OPEN ACCESS Freely available online

ISSN: 2153-0602
Journal of
Data Mining in
Genomics & Proteomics

Commentary

AI-Driven Feature Selection for Identifying Prognostic Genes in Breast Cancer

Rebecca Smith*

Department of Oncology Data Science, University of Cambridge, Cambridge, UK

DESCRIPTION

Breast cancer remains one of the most prevalent and deadly malignancies affecting women globally. Despite significant advances in screening and treatment, accurately predicting disease prognosis and personalizing therapy continues to pose a substantial challenge due to the heterogeneity of tumor biology. The advent of high-throughput genomic technologies, such as RNA sequencing and microarrays, has enabled researchers to profile gene expression patterns at an unprecedented scale. However, the complexity and sheer volume of genomic data necessitate robust computational tools to distill biologically meaningful information. In this context, Artificial Intelligence (AI) particularly Machine Learning (ML) has emerged as a transformative approach in biomedical data analysis. One of the most impactful applications of AI in cancer genomics is feature selection, the process of identifying a subset of relevant genes from large datasets that are predictive of patient outcomes. AI-driven feature selection offers a powerful means to identify prognostic genes in breast cancer, which can aid in risk stratification, treatment planning and the discovery of novel therapeutic targets. Feature selection is a critical step in genomic data analysis, especially when dealing with high-dimensional datasets where the number of features (genes) far exceeds the number of samples (patients). This high-dimensionality problem often leads to overfitting in traditional statistical models, reducing their generalizability. AI-driven feature selection techniques, including supervised learning algorithms and ensemble methods, help mitigate this issue by identifying the most informative genes associated with disease prognosis. These approaches not only improve the predictive accuracy of models but also enhance biological interpretability by narrowing down the gene list to those most relevant to cancer progression, metastasis, or recurrence.

Commonly used AI algorithms for gene selection in breast cancer include Random Forests, Support Vector Machines (SVM), LASSO regression and more recently, deep learning methods and hybrid ensemble models. These models assess the predictive value of each gene by measuring metrics such as information gain, mutual information, or feature importance scores derived from model performance. For instance, Random Forests, a tree-based ensemble method, calculate the importance of each gene based on how well it improves decision tree splits, providing a ranked list of genes that contribute to classifying patient survival outcomes or treatment response. LASSO (Least Absolute Shrinkage and Selection Operator), on the other hand, performs regression analysis while shrinking the coefficients of less important genes to zero, thus selecting only the most predictive ones. Deep learning techniques, especially autoencoders and Convolutional Neural Networks (CNNs), have also shown promise in unsupervised and semi-supervised feature selection. Autoencoders compress high-dimensional gene expression data into lower-dimensional representations, capturing the underlying structure of the data. By examining the encoded features or reconstruction errors, researchers can infer which genes carry the most prognostic information. CNNs, while traditionally used in image data, have been adapted for genomics by treating gene expression profiles as structured matrices, where the spatial relationships between genes (e.g., co-expression or pathway membership) can be leveraged to identify significant prognostic biomarkers.

AI-driven feature selection does not operate in isolation; it is often integrated with survival analysis models, such as Cox proportional hazards models, Kaplan-Meier survival curves and time-dependent AUC metrics. These statistical methods evaluate the prognostic power of selected genes by associating their expression levels with patient survival times, recurrence-free intervals, or disease-specific mortality. Combining machine learning-based selection with survival modeling results in more robust identification of gene signatures that can predict clinical outcomes in breast cancer patients with higher accuracy. Additionally, AI-based feature selection facilitates the subtype classification of breast cancer, such as luminal A, luminal B, HER2-enriched and triple-negative. These molecular subtypes differ significantly in prognosis and therapeutic response. AI models trained on large expression datasets, such as those from The Cancer Genome Atlas (TCGA) or METABRIC, can accurately classify patients into these subtypes using a refined

Correspondence to: Rebecca Smith, Department of Oncology Data Science, University of Cambridge, Cambridge, UK, Email: rsmith@cambridgeonc.uk

Received: 24-Feb-2025, Manuscript No. JDMGP-25-29283; **Editor assigned:** 26-Feb-2025, Pre QC No. JDMGP-25-29283 (PQ); **Reviewed:** 12-Mar-2025, QC No. JDMGP-25-29283; **Revised:** 18-Mar-2025, Manuscript No. JDMGP-25-29283 (R); **Published:** 26-Mar-2025, DOI: 10.35248/2153-0602.25.16.368

Citation: Smith R (2025). AI-Driven Feature Selection for Identifying Prognostic Genes in Breast Cancer. J Data Mining Genomics Proteomics.16: 368.

gene panel, allowing for more tailored treatment decisions. The integration of multi-omics data including genomics, proteomics and methylation data further enhances the power of feature selection, as AI can uncover interactions across different molecular layers that influence breast cancer outcomes. One of the key benefits of AI-driven feature selection is its potential for clinical translation. Gene panels identified through these methods can be validated in independent cohorts and eventually developed into commercial prognostic assays, such as Oncotype DX or MammaPrint. These tests guide clinicians in deciding whether adjuvant chemotherapy is necessary, helping to reduce overtreatment and improve quality of life for patients. Importantly, the reproducibility and transparency of AI models are essential for clinical acceptance. Hence, current research is focusing on model interpretability tools such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), which explain how selected genes contribute to predictions, thereby increasing the trustworthiness of AI outputs in a medical setting.