OPEN ACCESS Freely available online

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

Opinion Article

Multi-Omics Data Integration Using Ensemble Learning for Precision Oncology

Elizabeth Turner[*]

Department of Computational Oncology, University of Oxford, Oxford, UK

DESCRIPTION

In the era of precision medicine, oncology has become one of the most advanced frontiers for applying integrative and computational approaches to understand cancer complexity and guide individualized treatment strategies. Cancer is a multifactorial disease involving dynamic interactions among genetic, epigenetic, transcriptomic, proteomic and metabolomic layers. Relying on a single data modality often fails to capture the heterogeneous and multidimensional nature of tumor biology. Hence, multi-omics data integration has emerged as a transformative strategy in precision oncology, offering a comprehensive molecular portrait of tumors. However, integrating these complex, high-dimensional datasets poses significant analytical challenges. This is where ensemble learning, a powerful branch of machine learning that combines multiple predictive models to achieve better accuracy and robustness, offers a promising solution to unlock the full potential of multi-omics data in cancer diagnosis, prognosis and treatment optimization. Multi-omics data refers to the collective measurement of different molecular levels, such as genomics (DNA mutations and copy number alterations), epigenomics (DNA methylation and histone modifications), transcriptomics (gene expression), proteomics (protein abundance and modifications) and metabolomics (metabolic profiles). Each omics layer provides unique but complementary information. For example, genomic mutations may reveal cancer drivers, while transcriptomic changes show functional gene activity and proteomics uncovers pathway regulation. When analyzed in isolation, each omics layer only tells part of the story. Integrating them synergistically allows researchers to reconstruct a more complete picture of tumor behavior, enabling more accurate patient stratification and treatment prediction.

One of the most significant obstacles in multi-omics integration is the heterogeneity and dimensionality of the data. Different omics datasets often vary in scale, distribution, measurement technology and missing data patterns. Direct concatenation or simplistic fusion of datasets can lead to model overfitting or biased predictions. Ensemble learning addresses this challenge by combining the strengths of multiple learning algorithms or models, thereby enhancing generalizability and robustness. The idea is that a committee of weak or diverse learners, when strategically combined, can outperform any single model. In precision oncology, ensemble learning can integrate predictions derived separately from different omics layers, using methods such as stacking, bagging, boosting, or hybrid ensembles. Stacked generalization (stacking) is particularly well-suited for multi-omics integration. In this approach, individual base models (e.g., random forests for transcriptomics, support vector machines for proteomics and neural networks for genomics) are trained separately on each data type. Their outputs are then used as inputs to a meta-learner, which learns how to best combine them for the final prediction. This hierarchical model enables the ensemble to capture both data-specific patterns and cross-omics interactions. Boosting algorithms, such as Gradient Boosting Machines (GBM) or XGBoost, can also be adapted for multi-omics settings by sequentially training models to correct the errors of prior ones, thereby refining the integration process.

Ensemble learning also facilitates robust feature selection, a critical step in high-dimensional omics analysis. Identifying relevant biomarkers from thousands of variables reduces model complexity and enhances interpretability. Techniques like random forests naturally rank features by importance, while LASSO and Elastic Net regularization help in selecting stable and non-redundant predictors. By applying feature selection independently within each omics domain and combining the top-ranked features in ensemble models, researchers can prioritize the most informative molecular markers for downstream validation and clinical application. In cancer subtyping, ensemble learning has been applied to multi-omics datasets to discover molecular subgroups with distinct prognoses and therapeutic responses. For example, integrating transcriptomics, methylation and copy number variation data using ensemble clustering methods like Similarity Network Fusion (SNF) or multi-view spectral clustering has revealed robust cancer subtypes in glioblastoma, breast cancer and colorectal cancer. These subtypes often correlate with clinical

Correspondence to: Elizabeth Turner, Department of Computational Oncology, University of Oxford, Oxford, UK, Email: eturner@oxfordonc.uk

Received: 24-Feb-2025, Manuscript No. JDMGP-25-29289; **Editor assigned:** 26-Feb-2025, Pre QC No. JDMGP-25-29289 (PQ); **Reviewed:** 12-Mar-2025, QC No. JDMGP-25-29289; **Revised:** 18-Mar-2025, Manuscript No. JDMGP-25-29289 (R); **Published:** 26-Mar-2025, DOI: 10.35248/2153-0602.25.16.373

Citation: Turner E (2025). Multi-Omics Data Integration Using Ensemble Learning for Precision Oncology. J Data Mining Genomics Proteomics. 16: 373.

outcomes and may inform decisions about immunotherapy, chemotherapy, or targeted drugs.

Furthermore, ensemble learning models can be deployed for survival prediction and treatment response modeling. In such cases, the integration of gene expression with mutation and drug sensitivity profiles enhances the ability to predict patient-specific responses to targeted therapies. Models trained on large datasets like The Cancer Genome Atlas (TCGA) or Genomics of Drug Sensitivity in Cancer (GDSC) can be fine-tuned using transfer learning techniques to adapt to new clinical cohorts. This adaptability makes ensemble models particularly valuable for translating computational predictions into clinical decision-making. Interpretability and clinical relevance are crucial considerations. Ensemble learning, while powerful, can become opaque as model complexity increases. To counteract this, interpretable ensemble models such as decision tree ensembles (e.g., gradient boosting trees) or attention-based mechanisms can be used to trace how omics features contribute to predictions. Explainable AI tools, like SHAP values or LIME (Local Interpretable Model-Agnostic Explanations), are increasingly used to identify the most influential features driving a model's prediction for a particular patient.