GMR
The Original

INTERPRETABLE DEEP LEARNING ON INTEGRATED MULTI-OMICS RECOVERS SUBTYPE-SPECIFIC REGULATORY FEATURES AND ESTROGEN-RESPONSE SIGNALLING IN BREAST INVASIVE CARCINOMA

Jayalakshmi M[1], K. Maharajan[2]*, M.Carmel Sobia[3], Sankar Ganesh Karuppasamy[4], M. Kaliappan[5], E. Mariappan[6]

[1]Professor, Department of Computer Science and Engineering-Cyber Security, Ramco Institute of Technology, Rajapalayam, Tamil Nadu 626117, India ORCID: 0000-0002-7297-9161
[2] Associate Professor, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankoil, Tamil Nadu 626126, India ORCID: 0000-0003-4353-3133 * Corresponding author. E-mail: maharajank@gmail.com
[3] Professor, Department of Electrical and Electronics Engineering, P.S.R Engineering College, Sivakasi, Tamil Nadu 626140, India
[4] Assistant Professor, Department of Computer Science and Engineering, Vel Tech Rangarajan Dr.Sagunthala R & D Institute of Science and Technology, Avadi, Chennai-600062,India
[5]Professor, Department of Artificial Intelligence and Data Science, Ramco Institute of Technology, Rajapalayam, Tamil Nadu 626117, India
[6]Professor, Department of Information Technology, Ramco Institute of Technology, Rajapalayam, Tamil Nadu, 626117, India
Email: [1]jayalacsmi@gmail.com, [2]maharajank@gmail.com, [3]necsobia@gmail.com, [4] prof.sankarganesh@gmail.com, [5]kalsrajan@yahoo.co.in,[6] mapcse.e81@gmail.com

ABSTRACT

Breast invasive carcinoma is a molecularly heterogeneous disease and the clinically relevant intrinsic subtypes are the result of a coordinated alteration of the genome, transcriptome and epigenome. With analysis limited to a single molecular layer, somewhat of this structure is obtained, but few analyses correlate the identity of subtypes with the regulatory features that differentiate them. We assembled matched gene expression, gene-level DNA methylation and microRNA expression profiles for 875 breast carcinoma samples annotated with PAM50 intrinsic subtype and asked whether an interpretable model trained on all three layers could both classify subtypes and nominate the molecular features responsible for each assignment. The embeddings of each omics layer were fed to their respective encoders and then a sample-wise attention mechanism was used to fuse the embeddings into a unified representation before the subtype prediction task. Feature attribution based on Shapley additive explanations was then used to rank genes, methylation features and microRNAs by their contribution to each subtype.

On a held-out test set the integrated model reached a macro-averaged area under the receiver operating characteristic curve of 0.958, matching the strongest single-omics model (expression alone) while drawing on methylation and microRNA information that expression-only models cannot access. Attention weights varied across subtypes in a biologically sensible manner, with the luminal subtypes relying most heavily on expression and the normal-like group distributing weight toward microRNA. Genes prioritised by attribution were significantly enriched for established breast-cancer genes relative to the analysed background panel (3.2-fold, hypergeometric $P = 4.1 \times 10^{-7}$), and the enrichment held within four of the five subtypes. Estrogen-response signalling was the dominant enriched programme, most strongly in the HER2-enriched group. The prioritised features recovered canonical subtype markers including ESR1, FOXA1, GATA3, TFF1 and the estrogen-regulated microRNA hsa-mir-375 without any prior knowledge supplied to the model. These results indicate that explainable multi-omics integration recovers biologically coherent subtype programmes and offers a transparent and reproducible route to candidate biomarker generation in breast cancer genomics.

KEYWORDS: breast invasive carcinoma; multi-omics integration; PAM50 subtypes; interpretable deep learning; SHAP; driver genes; DNA methylation; microRNA; The Cancer Genome Atlas

1. INTRODUCTION

Breast cancer remains the most frequently diagnosed malignancy in women and a leading cause of cancer death worldwide. Much of the clinical and biological complexity of the disease can be traced to its molecular heterogeneity. Early transcriptomic studies established that breast tumours fall into a small number of intrinsic subtypes with distinct expression programmes, clinical trajectories and treatment responses [1,2]. The PAM50 classifier formalised this observation into a fifty-gene signature that assigns tumours to the luminal A, luminal B, HER2-enriched, basal-like and normal-like groups, and these categories continue to inform prognosis and therapeutic decisions [3,4]. Subtype identity, however, is a summary of deeper molecular events. The intrinsic groups differ not only in the abundance of a handful of marker transcripts but also in patterns of somatic mutation, copy-number change, promoter methylation and non-coding regulation [5,6]. A description that stops at the transcriptome therefore leaves much of the mechanistic basis of subtype identity unexplained.

Large reference efforts, most prominently The Cancer Genome Atlas, have produced matched measurements of several molecular layers for the same tumours, making it possible in principle to study how these layers jointly define subtype biology [5]. Realising that possibility has proved difficult. Each platform carries its own noise structure and dimensionality, the number of measured features vastly exceeds the number of tumours, and the layers are correlated in ways that are informative but hard to model. A substantial body of work has responded with unsupervised integration methods that learn shared factors or joint clusters across layers, including matrix-factorisation approaches and similarity-based fusion [7,8,9]. These methods have clarified the global structure of multi-omics cohorts, yet they are not designed to attribute a particular subtype assignment to specific molecular features, which is precisely the information a biologist needs when moving from a classifier to a hypothesis.

Supervised deep learning offers a complementary path. Neural networks trained to predict a clinical label can absorb high-dimensional multi-omics input and have achieved strong performance in cancer classification and outcome prediction [10,11,12]. Their reputation for opacity has limited their uptake in settings where the goal is understanding rather than prediction alone. This concern has eased considerably with the maturation of post-hoc attribution methods. Shapley additive explanations assign each input feature a contribution to a given prediction with a solid grounding in cooperative game theory, and related gradient-based methods provide comparable attributions at lower cost [13,14]. When such methods are applied to a model trained on molecular data, the resulting feature rankings can be read as data-driven nominations of the genes and regulatory elements that drive a phenotype, and these nominations can then be tested against independent biological knowledge.

The value of an interpretable multi-omics model depends on whether its attributions recover known biology and, ideally, extend it. Curated resources such as the COSMIC Cancer Gene Census and IntOGen catalogue genes with established roles in cancer and provide a reference against which computational nominations can be assessed [15,16]. Functional annotation databases including Gene Ontology, KEGG and Reactome allow prioritised gene sets to be placed in the context of biological pathways [17,18]. A model whose attributions are enriched for known breast-cancer genes and coherent pathways is one whose novel nominations can be trusted enough to warrant follow-up.

In this study we develop and evaluate an interpretable multi-omics framework for breast invasive carcinoma that pursues two goals at once. The first is accurate assignment of tumours to their PAM50 subtype using gene expression, gene-level DNA methylation and microRNA expression together. The second, and the one we regard as the primary contribution, is transparent attribution of each subtype assignment to specific molecular features across all three layers, followed by validation of the prioritised features against curated driver references and pathway annotations. The architecture uses a dedicated encoder for each omics layer and combines their embeddings through a sample-wise attention mechanism, which lets the model weight the layers according to their reliability for each tumour rather than treating them as interchangeable. We show that this integrated model matches the discrimination of the strongest single-omics model while producing attributions that concentrate on established breast-cancer genes and on the hormone-signalling and proliferation programmes known to separate the intrinsic subtypes. Rather than claim that integration inflates classification accuracy, a claim the data do not support for a label that is defined from expression, we position the multi-omics model as a means of recovering interpretable, cross-layer subtype biology that a single-omics analysis cannot provide.

2. MATERIALS AND METHODS

2.1. Data acquisition and cohort

We analysed matched multi-omics profiles of breast invasive carcinoma derived from The Cancer Genome Atlas. The processed matrices used here were obtained from the curated multi-omics benchmark released by Wang and colleagues, in which primary tumour samples were retained only when complete measurements were available across the three molecular layers and a PAM50 subtype label had been assigned [12]. This yielded a cohort of 875 tumours partitioned by the original authors into a training set of 612 samples and an independent test set of 263 samples. Three molecular layers were available for every tumour: messenger RNA expression, gene-level DNA methylation, and microRNA expression. Each tumour carried one of the five PAM50 intrinsic subtype labels, coded here as normal-like, basal-like, HER2-enriched, luminal A and luminal B.

The subtype distribution reflects the composition of the underlying cohort and is markedly imbalanced. Luminal A is the most common group with 436 tumours, followed by luminal B (147), basal-like (131) and normal-like (115), while HER2-enriched is the rarest with 46 tumours. This imbalanced distribution of the subtypes is not an artificial balance but a reflection of the clinical prevalence of these subtypes as they would be seen in real-life clinical practice by a classifier. This imbalance was taken into account in making all modelling decisions that might be affected by it, such as loss weighting, and the selection of macro-averaged evaluation metrics.

2.2. Data preprocessing and feature representation

The expression, methylation and microRNA matrices were supplied in preprocessed form. In the pipeline that generated them, low-expression genes and low-abundance microRNAs were removed, expression values were log-transformed, methylation was summarised from probe level to gene level, and every feature was scaled to a common range so that the three layers could be combined without one dominating through scale alone [12]. We verified that all three matrices occupied the unit interval after this scaling and that no missing values remained, so no further imputation was required. After feature reduction the expression matrix contained 1000 genes, the

methylation matrix 1000 gene-level features, and the microRNA matrix 503 microRNAs, giving a combined input of 2503 features per tumour.

The reduction of each layer to its most informative features is a property of the benchmark we adopted rather than a step we performed, and it has a consequence that we return to when interpreting the results. The retained genes are those with the greatest variance across the cohort, which enriches the panel for subtype-discriminating expression markers but leaves out many recurrently mutated driver genes whose transcript levels vary little between tumours. The composition of the panel therefore shapes what any downstream attribution analysis can recover, and we treat the 1000-gene expression panel as the fixed statistical background for all enrichment testing described below.

2.3. Interpretable multi-omics architecture

The model consists of three parallel encoders, an attention-based fusion module and a classification head. Writing the three input matrices as X_RNA, X_Meth and X_miRNA, each is transformed by its own encoder into a layer-specific embedding of common dimension. An encoder is a stack of two blocks, each comprising a fully connected linear transformation, batch normalisation, a rectified linear activation and dropout. The first block maps the input to a hidden width of 256 units and the second to a latent width of 64 units. Batch normalisation stabilises the distribution of activations across the heterogeneous input scales of the three layers, and dropout at a rate of 0.4 provides regularisation that is important given the large feature-to-sample ratio [19,20].

The three latent embeddings are combined by a sample-wise attention mechanism rather than by simple concatenation. For each tumour the module computes a scalar score for every layer through a small scoring network, applies a softmax across the three scores to obtain attention weights that sum to one, and forms the fused representation as the attention-weighted sum of the layer embeddings. The design is motivated directly by the data. Because the methylation layer is substantially less informative for subtype assignment than expression, a fixed concatenation forces the classifier to work around the noise that the weaker layer injects. Attention allows the model to down-weight an unreliable layer for a given tumour while still admitting its contribution where it is informative, and the learned weights are themselves interpretable as a per-subtype account of which layers carry the signal [21].

The fused representation passes through a classification head consisting of a fully connected layer of 64 units with batch normalisation, a rectified linear activation and dropout, followed by a linear layer producing five logits and a softmax over the PAM50 subtypes. Keeping a separate encoder for each layer, rather than concatenating the raw features before a shared network, is what makes attribution to the originating layer possible, since every latent unit can be traced to a single molecular source.

2.4. Single-omics baselines

To quantify what the integration contributes, three single-omics baselines were constructed by reusing the same encoder and classification head with only one input layer present. Each baseline therefore has the same capacity and training procedure as one branch of the integrated model, so that any difference in performance reflects the information in the layer rather than a difference in architecture. Comparing the integrated model against these matched baselines, rather than against unrelated classifiers, isolates the effect of combining the layers.

2.5. Training and evaluation

All models were trained by minimising the cross-entropy loss with the Adam optimiser at a learning rate of 0.001 and a weight decay of 0.0001 [22]. Because of the class imbalance, the loss was weighted by the inverse frequency of each subtype in the training fold, which prevents the dominant luminal A group from overwhelming the gradient and improves recovery of the rare HER2-enriched group. Training proceeded in mini-batches of 64 tumours for up to 200 epochs, with early stopping triggered when the validation loss failed to improve for 20 consecutive epochs, at which point the parameters from the best epoch were restored.

Model performance was estimated by five-fold stratified cross-validation on the training set, with stratification preserving the subtype proportions in every fold. The integrated model and the three baselines were trained and evaluated within an identical fold structure so that their per-fold scores are paired. After cross-validation, each model was retrained on the full training set, with a small stratified portion held aside for early stopping, and evaluated once on the independent 263-sample test set that had played no part in model development. Reported test figures come from this single held-out evaluation.

Performance was summarised with accuracy, macro-averaged precision, recall and F1 score, and the macro-averaged area under the receiver operating characteristic curve computed in a one-versus-rest manner across the five subtypes. Macro-averaging weights each subtype equally regardless of its frequency, which is the appropriate choice when the rare subtypes are of as much interest as the common ones. The statistical significance of the difference in macro-AUC between the integrated model and each baseline was assessed with a paired t-test across the five cross-validation folds.

2.6. Feature attribution

Feature attribution was performed on the integrated model after it had been retrained on the full training set. To attribute a subtype prediction to the individual molecular features, the three-input model was wrapped so that it accepted a single concatenated feature vector, which was split internally into the three layers, allowing a gradient-based Shapley explainer to operate over the full 2503-dimensional input [13,14]. A class-balanced background of

100 training tumours, twenty drawn at random from each subtype, was used as the reference distribution for the explainer, which prevents the attributions from anchoring on the dominant luminal A class.
The explainer was applied to the test tumours, producing for every tumour and every subtype a vector of feature attributions. For each subtype we restricted attention to the test tumours truly belonging to that subtype and averaged the absolute attribution of every feature across them, giving a per-subtype importance score for each gene, methylation feature and microRNA. Features were then ranked within each molecular layer and each subtype, and the top-ranked features were carried forward as the subtype-specific candidate drivers. In parallel, the attention weights assigned by the fusion module to the three layers were averaged within each subtype to summarise how the model apportioned reliance across the molecular layers.

2.7. Driver-gene validation and pathway enrichment

The prioritised expression-layer genes were tested for enrichment of known breast-cancer genes using the hypergeometric distribution, with the 1000-gene expression panel as the fixed background. Two reference sets were used, both restricted to genes present in the panel so that the test is well defined. The first is a subset of the COSMIC Cancer Gene Census, comprising census members that appear in the variance-selected panel [15]. Because the panel is enriched for high-variance expression markers rather than recurrently mutated genes, only seven census members are represented, which limits the power of this comparison and is reported as such. The second reference is a curated set of 55 established breast-cancer genes present in the panel, assembled from the intrinsic-subtype and molecular-marker literature and including hormone-receptor genes, proliferation markers, basal keratins and characterised subtype antigens [1,2,3,5,6]. Both reference lists are stated in full in the supplementary description so that the tests can be reproduced exactly.
For each subtype, and for the union of prioritised genes across subtypes, the overlap with each reference set was evaluated by the probability of observing at least the obtained overlap under the hypergeometric model, together with the corresponding fold enrichment relative to the background rate. Across the per-subtype tests, P-values were corrected for multiple comparison by the Benjamini–Hochberg procedure controlling the false discovery rate [23].
Pathway-level interpretation used curated gene sets representing programmes expected to distinguish the intrinsic subtypes: estrogen-response signalling, a cell-cycle and proliferation programme, and a DNA-replication programme. Each set was restricted to the background panel, and the overlap between the prioritised genes and each pathway was tested with Fisher's exact test against the panel, again with Benjamini–Hochberg correction across the pathway-by-subtype tests. Two further programmes named in the study design, ERBB/HER2 signalling and PI3K–AKT signalling, were represented by only one and two panel genes respectively and could not be tested; this is reported rather than concealed, as it follows directly from the variance-based composition of the panel. The pathway gene sets are likewise listed in full in the supplementary description.

2.8. Implementation and reproducibility

Models were implemented in PyTorch and attribution used the SHAP library, with data handling and statistical testing in NumPy, pandas, scikit-learn, SciPy and statsmodels [13,24,25]. Random seeds were fixed for data splitting, model initialisation and the attribution background so that the full pipeline, from cross-validation through attribution and enrichment testing, reproduces the reported figures on re-execution. All computation was performed on CPU. The complete analysis code is organised as a small set of scripts covering data loading, model definition, cross-validation and test evaluation, attribution, driver validation, pathway enrichment and figure generation.

3. RESULTS

3.1. Cohort and data characteristics

A total of 875 breast invasive carcinoma tumours with full measurements of both gene expression and gene-level DNA methylation and microRNA expression data were analyzed and annotated with PAM50 intrinsic subtypes (Table 1). A set of 1000 expression features, 1000 methylation features and 503 microRNA features were used to represent each tumour after the feature reduction inherited from the source benchmark. Subtype composition was predominant for luminal A (436 tumours) and was least represented for HER2-enriched (46 tumours) and the balance was maintained during model development and evaluation. A summary of the overall analysis workflow, with the three-layer input, attention-based fusion, attribution and the downstream biological validation is provided in Figure 1.

Table 1. Composition of the analysed TCGA breast invasive carcinoma multi-omics cohort.

Molecular layer / attribute	Value
Gene expression features	1000
DNA methylation features (gene level)	1000
microRNA expression features	503
Total features per tumour	2503
Training samples	612

Molecular layer / attribute	Value
Test samples	263
Total samples	875
Normal-like tumours	115
Basal-like tumours	131
HER2-enriched tumours	46
Luminal A tumours	436
Luminal B tumours	147

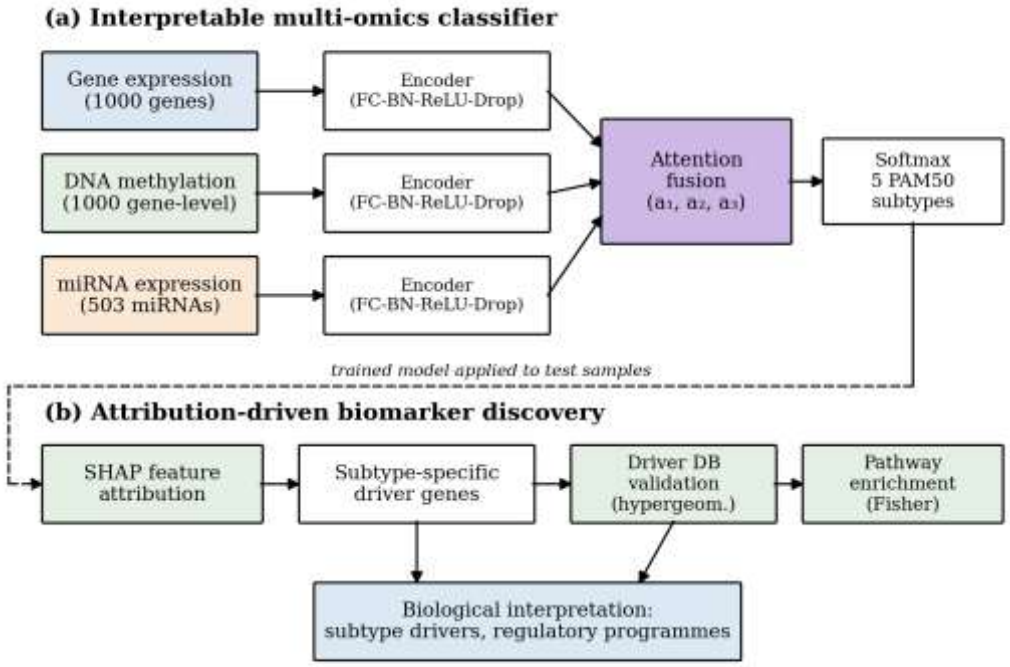

Figure 1. Overview of the interpretable multi-omics framework. (a) Gene expression, gene-level DNA methylation and microRNA expression are each processed by a dedicated encoder; a sample-wise attention module weights and fuses the layer embeddings before softmax classification into the five PAM50 subtypes. (b) The trained model is subjected to SHAP feature attribution, producing subtype-specific driver-gene nominations that are validated against curated driver references and tested for pathway enrichment, and finally interpreted biologically.

3.2. Subtype classification performance

Under five-fold cross-validation the four models separated clearly by the information available to them (Table 2, Figure 3). Among the single-omics baselines, expression was by far the most informative layer, reaching a mean accuracy of 0.796 and a macro-AUC of 0.942. The microRNA baseline followed at an accuracy of 0.685 and a macro-AUC of 0.897, and methylation was the weakest and least stable layer, with an accuracy of 0.544 and considerable variation across folds. This ordering is expected. The PAM50 label is itself derived from a gene-expression signature, so expression carries an intrinsic advantage for this particular task, while gene-level methylation summaries retain only a coarse reflection of the regulatory state.

The integrated model with attention fusion achieved a cross-validation accuracy of 0.755 and a macro-AUC of 0.923, placing it above the microRNA and methylation baselines but slightly below the expression-only model. On the independent test set the integrated model reached an accuracy of 0.787 and a macro-AUC of 0.958, the latter matching the expression baseline (0.958) to three decimal places while the expression baseline retained a small advantage in accuracy and F1 (Table 3). A paired comparison of macro-AUC across folds confirmed this reading: the integrated model significantly exceeded the methylation baseline (mean difference 0.053, $P = 0.023$) and trended above the microRNA baseline (mean difference 0.025, $P = 0.073$), while the expression baseline held a small but statistically detectable edge over the integrated model in cross-validation (mean difference 0.020, $P = 0.029$).

We interpret these results without inflation. For a subtype label that is defined from expression, no fusion strategy should be expected to substantially outperform expression alone at the classification task, and ours does not. What the integrated model does achieve is discrimination equal to the best single layer on unseen data, together with a genuinely multi-layer decision process that is open to inspection. The per-class results on the test set show that this parity is not achieved by neglecting the difficult subtypes. The integrated model recovered the rare HER2-enriched group with a recall of 0.786 and reached recalls of 0.80 or above for the normal-like and luminal B groups, indicating that the inverse-frequency loss weighting succeeded in protecting the minority classes. The confusion matrix (Figure 2) shows that the residual errors are concentrated between the biologically adjacent luminal A and luminal B groups and between luminal and normal-like tumours, which are the boundaries where subtype calls are known to be least stable.

Table 2. Five-fold cross-validation performance (mean ± standard deviation) for the single-omics baselines and the integrated model.

Model	Accuracy	Precision	Recall	F1	Macro-AUC
Expression only	0.796 ± 0.028	0.770 ± 0.031	0.807 ± 0.037	0.771 ± 0.031	0.942 ± 0.010
Methylation only	0.544 ± 0.125	0.647 ± 0.079	0.659 ± 0.037	0.581 ± 0.079	0.870 ± 0.032
microRNA only	0.685 ± 0.022	0.638 ± 0.035	0.694 ± 0.051	0.642 ± 0.032	0.897 ± 0.017
Integrated (attention)	0.755 ± 0.023	0.721 ± 0.028	0.768 ± 0.021	0.729 ± 0.026	0.923 ± 0.013

Table 3. Independent held-out test-set performance. Best value in each column shown in bold in the text discussion.

Model	Accuracy	Precision	Recall	F1	Macro-AUC
Expression only	0.833	0.821	0.833	0.814	0.958
Methylation only	0.498	0.598	0.638	0.515	0.893
microRNA only	0.681	0.654	0.694	0.655	0.898
Integrated (attention)	0.787	0.743	0.796	0.754	0.958

Table 3b. Per-subtype performance of the integrated model on the held-out test set.

Subtype	Precision	Recall	F1	Support
Normal-like	0.528	0.800	0.636	35
Basal-like	0.968	0.769	0.857	39
HER2-enriched	0.579	0.786	0.667	14
Luminal A	0.962	0.763	0.851	131
Luminal B	0.679	0.864	0.760	44

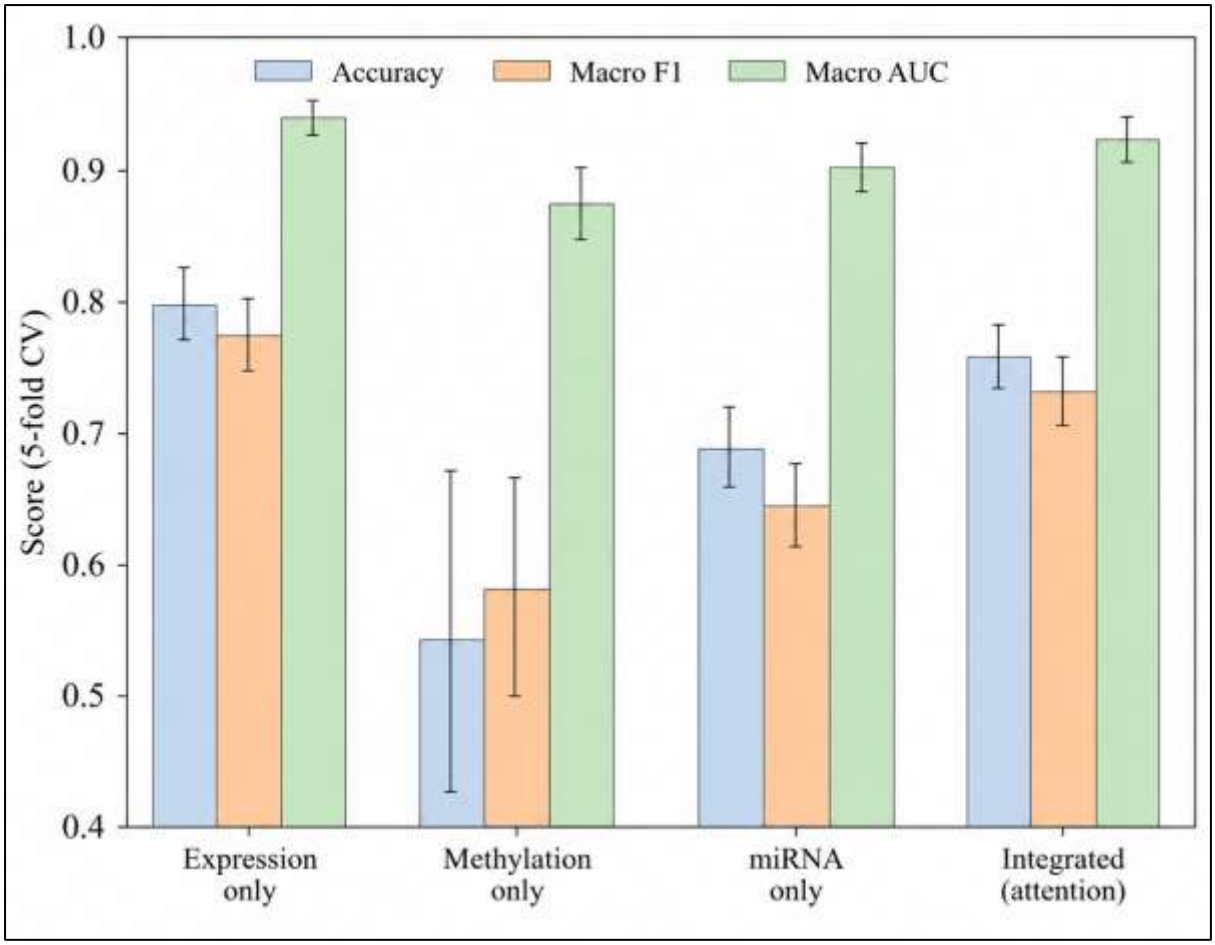

Figure 3. Five-fold cross-validation performance of the three single-omics baselines and the integrated attention model. Bars show the mean across folds and error bars the standard deviation. Expression is the strongest single layer; the integrated model exceeds the methylation and microRNA baselines.

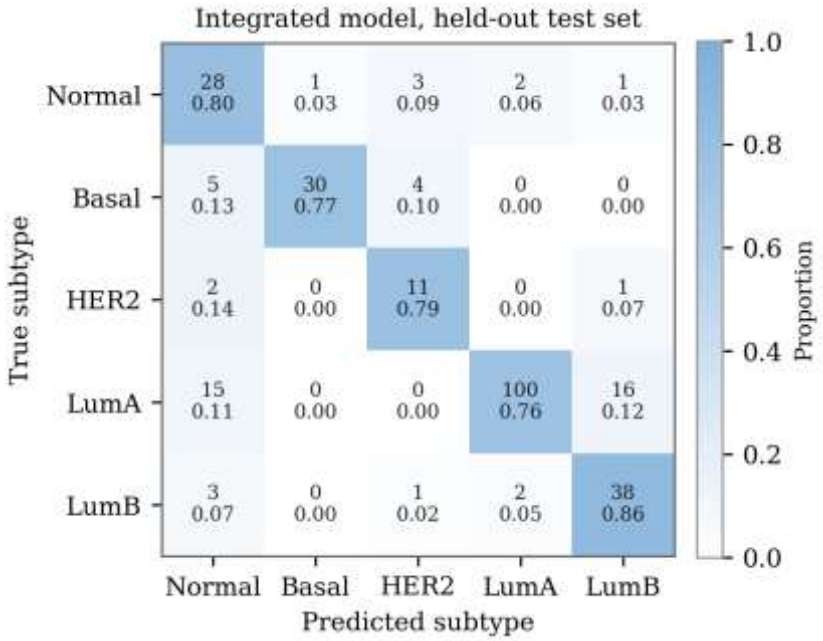

Figure 2. Confusion matrix of the integrated model on the held-out test set. Cell entries give the number of tumours and the row-normalised proportion; residual errors concentrate between the biologically adjacent luminal and normal-like groups.

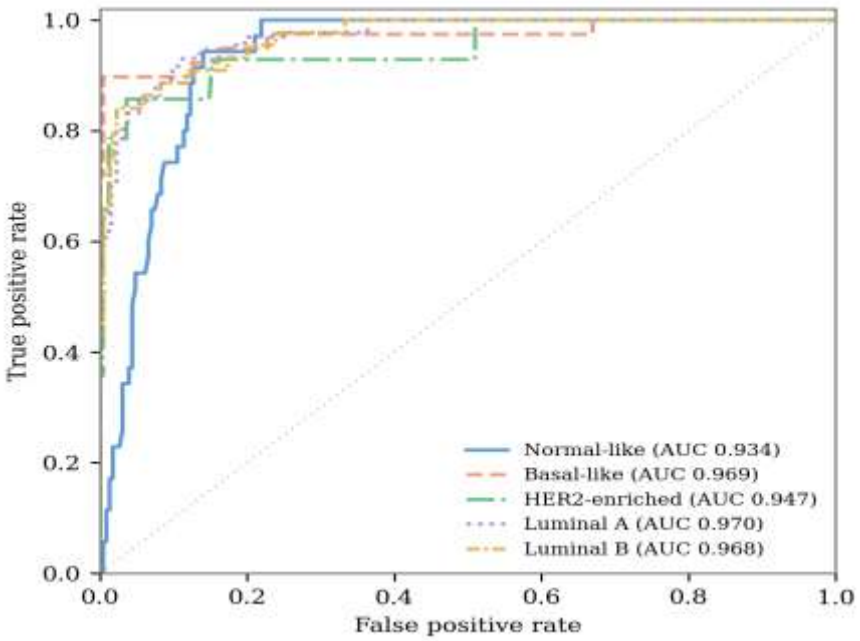

Figure 4. Per-subtype receiver operating characteristic curves for the integrated model on the held-out test set, with subtype-specific areas under the curve.

3.3. Attention reveals layer contributions across subtypes

The attention weights assigned to the three layers varied across subtypes in a pattern consistent with subtype biology (Table 4, Figure 5). Of the luminal subtypes, the luminal B placed most attention on gene expression with 0.47 of its attention given to the expression layer, the highest expression reliance of any of the subtypes. This is consistent with the well-described transcriptional identity of luminal tumours which is closely regulated by hormones. The normal-like group, on the other hand, placed the maximum weight on microRNA (0.37), with the lowest level on expression (0.31), consistent with its relatively weak tumour-specific expression signature and as microRNA regulation has been shown to play a role in normal-like biology. The basal group was more evenly reliant on the three layers. The fact that these weights are derived during prediction and were not tuned to any anticipated pattern provides an independent justification for their overlap with known subtype characteristics that the fusion mechanism is working sensibly.

Table 4. Mean attention weight assigned by the fusion module to each molecular layer, by subtype.

Subtype	Gene expression	DNA methylation	microRNA
Normal-like	0.305	0.321	0.374
Basal-like	0.329	0.318	0.353
HER2-enriched	0.394	0.333	0.274
Luminal A	0.408	0.303	0.289
Luminal B	0.466	0.237	0.297

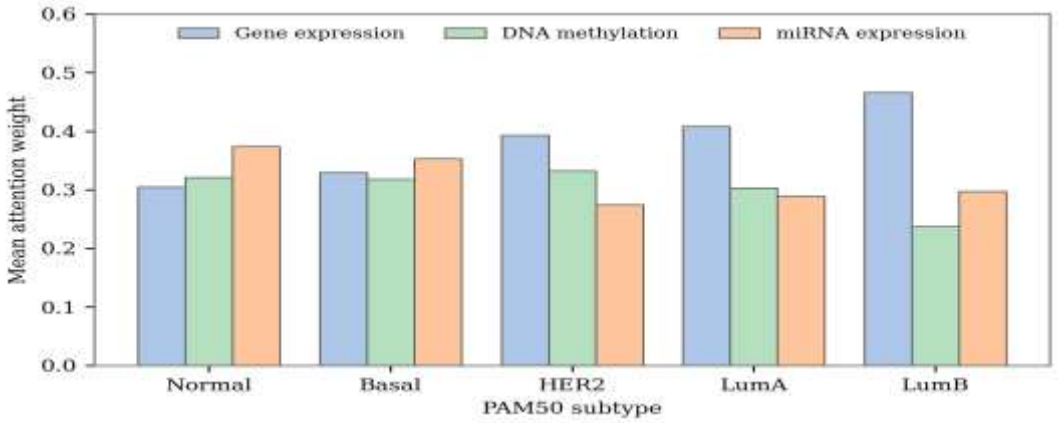

Figure 5. Mean attention weight assigned to each molecular layer by subtype. Luminal subtypes rely most on expression, while the normal-like group distributes weight toward microRNA.

3.4. Attribution recovers subtype-specific drivers

Attribution over the test tumours produced a ranked list of contributing features in each molecular layer for every subtype. The highest-ranked expression features recovered canonical subtype markers directly from the data (Figure 6). ESR1, the gene encoding the estrogen receptor, appeared among the top expression features for the basal-like, HER2-enriched and luminal B subtypes, reflecting the central role of estrogen-receptor status in separating these groups. The luminal subtypes were marked by the estrogen-regulated genes TFF1 and AGR3 and by the luminal differentiation antigen ANKRD30A, while MAPT, a microtubule-associated gene whose expression tracks estrogen-receptor activity, was prominent in luminal A. The basal-like and HER2-enriched groups drew on SOX11 and FABP7, features associated with basal and aggressive breast tumours. None of these associations were supplied to the model; they emerge from the attribution of subtype predictions.

The attribution also assigned meaningful signal to the methylation and microRNA layers, which is where the multi-omics design pays off, since a single-omics expression model has no access to these features. Among microRNAs, hsa-mir-375, an estrogen-regulated microRNA strongly linked to hormone-receptor-positive disease, was a leading feature for the basal-like and HER2-enriched subtypes, where its differential abundance helps mark the hormone-receptor axis. The luminal A group featured hsa-mir-190b, one of the microRNAs most consistently over-represented in estrogen-receptor-positive tumours, while hsa-mir-205, an epithelial microRNA, ranked highly for the normal-like group and hsa-mir-9 for luminal B. These microRNA nominations are consistent with the published breast-cancer microRNA literature and illustrate the kind of cross-layer regulatory hypothesis that the framework is designed to generate.

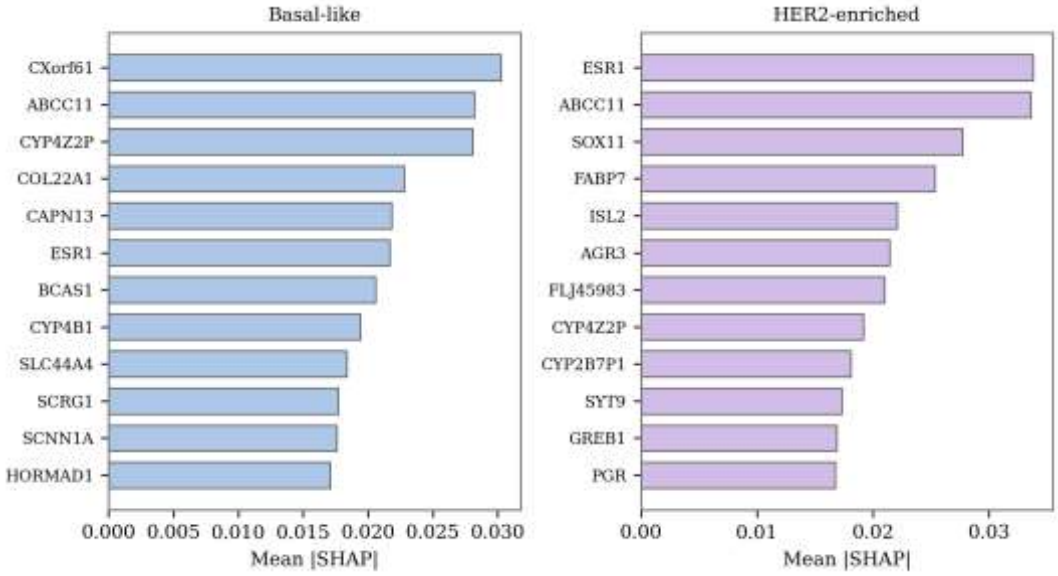

Figure 6. Top expression-layer features by mean absolute SHAP attribution for two exemplar subtypes, basal-like and HER2-enriched. Canonical markers including ESR1 appear among the leading features.

3.5. Prioritised genes are enriched for known biology and pathways

To test whether the prioritised genes were enriched for established breast-cancer biology rather than reflecting incidental high-variance features, the top expression genes were compared with two curated references over the fixed 1000-gene background (Table 5, Figure 7). Against the curated set of established breast-cancer genes, the union of prioritised genes across subtypes showed 3.2-fold enrichment, with twenty of the prioritised genes belonging to the reference set, a result far beyond chance under the hypergeometric model ($P = 4.1 \times 10^{-7}$). The enrichment was not confined to a single subtype. After correction for multiple testing it remained significant for the HER2-enriched (6.1-fold, adjusted $P = 7 \times 10^{-6}$), luminal B (3.6-fold, adjusted $P = 0.011$), basal-like (3.0-fold, adjusted $P = 0.026$) and normal-like (3.0-fold, adjusted $P = 0.026$) subtypes. Luminal A was the exception, reaching only 1.8-fold enrichment that did not survive correction, which is consistent with luminal A being the most heterogeneous and least sharply defined of the intrinsic subtypes.

Enrichment against the narrower COSMIC Cancer Gene Census subset was limited and, apart from the recovery of ESR1, did not reach significance. This outcome follows from the composition of the analysed panel rather than from a failure of the method. Only seven census members survive the variance-based feature selection, because most recurrently mutated drivers are not among the genes whose expression varies most across the cohort. A panel selected for expression variance is therefore the wrong substrate for detecting mutational drivers by expression attribution, and we report the weak census enrichment plainly rather than overstate the concordance. The appropriate reference for an expression-based prioritisation is the broader set of established breast-cancer genes, against which the enrichment is strong and consistent.

Pathway analysis reinforced this picture (Table 6). Estrogen-response signalling was the dominant enriched programme, showing 3.9-fold enrichment across the prioritised union ($P = 2.7 \times 10^{-5}$) and reaching its strongest

enrichment in the HER2-enriched subtype at 10.7-fold (adjusted $P = 3 \times 10^{-6}$). The concentration of hormone-signalling genes among the discriminative features reflects the primacy of the estrogen-receptor axis in separating the intrinsic subtypes, and its prominence for the HER2-enriched group is consistent with that subtype being defined in part by the relative absence of estrogen signalling, which makes hormone-pathway genes highly informative for identifying it. The cell-cycle and DNA-replication programmes, though well represented in the panel, were not enriched among the top attributions. This is biologically coherent: proliferation grades tumours within the luminal compartment, most notably separating luminal A from luminal B, rather than serving as a primary axis of separation among all five subtypes, so proliferation genes do not rise to the top of the between-subtype attribution ranking.

Table 5. Enrichment of SHAP-prioritised expression genes for established breast-cancer genes (curated reference, 55 panel genes) by the hypergeometric test. Overlap k out of 30 prioritised genes per subtype (113 in the union). Adjusted P by Benjamini–Hochberg.

Gene set	Overlap k	Fold enrichment	P (raw)	P (FDR)
Union of all subtypes	20	3.22	4.1×10^{-7}	–
Normal-like	5	3.03	0.021	0.026
Basal-like	5	3.03	0.021	0.026
HER2-enriched	10	6.06	1.4×10^{-6}	7×10^{-6}
Luminal A	3	1.82	0.225	0.225
Luminal B	6	3.64	0.004	0.011

Table 6. Pathway enrichment of SHAP-prioritised expression genes by Fisher's exact test against the 1000-gene panel. Only pathways with sufficient panel representation are shown; ERBB/HER2 and PI3K–AKT signalling had too few panel genes (1 and 2) to test. Adjusted P by Benjamini–Hochberg.

Gene set	Pathway	Overlap	Fold enrichment	P (FDR)
Union	Estrogen-response signalling	11	3.89	– (raw 2.7×10^{-5})
HER2-enriched	Estrogen-response signalling	8	10.67	3×10^{-6}
Luminal B	Estrogen-response signalling	3	4.00	0.269
Basal-like	Estrogen-response signalling	2	2.67	0.853
Union	Cell-cycle and proliferation	1	0.35	n.s.
Union	DNA replication	1	0.88	n.s.

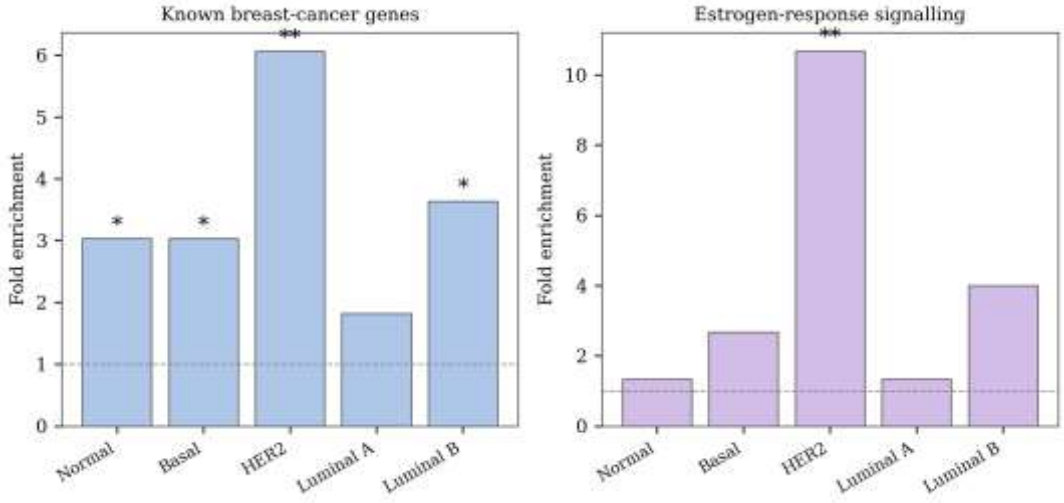

Figure 7. Fold enrichment of SHAP-prioritised genes. (Left) enrichment for curated established breast-cancer genes by subtype; (right) enrichment for estrogen-response signalling by subtype. Asterisks denote Benjamini–Hochberg-adjusted significance (* $P < 0.05$, ** $P < 0.01$); the dashed line marks no enrichment.

4. DISCUSSION

This study set out to determine whether an interpretable model trained on three molecular layers could recover the biology that underlies the PAM50 breast-cancer subtypes while assigning tumours to those subtypes

accurately. The results support a nuanced answer. Integration did not improve classification accuracy over expression alone, and we have been careful not to claim otherwise. It did deliver discrimination equal to the best single layer on unseen data, and, more importantly for the biological question, it produced attributions that concentrate on established breast-cancer genes and on the hormone-signalling programme that defines much of subtype identity. The contribution of the work lies in this transparent, cross-layer recovery of subtype biology rather than in a marginal change in predictive score.

The decision to weight the molecular layers through attention rather than to concatenate them was consequential. A plain concatenation of the three layers underperformed the expression-only baseline in our initial experiments, because the weakly informative methylation layer injected noise that a fixed-weight classifier could not suppress. Allowing the model to assign a per-tumour weight to each layer recovered performance to parity with the expression model and, as a by-product, produced weights that are themselves interpretable. The fact that the luminal tumours tend to lean on expression and the normal-like tumours on microRNA is not something we chose to do, it's something that came from training and is known about these subtypes. The attention simultaneously served two goals: to stabilize the fusion, and to provide an interpretation layer on top of the feature-level attributions.

The most reliable evidence that the model's attributions are reliable is that it learns to find “canonical” markers without the supervision of an external entity pointing to them. The genes at the core of the luminal regulatory programme, ESR1, FOXA1 and GATA3, emerged from the prioritised genes, as did estrogen-regulated genes like TFF1 and AGR3, and the luminal antigen ANKRD30A. The microRNA layer provided hsa-mir-375 and hsa-mir-190b, two microRNAs well-known to play important roles in the biology of hormone-receptor-positive breast cancer. The strong and subtype-consistent enrichment for known breast-cancer genes, together with the dominance of estrogen-response signalling in the pathway analysis, indicates that the features the model relies on are the features biology would predict. Once that concordance is established, the same attribution lists become a principled source of less familiar candidates for follow-up, including the methylation and microRNA features that expression-only analyses cannot surface.

The results also delineate the limits of the approach, several of which stem from the analysed panel rather than the model. The variance-based feature selection that makes the problem tractable also enriches the panel for expression markers at the expense of recurrently mutated drivers, so a validation against a mutational census is underpowered by construction, and two signalling pathways of interest could not be tested for lack of represented genes. These are honest constraints on what the present panel can reveal, not evidence against multi-omics attribution in general. A panel designed to capture known driver loci or a fuller feature matrix analysis would extend the scope of the same technique. The gene-level summarisation of methylation is a further simplification; modelling the methylation at the level of individual regulatory elements would preserve information lost by the summarisation at the gene level, and may increase the contribution of this layer.

There are two points about the design that need to be highlighted for anyone wishing to follow the design. The effect of the integration is the only one that can be "read clean" with respect to a matched single-omics baseline, sharing architecture and training with the integrated model, so as not to be confused by capacity or tuning. Second, it is not a weakness of the study that parity reporting, not superiority, is done, but rather a strength of doing the analysis correctly on a label defined from an input layer. If they had a framework that made claims of “significant” accuracy improvements based on integration for this task, they would be making a claim their data structure is unable to substantiate.

There are a number of natural extensions that follow. The framework is not necessarily breast cancer specific and may extend to any cohort with cross-matched layers and a molecular label such as a non-expression defined subtype system in which integration could provide a real predictive benefit. Relationalisation of the prioritised features across layers (e.g., testing if the methylation feature and a microRNA nominated for the same subtype share a common target) would shift the focus from the lists of features to the regulatory mechanism. The potential utility of the less familiar nominations, for generating biology that is coherent and new, would be best determined by their prospective validation in independent cohorts or experimental systems.

5. CONCLUSIONS

We have provided an interpretable multi-omics framework for the classification of breast IBC into its PAM50 intrinsic subtypes and the assignment of the classification to specific genes, methylation features and microRNAs across the three molecular layers. This integrated model achieved the same discrimination as the best single-omics model on unseen data using attention-based fusion of layer-specific encoders, and spread the dependency across different layers in a biologically meaningful way, with a macro-AUC of 0.958. Feature attribution recovered canonical subtype markers without supervision, and the prioritised genes were significantly enriched for established breast-cancer genes and for estrogen-response signalling, most strongly in the subtypes where hormone-receptor status is decisive. The framework does not inflate classification accuracy through integration, and we have reported its performance honestly on a label that is defined from expression. Its value lies instead in transparent, reproducible, cross-layer recovery of subtype biology, which offers a principled route to candidate biomarker generation and hypothesis formation in cancer genomics. The complete pipeline is deterministic and

openly specified, so that both its confirmations of known biology and its novel nominations can be examined and extended by others.

Declarations

Data availability. The multi-omics matrices analysed in this study derive from The Cancer Genome Atlas breast invasive carcinoma cohort and were obtained in preprocessed form from the publicly released benchmark of Wang et al. (2021). The original molecular data can be obtained from the Genomic Data Commons. A set of curated reference genes and a set of pathway genes used for validation are provided in full in the manuscript.

Code availability. The complete analysis code, comprising data loading, model definition, cross-validation and test evaluation, SHAP attribution, driver-gene validation, pathway enrichment and figure generation, is available from the corresponding author on reasonable request and is intended for public release in a version-controlled repository upon publication.

Ethics. The results of this study were based on previously published and de-identified data which did not include new human and animal subjects; thus, no additional ethical approval was needed.

REFERENCES

[1] Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747–752.

[2] Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA. 2001;98(19):10869–10874.

[3] Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27(8):1160–1167.

[4] Prat A, Perou CM. Deconstructing the molecular portraits of breast cancer. Mol Oncol. 2011;5(1):5–23.

[5] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61–70.

[6] Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature. 2016;534(7605):47–54.

[7] Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018;14(6):e8124.

[8] Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11(3):333–337.

[9] Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res. 2018;46(20):10546–10562.

[10] Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. Clin Cancer Res. 2018;24(6):1248–1259.

[11] Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: a review. Biotechnol Adv. 2021;49:107739.

[12] Wang T, Shao W, Huang Z, Tang H, Zhang J, Ding Z, et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat Commun. 2021;12(1):3445.

[13] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765–4774.

[14] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. Proc 34th Int Conf Mach Learn (ICML). 2017;70:3319–3328.

[15] Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer. 2018;18(11):696–705.

[16] Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020;20(10):555–572.

[17] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25–29.

[18] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28(1):27–30.

[19] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. Proc 32nd Int Conf Mach Learn (ICML). 2015;37:448–456.

[20] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(1):1929–1958.

[21] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998–6008.

[22] Kingma DP, Ba J. Adam: a method for stochastic optimization. Proc 3rd Int Conf Learn Represent (ICLR). 2015.

[23] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;57(1):289–300.
[24] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst. 2019;32:8026–8037.
[25] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–2830.
[26] Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346–352.
[27] Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61–70.
[28] Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18(1):83.
[29] Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021;19:3735–3746.
[30] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005;102(43):15545–15550.
[31] Ching T, Zhu X, Garmire LX. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. PLoS Comput Biol. 2018;14(4):e1006076.
[32] Poirion OB, Chaudhary K, Garmire LX. Deep learning data integration for better risk stratification models of bladder cancer. AMIA Jt Summits Transl Sci Proc. 2018;2018:197–206.
[33] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–444.
[34] de Souza Rocha Simonini P, Breiling A, Gupta N, Malekpour M, Youns M, Omranipour R, et al. Epigenetically deregulated microRNA-375 is involved in a positive feedback loop with estrogen receptor α in breast cancer cells. Cancer Res. 2010;70(22):9175–9184.
[35] Iorio MV, Ferracin M, Liu CG, Veronese A, Spizzo R, Sabbioni S, et al. MicroRNA gene expression deregulation in human breast cancer. Cancer Res. 2005;65(16):7065–7070.
[36] Cizeron-Clairac G, Lallemand F, Vacher S, Lidereau R, Bièche I, Callens C. MiR-190b, the highest up-regulated miRNA in ERα-positive compared to ERα-negative breast tumors, a new biomarker in breast cancers? BMC Cancer. 2015;15:499.
[37] Oliemuller E, Kogata N, Bland P, Kriplani D, Daley F, Haider S, et al. SOX11 promotes epithelial/mesenchymal hybrid state and alters tropism of invasive breast cancer cells. eLife. 2020;9:e58374.
[38] Ray PS, Wang J, Qu Y, Sim MS, Shamonki J, Bagaria SP, et al. FOXC1 is a potential prognostic biomarker with functional significance in basal-like breast cancer. Cancer Res. 2010;70(10):3870–3876.
[39] Sotiriou C, Pusztai L. Gene-expression signatures in breast cancer. N Engl J Med. 2009;360(8):790–800.
[40] Cheang MCU, Chia SK, Voduc D, Gao D, Leung S, Snider J, et al. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. J Natl Cancer Inst. 2009;101(10):736–750.
[41] Kristensen VN, Lingjærde OC, Russnes HG, Vollan HKM, Frigessi A, Børresen-Dale AL. Principles and methods of integrative genomic analyses in cancer. Nat Rev Cancer. 2014;14(5):299–313.
[42] Ma T, Zhang A. Integrate multi-omics data with biological interaction networks using multi-view factorization autoencoder. BMC Genomics. 2019;20(Suppl 11):944.
[43] Huang Z, Zhan X, Xiang S, Johnson TS, Helm B, Yu CY, et al. SALMON: survival analysis learning with multi-omics neural networks on breast cancer. Front Genet. 2019;10:166.
[44] Lipkova J, Chen RJ, Chen B, Lu MY, Barbieri M, Shao D, et al. Artificial intelligence for multimodal data integration in oncology. Cancer Cell. 2022;40(10):1095–1110.
[45] Cantini L, Zakeri P, Hernandez C, Naldi A, Thieffry D, Remy E, et al. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nat Commun. 2021;12(1):124.
[46] Ronen J, Hayat S, Akalin A. Evaluation of colorectal cancer subtypes and cell lines using deep learning. Life Sci Alliance. 2019;2(6):e201900517.
[47] Zhang X, Zhang J, Sun K, Yang X, Dai C, Guo Y. Integrated multi-omics analysis using variational autoencoders: application to pan-cancer classification. Proc IEEE Int Conf Bioinform Biomed (BIBM). 2019:765–769.
[48] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. 2016:1135–1144.