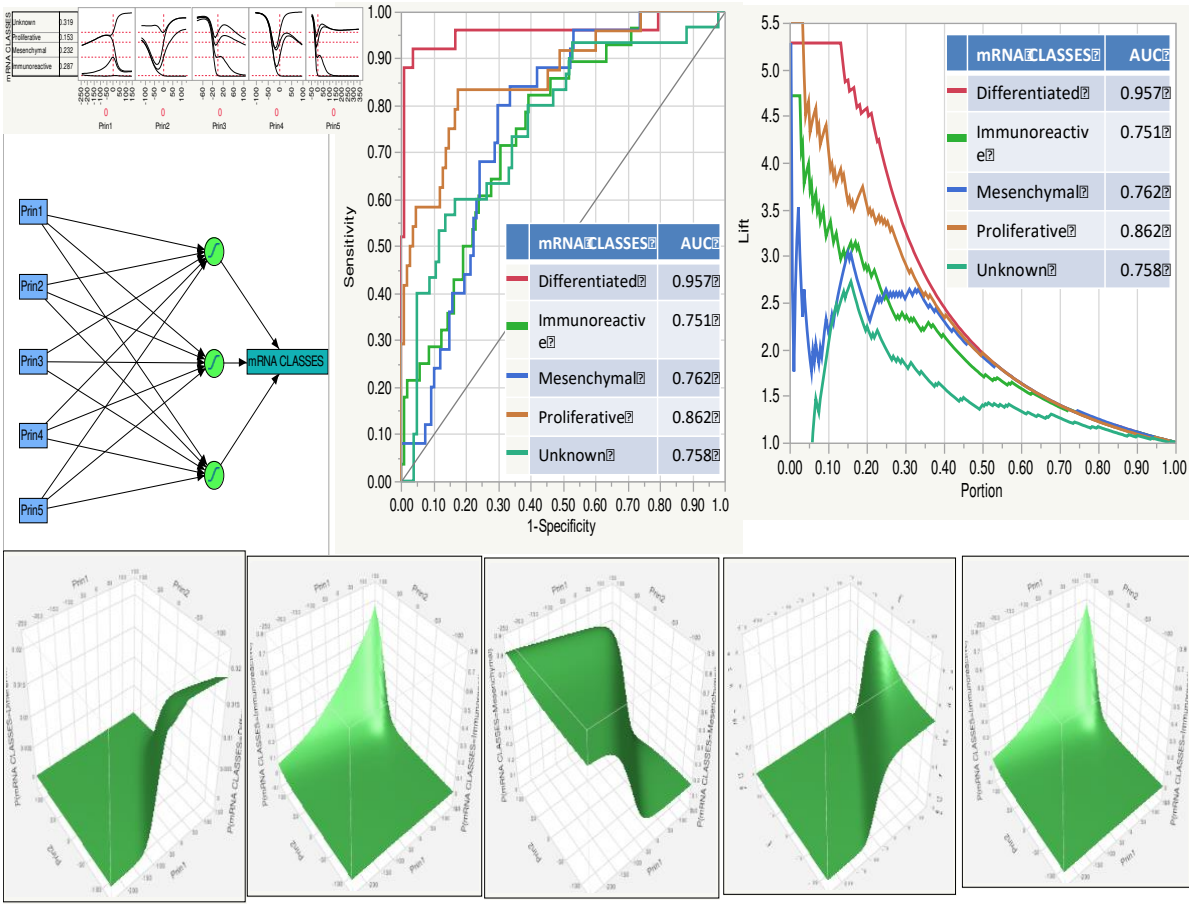

Figure 4. Top selected important proteins using NN (three layers, 10 hidden nodes) for mRNA classes prediction

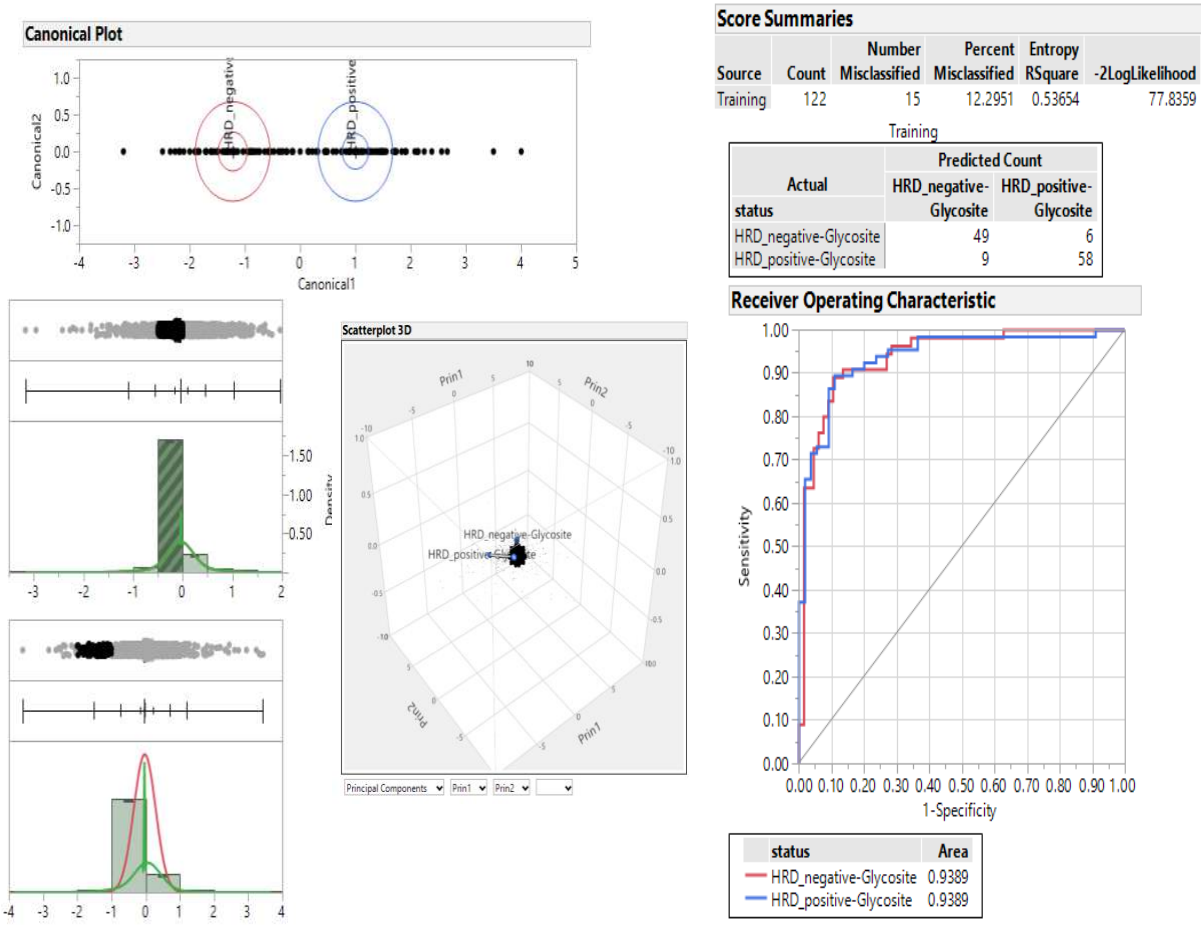

Figure 5. Top selected important proteins from HEFs for HRD classes' prediction compared with discriminant analysis: Red and green dots represent OV HRD Positive versus negative and normal samples, respectively.

Table 1. Performance comparison for different ML approaches with HEFS when Interpretation set (34 biomarkers) is applied for HRD classes

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 for different ML approaches with HEFS when Prediction set (9 biomarkers) is applied for HRD classes

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 for different ML with HEFS for different set of variables are applied for multiple mRNA classes

	Interpretation set (197 biomarkers)			Prediction set (20biomarkers)		
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 mlWeight	0.61	(0.5, 0.7)	0.51	0.49	(0.3, 0.5)	0.36
Decay	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

Table 4. Final ranking of top 8 proteins (out of 9606) selected by superior ML method (Treebag) with HEFS using prediction set for binary HDR 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

Table 5. Top 20 proteins (out of 9606) selected by superior ML method (wsrf) using prediction set for 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.sup
pressor.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;.zinc.finger.protein.260;.zinc.finger.protein.260;.zinc.finger.protein.26
0
family.with.sequence.similarity.186,.member.A
sushi.domain-containing.protein.1.precursor
adenomatous.polyposis.coli.protein.isoform.b;.adenomatous.polyposis.coli.protein.isoform.a;
.adenomatous.polyposis.coli.protein.isoform.b
leucine-rich.repeat.transmembrane.protein.FLRT3.precursor;.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

Figures

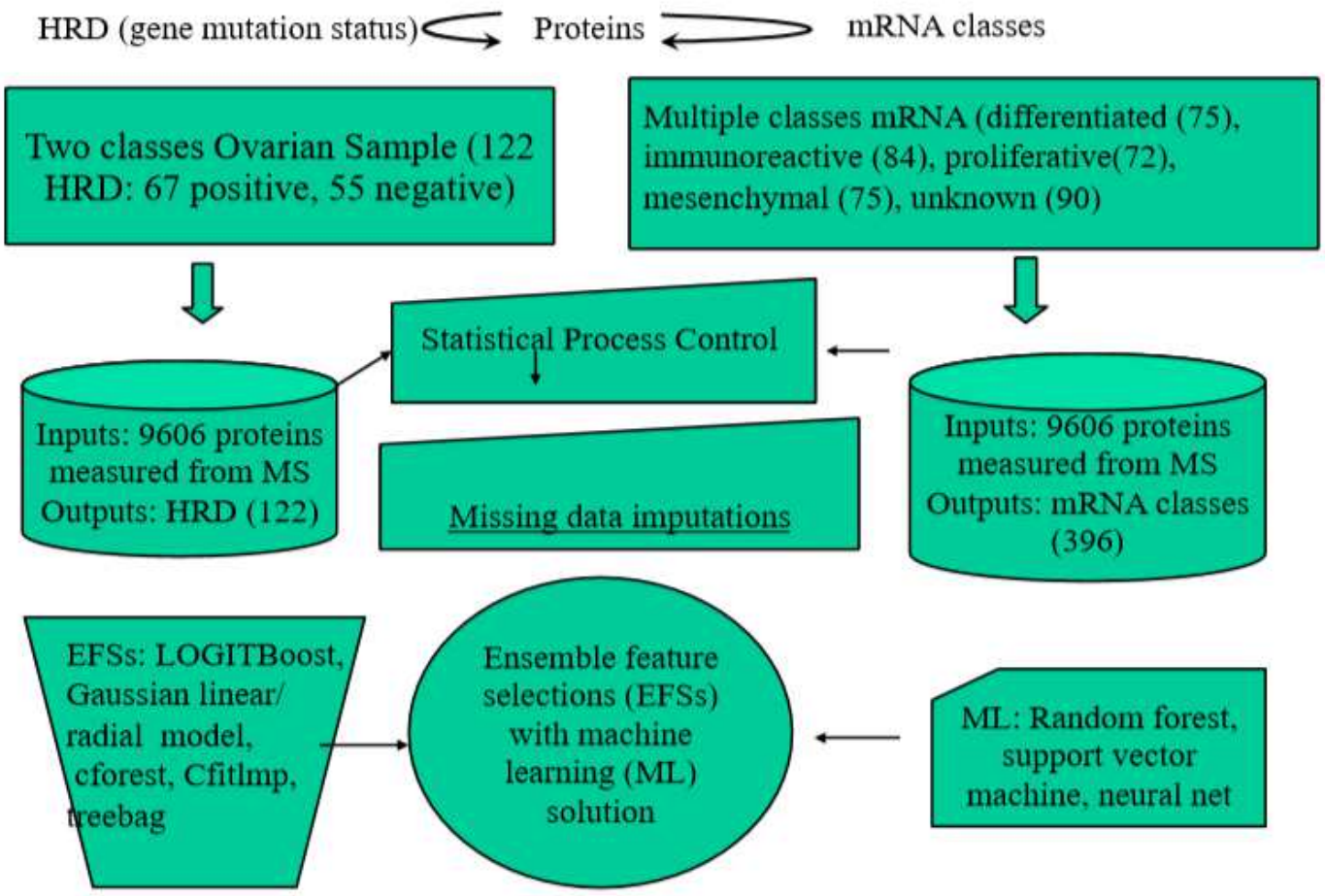

Figure 1

Machine learning Solutions with Ensemble Feature Selections for Proteogenomics Data Prediction

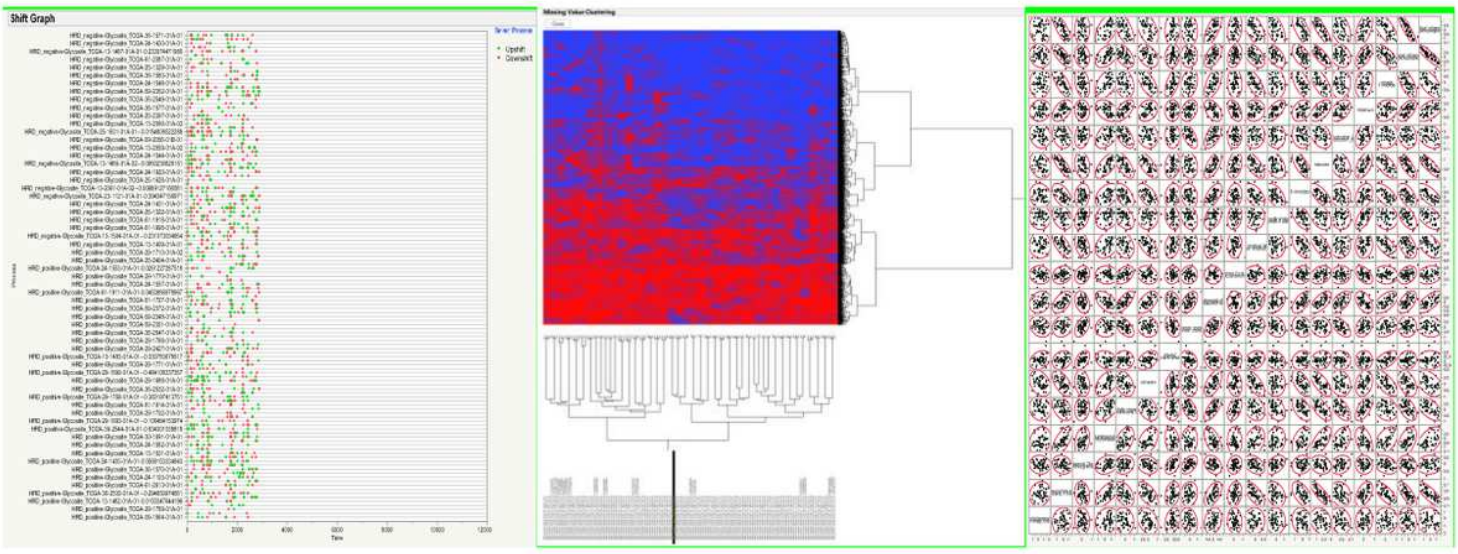

Figure 2

Left: Statistical process control for quality examination; Middle: Missing patterns examinations indicates between. 15% - 33% expression missing occurred; Right: Multivariate scatter analysis shows the high correlations among the measured proteins

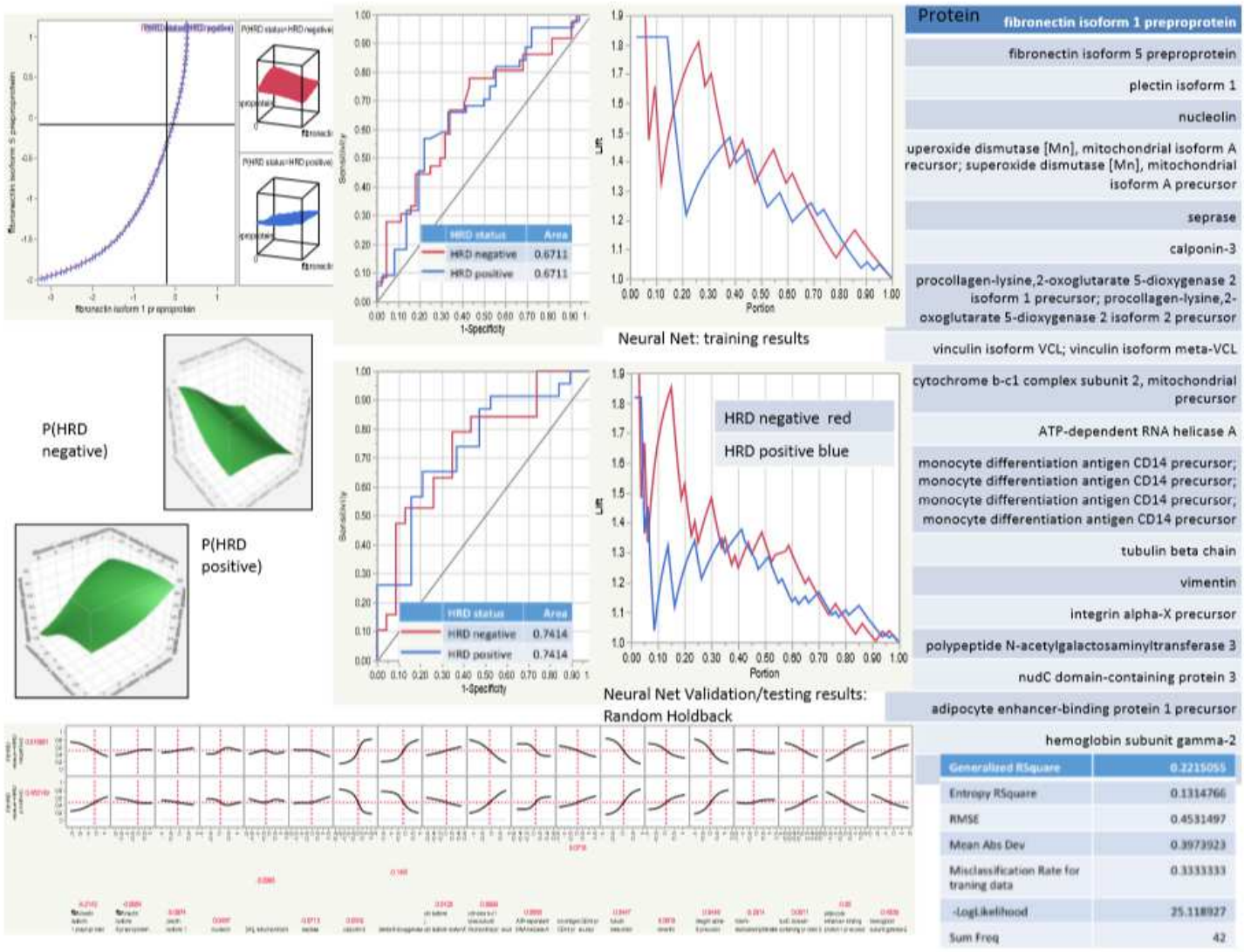

Figure 3

Top selected proteins using mlpML (NN: three layers, 10 hidden nodes) for HRD two classes prediction: Red and blue represent HRD Positive versus negative, respectively.

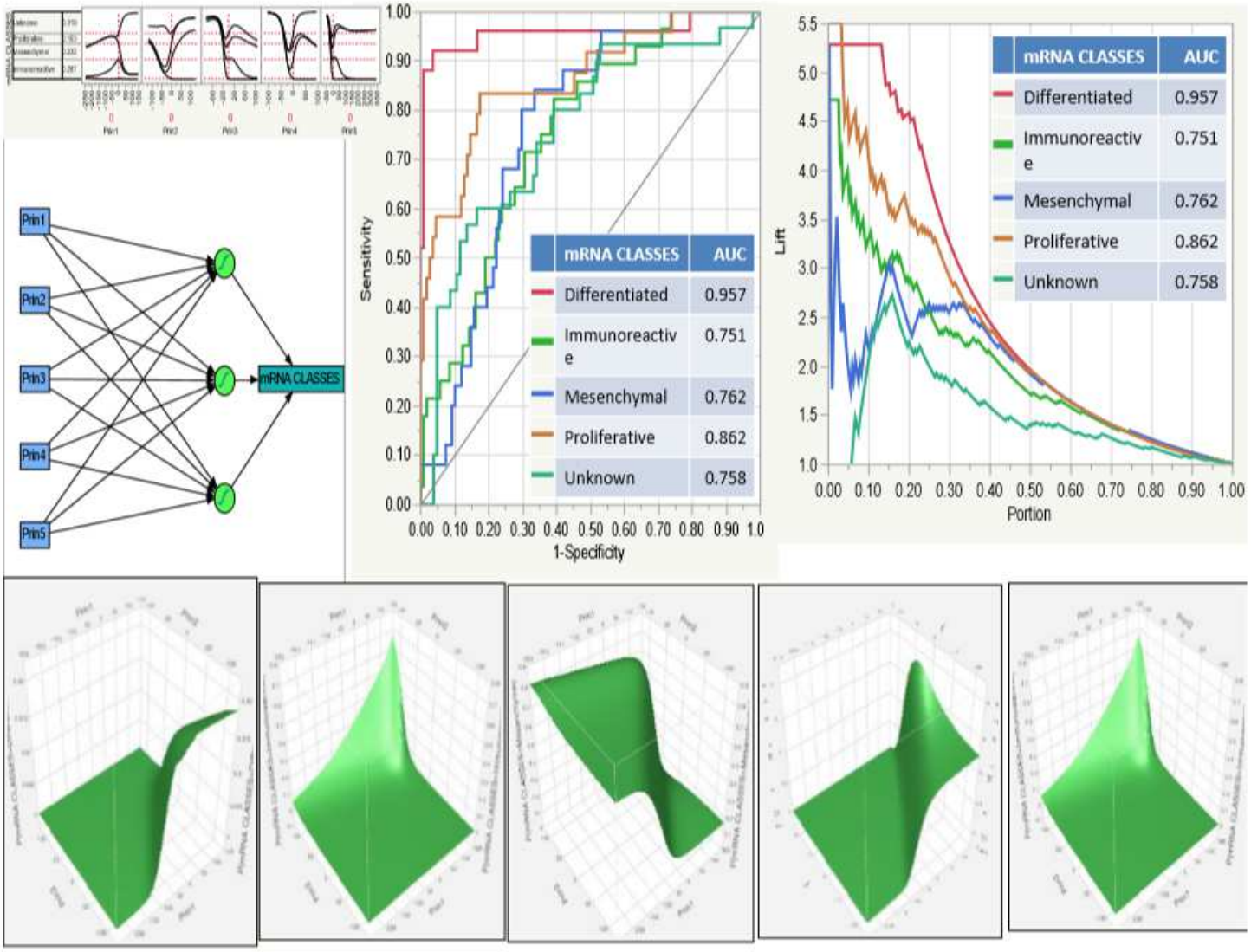

Figure 4

Top selected important proteins using NN (three layers, 10 hidden nodes) for mRNA classes prediction

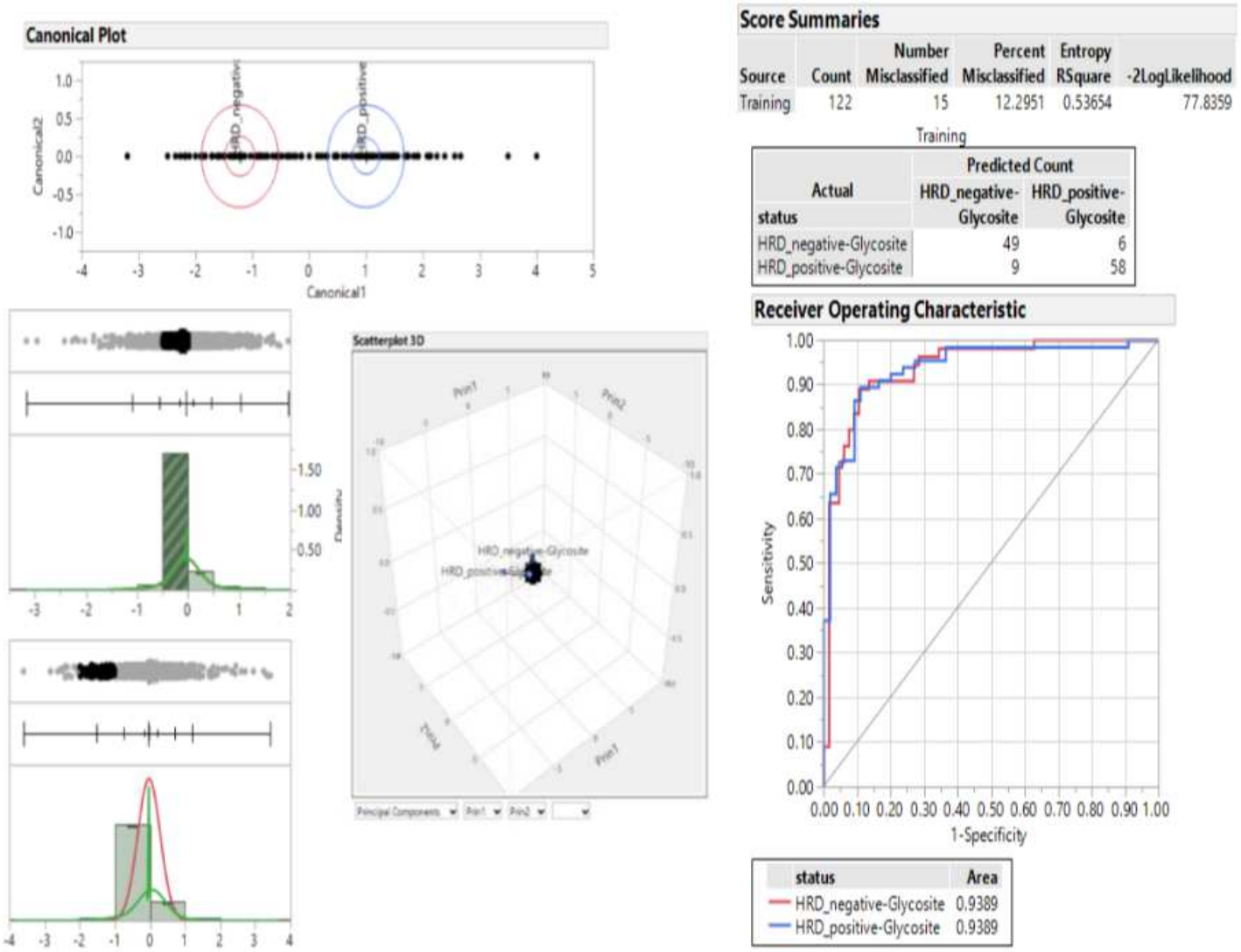

Figure 5

Top selected important proteins from HEFs for HRD classes' prediction compared with discriminant analysis: Red and green dots represent OV HRD Positive versus negative and normal samples, respectively.