Computational and Structural Biotechnology Journal 27 (2025) 4505–4516

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

Multimodal fusion strategies for survival prediction in breast cancer: A comparative deep learning study

Aurora Sucre [a,b,c,*], Xabier Calle Sánchez [a,b], Laura Valeria Perez-Herrera [a,b], María dM Vivanco [d], María Jesús García-González [a,b,e], Karen López-Linares [a,b], Borja Calvo [c], Alba Garin-Muga [a,b]

[a] Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, Donostia-San Sebastián 20009, Spain
[b] Biogipuzkoa Health Research Institute, E-Health Group, Donostia-San Sebastián 20014, Spain
[c] Department of Computer Science and Artificial Intelligence, University of the Basque Country (EHU), Donostia-San Sebastián 20018, Spain
[d] Cancer Heterogeneity Lab, CIC bioGUNE, Basque Research and Technology Alliance (BRTA), Derio 48160, Spain
[e] Universitat Politècnica de València, Valencia 46022, Spain

ARTICLE INFO

Keywords:
Multiomics
Deep learning
Neural networks
Multimodal fusion
Breast cancer
Survival prediction

ABSTRACT

Accurate survival prediction in breast cancer remains a key challenge in oncology, requiring models that can integrate diverse clinical, molecular, and imaging data sources to guide breast cancer management. While recent deep learning models have explored multimodal integration for cancer survival prediction, their generalizability to unseen data remains limited. In this study, we developed and optimized unimodal and multimodal models for breast cancer survival prediction, systematically assessing our optimized early and late integration strategies and their impact on out-of-sample generalization performance. We integrated clinical variables, somatic mutations, RNA expression, copy number variation, miRNA expression, and histopathology images from The Cancer Genome Atlas breast cancer dataset. Across all modality combinations, late fusion models consistently outperformed early fusion approaches and late and intermediate benchmark methods, with the combination of omics and clinical data yielding the highest test-set concordance indices. Explainability analyses showed that our models captured biologically relevant features associated with patient survival. These findings highlight the value of late-fusion multimodal deep learning frameworks for robust and explainable survival prediction in breast cancer.

1. Introduction

Breast cancer remains the most diagnosed malignancy and a leading cause of cancer-related mortality in women worldwide [1]. Accurate survival prediction is essential for guiding treatment decisions and improving patient outcomes. While traditional prognostic models rely heavily on clinicopathological variables such as age, tumor stage, and molecular subtype [2], these clinical features alone do not fully account for the molecular and microenvironmental heterogeneity that drives patient-specific outcomes.

Recent advances in high-throughput molecular profiling and digital pathology have created new opportunities for improving survival prediction. Large-scale initiatives such as The Cancer Genome Atlas (TCGA), provide publicly available datasets that integrate clinical, omics, and histopathology imaging data for thousands of breast cancer patients [3]. Leveraging these multimodal datasets, several deep learning models, including MCAT [4], PORPOISE [5], and MGCT [6], have been developed to predict survival.

However, while these models have demonstrated strong training performance, they encounter significant challenges in generalizing effectively to unseen data within the same cohort. Recurrent limitations across the broader literature include susceptibility to overfitting (particularly with limited sample sizes), difficulties in effectively integrating heterogeneous data modalities, and the use of single data splits for validation, which can increase the variability in the estimation of the true performance [7–9].

Within this context, a crucial unresolved question in multimodal integration is how to identify the optimal fusion strategy that reliably enhances model generalization. Early fusion approaches, which simply concatenate features from all modalities at the input layer, risk

* Corresponding author at: Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, Donostia-San Sebastián 20009, Spain.
E-mail address: amsucre@vicomtech.org (A. Sucre).

https://doi.org/10.1016/j.csbj.2025.10.038
Received 8 August 2025; Received in revised form 19 October 2025; Accepted 20 October 2025
Available online 21 October 2025

overfitting and poor gradient flow in high-dimensional settings [10]. While methods employing later-stage (e.g., intermediate or late) fusion are established [4–6] to combine modality-specific predictions at decision level, their relative impact on model generalizability and optimal implementation for breast cancer survival prediction remain less thoroughly assessed.

In this study, we systematically evaluate unimodal and multimodal deep learning models for overall survival prediction in breast cancer, utilizing a comprehensive TCGA-derived dataset that includes clinical, omics, and histopathology imaging data (Fig. 1). Our approach directly addresses the identified generalization limitations by focusing on enhancing modelling robustness through three key steps: (1) implementing a rigorous validation framework based on cross-validation and a fixed, held-out test set to ensure an unbiased assessment of out-of-sample performance; (2) performing a systematic comparison of early and late fusion to identify the most robust integration strategy; and (3) benchmarking our models against established late and intermediate multimodal frameworks (MCAT, PORPOISE, MGCT) under consistent conditions to rigorously assess generalization capabilities. Furthermore, we apply explainability methods to elucidate the molecular features driving survival predictions, aiming to confirm biologically relevant signals and enhance model plausibility.

2. Materials and methods

2.1. Data acquisition and preprocessing

2.1.1. Patient cohort and outcome definition

Data was obtained from TCGA, a publicly available repository of molecular and clinical profiles across cancer types [3]. We focused on breast cancer patients with matched data across multiple modalities. The dataset included tabular data comprising both clinical variables and genomic profiles (somatic single nucleotide variants (SNVs), RNA gene expression (RNA-seq), gene-level copy number variations (CNVs), and microRNA expression (miRNA)), downloaded from the UCSC XENA browser [11], as well as histopathology images derived from H&E-stained whole-slide images, downloaded from the TCGA portal. Further details can be found in **SM1.a**.

Only primary tumor samples from female patients were kept; normal tissue and male patient samples were excluded. The survival data for the cohort was established by calculating the survival months from days to death or to the last follow-up, and censorship was derived from vital status. The primary outcome was defined as a discretized survival endpoint, as described in [5]. To create this endpoint, the continuous survival time was binned into four categories based on the observed survival distribution.

2.1.2. Clinical data

The chosen clinical features consisted of patient-level variables extracted from TCGA, including demographic data (age at index, referring to the moment the patient was enrolled in the study; age of diagnosis, referring to when breast cancer was initially diagnosed; and race), tumor subtype (PAM50, OncoTree), and pathological features (stage). Preprocessing included one-hot encoding of categorical variables (race, OncoTree and PAM50) and imputation of missing values for age at index and age at diagnosis. In 1.5 % of patients, where only one of these values was missing, it was imputed using the other available value from the same individual. The final clinical matrix contained 27 features from 1084 patients (Table 1).

2.1.3. Omics data

Omics data included four molecular modalities derived from TCGA: SNV, RNA-seq, CNVs and miRNA. Distinct filtering strategies were employed for each modality to maximize feature relevance while managing dimensionality:

- Somatic SNV data were processed into a binary sample-by-gene matrix indicating the presence of mutations. A low 1 % mutation frequency threshold was chosen to ensure the retention of sufficient genetic variability for model training.
- RNA-seq data, reported in FPKM units, were transposed to a sample-by-gene format and restricted to genes annotated as cancer-related in the CGN MSigDB gene set [12,13], because initial frequency-based

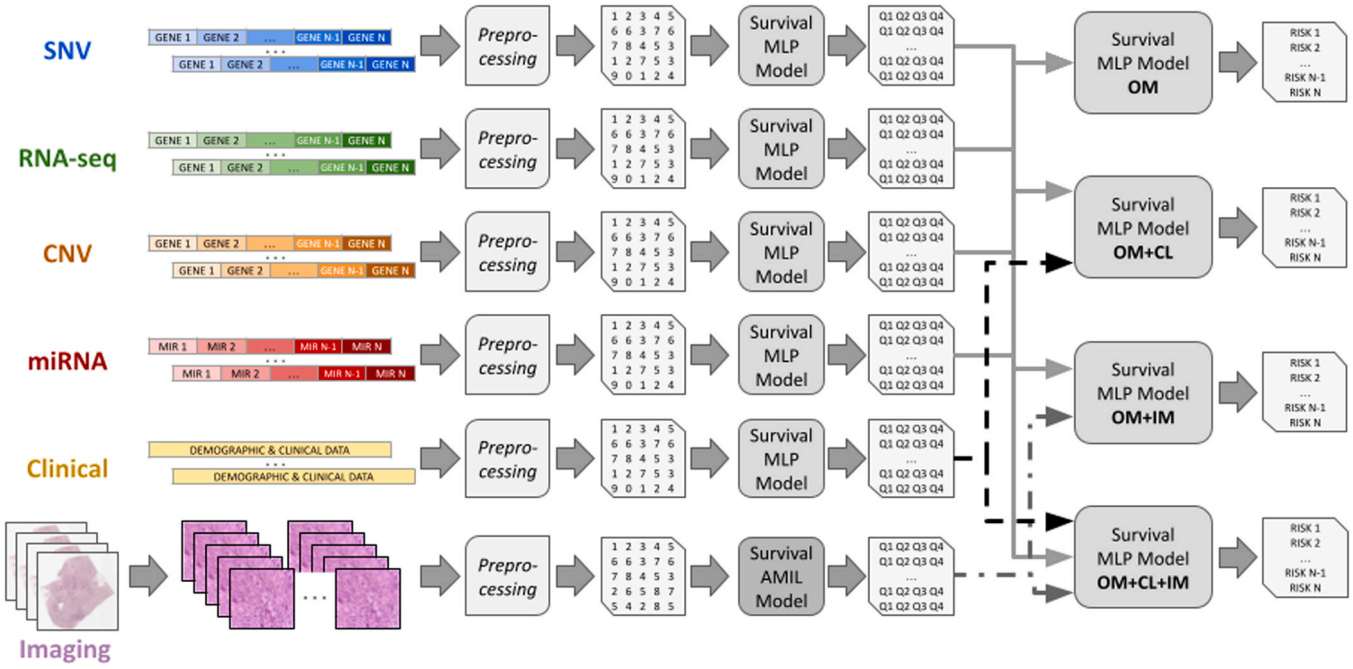

Fig. 1. Model architecture: architecture implemented for survival prediction using unimodal (single data modality) and multimodal (combined clinical, omics, and/or imaging) data. Omics data include SNVs, RNA-seq, CNVs, and miRNA data. Imaging data consisted of hematoxylin and eosin–stained whole-slide images. In the figure, CL = clinical data, OM = omics data, IM = imaging data, MLP = multilayer perceptron, AMIL = attention-based multiple instance learning.

Table 1
Model data: Number of patients, features and dataset splits for each data source. Data sources include clinical data (CL), omics data (OM) and imaging (IM). Dataset splits include training (TRAIN), validation (VAL), and test (TEST) sets.

DATASET		PROCESSED DATA		SPLIT SETS (PATIENTS)		
		PATIENTS	FEATURES	TRAIN	VAL	TEST
CL	Clinical	1084	27	728	181	175
OM	Somatic SNV	967	1874	634	158	175
	RNA Expression	1093	4863	735	183	175
	Gene-Level CNV	1035	4868	688	172	175
	miRNA Expression	1076	981	721	180	175
IM	Imaging data	872	-	558	139	175
LATE INTEGRATION (L)	OM (L)	735	$4 \times 4 = 16$	448	112	175
	OM+CL (L)	726	$5 \times 4 = 20$	441	110	175
	OM+IM (L)	675	$5 \times 4 = 20$	400	100	175
	CL+IM (L)	861	$2 \times 4 = 8$	549	137	175
	OM+CL+IM (L)	666	$6 \times 4 = 24$	393	98	175
EARLY INTEGRATION (E)	OM (E)	735	12,586	448	112	175
	OM+CL (E)	726	12,613	441	110	175
	OM + CL + IM (E)	666	13,636	393	98	175

filtering was insufficient for feature reduction, so a biologically informed approach was adopted.

- CNV data were also transposed to a sample-by-gene format and normalized to a range of –2 – 2, with gene-level scores aggregated across isoforms and filtered to remove invariant features and to only keep cancer-related genes according to the CGN MSigDB gene set, following the same logic as for RNA-seq data filtering.
- miRNA expression profiles were processed following the same approach and filtered to retain only miRNAs altered in at least 10 % of the cohort. The higher threshold was chosen to prioritize features with stable and reliable expression signals, by excluding transcripts with low detection across most samples.

After preprocessing, the final genomic datasets ranged in feature size from 981 to 4868 and included between 967 and 1093 patients, depending on the modality (Table 1).

2.1.4. Image data

Imaging data consisted of hematoxylin and eosin (H&E)–stained whole-slide images (WSI) from TCGA [3], annotated at image-level. They were processed using the CLAM pipeline [14] for automated tissue segmentation followed by 512×512 patch extraction. Each extracted patch was encoded into a 1024-dimensional feature vector using pretrained deep neural networks. Further details can be found in **SM2**. The final dataset included slides from 872 patients (Table 1).

2.1.5. Data integration, splitting and cross-validation strategy

Integrated datasets, varying from 666 to 861 patients, were generated (Table 1). Their sizes varied based on the number of samples available across the merged modalities; not every patient has information available for all considered data modalities.

To ensure consistent model evaluation and comparability across experiments, we defined a fixed test set comprising 175 patients, corresponding to 20 % of the smallest unimodal dataset. This test set was held out for final performance assessment and was not used during training or hyperparameter tuning. Per modality, the remaining samples were partitioned into five stratified folds for cross-validation, maintaining balanced representation of outcome categories within each fold (Table 1). Model development and optimization were conducted exclusively within the training folds.

2.2. Predictive modeling framework

We implemented supervised models to predict overall survival based on unimodal and multimodal data inputs, as illustrated in Fig. 1. A robust framework was established for model optimization, training and evaluation. The central goal was to minimize a discretized survival log-likelihood function based on the PORPOISE methodology [5]. The concordance index (C-Index) and Integrated Brier Score (IBS) were used as the primary evaluation metrics.

2.2.1. Survival loss function

For survival prediction, our three chosen benchmark models [4–6] adopted a discretized survival modeling approach that accounts for both right-censored and uncensored cases, based on the methodology proposed in the PORPOISE [5] manuscript. Therefore, we followed this same approach and employed its loss function to ensure a fair comparison. The objective during training is to minimize the negative log-likelihood of the observed survival outcomes.

The first step involves discretization, dividing continuous survival times into four discrete intervals (bins), defined by the quartiles of the observed survival distribution. This reformulates the continuous time-to-event prediction problem into a sequence of conditional survival probability estimations, where the model predicts the likelihood of surviving each interval and the probability of the event (death) occurring within it. This process naturally accounts for right-censored patients, for whom the true event times are unknown, by considering survival up to their last observed interval.

For each patient, the model estimates the hazard function (the conditional probability of experiencing the event in a given interval, depending on survival up to that point) and the survival function (the cumulative probability of not experiencing the event across all previous intervals). Therefore, a higher hazard probability implies a greater immediate risk of death and, correspondingly, a lower survival probability.

The loss function rewards different outcomes based on censorship status. For uncensored patients, the loss rewards high survival probabilities up to the interval preceding the event and high hazard probabilities at the correct event interval. For censored patients, however, the loss rewards high survival probabilities only up to the censoring interval without penalizing predictions beyond it due to unobserved outcomes.

The final objective function, denoted as L_{surv}, minimizes the base likelihood term L while incorporating an additional term emphasizing uncensored cases, weighted by an adaptable factor β:

$$L_{surv} = (1-\beta) \times L + \beta \times L_{uncensored}$$

where L is defined as:

$$L = -c_j \times \log\left(f_{surv}\left(Y_j | h_{bag_j}\right)\right) - (1-c_j) \times \left[\log\left(f_{surv}\left(Y_j - 1 | h_{bag_j}\right)\right) + \log\left(f_{hazard}\left(Y_j | h_{bag_j}\right)\right)\right]$$

where c_j is the censoring status for patient j, Y_j its assigned survival bin and $h_{bag\,j}$ its feature vector.

The parameter β controls the influence of uncensored samples, which

provide stronger supervision since their true event times are known. Increasing β increases the contribution of these cases, thereby stabilizing convergence and preventing the model optimization from being dominated by less informative censored samples in high-censoring ratio datasets.

2.2.2. Model architecture and optimization

Unimodal and multimodal supervised models were implemented to predict overall survival. Individual models from tabular data and multimodal models employed a shared architecture template consisting of fully connected feedforward neural networks. Individual image models, however, utilized convolutional networks and transformer architectures for feature extraction.

The specific depth, width, and training configuration were adapted for each scenario through hyperparameter optimization. Hyperparameters were tuned using the Optuna framework [15] with nested cross-validation on the training folds (Hyperparameter ranges are described in **SM1.b**).

Model training and validation were then performed using PyTorch [16], following a stratified five-fold cross-validation scheme, with a fixed test set held out for final evaluation.

2.2.3. Unimodal models

Unimodal models were first trained independently for each data type. Tabular data models were built using individual matrices for clinical variables, SNVs, RNA-seq, CNV, and miRNA. These models shared a common architecture template, based on fully connected feedforward neural networks. For histopathology images, an attention-based multiple instance learning (MIL) approach was used to predict survival using the WSIs. The MIL framework included a feature extractor where two encoders were tested: ResNet-50 general-purpose model [17], and UNI pathology foundation model [18]; then, an attention mechanism for slide-level aggregation; and finally, a fully connected layer to predict survival categories.

2.2.4. Multimodal models

Integrative models were constructed to combine information across modalities. For omics-only integration (OM), two strategies were employed. Early integration involved concatenating feature matrices from the four omics types into a single input vector for model training. Conversely, late integration used the optimized unimodal models to generate output logits (unnormalized scores from the last model layer), which were concatenated and passed to a secondary fusion network. Both approaches were also evaluated across different data combinations to incorporate clinical (CL) and imaging (IM) data (e.g., omics plus clinical, omics plus imaging features, or full integration). Input dimensionality varied across configurations, and model capacity was scaled accordingly during optimization.

2.2.5. Benchmark models

For comparative analysis, we replicated three well-established multimodal deep learning approaches that integrate clinical, omics and imaging data: MCAT, PORPOISE, and MGCT. These were chosen for their relevance in the field of multiomics integration for survival prediction. For each model we used parameter values proposed by their authors or default settings when unavailable. To ensure a fair comparison, all these models were trained and validated using the same TCGA dataset, fixed test set, and cross-validation strategy as our models.

PORPOISE is a late-fusion multimodal deep learning framework which integrates whole-slide images (WSIs) of histopathology with multi-omics data, such as mutation status (SNV), copy-number variation (CNV) and RNA expression profiles, to predict patient survival. The histopathology branch employs a multiple instance learning pipeline with an attention mechanism (AMIL) based on CLAM to extract and aggregate patch-level features extracted by a pre-trained ResNet-50 into a slide-level representation. The omics branch processes each molecular modality through a self-normalizing neural network (SNN) to generate compact feature representations. These image- and omics-level embeddings are then combined via a Kronecker product, based fusion layer, which enables the joint modelling of cross-modal interactions between molecular and morphological patterns.

MGCT uses a similar multimodal pipeline as PORPOISE, based on the same CLAM-based workflow for histopathology images and a self-normalizing network (SNN) for multi-omics features. However, the main distinction lies in its biological prior and fusion strategy (intermediate). Gene expression features are grouped into six functional categories to reflect pathway-level organization. Additionally, the model introduces an intermediate, mutual-guided cross-modality attention mechanism that enables bidirectional information exchange between histology and omics representations. This design enables each modality to guide the feature learning of the other, resulting in more coherent and biologically aligned multimodal representations.

MCAT processes each modality similarly to MGCT but integrates them via an intermediate genomic-guided co-attention mechanism. In this transformer-based approach, genomic embeddings act as queries and histopathology (WSI) embeddings act as both keys and values. This allows the model to generate genomic-guided visual representations. Ideally, this fusion strategy enables pathology features to be modulated by genomic context, improving alignment between molecular and morphological information.

2.2.6. Model evaluation metrics

Model performance was evaluated using two distinct metrics: the concordance index (C-Index) for discrimination and the Integrated Brier Score (IBS) for calibration. All metrics were computed as five independent values generated by the 5-fold cross-validation scheme (one value per fold).

The C-Index, a measure of the model's ranking accuracy, was computed for training, validation and test sets. To apply this metric to our multi-category output, a single scalar risk score (η) was calculated per patient using the negative of the estimated Restricted Mean Survival Time (RMST):

$$\eta = -\sum_{i=1}^{4} P(Survival > \quad t_i)$$

This scalar risk score was then used as input in the standard time-dependent C-Index formulation provided by the scikit-survival's concordance_index_censored function [20], along with the real survival times and event information.

The IBS, a measure of prediction error and calibration, was calculated solely for the fixed test set using the IBS function from the SurvMetric R package [21]. For the calculation, the full matrix of predicted survival probabilities, the observed survival times and the discretization time thresholds were employed, providing a continuous metric of prediction accuracy over the full observation period.

For each performance metric (C-Index and IBS), the final results are reported as the mean and standard deviation (SD) of the five values obtained from the cross-validation folds. The SD was normalized by N-1 as recommended to obtain the sample SD (rather than the population SD). Finally, to facilitate the assessment of model performance distribution, violin plots were generated for all key metrics.

2.3. Model explainability and biological analysis

To interpret model behavior and assess the prognostic relevance of molecular features, we applied SHAP (SHapley Additive exPlanations) to the late integration models and to unimodal models trained on RNA-seq and miRNA data, using the DeepExplainer from the SHAP Python package [19]. SHAP values were used to quantify the contribution of each input feature to survival predictions. Unimodal features were ranked by their average absolute SHAP value, and the 10 top-ranked

genes and miRNAs were selected.

For each modality, we assigned a label to each top feature per sample, classifying its expression as overexpressed (>Q75), under-expressed (<Q25) or normal (>Q25, <Q75). Next, for each modality and sample, we calculated the total count of features classified as overexpressed or under-expressed among the selected top-ranked features. Based on the relative prevalence of these over- and under-expressed features, patients were stratified into three groups: predominantly overexpressed, predominantly under-expressed, and balanced profiles. Kaplan–Meier survival analysis and Cox proportional hazards regression were then performed on these groups using the survival R library [22], to evaluate the prognostic significance of aggregated expression patterns. This analysis enabled us to assess whether features identified as important by the models were also associated with survival-relevant stratification in the patient cohort.

In addition, we performed an enrichment analysis on the top features using the ClusterProfiler R library [23] to discover which terms from gene (GO), pathways (KEGG) and disease ontologies (HPO, HDO, NCG, DisGeNet) were significantly associated with our top genes and miRNA. To find miRNA-related terms, we first identified their target genes and performed enrichment analysis on those. This analysis allowed us to evaluate the biological relevance and functional implications of our models' top features.

Further details regarding the explainability analysis pipeline can be found in **SM2**.

3. Results

3.1. Overall survival distribution

The Overall Survival (OS) distribution for the patient cohort is represented using the Kaplan-Meier curves in Fig. 2. This figure includes separate survival curves for the full cohort and the held-out test set. The cohort was drawn from the TCGA initiative, including female breast cancer patients. Overall survival time and censorship status were obtained from this same source for model training and evaluation.

The median survival time was 27 months for the full cohort and 29 months for the test set. The overall event rate (proportion of death events) was 13.83 % in the full cohort and 12 % in the test set. The close alignment between the survival curves of the two groups confirms that the test set is statistically representative of the overall population, which is vital for ensuring an unbiased assessment of model generalization.

3.2. Model optimization

Extensive hyperparameter optimization was performed to ensure fair and effective comparisons across data types and integration strategies. The process revealed distinct architectural preferences depending on the input modality and fusion strategy, as summarized in Table 2.

Unimodal models generally favored deeper architectures, with optimal depths ranging from one to seven layers. The gene-level CNV model selected the deepest network (seven layers), while the miRNA model was optimized with a single-layer configuration. Among integrative models, no single consistent depth preference emerged. Early fusion models for tabular data (omics-only, omics+clinical) favored deeper architectures with five layers, whereas the corresponding late fusion models performed best with shallow configurations of a single layer. In contrast, the early and late integration models including all modalities (omics+clinical+imaging) were optimized with a 2-layer configuration while the other late integration models including imaging data (omics+imaging, clinical+imaging) favor deeper architectures.

Across models, hidden layers were typically wide (selecting over 100 neurons per layer), though some models favored narrower configurations. Learning rates spanned a broad range from 1e-5–1e-2 without a clear pattern tied to modality. A batch size of 64 was selected for most models. No consistent preference was observed for any specific

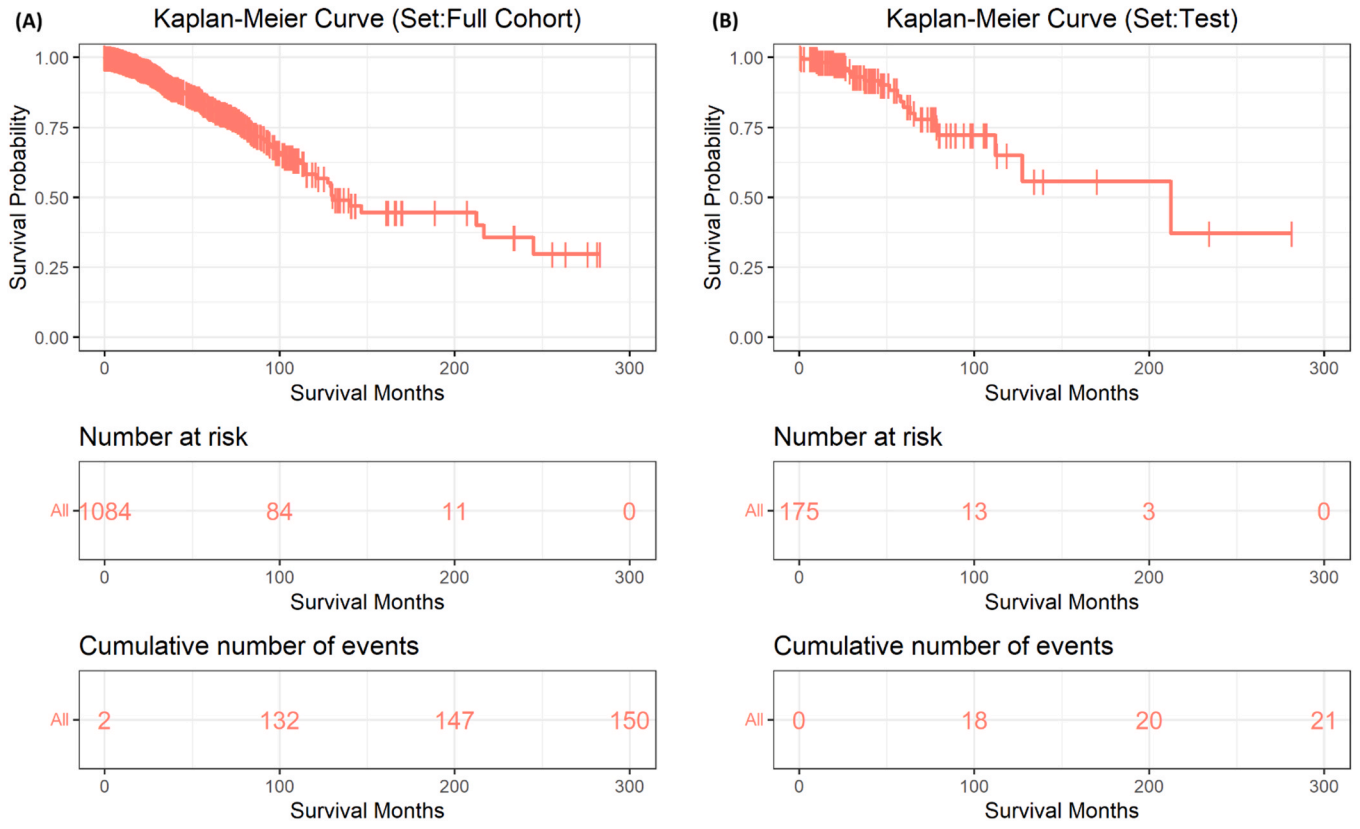

Fig. 2. Overall survival distribution: Kaplan-Meier curves representing the overall survival for (A) the entire study cohort and (B) the held-out test set.

Table 2
Optuna hyperparameters: optimized hyperparameters across model architectures. In the figure, CL = clinical data, OM = omics data, IM = imaging data, NL = number of layers, NN = number of neurons per layer, LR = learning rate, BS = batch size, OPT = optimizer, β = loss function weighting factor.

SUBSET	MODEL	NL	NN	LR	BS	OPT	β
CL	Clinical	3	140, 168, 144	3.36e−3	16	Adam	0.5
OM	SNV	6	248, 156, 92, 120, 236, 224	1.28e−2	64	SGD-M	0.5
	RNA-seq	6	164, 224, 120, 256, 112, 136	7.48e−4	16	Adam	0.5
	CNV	7	228, 132, 248, 256, 56, 228, 68	8.32e−2	32	SGD-M	0.5
	miRNA	1	152	2.09e−3	16	SGD	0.5
LATE INTEGRATION (L)	OM (L)	1	100	4.45e−2	64	SGD-M	0.5
	OM+CL (L)	1	244	1.27e−2	64	SGD	0.5
	OM+IM (L)	7	176, 200, 120, 80, 108, 76, 48	7.24e−4	64	RMSprop	0.5
	CL+IM (L)	5	116, 244, 72, 108, 184	1.36e−3	64	RMSprop	0.5
	OM+CL+IM (L)	2	208, 44	8.36e−2	64	SGD-M	0.5
EARLY INTEGRATION (E)	OM (E)	5	128, 20, 132, 120, 120	1.31e−2	64	SGD	0.5
	OM+CL (E)	5	20, 244, 120, 160, 180	7.88e−5	64	Adam	0.5
	OM+CL+IM (E)	2	128, 216	4.91e−5	64	RMSprop	0.5

optimizer (Adam, SGD, SGD with momentum, and RMSprop were all represented). The β parameter used to weigh the loss component was consistently set to 0.5 in all optimal models.

3.3. Model performance

The performance of all models was evaluated using the C-Index (discrimination) and the IBS (calibration). Results are presented as the mean ± Standard Deviation (SD) of the 5-fold cross-validation runs (Table 3) with statistical significance assessed via the visual comparison of metric distributions using violin plots (Fig. 3). We first examined unimodal performance across all input types and then assessed the effect of integrating multiple data modalities. Finally, we compared our models to established multimodal benchmarks under consistent evaluation conditions.

3.3.1. Discrimination power

Two feature extractors were evaluated in the MIL framework to predict survival using imaging data alone (Details on the feature extractors can be found in **SM3**). The general-purpose encoder produced the best test results (0.649 ± 0.034) and was thus chosen for integration.

Evaluation of unimodal models revealed high variability in performance and stability, with training mean C-indices ranging from 0.557 to 0.739, and testing mean C-indices between 0.485 and 0.697 (Table 3). The model based on SNVs, despite showing the highest training score (0.739 ± 0.178), also demonstrated the largest performance reduction upon testing (0.485 ± 0.085), suggesting significant overfitting. Conversely, the clinical data model maintained a stable performance with the best testing score (train: 0.730 ± 0.043; test: 0.697 ± 0.051), followed closely by the miRNA model (test: 0.668 ± 0.028). Models based on RNA-seq, miRNA, and imaging showed a more consistent out-of-sample generalization compared to the more volatile SNV and CNV models.

The integration strategy had a marked effect on generalization and stability. As displayed in Table 3 and visually confirmed by the violin plot (Fig. 3-A), late integration models consistently and significantly outperformed their early integration counterparts.

Among the late integration approaches, the omics+clinical model achieved the highest testing score (0.740 ± 0.002), slightly exceeding the full OM+CL+IM model (0.705 ± 0.009). The OM+CL model demonstrated the highest statistical stability, reflected by its narrowest distribution in the violin plot, not overlapping with any other model tested. The OM-only and OM+CL+IM late models followed, showing similarly low variability.

In contrast, early integration models yielded lower overall C-index averages, and their distributions were significantly wider. This pattern of poor performance and high variability across all early integration results confirms the methodological superiority, in terms of

Table 3
Model performance: survival prediction performance (C-Indexes & IBS) across data modalities and integration approaches for training (Train), validation (Val) and testing (Test). State of the art published metrics are given as reference (Ref). In the figure, CL = clinical data, OM = omics data, IM = imaging data., IBS=Integrated Brier Score, AVG=Average, SD=Standard Deviation.

SUBSET	MODEL	C-Index (Ref)	C-Index (Train) *AVG* ± *SD*	C-Index (Val) *AVG* ± *SD*	C-Index (Test) *AVG* ± *SD*	IBS (Test) *AVG* ± *SD*
SOA	MCAT	0.580	0.952 ± 0.009	0.557 ± 0.098	0.658 ± 0.051	0.398 ± 0.068
	PORPOISE	0.628	0.961 ± 0.010	0.528 ± 0.131	0.546 ± 0.014	0.290 ± 0.052
	MGCT	0.608	0.949 ± 0.041	0.473 ± 0.110	0.552 ± 0.016	0.571 ± 0.095
CL	Clinical	N/A	0.730 ± 0.043	0.716 ± 0.094	0.697 ± 0.051	0.220 ± 0.020
OM	SNV	N/A	0.739 ± 0.178	0.538 ± 0.067	0.485 ± 0.085	0.243 ± 0.016
	RNA-seq	N/A	0.627 ± 0.087	0.603 ± 0.066	0.642 ± 0.029	0.223 ± 0.020
	CNV	N/A	0.557 ± 0.114	0.516 ± 0.103	0.49 ± 0.031	0.223 ± 0.028
	miRNA	N/A	0.645 ± 0.054	0.577 ± 0.038	0.668 ± 0.028	0.271 ± 0.065
IM	Imaging data	N/A	0.628 ± 0.026	0.558 ± 0.131	0.546 ± 0.014	0.300 ± 0.047
LATE INTEGRATION (L)	OM (L)	N/A	0.588 ± 0.046	0.628 ± 0.084	0.610 ± 0.007	0.205 ± 0.032
	OM+CL (L)	N/A	0.726 ± 0.021	0.727 ± 0.073	0.740 ± 0.002	0.218 ± 0.023
	OM+IM (L)	N/A	0.611 ± 0.038	0.640 ± 0.058	0.605 ± 0.018	0.206 ± 0.028
	CL+IM (L)	N/A	0.733 ± 0.032	0.744 ± 0.058	0.617 ± 0.046	0.224 ± 0.038
	OM+CL+IM (L)	N/A	0.747 ± 0.011	0.749 ± 0.051	0.705 ± 0.009	0.227 ± 0.044
EARLY INTEGRATION (E)	OM (E)	N/A	0.565 ± 0.057	0.487 ± 0.078	0.557 ± 0.047	0.230 ± 0.013
	OM+CL (E)	N/A	0.709 ± 0.112	0.566 ± 0.052	0.619 ± 0.047	0.223 ± 0.011
	OM+CL+IM (E)	N/A	0.663 ± 0.080	0.468 ± 0.09	0.600 ± 0.04	0.252 ± 0.065

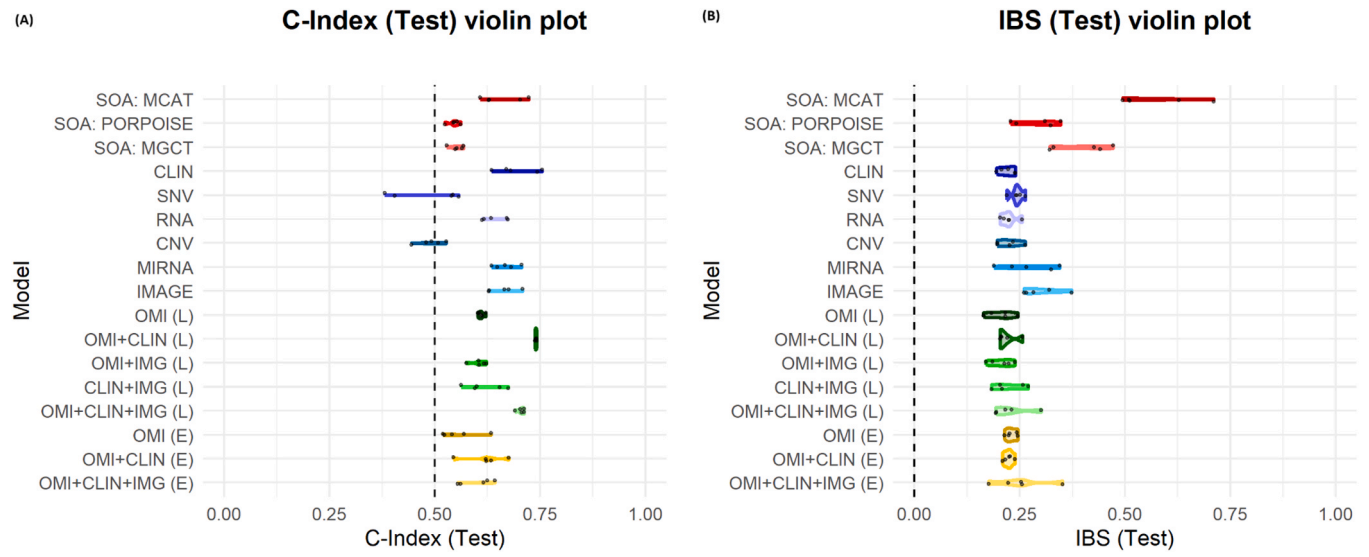

Fig. 3. Violin plots: distribution of the 5-fold performance metrics values calculated using the test set across all evaluated models: (A) C-Index and (B) IBS.

discriminatory power, of separating modality-specific learning before fusion.

Benchmarking relative to well-established multimodal approaches (MCAT, PORPOISE, and MGCT) revealed that all three demonstrated high training performance (mean C-indices > 0.936) but exhibited limited out-of-sample generalization with a significant decrease in testing performance (MCAT: 0.685 ± 0.051; PORPOISE: 0.546 ± 0.014; MGCT: 0.543 ± 0.012). As shown in Table 3 and Fig. 3-A, the late integration models developed in this study consistently outperformed all benchmarks. Our late integration OM+CL model showed clear superiority over the best benchmark (MCAT) in terms of average performance and stability. While MCAT held the best benchmark average, its distribution was the widest, indicating statistical instability. In contrast, PORPOISE and MGCT showed narrower, highly overlapping distributions, but were significantly lower than MCAT and all our late-integration models. These results indicate improved out-of-sample generalization in the proposed late-fusion models relative to existing methods.

3.3.2. Calibration power

The analysis of the IBS complements the C-Index by assessing model precision and stability. The visual analysis of the IBS distributions (Fig. 3-B) shows that, in terms of calibration, there is no single superior structural approach, as the precision and stability of our models are too similar to be ranked definitively.

The benchmark models showed a significantly reduced calibration power in comparison with almost every optimal model trained in this study. The MCAT benchmark exhibited the largest mean IBS (0.398 ± 0.068) and the widest distribution of all models, indicating the least reliable calibration. The other benchmark models (PORPOISE, MGCT) were better than MCAT, but worse than all late integration, most unimodal and most early integration models.

Regarding our unimodal models, the clinical model exhibited the lowest IBS value (0.220 ± 0.020), which was slightly improved by three late integration models: OM (0.205 ± 0.032), OM+IM (0.206 ± 0.028) and OM+CL (0.218 ± 0.023). However, the distributional overlap in the violin plots (Fig. 3B) confirms that the gain in calibration power achieved by the best models is not statistically significant over any other.

The early integration models showed varying stability. Surprisingly, the OM (E) and OM+CL (E) models displayed narrower distributions, suggesting that, while their average precision was slightly lower, simple feature concatenation provided an unexpected stronger calibration stability.

Despite the near-total overlap, the consistency of low average IBS scores in late integration supports their beneficial effect on predictive precision over other approaches.

3.4. Model explainability

To assess the interpretability and biological plausibility of our models, we applied SHAP to quantify feature contributions to survival predictions. We analyzed both multimodal and unimodal models and focused the subsequent assessment on the top-performing RNA-seq and miRNA models.

SHAP analysis on multimodal models consistently highlighted the clinical component as the dominant contributor (Fig. 4-A), aligning with its known prognostic relevance. For unimodal models, we focused on the top-performing models (RNA and miRNA), selecting their top 10 highest-contributing features (shown in Figs. 4-B and 4-C, listed in **SM4.a** and **SM4.b**) for deeper investigation.

Survival analysis using Kaplan-Meier curves [24] revealed distinct patterns (Fig. 5). Samples with more under-expressed than overexpressed genes showed better survival, while those with more overexpressed than under-expressed miRNA also exhibited improved survival. This highlights an inverse relationship in their prognostic impact.

The top 10 genes from SHAP analysis, along with the target genes of the top 10 miRNA (listed in **SM4.c**), were significantly associated with biological processes and mechanisms relevant to breast cancer, including cell proliferation, metabolism and tumor plasticity, among others (Detailed in **SM4.d**, **SM4.e** and **SM4.f**). This suggests a link between these features and disease development and prognosis. For miRNA target analysis (results shown in Fig. 6), a substantial 193 out of 280 enriched GO terms can be associated with the disease. Additionally, all pathway and disease ontology enrichment analyses specifically highlighted breast cancer. Notably, in the Network of Cancer Genes (NCG) analysis, triple negative breast cancer, which represents the most aggressive breast cancer subtype, emerged as the most significantly enriched term.

3.5. Translational web platform

To maximize the accessibility of the results of this study, a functional web application has been developed and published. This tool is not intended for immediate clinical use, as the embedded model requires further external validation, but serves as a valuable example of the translational potential and utility of the implemented rigorous

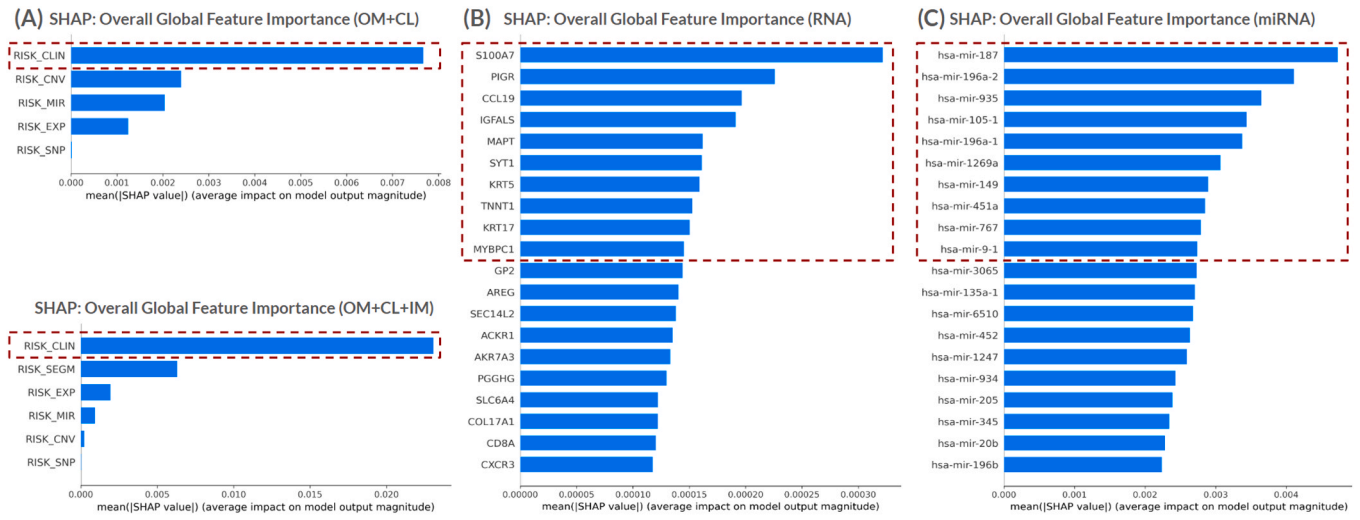

Fig. 4. SHAP results: top SHAP features from (A) late integration models, (B) RNA model and (C) miRNA model, highlighting the top 10 features (B and C) selected for deep assessment.

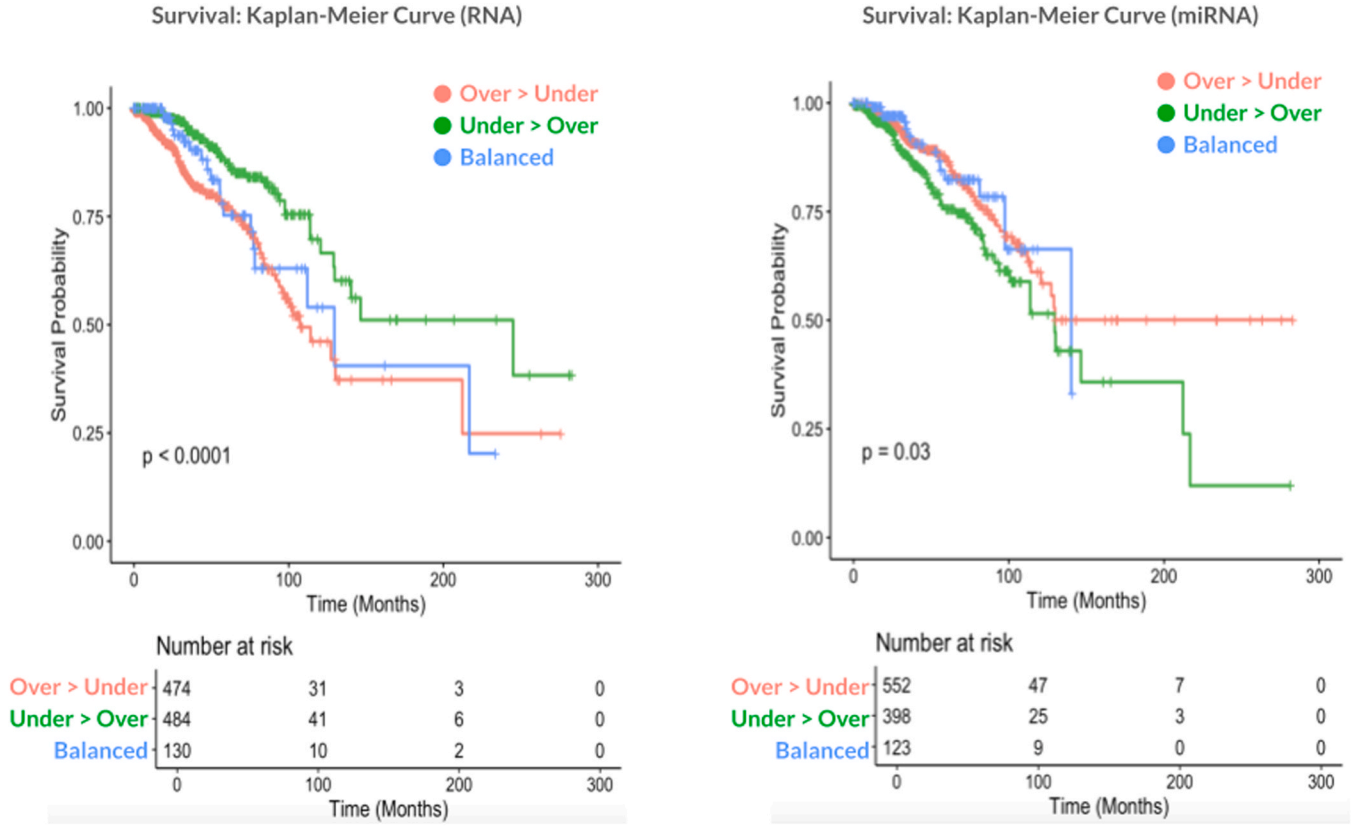

Fig. 5. Prognostic significance of feature expression patterns: Kaplan-Meier survival curves comparing patient survival patterns of stratified groups defined based on the relative expression of the top 10 most relevant features from (A) RNA-seq and (B) miRNA unimodal models. In the figure: Over>Under (Predominantly over-expressed), Under>Over (Predominantly underexpressed), Balanced (Balanced expression).

framework to both researchers and clinicians.

The web application allows users to upload a single patient's multiomics data (including SNVs, RNA-seq expression, miRNA expression, CNVs and clinical data) and perform the following actions:

- Predict a relative risk score and place it within the TCGA test cohort reference distribution.
- Visualize expression values for selected candidate genes and miRNAs, compared against the TCGA test cohort distribution.
- Visualize the individual's mutation pattern on a PCA projection alongside the TCGA test cohort.
- Compare the individual's subtype and age against their distributions in the TCGA test cohort.

This web tool is publicly available at https://brca-model.vicomtech.

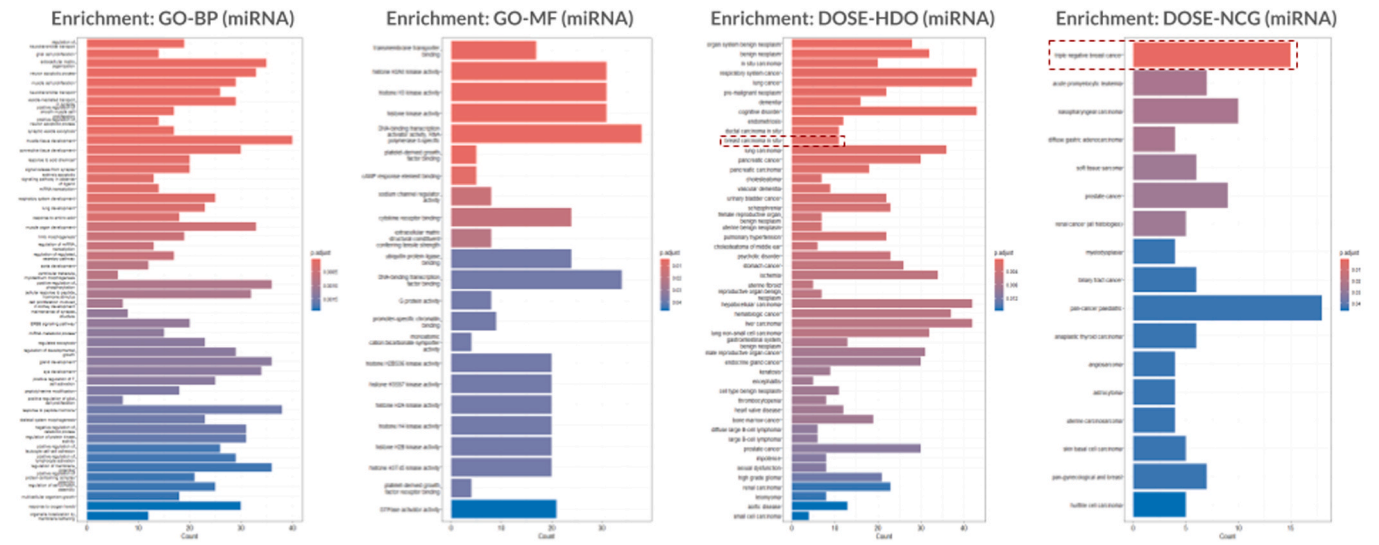

Fig. 6. Enrichment results: top 10 miRNA target genes enrichment analysis results, including Gene Ontology – Biological Process (GO-BP), Gene Ontology – Molecular Function (GO-MF), Disease Ontology Semantic and Enrichment - Human Disease Ontology (DOSE-HDO) and Network of Cancer Genes (NCG) ontologies.

org.

4. Discussion

4.1. Stability assessment

Significant disparities were identified in the C-Index variance metrics (±SD) across models. For instance, the lowest C-Index variance (SD=0.002), associated with the OM+CL late-fusion model is ~40 times smaller than the largest variance (SD=0.085) associated with the unimodal SNV model.

These large SD differences are a direct reflection of model stability during training, which is heavily dependent on the structure and volume of the input data. Models exhibiting the largest variability (unimodal and early integration models) receive a large and heterogeneous array of raw input features. This forces the network to simultaneously process low-level noise and extract high-level prognostic signals, making its performance highly sensitive to the composition of each training fold, thus resulting in high SD.

Conversely, the late integration models exhibit considerably smaller SD values because their input consists of shorter pre-processed prediction vectors obtained from the unimodal models. This structural difference minimizes the direct influence of raw input noise. This analysis reveals that the late fusion of multimodal data acts as a robust structural regularizer, significantly reducing performance variability. This process ensures the model converges to a highly similar solution regardless of the specific data found in each cross-validation fold, thereby improving the model's robustness.

4.2. Performance hierarchy

The discrimination performance (C-Index) provides critical guidance for designing future multimodal architectures. The core finding of this study is that late fusion is the methodologically superior strategy for complex data fusion. This is supported by the consistent outperformance of the late-fusion models over their early integration counterparts across all modality combinations. Early fusion's failure to match this performance is primarily attributable to the high-dimensionality challenges inherent when concatenating heterogeneous raw features at the input layer, making the training process inefficient.

The analysis also confirmed the importance of clinical data; the clinical unimodal model was the best in its category, and the late-fusion OM+CL configuration achieved the highest overall average C-Index and superior out-of-sample generalization. This model's strength is based on its statistically significant superiority over the other models, highlighting the non-redundant prognostic value of integrating clinical and omics data. Furthermore, while the IBS did not show statistically significant superiority in calibration power, the OM+CL model maintained an average IBS among the top three models, confirming its consistently reliable precision.

The optimal integration strategy relies on key advantages: late fusion effectively mitigates the risk of overfitting associated with high-dimensional feature concatenation [7,8] and enables each modality-specific subnetwork to learn tailored representations [7], which is particularly relevant when integrating omics and imaging modalities, where the joint feature dimensionality can easily exceed thousands of variables. Additionally, late fusion provides inherent robustness to data heterogeneity and missing values [7].

The superior performance is further supported by our rigorous validation methodology: the implementation of a strictly hold-out test set and nested cross-validation for hyperparameter tuning ensured robust performance estimation and minimized the risk of selection bias, providing a more reliable assessment of true model generalization.

Finally, the poor out-of-sample generalization of the established late and intermediate benchmark models (MCAT, PORPOISE, MGCT), exhibited by its significant performance drop from training to testing, justifies the need for the optimized and highly stable architectures developed in this study. In conclusion, the performance hierarchy of this study advocates for simpler frameworks that prioritize modality-specific learning before the final fusion step.

Our findings align with previous TCGA-wide survival modeling studies where late fusion consistently outperformed early fusion strategies, supporting the generalizability of this design choice across different cancer types and multimodal configurations [7,8,25].

4.3. Biological assessment

Beyond predictive performance and generalization, we confirmed that the features driving our models predictions are clinically and biologically relevant.

Applying SHAP and survival analysis demonstrated that the top-ranked molecular features drive distinct survival patterns across patient groups. Additionally, enrichment analysis further linked these features to specific aggressive molecular subtypes (i.e. triple negative

breast cancer) and key breast cancer mechanisms, such as tumor plasticity and cell proliferation.

For example, S100A7, one of the top 10 genes, can promote breast cancer growth and metastasis by fostering an immunosuppressive tumor microenvironment [26]. KRT5 is highly expressed in basal-like breast cancer, which is associated with poor prognosis. Cells expressing high levels of KRT5 exhibit strong cancer stem cell properties, which have been associated with endocrine therapy resistance [27] and poor prognosis [28]. Furthermore, KRT5 and KRT17 are dimeric partners and regulate cell signaling, also implicating KRT17 as an oncoprotein that correlates with shorter patient survival [29].

This interpretability layer is crucial as it transforms the model from a 'black box' predictor into a tool for identifying novel therapeutic and prognostic biomarkers.

4.4. Limitations and future work

This study provides a rigorous comparative framework with encouraging results, but several limitations must be acknowledged to contextualize our findings and guide future research. These constraints pertain to the architectures chosen for complex modalities, the assumptions underlying our survival model, and the scope of our data for external generalization.

4.4.1. Imaging models

Our evaluation process suggested that the integration of imaging data showed limited overall improvement, which we attribute to challenges within the imaging models themselves. Generalization failure suggests potential overfitting, possibly due to site-specific biases or the constraints of the Attention-based Multiple Instance Learning (AMIL) approach in extracting survival-relevant context from small patches.

In order to address these limitations, we will evaluate different alternatives to fine-tune the existing encoders for the specific survival prediction task and investigate alternative architectures, such as graph-based models, which may better capture spatial context and inter-patch relationships within whole slide images.

4.4.2. Discretized survival approach

This study implemented a discretized survival modelling approach, which is unconventional in this context compared to established continuous-time methods. This decision was made for two key reasons:

- Benchmarking consistency: One of our main goals was to conduct a fair and precise comparison against three established benchmark models, all of which utilize this same discrete approach.
- Information richness: We considered that obtaining four intermediate metrics per individual model, rather than a single risk score, provides the late-integration models with significantly richer, more granular temporal insights from each subnetwork. This design allows the final models to better leverage the information learned by each individual modality.

Nevertheless, adapting our modelling framework to a continuous-time survival model is entirely feasible and represents a valuable direction for future work. This modification would require several key adjustments, including changing the loss function and the final neural network layer, providing continuous survival time rather than discretized bin information as input, and adapting the intermediate metrics generated from the unimodal models to use in the late integration models. The main challenge would be selecting an appropriate intermediate metric for fusion that does not oversimplify the prognostic information, which could compromise the performance of the late integration model.

4.4.3. Data availability and external validation

This study acknowledges the limitation of relying on a single data source (~1000 patients), as this may restrict the generalizability of the models to broader patient populations. Nevertheless, analysis of the current literature [30] confirms that this sample size is consistent with the state-of-the-art for multi-modal deep learning studies in breast cancer, due to the difficulty in finding large public cohorts with full patient overlap across all modalities of interest (omics, clinical, imaging).

To mitigate the challenges associated with a smaller sample size and ensure the robustness of our results, several specific measures were applied in this study:

- We chose simpler multi-layer perceptron (MLP)-based architectures to ensure a better fit between model complexity and the available sample size, thereby reducing the risk of overfitting to noisy data.
- We favored late integration strategies to prevent the model from learning complex noise from all modalities at the input layer and to promote independent learning within each specialized network.
- We followed a rigorous validation strategy to ensure an unbiased estimate of out-of-sample performance.

Additionally, we recognize that external validation would substantially strengthen this study; however, pursuing this goal was restricted by critical constraints associated with alternative public datasets that were identified for this matter: METABRIC [31] and CPTAC [32].

- The METABRIC cohort was discarded due to differences in the quantification and technological platforms used for several data types, preventing a fair and consistent evaluation against our TCGA-trained models.
- The CPTAC initiative is a proteomic extension of the TCGA project and could not be considered an independent dataset for true external validation, as it utilizes a subset of the same patient samples.

Nonetheless, we believe that the results obtained in this study justify future efforts that focus on validating the findings on larger, independent datasets once available.

5. Conclusions

In this study, we investigated the potential of multimodal deep learning for overall survival prediction in breast cancer, leveraging a comprehensive dataset from TCGA that included clinical, omics, and histopathology imaging data. Our primary goal was a systematic comparison of fusion strategies to determine the most robust predictive approach.

The results conclusively demonstrated that the late-fusion approach is the superior modeling strategy, achieving both the highest discriminatory power and structural stability. The OM+CL late fusion model outperformed all early fusion and established late and intermediate multimodal benchmark models in discriminatory power, yielding the best overall C-Index (0.740 ± 0.002) and a distribution that did not overlap with any other model, confirming its statistically significant superiority in discriminatory power. This result, coupled with the model's significantly low standard deviation (SD=0.002), demonstrates a structural stability that surpasses all other models. This superiority validates that the late fusion strategy effectively mitigates the risk of overfitting and enables each modality-specific subnetwork to learn tailored representations [7,8].

The benefit of our work is two-fold, offering methodological guidance for computational researchers and delivering crucial insights for clinicians and researchers.

Methodologically, our study offers clear, evidence-based guidance on how to integrate multi-omics data in predictive models. The key finding is that late fusion consistently outperforms early and intermediate fusion, demonstrating it is the most suitable structural design for stable deep learning models in this field. Clinically and biologically, our

work delivers practical information regarding prognostic feature importance that benefits both data collection and molecular assessment; it helped identify key genes and pathways strongly associated with breast cancer mechanisms and aggressive subtypes. This explainable layer could help clinicians and researchers prioritize the most influential molecular features in future studies.

In conclusion, our findings highlight the value of carefully designing multimodal deep learning approaches for survival prediction in breast cancer. By combining rigorous data preprocessing, late-fusion integration strategies, and strong validation procedures, we achieved models that not only generalize well but also capture biologically meaningful patterns. These results underscore the potential of integrating diverse clinical, molecular, and imaging data to improve prognostic modeling in oncology.

CRediT authorship contribution statement

Aurora Sucre: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Karen López-Linares:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Borja Calvo:** Writing – review & editing, Supervision, Conceptualization. **Alba Garin-Muga:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Xabier Calle Sánchez:** Writing – original draft, Visualization, Supervision, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Perez-Herrera Laura:** Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Vivanco María:** Writing – review & editing, Project administration, Funding acquisition. **García-González María:** Writing – original draft, Methodology, Conceptualization.

Ethical statement

The data exploited in this study was obtained from The Cancer Genome Atlas (TCGA) Research Network, which publicly provides de-identified patient data. The original collection, processing, and public release of TCGA data adhered to strict ethical and legal guidelines, including comprehensive ethical review by Institutional Review Boards and the acquisition of informed consent from all participants for research purposes and data sharing.As this study constitutes a secondary analysis of publicly available, de-identified data from TCGA, it was determined to be exempt from additional ethical committee review by our institution. All data was used strictly as intended by the original data providers and in accordance with TCGA data use policies.

Funding

This work has been partially funded by the Basque Government ELKARTEK Program, within the BG24 Project (KK-2024/00019), granted to AS, XCS, LVP, MdMV, MJG, KL and AG. This project focuses on the exploration and characterization of molecular factors in breast cancer and its innovative applications in precision oncology.

B. Calvo acknowledges partial support by the Research Groups 2022–2025 (IT1504-22) from the Basque Government, and the PID2022-137442NB-I00 research project from the Spanish Ministry of Science.

The funding sources were not involved in the design of this study.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used GEMINI and ChatGPT in order to improve the readability of certain sections. After using these services, the authors reviewed and edited the content as needed, and they take full responsibility for the content of the published article.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Aurora Sucre reports financial support was provided by Basque Government. Xabier Calle Sanchez reports financial support was provided by Basque Government. Laura Valeria Perez-Herrera reports financial support was provided by Basque Government. Maria dM Vivanco reports financial support was provided by Basque Government. Maria Jesus Garcia-Sanchez reports financial support was provided by Basque Government. Karen Lopez-Linares reports financial support was provided by Basque Government. Alba Garin-Muga reports financial support was provided by Basque Government. Borja Calvo reports financial support was provided by Basque Government. Borja Calvo reports financial support was provided by Spanish Ministry of Science. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2025.10.038.

Data availability

The results published in this article are based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. All data employed during this study is available for public use, either on the GDC data portal (https://portal.gdc.cancer.gov/) or on the UCSC Xena Data Browser (https://xenabrowser.net/).

The code and scripts necessary for data preprocessing, hyperparameter optimization, model training and evaluation, and result analysis are publicly available in a dedicated repository on GitHub (https://github.com/amsucre/brca-survival). A translational tool that incorporates the best final predictive model is accessible via a public web application at https://brca-model.vicomtech.org.

References

[1] Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA Cancer J Clin 2024; 74(1):12–49. https://doi.org/10.3322/caac.21820.

[2] Harbeck Nadia, Penault-Llorca Frédérique, Cortes Javier, Gnant Michael, Houssami Nehmat, et al. Breast cancer. Nat Rev Dis Prim 2019;5(1):66. https://doi.org/10.1038/s41572–019-0111–2.

[3] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature 2012;490(7418):61–70. https://doi.org/10.1038/nature11412.

[4] Chen RJ, Lu MY, Weng W, Chen TY, Williamson DF, et al. Multimodal Co-Attention transformer for survival prediction in gigapixel whole slide images. ICCV 2021: 3995–4005. https://doi.org/10.1109/ICCV48922.2021.00398.

[5] Chen RJ, Lu MY, Williamson DFK, Chen TY, Lipkova J, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell 2022;40(8): 865–878.e6. https://doi.org/10.1016/j.ccell.2022.07.004.

[6] Liu M, Liu Y, Cui H, Li C, Ma J. MGCT: Mutual-guided cross-modality transformer for survival outcome prediction using integrative histopathology-genomic features. IEEE BIBM Proceedings; 2023. p. 1306–12. https://doi.org/10.1109/BIBM58861.2023.10385897.

[7] Nikolaou N., Salazar D., RaviPrakash H., Gonçalves M., Mulla R., et al. (2024) Quantifying the advantage of multimodal data fusion for survival prediction in cancer patients. bioRxiv. doi:10.1101/2024.01.08.574756. - Unpublished results (Preprint).

[8] Nikolaou N, Salazar D, RaviPrakash H, Gonçalves M, Mulla R, et al. A machine learning approach for multimodal data fusion for survival prediction in cancer patients. NPJ Precis Oncol 2025;9(1):128. https://doi.org/10.1038/s41698–025-00917–6.

[9] Zhao L, Dong Q, Luo C, Wu Y, Bu D, et al. DeepOmix: a scalable and interpretable multi-omics deep learning framework and application in cancer survival analysis. Comput Struct Biotechnol J 2021;19:2719–25. https://doi.org/10.1016/j.csbj.2021.04.067.

[10] Li Y, Daho MEH, Conze P, Zeghlache R, Boité HL, et al. A review of deep learning-based information fusion techniques for multimodal medical image classification. Comput Biol Med 2024;177. https://doi.org/10.1016/j.compbiomed.2024.108635.
[11] Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, et al. Visualizing and interpreting cancer genomics data via the xena platform. Nat Biotechnol 2020;38(6):675–8. https://doi.org/10.1038/s41587–020-0546–8.
[12] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert Benjamin L, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. PNAS 2005;102(43):15545–50. https://doi.org/10.1073/pnas.0506580102.
[13] Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, et al. Molecular signatures database (MSigDB) 3.0. Bioinformatics 2011;27(12):1739–40. https://doi.org/10.1093/bioinformatics/btr260.
[14] Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, et al. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng 2021;5(6):555–70. https://doi.org/10.1038/s41551–020-00682-w.
[15] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. KDD Proc 2019:2623–31. https://doi.org/10.1145/3292500.3330701.
[16] Ansel J, Yang E, He H, Gimelshein N, Jain A, et al. PyTorch 2: faster machine learning through dynamic python bytecode transformation and graph compilation. Apr 27 ASPLOS Proc 2024:929–47. https://doi.org/10.1145/3620665.3640366.
[17] He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian. Deep residual learning for image recognition. CVPR 2016:770–8. https://doi.org/10.1109/CVPR.2016.90.
[18] Chen RJ, Ding T, Lu MY, Williamson DFK, Jaume G, et al. Towards a general-purpose foundation model for computational pathology. Nat Med 2024;30(3):850–62. https://doi.org/10.1038/s41591–024-02857–3.
[19] Lundberg S, Lee S. A unified approach to interpreting model predictions. NIPS Proc 2017:4768–77. https://doi.org/10.5555/3295222.3295230.
[20] Pölsterl S. scikit-survival: a library for Time-to-Event analysis built on top of scikit-learn. J Mach Learn Res 2020;21(212):1–6.
[21] Zhou H, Cheng X, Wang S, Zou Y, Wang H. SurvMetrics: predictive evaluation metrics in survival analysis 2025. https://doi.org/10.32614/CRAN.package.SurvMetrics.
[22] Therneau TM, Grambsch PM. Modeling survival data: extending the cox model. Springer; 2000.
[23] Yu G, Wang L, Han Y, He Q. Clusterprofiler: an r package for comparing biological themes among gene clusters. OMICS 2012;16(5):284–7. https://doi.org/10.1089/omi.2011.0118.
[24] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53(282):457–81. https://doi.org/10.1080/01621459.1958.10501452.
[25] Leng D, Zheng L, Wen Y, Zhang Y, Wu L, et al. A benchmark study of deep learning-based multi-omics data fusion methods for cancer. Genome Biol 2022;23(1):171. https://doi.org/10.1186/s13059–022-02739–2.
[26] Mishra S, Charan M, Shukla RK, Agarwal P, Misri S, et al. cPLA2 blockade attenuates S100A7-mediated breast tumorigenicity by inhibiting the immunosuppressive tumor microenvironment. J Exp Clin Cancer Res 2022;41(1):54. https://doi.org/10.1186/s13046–021-02221–0.
[27] Aurrekoetxea-Rodriguez I, Lee SY, Rábano M, Gris-Cárdenas I, Gamboa-Aldecoa V, et al. Polyoxometalate inhibition of SOX2-mediated tamoxifen resistance in breast cancer. Cell Commun Signal 2024;22(1):425. https://doi.org/10.1186/s12964–024-01800-w.
[28] McGinn O, Ward AV, Fettig LM, Riley D, Ivie J, et al. Cytokeratin 5 alters β-catenin dynamics in breast cancer cells. Oncogene 2020;39(12):2478–92. https://doi.org/10.1038/s41388–020-1164–0.
[29] Baraks G, Tseng R, Pan C, Kasliwal S, Leiton CV, et al. Dissecting the oncogenic roles of keratin 17 in the hallmarks of cancer. Cancer Res 2022;82(7):1159–66. https://doi.org/10.1158/0008–5472.can-21–2522.
[30] Nakach FZ, Idri A, Goceri E. A comprehensive investigation of multimodal deep learning fusion strategies for breast cancer classification. Artif Intell Rev 2024;57(327). https://doi.org/10.1007/s10462–024-10984-z.
[31] Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, et al. The genomic and transcriptomic architecture of 2,000 breast tumors reveals novel subgroups. Nature 2012;486(7403):346–52. https://doi.org/10.1038/nature10983.
[32] Ellis MJ, Gillette M, Carr SA, Paulovich AG, Smith RD, et al. Connecting genomic alterations to cancer biology with proteomics: the NCI clinical proteomic tumor analysis consortium. Cancer Discov 2013;3:1108–12.