Identification of Potential Biomarkers for Breast Cancer through Machine Learning approach

Nida Sattar , Itrat Zehrh, Department of Pathology & Lab Medicine
Aga Khan University & Hospital, Karachi, Pakistan.

INTRODUCTION

One of the main causes of rising women mortality rates worldwide is breast cancer. Radiologists find it challenging to accurately detect breast cancer due to the complicated nature of the disease, which includes microcalcification and lumps in the cells.

Consequently, radiologists are assisted in diagnosing cancer cells by a variety of computer-aided diagnostic (CAD) systems that have been previously established. However, it is essential to enhance the current diagnostic systems because of the inherent hazards connected to a delayed or inaccurate diagnosis.

In this scenario, the integration of gene expression data & machine learning has emerged as a promising tool for the accurate and early diagnosis of breast cancer.

Machine learning algorithms are being used to identify risk variables and provide early prediction by providing efficient means of extracting valuable information from large, complicated databases.

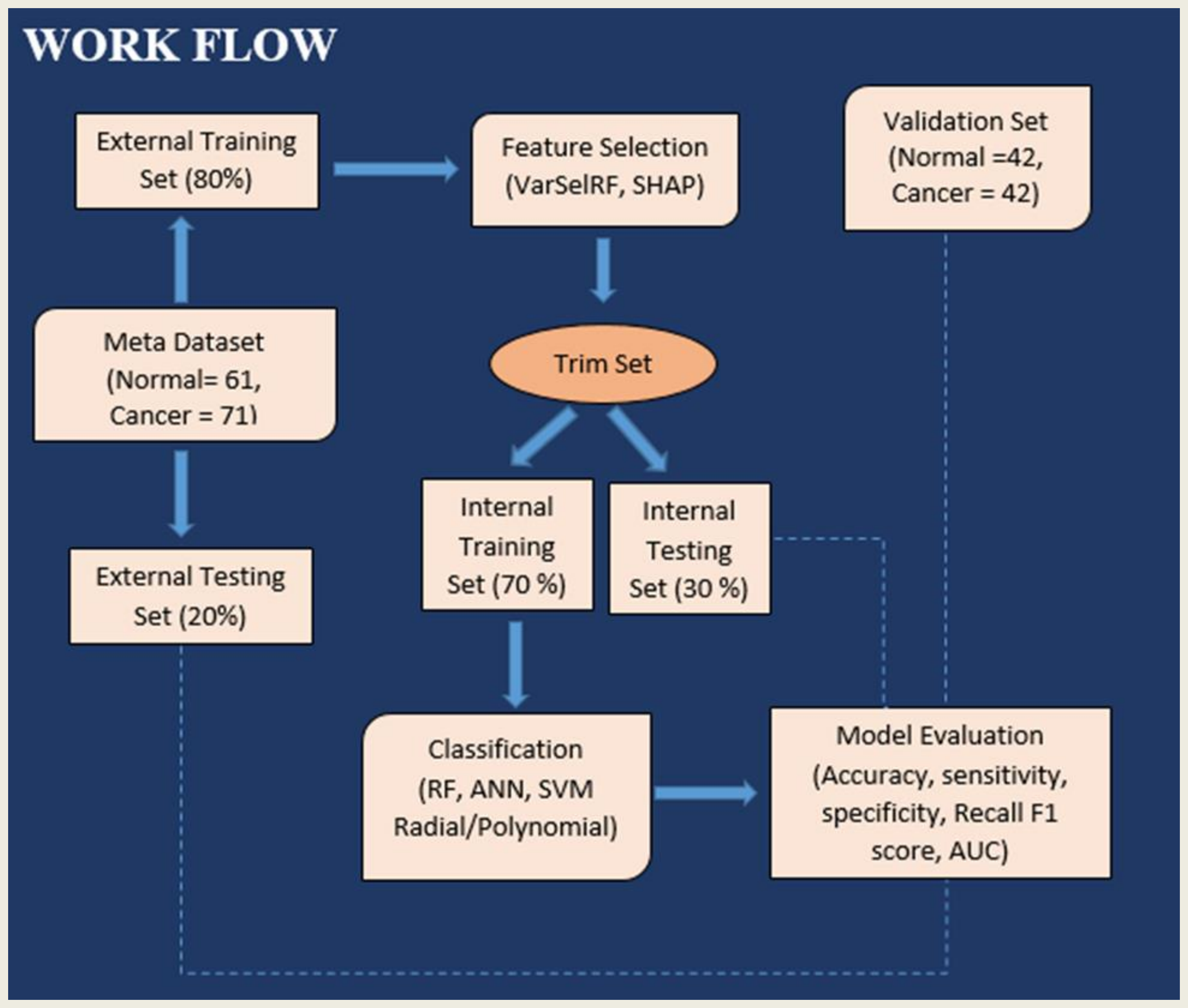
AIM & OBJECTIVE

We aimed to predict significant genes for breast cancer detection by using machine learning methods.

METHODS AND MATERIALS

In this study, the gene expression data were obtained from the NCBI GEO repository. Feature selection methods, Random Forests (VarSelRF) and SHAP (SHapley Additive explanations) were employed to identify biomarkers. For classification, machine learning methods such as Random Forest, Artificial Neural Networks (ANN), and Support Vector Machines (SVM polynomial and radial) were used. The following evaluation metrics were utilized to assess the performance of the classifier: accuracy, specificity, sensitivity, F1 score, precision, recall, and AUC(ROC). Furthermore, expression analysis of significant genes was also generated by Metascape.

WORK FLOW



External Training Set (80%)

Feature Selection (VarSelRF, SHAP)

Validation Set (Normal =42, Cancer = 42)

Meta Dataset (Normal= 61, Cancer = 71)

Trim Set

Internal Training Set (70 %)

Internal Testing Set (30 %)

External Testing Set (20%)

Classification (RF, ANN, SVM Radial/Polynomial)

Model Evaluation (Accuracy, sensitivity, specificity, Recall F1 score, AUC)

Dataset Summary	GEO ID	Normal	Cancer
Meta Dataset	GSE20711	2	2
	GSE3744	7	7
	GSE42568	17	17
	GSE50567	6	6
	GSE5764	20	10
	GSE7904	19	19
Validation set	GSE10780	42	42

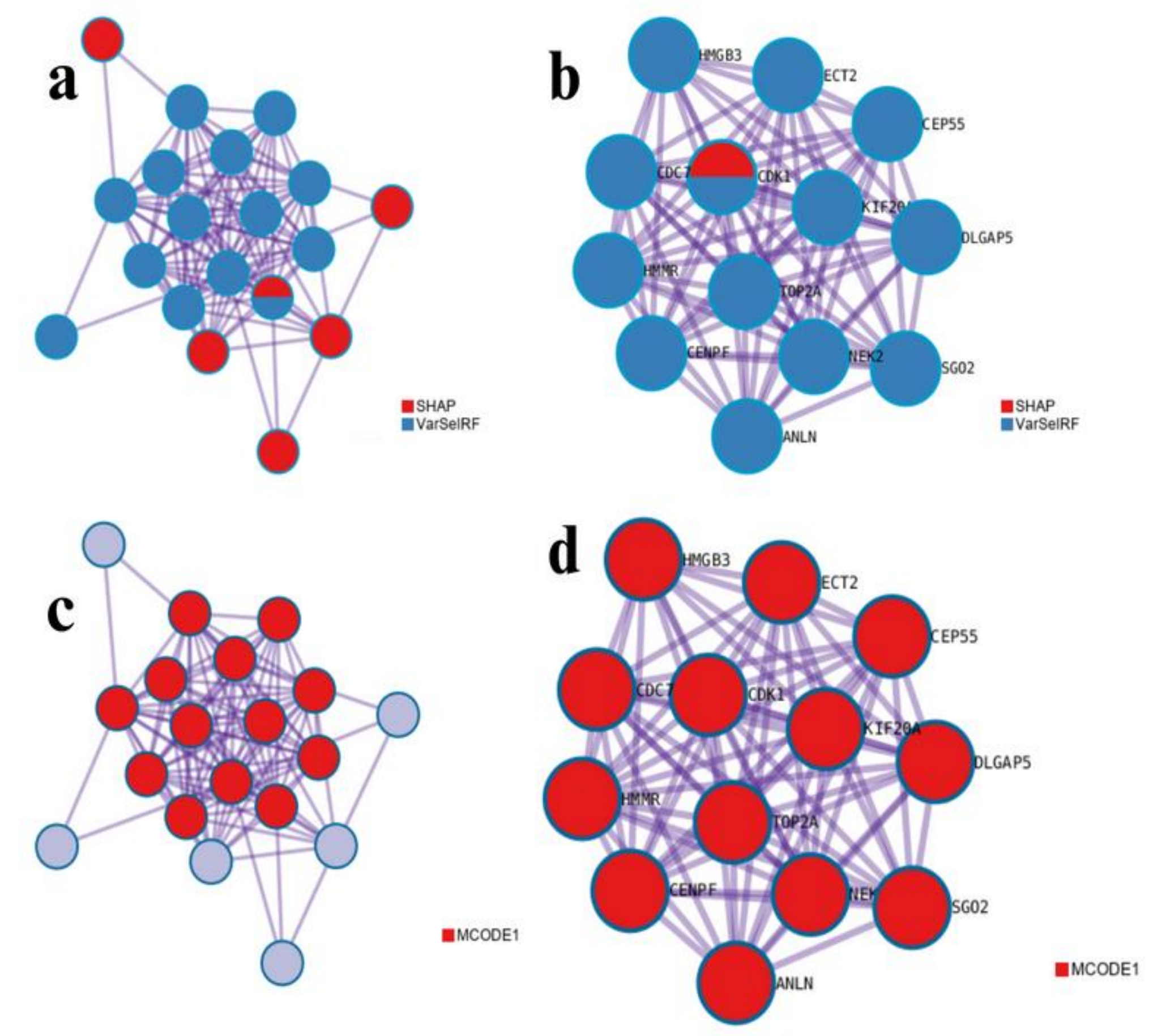


Figure 4. Protein-protein Interaction Enrichment Analysis, (a) All lists merged(Full Connection), (b) All lists merged(Keep MCODE Nodes Only), Colored by Counts, (c) All lists merged (Full Connection), (d) All lists merged (Keep MCODE Nodes Only), Colored by Cluster

RESULTS

As a result, feature selection methods identify 21 top genes (VarSelRF = 14 & SHAP = 08) in total and 01 gene (CDK1) is common in both. Out of these 21 genes Metascape validated 05 genes (CDK1, CENPF, HMMR, TOP2A & MSI2) involved cancer-related diseases in which HMMR recognized as breast cancer related gene.

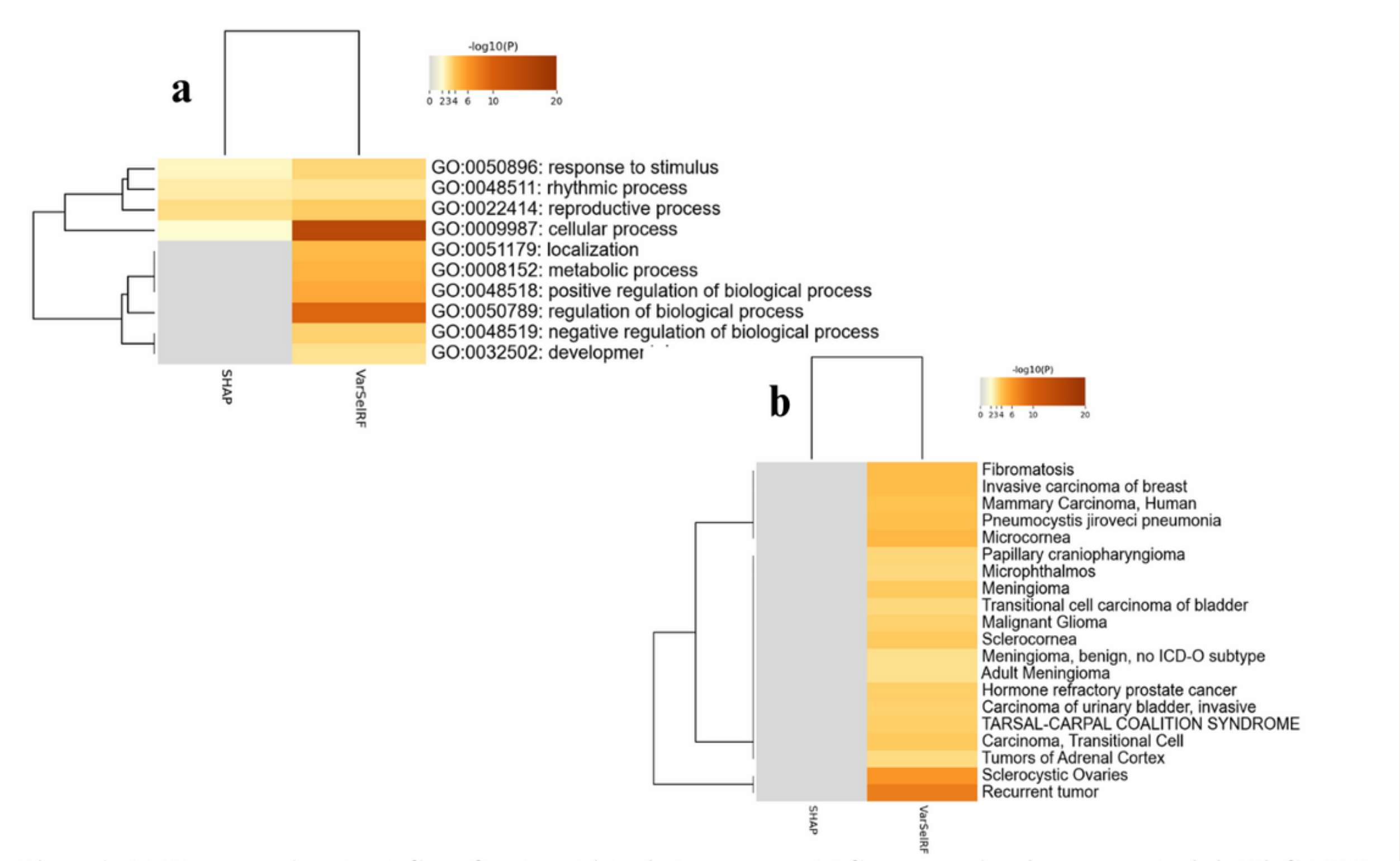


Figure 2. (a) Heatmap of top-level Gene Ontology biological processes, (b) Summary of enrichment analysis in DisGeNET, colored by p-values

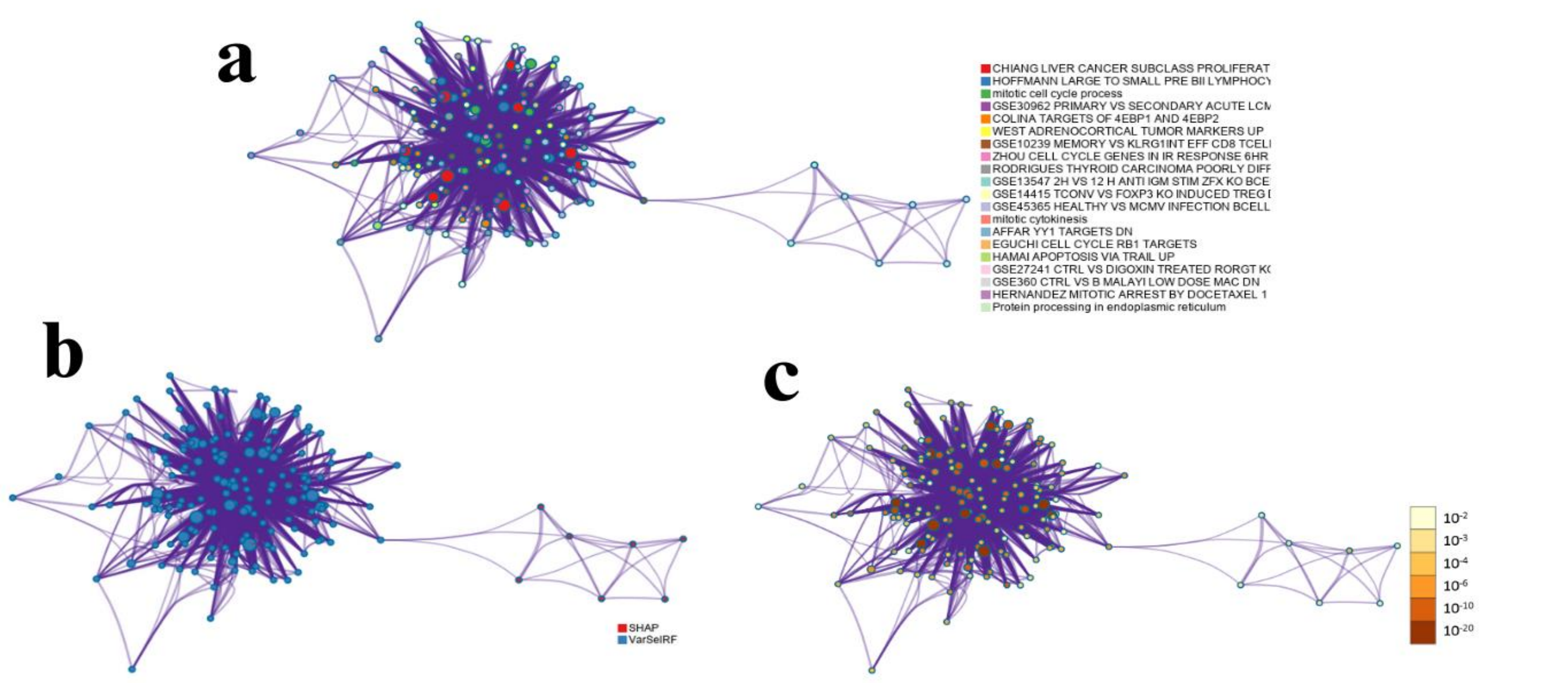


Figure 3. Network of enriched terms: (a) colored by cluster ID, (b) colored by identified feature selection methods (c) colored by p-value,

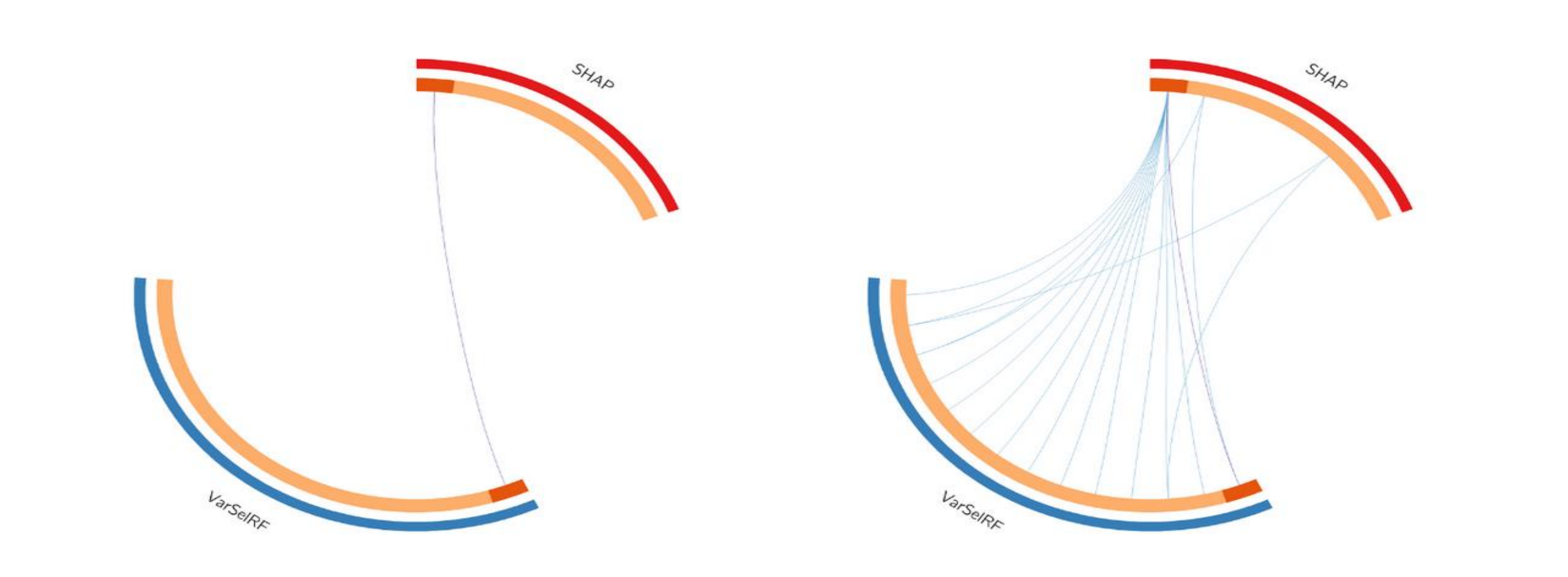


Figure 1. Overlap genes between VarSelfRF and SHAP identified genes: (a) only at the gene level, where purple curves link identical genes; (b) including the shared term level, where blue curves link genes that belong to the same enriched ontology term. The inner circle represents gene lists, where hits are arranged along the arc. Genes that hit multiple lists are colored in dark orange, and genes unique to a list are shown in light orange.

CONCLUSIONS

The findings showed that feature selection and Machine Learning may identify possible biomarkers for early diagnosis of breast cancer.
Clinical validation of these genes with the integration of clinical and demographic data will be considered in the future.

REFERENCES

Tabl, A. A., Alkhateeb, A., ElMaraghy, W., Rueda, L., & Ngom, A. (2019). A machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer. *Frontiers in genetics*, *10*, 256.

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., ... & Chanda, S. K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature communications*, *10*(1), 1523.