npj | precision oncology

Article

Published in partnership with The Hormel Institute, University of Minnesota

https://doi.org/10.1038/s41698-025-00917-6

A machine learning approach for multimodal data fusion for survival prediction in cancer patients

Nikolaos Nikolaou[1,2], Domingo Salazar[1], Harish RaviPrakash[3], Miguel Gonçalves[1], Rob Mulla[3], Nikolay Burlutskiy[1], Natasha Markuzon[3] ✉ & Etai Jacob[3] ✉

Technological advancements of the past decade have transformed cancer research, improving patient survival predictions through genotyping and multimodal data analysis. However, there is no comprehensive machine-learning pipeline for comparing methods to enhance these predictions. To address this, a versatile pipeline using The Cancer Genome Atlas (TCGA) data was developed, incorporating various data modalities such as transcripts, proteins, metabolites, and clinical factors. This approach manages challenges like high dimensionality, small sample sizes, and data heterogeneity. By applying different feature extraction and fusion strategies, notably late fusion models, the effectiveness of integrating diverse data types was demonstrated. Late fusion models consistently outperformed single-modality approaches in TCGA lung, breast, and pan-cancer datasets, offering higher accuracy and robustness. This research highlights the potential of comprehensive multimodal data integration in precision oncology to improve survival predictions for cancer patients. The study provides a reusable pipeline for the research community, suggesting future work on larger cohorts.

Improving survival predictions (such as overall survival [OS] and progression-free survival) in cancer patients is a crucial step in the effort to achieve biological insights and assist clinicians in making more informed clinical decisions. Recent advances in high-throughput sequencing technologies and other molecular assays (such as genomic, transcriptomic, epigenomic, and proteomic methods) have provided a breadth of independent measurements from patients.

Comprehensive integrated analysis of multi-omics data can be used to discover the complex mechanisms underlying cancer development and progression. Training predictive models using information from multiple sources can lead to improved model predictions, and many examples, including biological[1–3] and nonbiological[4–6] applications, are reported in the literature. Different data modalities can provide complementary information about patient outcomes, including OS. When modalities are correlated, they can help to reduce the variance in these predictions by producing more robust models, which is especially useful when working with data with a low signal-to-noise ratio or a high degree of missingness[5,7,8].

Combining data from multiple modalities is challenging, however, and the optimal way to achieve it is largely problem-specific[9–13]. Beyond the additional computational and memory burden incurred by increasing the dimensions of the feature space (i.e., biomarkers that constitute model inputs, or independent variables), introducing additional data modalities for a given patient sample also has statistical implications. Because the sample size (number of patients) is fixed, increasing the size of the feature space leads to an increased risk of overfitting and thus the need for regularization. Yet another complication is data heterogeneity. Different modalities might consist of different data types, such as imaging, time-series, text, and tabular data, and often require specific preprocessing and expertise in analysis, modeling, and interpretation. The need for extensive and coherent comparisons across the various multimodal methods is raised repeatedly in several review studies[1,3,14–16].

Multimodal data fusion in "omics" data sets usually suffers from low sample size to feature space ratios. In addition, most individual features are irrelevant or only weakly relevant to the outcome (low signal-to-noise ratio). Some modalities suffer from sparsity of the signal (e.g., mutations) or high degrees of missingness (clinical data), whereas others require batch normalization (gene expression). In addition, the presence of intermodality and intramodality correlations is high[17–20]. These challenges downgrade the potential of multimodal data fusion to add value for biological applications by increasing the likelihood of multimodal-based model overfitting.

[1]Oncology Data Science, Oncology R&D, AstraZeneca, Cambridge, UK. [2]Department of Physics & Astronomy, University College London, London, UK. [3]Oncology Data Science, Oncology R&D, AstraZeneca, Waltham, MA, USA. ✉e-mail: natasha.markuzon@astrazeneca.com; etai.jacob@astrazeneca.com

The existing literature on data fusion methods using multi-omics data for the prediction of OS in cancer patients, although quite extensive[1,3,14–16], is characterized by several shortcomings. These include greater emphasis on unsupervised feature reduction methods, linear predictive modeling for survival prediction, modality fusion approaches that ignore the properties commonly characterizing omics data, and a lack of standardized evaluation approaches that would allow comparison across different methods. The need for extensive and coherent comparisons across the various multimodal methods is a common theme of several reviews[1,3,14–16]. Here we summarize the four points we identified as key components in the construction and evaluation of data fusion strategies: dimensionality reduction, survival models, fusion strategies, and evaluation.

The broad term "dimensionality reduction" encompasses any technique that reduces the size of the feature space. It can refer to "feature selection" techniques that return a subset of the original feature dimensions or to "feature extraction" techniques that return a new, smaller feature set consisting of dimensions that are functions of subsets of the original features. Bioinformatics data sets typically have a very low ratio of samples to number of dimensions, making dimensionality reduction critical for protecting survival models against overfitting. Although typically a relatively small subset of genomic features contributes to survival prediction[21,22], only a few dimensionality methods have been explored in the context of survival analysis in the literature, which concentrates mostly on unsupervised methods such as principal component analysis or autoencoders[23]. Supervised feature selection methods are limited to the use of univariate Cox proportional hazards (PH) models or training multivariate Cox PH models with Lasso regression (L1-regularization) to impose sparsity [21,24,25]. These methods have the advantage of accounting for right-censoring of the data, but are quite slow and scale poorly as the feature space increases. They are incapable of accounting for nonlinear correlations with the target (OS time) or for interactions among variables. As such, they limit opportunities for downstream analysis. A range of feature selection methods like Spearman correlation[26] and various information-theoretic approaches[27–31] can address some of these issues.

We expected the outcome of a survival model (OS time) to be nonlinearly related to its inputs (biomarkers from the various modalities). Yet the literature largely relies on linear Cox PH models[22,24,32]. More recently, deep learning models have also been examined in a few relevant works[33–36], but several nonlinear alternatives, like gradient boosting[37] or random forests[38], are absent from most comparisons. These methods have demonstrated success in survival modeling[39–42] and, although more flexible than linear models, they are equipped with inductive biases that enforce regularization with much less hyperparameter tuning than deep neural networks. Consequently, gradient boosting, random forest, and heterogeneous ensembles[43–45] typically outperform deep neural networks on tabular data[46–48] like multi-omics features. Interestingly, although some of these methods have been widely adopted in other biomedical[49,50] and non-biomedical[51–53] settings, we were not able to find much information on their use in survival analysis.

The success of early fusion in other multimodal settings[6,54–56] has led to recent approaches for modeling predictions of OS in cancer patients with the use of early and intermediate fusion strategies (i.e., data-level fusion). These settings differ from the typical bioinformatics setting, the most important differentiating factor being the number of available data points relative to the input dimensions. For instance, the training set used by Jain et al.[56] contains 1.8–6 billion data points (for the different modalities), whereas The Cancer Genome Atlas (TCGA) data set involves sample sizes on the order of 10–10^3, depending on cancer type. In the former multimodal application, the feature space is on the order of 10^3, whereas for TCGA the feature space is on the order of 10^5. Thus, in the case of multimodal models trained on TCGA[57], late fusion methods (i.e., prediction-level fusion) present an opportunity to outperform early fusion approaches, due to increased resistance to overfitting[16,58], ease of addressing data heterogeneity, and the ability to more naturally weigh each modality based on its informativeness of OS without being affected by the highly imbalanced dimensionalities across modalities[3,16].

Many multimodal models for OS prediction that have been reported in the literature suffer from compromised evaluation practices[1]. Many fail to account for the considerable uncertainty arising from different training-test set splits of the data, either by omitting multiple splits altogether[21,22,24,34,35,58–60], or, even if not, by only reporting average C-indices over them without the accompanying CIs[32,36,61]. Most lack comparisons spanning different modality combinations or different data fusion approaches. Many even lack direct comparisons against unimodal approaches[21,22,24,25,33,34]. Finally, some works propose multimodal approaches in name only, reporting the learning algorithm data from multiple modalities (early fusion) that produced a unimodal model (which used features from only one modality). Characteristic examples include Chai et al.[33], where the model ultimately only uses mRNA gene expression features, and Wulczyn et al.[62], where unimodal models outperform the multimodal alternative. Any study that does not report the features of the final model, the contribution of each modality, or compare performance against unimodal models risks encountering this issue. As a result, most works do not provide a clear answer to whether modalities should be combined and, if so, which ones to use and how to combine them.

Results

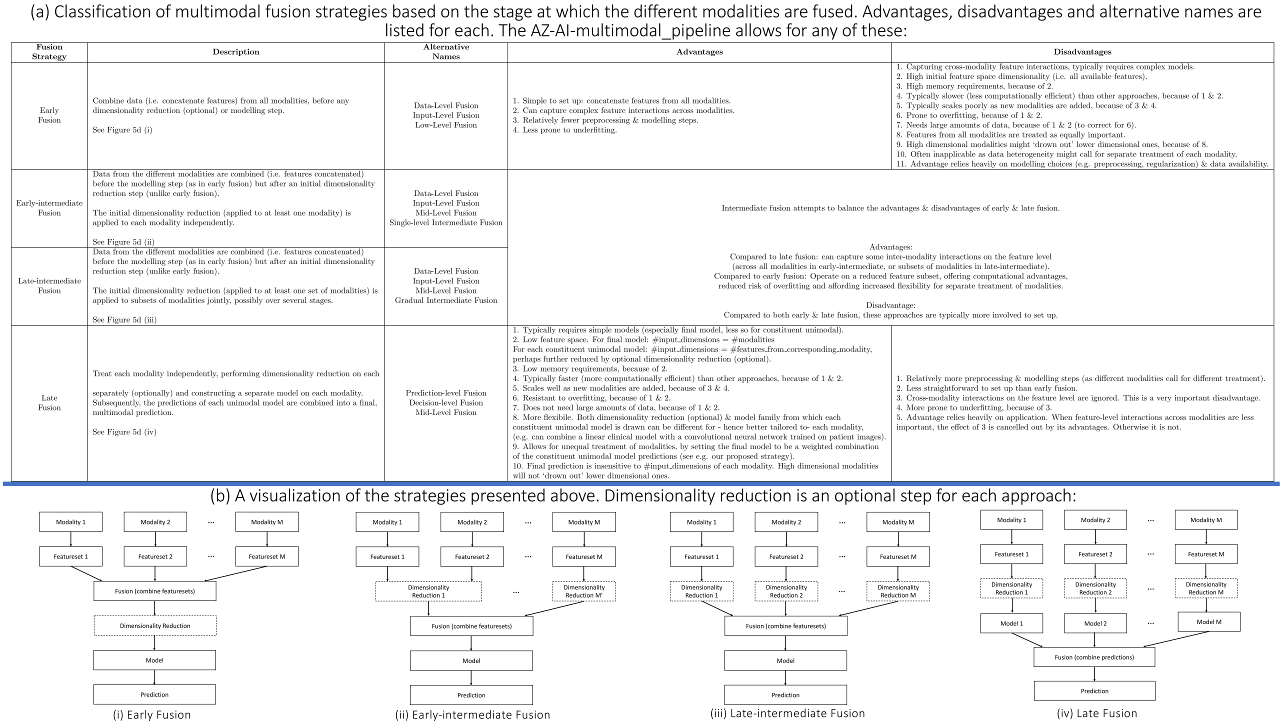
Multimodal data integration pipeline

The AstraZeneca–artificial intelligence (AZ-AI) multimodal pipeline, developed in the context of this work by the AstraZeneca Oncology Data Science Team, is a Python library for multimodal feature integration and survival prediction. It can be used to preprocess and reduce the dimensionality of tabular data sets (unimodal or multimodal) and to train and evaluate survival models. Its functionalities include several preprocessing and imputation options, flexibility regarding when to integrate modalities (Fig. 1), a range of feature reduction approaches (Table 1) and survival modeling methods (Table 2), and rigorous evaluation (as described in "Pipeline overview"), including the option to report the models' feature importance. An outline of the pipeline is provided in Fig. 2.

The library can be used to replicate and extend the results presented in this paper. It has already been successfully used for constructing state-of-the-art early integration multimodal models combining clinical and radiological features for OS prediction in lung cancer patients in a forthcoming paper by Patwardhan et al.[63] The setting examined in this paper is quite different from that of those investigators. As a result, different methods proved more successful in each application. In the work by Patwardhan et al., the data consist of only two modalities, the number of total features was on the order of 10^2–10^3, and the number of data points was on the order of 10^3, all being complete cases (no need for data imputation). This paper examines data sets of four to seven modalities, with missing entries and the total number of features on the order of 10^3–10^5, and the number of data points being 10–10^3, depending on cancer type (approaching 10^4 only for pan-cancer models). As a result, the risk of overfitting here is considerably higher. In the setting of Padwardhan et al., early and intermediate fusion strategies and nonlinear and nonmonotonic feature selection methods (mutual information) worked best. In our setting, late fusion strategies and linear or monotonic feature selection methods (Pearson and Spearman correlation) outperformed the other approaches. In both cases, ensemble survival models outperformed single models. This comparison demonstrates that different approaches to multimodal fusion are better suited to different settings. Simpler feature selection methods and late fusion are more suitable for situations in which the risk of overfitting is high, forming ensembles of multiple survival models is always beneficial, and our AZ-AI multimodal pipeline is flexible enough to provide a solid multimodal fusion solution in different settings.

Results on using late fusion in TCGA cancer patients

Throughout this paper, we report the average test set C-index and 95% confidence interval (CI) across 10 runs. The experimental setup, proposed late fusion strategy, and details on the evaluation and the data set are all described in Methods ("Data"). Unless stated otherwise, the results



(a) Classification of multimodal fusion strategies based on the stage at which the different modalities are fused. Advantages, disadvantages and alternative names are listed for each. The AZ-AI-multimodal_pipeline allows for any of these:

Fusion Strategy	Description	Alternative Names	Advantages	Disadvantages
Early Fusion	Combine data (i.e. concatenate features) from all modalities, before any dimensionality reduction (optional) or modelling step. See Figure 5d (i)	Data-Level Fusion Input-Level Fusion Low-Level Fusion	1. Simple to set up: concatenate features from all modalities. 2. Can capture complex feature interactions across modalities. 3. Relatively fewer preprocessing & modelling steps. 4. Less prone to underfitting.	1. Capturing cross-modality feature interactions, typically requires complex models. 2. High initial feature space dimensionality (i.e. all available features). 3. High memory requirements, because of 2. 4. Typically slower (less computationally efficient) than other approaches, because of 1 & 2. 5. Typically scales poorly as new modalities are added, because of 3 & 4. 6. Prone to overfitting, because of 1 & 2. 7. Needs large amounts of data, because of 1 & 2 (to correct for 6). 8. Features from all modalities are treated as equally important. 9. High dimensional modalities might 'drown out' lower dimensional ones, because of 8. 10. Often inapplicable as data heterogeneity might call for separate treatment of each modality. 11. Advantage relies heavily on modelling choices (e.g. preprocessing, regularization) & data availability.
Early-intermediate Fusion	Data from the different modalities are combined (i.e. features concatenated) before the modelling step (as in early fusion) but after an initial dimensionality reduction step (unlike early fusion). The initial dimensionality reduction (applied to at least one modality) is applied to each modality independently. See Figure 5d (ii)	Data-Level Fusion Input-Level Fusion Mid-Level Fusion Single-level Intermediate Fusion	Intermediate fusion attempts to balance the advantages & disadvantages of early & late fusion. Advantages: Compared to late fusion: can capture some inter-modality interactions on the feature level (across all modalities in early-intermediate, or subsets of modalities in late-intermediate). Compared to early fusion: Operate on a reduced feature subset, offering computational advantages, reduced risk of overfitting and affording increased flexibility for separate treatment of modalities. Disadvantage: Compared to both early & late fusion, these approaches are typically more involved to set up.	
Late-intermediate Fusion	Data from the different modalities are combined (i.e. features concatenated) before the modelling step (as in early fusion) but after an initial dimensionality reduction step (unlike early fusion). The initial dimensionality reduction (applied to at least one set of modalities) is applied to subsets of modalities jointly, possibly over several stages. See Figure 5d (iii)	Data-Level Fusion Input-Level Fusion Mid-Level Fusion Gradual Intermediate Fusion		
Late Fusion	Treat each modality independently, performing dimensionality reduction on each separately (optionally) and constructing a separate model on each modality. Subsequently, the predictions of each unimodal model are combined into a final, multimodal prediction. See Figure 5d (iv)	Prediction-level Fusion Decision-level Fusion Mid-Level Fusion	1. Typically requires simple models (especially final model, less so for constituent unimodal). 2. Low feature space. For final model: #input_dimensions = #modalities For each constituent unimodal model: #input_dimensions = #features_from_corresponding_modality, perhaps further reduced by optional dimensionality reduction (optional). 3. Low memory requirements, because of 2. 4. Typically faster (more computationally efficient) than other approaches, because of 1 & 2. 5. Scales well as new modalities are added, because of 3 & 4. 6. Resistant to overfitting, because of 1 & 2. 7. Does not need large amounts of data, because of 1 & 2. 8. More flexibile. Both dimensionality reduction (optional) & model family from which each constituent unimodal model is drawn can be different for - hence better tailored to- each modality, (e.g. can combine a linear clinical model with a convolutional neural network trained on patient images). 9. Allows for unequal treatment of modalities, by setting the final model to be a weighted combination of the constituent unimodal model predictions (see e.g. our proposed strategy). 10. Final prediction is insensitive to #input_dimensions of each modality. High dimensional modalities will not 'drown out' lower dimensional ones.	1. Relatively more preprocessing & modelling steps (as different modalities call for different treatment). 2. Less straightforward to set up than early fusion. 3. Cross-modality interactions on the feature level are ignored. This is a very important disadvantage. 4. More prone to underfitting, because of 3. 5. Advantage relies heavily on application. When feature-level interactions across modalities are less important, the effect of 3 is cancelled out by its advantages. Otherwise it is not.

(b) A visualization of the strategies presented above. Dimensionality reduction is an optional step for each approach:

Modality 1
Modality 2
...
Modality M
Featureset 1
Featureset 2
Featureset M
Fusion (combine featuresets)
Dimensionality Reduction
Model
Prediction
(i) Early Fusion
Dimensionality Reduction 1
Dimensionality Reduction M'
(ii) Early-intermediate Fusion
Dimensionality Reduction 2
Dimensionality Reduction M
(iii) Late-intermediate Fusion
Model 1
Model 2
Model M
Fusion (combine predictions)
(iv) Late Fusion

Fig. 1 | Summary of late, early, and intermediate multimodal data fusion strategies. **a** Description of strategies and their advantages, disadvantages, and alternative names used in the literature. **b** Visual explanation.

Table 1 | Dimensionality reduction methods included in the pipeline

Dimensionality reduction method	Subset of original dimensions	Nonlinear (target)	Supervised	Accounts for feature interactions	Accounts for censoring	Incorporates domain knowledge
Filter correlated features (Pearson)	Y	N	N	Y[a]	N	N
Filter low-value variance	Y	N	N	N	N	N
Linear correlation (Pearson)	Y	N	Y	N	N	N
Monotonic correlation (Spearman)	Y	Y[b]	Y	N	N	N
Univariate linear Cox PH models	Y	N	Y	N	Y	N
MIM	Y	Y	Y	N	N	N
JMI	Y	Y	Y	Y[c]	N	N
CMIM	Y	Y	Y	Y	N	N
Pathways	N[d]	Y	N	N	N	Y
Relevant gene lists	Y	Y	N[e]	N	N	Y
Autoencoders	N	Y	N	Y[f]	N	N

[a] Pairwise interactions, linear, not with regard to target.
[b] Monotonic.
[c] Pairwise interactions, nonlinear, with regard to target.
[d] Biologically meaningful.
[e] Presumed relevant for overall survival.
[f] Dependent on architecture, not with regard to target.
CMIM conditional mutual information, *JMI* joint mutual information, *MIM* mutual information maximization.

correspond to the feature selection method being Spearman correlation with OS and the survival model being a heterogeneous ensemble of all models listed in Table 2.

Performance of multimodal versus unimodal models across different cancer types

For each cancer type in TCGA, we calculated the advantage (henceforth referred to as the "delta") of the multimodal approach over unimodal models as the difference in average C-index between the multimodal model and the best unimodal model:

$$d = C_{multimodal} - C_{best\ unimodal}$$

Figure 3 shows the d across all 33 TCGA cancer types. In 25 of 33 cancer types (76% of the independent data sets), the deltas were positive, supporting the hypothesis that late fusion multimodal models generally outperform unimodal models. This result is statistically significant ($P = 0.001$) under a Wilcoxon signed-rank test against the null hypothesis of

Table 2 | Survival modeling methods included in the pipeline

Survival model method	Nonlinear	Description	Implementation
CPH-L2	N	Cox PH model with Ridge (L2-norm) regularization	Scikit-survival[54]
CPH-EN	N	Cox PH model with Elastic Net (combined L1-norm and L2-norm) regularization	Scikit-survival[54]
RSF	Y	Random survival forest model[25,26]	Scikit-survival[54]
CPH-GB	Y	Gradient-boosted[1,15] Cox PH loss with regression trees as base learner. The partial likelihood of the PH model is optimized as described in Raschka[78].	Scikit-survival[54]
CLS-GB	Y	Gradient boosting[1,15] with component-wise least squares as base learner[21]	Scikit-survival[54]
DS	Y	Deep neural network trained under a Cox PH loss (DeepSurv)[31]	PyCox[34]
ENS	Y[a]	Heterogeneous ensemble of any combination of the above models weighted by their respective validation set performances, per Eq. 1	Custom

[a] Nonlinear, provided at least ONE base model with non-zero weight is nonlinear.

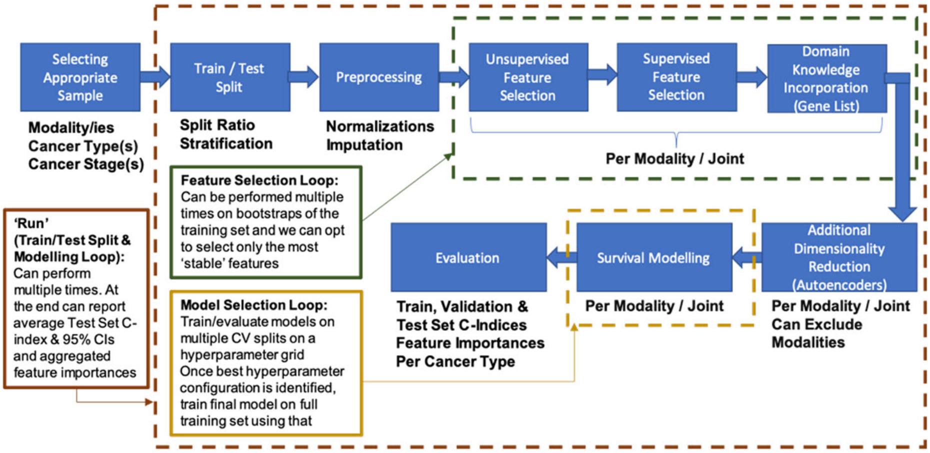

Selecting Appropriate Sample
Modality/ies
Cancer Type(s)
Cancer Stage(s)
Train / Test Split
Split Ratio
Stratification
Preprocessing
Normalizations
Imputation
Unsupervised Feature Selection
Supervised Feature Selection
Domain Knowledge Incorporation (Gene List)
Per Modality / Joint
Feature Selection Loop: Can be performed multiple times on bootstraps of the training set and we can opt to select only the most 'stable' features
'Run' (Train/Test Split & Modelling Loop): Can perform multiple times. At the end can report average Test Set C-index & 95% CIs and aggregated feature importances
Model Selection Loop: Train/evaluate models on multiple CV splits on a hyperparameter grid Once best hyperparameter configuration is identified, train final model on full training set using that
Evaluation
Train, Validation & Test Set C-Indices Feature Importances Per Cancer Type
Survival Modelling
Per Modality / Joint
Additional Dimensionality Reduction (Autoencoders)
Per Modality / Joint Can Exclude Modalities

Fig. 2 | Overview of the AZ-AI multimodal pipeline. Shown is a brief outline of the pipeline's main steps and functionalities. Fig. 1 shows a classification of multimodal fusion strategies. The pipeline allows for any of these, depending on which of its execution steps are run "per modality" or "jointly".

no multimodal advantage ($d = 0$). Aligned with our expectations of the impact of the sample size, the multimodal advantage d positively correlated with the training set size. The associated Pearson correlation $r(d,$ sample size$) = 0.361$ is statistically significant ($P = 0.039$). Finally, d also positively correlates with the performance of the best unimodal model. In other words, the better the best unimodal model, the larger the advantage of using it in multimodal models. This might seem counterintuitive, but it is largely due to our late fusion strategy, which, because it combines all unimodal models into a multimodal model, requires at least one strong base model to achieve good performance. That said, this finding might also indicate other mechanisms worth exploring, such as the presence of strong cross-modality correlations; if this is the case, then when one modality is predictive of OS, others would be as well, and combining them into multimodal models would further improve results. The associated Pearson correlation adjusted by weighting data sets by sample size $r_{adj}(d, C_{best\ unimodal}) = 0.365$ is statistically significant ($P = 0.029$).

Summarizing the above results, we found that multimodal models trained under our proposed late fusion strategy tended to outperform unimodal models across TCGA cancer types. We also found that the advantage of multimodal models increased with sample size and with the performance of unimodal models.

Comparing multimodal with unimodal models in NSCLC

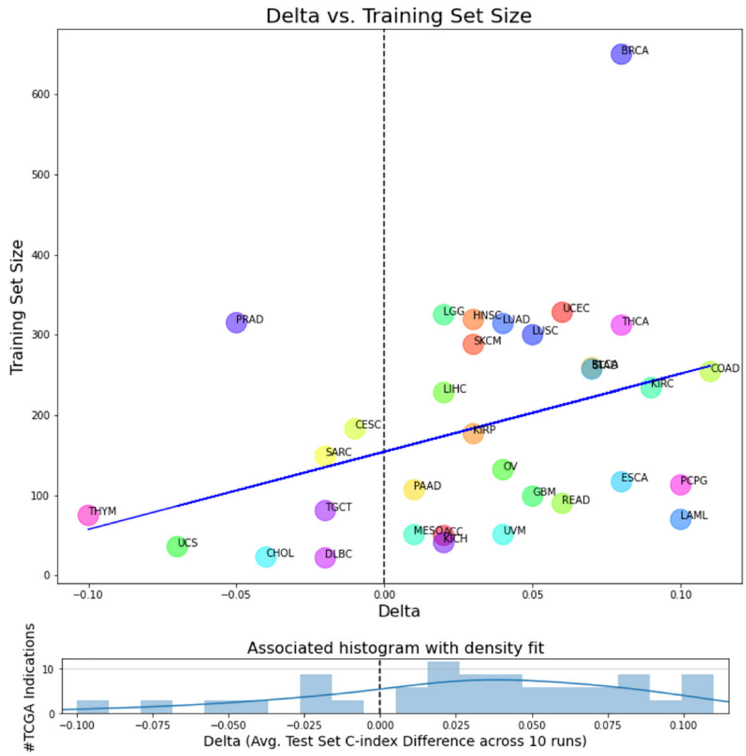
Within the TCGA data, non–small-cell lung carcinoma (NSCLC) was the most well-represented cancer type, with more than 1000 patients having a diagnosis of lung adenocarcinoma (LUAD) or lung squamous cell carcinoma (LUSC). We therefore discuss the NSCLC results in more detail. In addition, we included more modalities than we did for other TCGA cancer types, allowing a more comprehensive analysis.

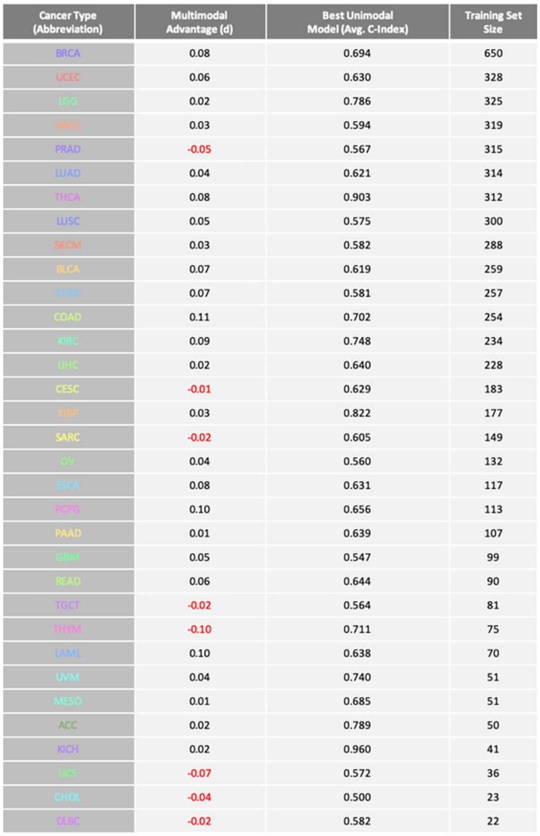
In Fig. 4, we compare the performance of unimodal models from each available modality against multimodal models trained under our proposed late fusion strategy (FUSED). Fig. 4a shows results for models trained jointly on NSCLC patients, and Fig. 4b, c show results on its two subtypes (LUAD and LUSC) separately (joint treatment of both subtypes of NSCLC is common in the literature[35,64]). Univariate Cox PH models were used for feature selection. Similar results obtained for all indications included in the TCGA data set are provided in Supplementary Fig. 1.

Regarding individual modalities, Fig. 4 suggests that each carried some signal for predicting OS, since unimodal models of any modality (except for mutations, in the case of LUSC patients) attained an average C-index of >0.5 on new data. Not all individual modalities were equally predictive of OS. For example, across all NSCLC patients, clinical and demographic features (CLIN) constituted the best individual modality, mainly due to including cancer stage, a strong predictor of OS. They are followed by gene expressions (EXP), with mutations (MUT) the weakest individual modality, possibly because of its binary treatment. The high variance across runs (which justifies our evaluation strategy) permits us to compare only the average performance of each modality.

Comparing unimodal models to the multimodal model, we observed a small advantage in terms of average C-index of the multimodal model over even the best individual modality (here: clinical features). We also found that the variance of the multimodal models was considerably reduced despite the increased feature space. The results applied to patients of both NSCLC subtypes, although the advantage of multimodal models was reduced, possibly due to the smaller sample size. These observations generally hold beyond NSCLC across all TCGA indications (see Supplementary Fig. 1).

The relative importance of each modality in the multimodal model (based on the weights calculated on the validation set) correlated well with





Cancer Type (Abbreviation)	Multimodal Advantage (d)	Best Unimodal Model (Avg. C-Index)	Training Set Size
BRCA	0.08	0.694	650
UCEC	0.06	0.630	328
LGG	0.02	0.786	325
HNSC	0.03	0.594	319
PRAD	-0.05	0.567	315
LUAD	0.04	0.621	314
THCA	0.08	0.903	312
LUSC	0.05	0.575	300
SKCM	0.03	0.582	288
BLCA	0.07	0.619	259
STAD	0.07	0.581	257
COAD	0.11	0.702	254
KIRC	0.09	0.748	234
LIHC	0.02	0.640	228
CESC	-0.01	0.629	183
KIRP	0.03	0.822	177
SARC	-0.02	0.605	149
OV	0.04	0.560	132
ESCA	0.08	0.631	117
PCPG	0.10	0.656	113
PAAD	0.01	0.639	107
GBM	0.05	0.547	99
READ	0.06	0.644	90
TGCT	-0.02	0.564	81
THYM	-0.10	0.711	75
LAML	0.10	0.638	70
UVM	0.04	0.740	51
MESO	0.01	0.685	51
ACC	0.02	0.789	50
KICH	0.02	0.960	41
UCS	-0.07	0.572	36
CHOL	-0.04	0.500	23
DLBC	-0.02	0.582	22

Fig. 3 | [Right] Table showing the advantage *d* of multimodal versus best unimodal model in terms of average C-index across all 33 TCGA cancer types. The table also lists the average C-index of the best unimodal model and the size of the training set per cancer type (entries listed in decreasing order). [Left, top] Multimodal advantage *d* versus training set size. The two quantities are positively correlated (blue trend line added for emphasis). Advantage *d* and training set size are also positively correlated. [Left, bottom] Associated histogram of advantage *d* with density fit shown. Dashed line denotes no advantage ($d = 0$). Multimodal models dominate the best unimodal ones in 25 of the 33 indications ($d > 0$).

the performance of each unimodal model (based on average C-index on the test set) (see Supplementary Fig. 2). Inspecting the weights also verified that the multimodal model did not reduce to a unimodal model (no weight was equal to 0, meaning all modalities were considered).

Exhaustive comparison of all modality combinations in NSCLC

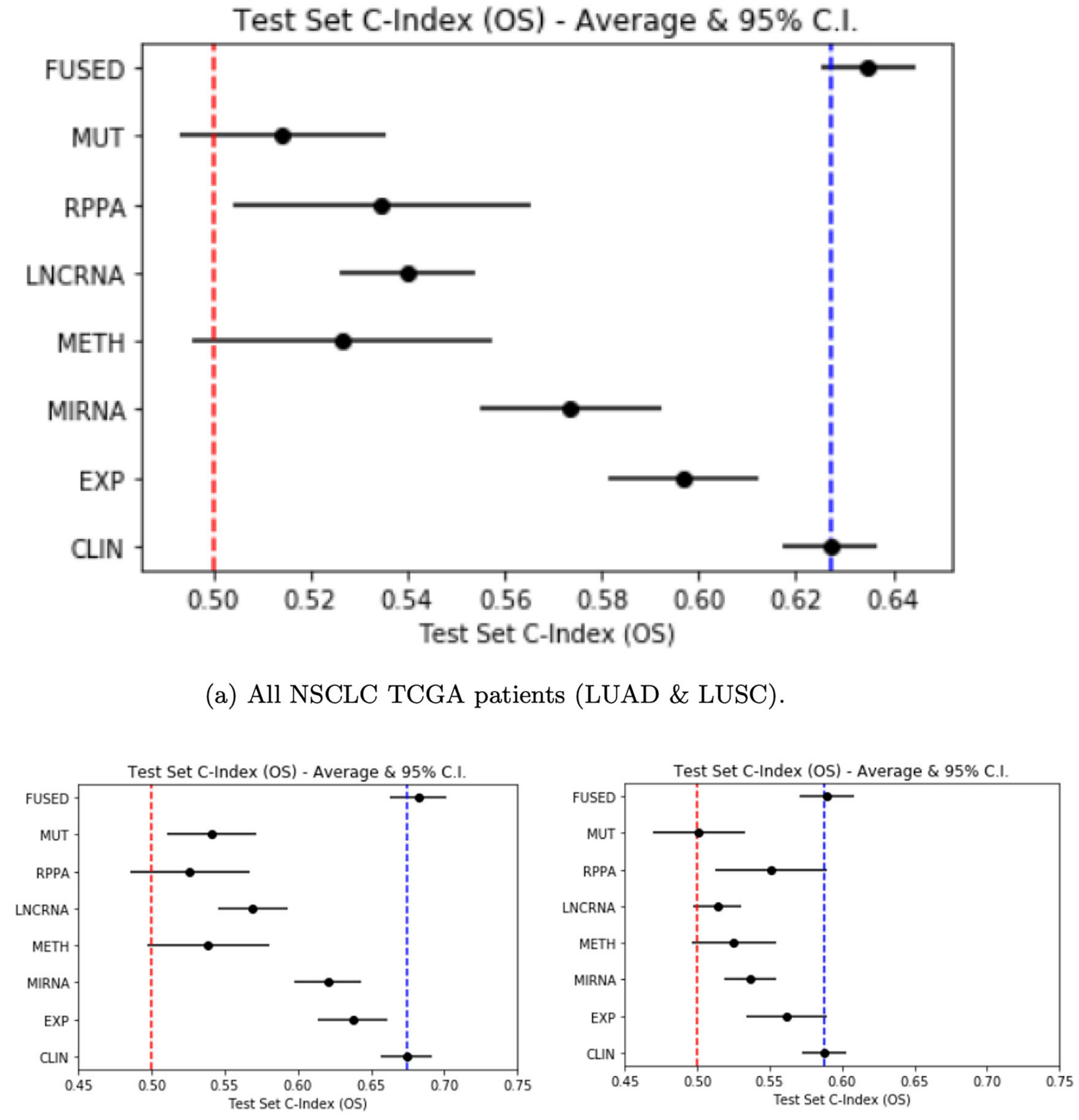
Figure 5 shows the average test set C-index for all possible modality combinations on NSCLC patients, under all models included in the study. The results suggest that adding more modalities leads to improved survival predictions (in the worst, best, and average cases). The performance improvement appeared to plateau after including the two to three most informative modalities, and adding more beyond this point conferred diminishing benefits. In addition, excluding the best individual modality (here: clinical) from the fused model resulted in a large drop in performance (≤4–7.5%, depending on the subpopulation) between combinations of modalities that did and did not include clinical features. Similar results on all cancer types for modalities EXP, MUT, reverse-phase protein array, and CLIN are included in Fig. 6. While results vary considerably by cancer type, the best-performing combinations typically—but not always—include more modalities, as the proposed late fusion approach limits the risk of overfitting if irrelevant features (modalities) are included. Moreover, the best-performing combinations typically—but not always—include the best-performing individual modalities. We observe this in the NSCLC results (Fig. 5), where excluding the most informative modality (CLIN) causes a noticeable drop in performance. On the other hand, we see that even though across cancer types EXP seems to be the highest ranked individual modality, it is outranked by combinations excluding it [CLIN, RPPA].

Exhaustive modality combination plots such as the one shown in Fig. 5 are a useful tool for determining whether adding modalities improves performance and by how much, as well as which modalities should be combined. Such a plot would be computationally intensive to produce for an early fusion strategy, but it can be efficiently obtained under a late fusion strategy, as all underlying unimodal models are already trained.

PANCANCER model results

Figure 7 shows the results of a comparison of the performance of unimodal models and multimodal models trained on the subset of all TCGA patients (PANCANCER). The multimodal model significantly outperformed all unimodal models ($P < 0.0001$), including the best-performing individual modality (clinical features), and exhibited decreased variance. For all TCGA stages and cancer types, multimodal (FUSED) models had an average (±standard deviation) test set C-index of 0.785 ± 0.005, whereas the best unimodal (CLIN) had an average test set C-index of 0.763 ± 0.007 across 10 runs. On this larger sample of patients, our proposed late fusion strategy clearly showed benefits over unimodal models. To our knowledge, these are the top-performing results attained by TCGA pan-cancer models[34,65]. In contrast to individual cancer types (e.g., NSCLC in Fig. 4), aggregating over all 33 indications, we found no noticeable differences on the modality level between early-stage patients (Fig. 4a) and late-stage patients (Fig. 4b) (additional details in Supplementary Fig. 3).

In these subpopulations, cancer stage (clinical feature) was removed. For both groups, clinical features and gene expressions were tied for best individual modalities for predicting OS. Among the clinical features, cancer type is an important indicator of OS. A likely explanation for gene



(a) All NSCLC TCGA patients (LUAD & LUSC).

(b) LUAD patients.

(c) LUSC patients.

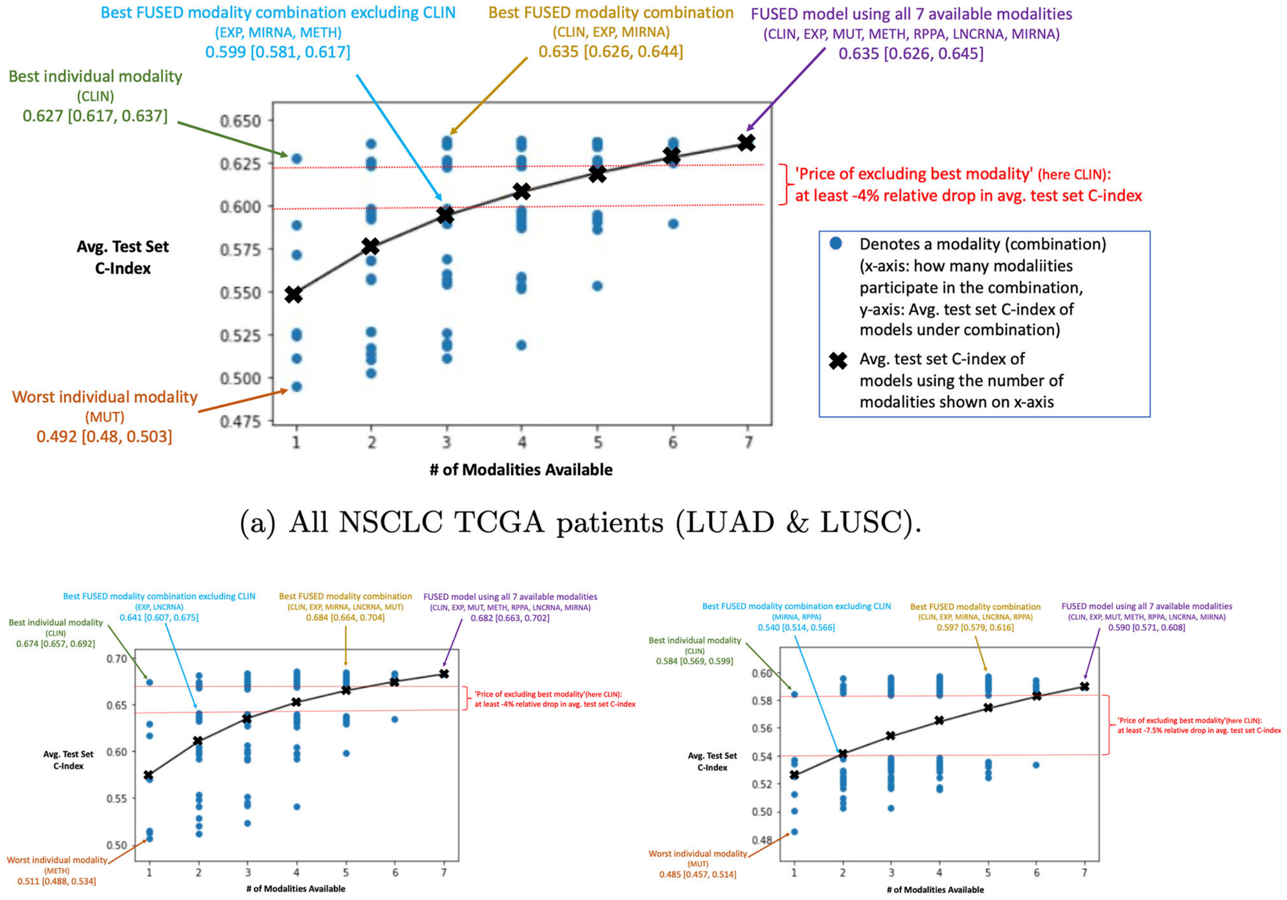
Fig. 4 | Performance of multimodal models (FUSED) versus unimodal models of each modality for NSCLC patients. Results shown for **a** all NSCLC patients, **b** LUAD patients only, and **c** LUSC patients only. The average test set C-index and 95% CI across 10 runs are reported. The red dashed line denotes random prediction performance (C-index = 0.5). The blue dashed line denotes the average C-index of the best individual modality (here: clinical features). Multimodal models outperformed all unimodal models on average and had lower variance. See Supplementary Fig. 1 for other TCGA cancer types.

expression being equally predictive is that, because different types of cancers affect different tissues and different tissues are characterized by distinctive gene expression profiles, gene expression is used by the model as a proxy for cancer type. This is a reminder that information from different modalities can overlap. Nevertheless, obtaining the same information from different data sources can decrease the multimodal model's uncertainty. This is another advantage of multimodal models, evidence for which is shown in these results.

Pan-cancer models are not uncommon in the multi-omics literature[34,65], as the larger sample size lowers the risk of overfitting (which increases as more modalities are added). The clinical usefulness of pan-cancer models is limited, however, as they are trained and evaluated on 33 different types of cancers. We included these results primarily to showcase the power of our proposed fusion strategy on a larger data set. However, there might still be benefits in the multimodal setting of training models on a pan-cancer level to make predictions on individual cancer types (see Supplementary Fig. 4).

Discussion

We observed a consistent advantage of multimodal data fusion in these experiments. Multimodal models outperformed the best unimodal models on 24 of 33 TCGA cancer types. These models exhibited reduced variance and better average predictive performance, showing clear superiority in larger sample sizes. We found that adding modalities yielded diminishing benefits but did not adversely affect performance in pan-cancer models. Finally, we found that in pan-cancer models, where the sample size was larger, the multimodal advantage was particularly pronounced. Under our proposed late fusion strategy, we concluded that we should combine any available omics data. We expect larger benefits with larger training sets and more informative individual modalities.



(a) All NSCLC TCGA patients (LUAD & LUSC).

(b) LUAD patients.

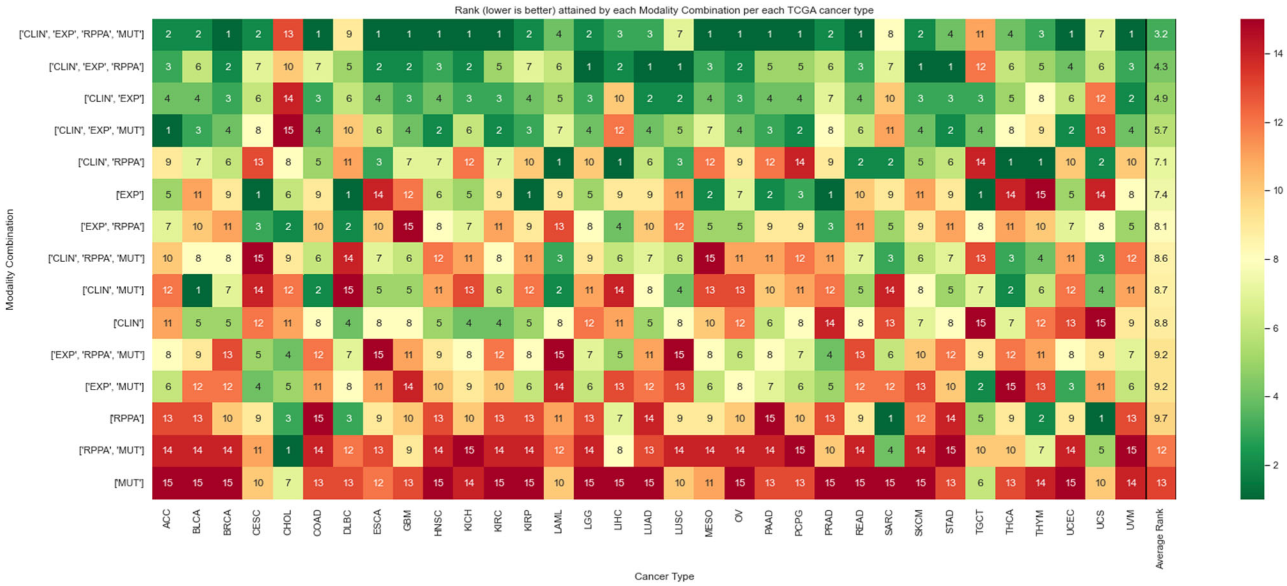
(c) LUSC patients.

Fig. 5 | Average test set C-index for each of the 27-1 possible modality combinations for NSCLC patients. C-index of each modality combination (blue points) on the subset of **a** all NSCLC TCGA patients (LUAD and LUSC), **b** LUAD patients only, and **c** LUSC patients only. The black crosses indicate the average C-index across all multimodal models trained on k modalities, where k is equal to the number shown on the x axes (black trend line added for emphasis). On average, the more modalities added, the better the resulting model. The red dashed lines mark the effect of not including the best individual modality in the multimodal fusion. Also shown is the average test set C-index and 95% CI for: (i) the worst individual modality (orange), (ii) the best individual modality (green), (iii) the best modality combination excluding the clinical features (light blue), (iv) the best modality combination (gold), and (v) the multimodal fusion of all seven modalities (purple). We note the diminishing benefit of adding more modalities and the high "price" of excluding the best individual modality (here: clinical features).

Across the different subpopulations examined, clinical features and gene expressions generally carried a stronger signal for predicting OS than other modalities (e.g., mutations). This could reflect the limitation of our modeling (e.g., binary encoding of mutations), but most likely it is due to the models being inherently more informative. For example, clinical features include variables like "cancer type" and "cancer stage," which are expected to be strong predictors of OS. We hypothesize that gene expression acts as a proxy for "cancer type" (different tissues have different gene expression profiles) and, possibly to some extent, "cancer stage." As a result, multimodal models that do not use these modalities tend to be inferior to those that do, and adding more modalities tends to confer only minor advantages (mostly in terms of reduced variance). This might reflect modeling limitations or could be evidence that the various omics modalities are largely redundant given gene expression. More uncorrelated sources of information, like imaging data, could also be beneficial, and early works, including those by Patwardhan et al.[63], Cheerla and Gevaert[34], and Wulczyn et al.[62] suggest that this is the case.

Other interesting findings include the observation that all modalities were used by the resulting multimodal models, and their relative importance roughly agreed with their predictiveness of OS (Supplementary Fig. 2). Our initial exploration into differences across cancer stages suggests that the role of protein expressions and mutations in late-stage NSCLC and breast invasive carcinoma (BRCA) patients in determining OS is greater than in early stages (Supplementary Figs. 3 and 4). Finally, we demonstrated an example of a multimodal pan-cancer model outperforming models trained on a single cancer type (Supplementary Fig. 1).

The pipeline described in "Pipeline Overview" above is in no way exhaustive but is quite extensive and can be easily adapted to include new methods. It would be especially useful for benchmarking and exploring general tendencies (e.g., early vs. late fusion). The proposed fusion strategy, however, does have limitations. As a late fusion method, it cannot capture cross-modality feature interactions. We did not choose it by an exhaustive search over the possibilities offered by the AZ-AI multimodal pipeline but rather for its simplicity, flexibility, and scalability, and because it yields consistently good results that demonstrate the added benefit of multimodal data fusion (as suggested by our results). On different data sets (e.g., with larger sample size–to–dimensionality ratio, higher signal-to-noise ratio, or higher degree of cross-modality interactions), we do not necessarily expect it to outperform other approaches, but its qualitative advantages (simplicity, flexibility, and scalability) will still hold. Bioinformatics data sets typically consist of various modalities, such as genomic, transcriptomic, epigenomic, proteomic, lifestyle, and phenotypic. Multimodal data fusion offers the potential to improve patient outcomes and gain novel insights by combining information across modalities. Multiple strategies can be used to integrate multimodal data, however, and they are highly problem-specific; identifying an appropriate approach for a given setting is challenging. The literature on



	ACC	BLCA	BRCA	CESC	CHOL	COAD	DLBC	ESCA	GBM	HNSC	KICH	KIRC	KIRP	LAML	LGG	LIHC	LUAD	LUSC	MESO	OV	PAAD	PCPG	PRAD	READ	SARC	SKCM	STAD	TGCT	THCA	THYM	UCEC	UCS	UVM	Average Rank
['CLIN', 'EXP', 'RPPA', 'MUT']	2	2	1	2	13	1	9	1	1	1	1	1	2	4	2	3	3	7	1	1	1	1	2	1	8	2	4	11	4	3	1	7	1	3.2
['CLIN', 'EXP', 'RPPA']	3	6	2	7	10	7	5	2	2	3	2	5	7	6	1	2	1	1	3	2	5	5	6	3	7	1	1	12	6	5	4	6	3	4.3
['CLIN', 'EXP']	4	4	3	6	14	3	6	4	3	4	3	3	4	5	3	10	2	2	4	3	4	4	7	4	10	3	3	3	5	8	6	12	2	4.9
['CLIN', 'EXP', 'MUT']	1	3	4	8	15	4	10	6	4	2	6	2	3	7	4	12	4	5	7	4	3	2	8	6	11	4	2	4	8	9	2	13	4	5.7
['CLIN', 'RPPA']	9	7	6	13	8	5	11	3	7	7	12	7	10	1	10	1	6	3	12	9	12	14	9	2	2	5	6	14	1	1	10	2	10	7.1
['EXP']	5	11	9	1	6	9	1	14	12	6	5	9	1	9	5	9	9	11	2	7	2	3	1	10	9	11	9	1	14	15	5	14	8	7.4
['EXP', 'RPPA']	7	10	11	3	2	10	2	10	15	8	7	11	9	13	8	4	10	12	5	5	9	9	3	11	5	9	11	8	11	10	7	8	5	8.1
['CLIN', 'RPPA', 'MUT']	10	8	8	15	9	6	14	7	6	12	11	8	11	3	9	6	7	6	15	11	11	12	11	7	3	6	7	13	3	4	11	3	12	8.6
['CLIN', 'MUT']	12	1	7	14	12	2	15	5	5	11	13	6	12	2	11	14	8	4	13	13	10	11	12	5	14	8	5	7	2	6	12	4	11	8.7
['CLIN']	11	5	5	12	11	8	4	8	8	5	4	4	5	8	12	11	5	8	10	12	6	8	14	8	13	7	8	15	7	12	13	15	9	8.8
['EXP', 'RPPA', 'MUT']	8	9	13	5	4	12	7	15	11	9	8	12	8	15	7	5	11	15	8	6	8	7	4	13	6	10	12	9	12	11	8	9	7	9.2
['EXP', 'MUT']	6	12	12	4	5	11	8	11	14	10	9	10	6	14	6	13	12	13	6	8	7	6	5	12	12	13	10	2	15	13	3	11	6	9.2
['RPPA']	13	13	10	9	3	15	3	9	10	13	10	13	13	11	13	7	14	9	9	10	15	10	13	9	1	12	14	5	9	2	9	1	13	9.7
['RPPA', 'MUT']	14	14	14	11	1	14	12	13	9	14	15	14	14	12	14	8	13	14	14	14	14	15	10	14	4	14	15	10	10	7	14	5	15	12
['MUT']	15	15	15	10	7	13	13	12	13	15	14	15	15	10	15	15	15	10	11	15	13	13	15	15	15	15	13	6	13	14	15	10	14	13

Fig. 6 | Ranking of each modality combination for each TCGA cancer type. Relative rank (1: best performing to 15: worst performing) attained by each modality combination per cancer type is based on the average C-index of the corresponding survival models across all train/test splits. In case of ties, combinations using fewer modalities are ranked higher, to favor more parsimonious solutions. Rows correspond to modality combinations, columns to cancer types, the last column showing the average rank attained by each modality combination across all cancer types. The modality combinations are ordered according to their average rank (top: best performing, bottom: worst performing).

predicting OS in cancer patients is fragmented and inconsistent, and several promising methods remain underexplored. A consistent and rigorous comparison of multimodal methods against one another and against unimodal models is also missing from the currently available literature. Such comparisons are necessary to determine whether different omics data sets should be combined and, if so, which ones should be used and how they should be combined.

To answer these questions, we implemented the AZ-AI multimodal pipeline, which allows us to train, evaluate, and compare a large number of multi-omics methods for survival analysis. We identified a late fusion strategy that consistently demonstrated the advantages of multimodal data fusion of CLINs, DNA mutations, gene-coding RNA expressions, long non-coding RNA expressions, microRNA expressions, methylations, and protein expressions for the prediction of OS in patients in TCGA. This advantage was consistently demonstrated across individual TCGA cancer types and on pan-cancer models. Multimodal models exhibited reduced variance and better average predictive performance, showing clear superiority in larger sample sizes. Finally, we found that adding modalities did not harm performance, although the benefits diminished. We therefore conclude that, with our proposed strategy, we should combine different omics data sets.

Clinical and demographic data, along with data on differential gene expression, were found to be the most informative modalities overall. Our initial explorations into differences between early- and late-stage NSCLC and BRCA patients suggest an increased role of protein expressions and mutations in patients with late-stage disease in determining OS. A promising future direction would be to further explore the relative importance of modalities for different patient subpopulations, especially in relation to a given treatment. Beyond the modality level, identifying key biomarkers that drive the prediction would require the use of model interpretability methods such as permutation feature importance (included in the AZ-AI multimodal pipeline) or Shapley Additive exPlanations (SHAP)[66].

Our results, along with those of Patwardhan et al.[63] support the claims of others[3,16,58] that late fusion is more successful than early fusion in settings where there is high signal-to-noise ratio and low sample size–to–dimensions ratio, due to a decreased risk of overfitting. A detailed exploration of problem characteristics such as these and others (e.g., degree of intramodality and intermodality correlations) in relation to an optimal fusion strategy, which is possible with the use of the AZ-AI multimodal pipeline, is left for future work.

Finally, it would be interesting to explore more ways of incorporating domain knowledge into the models. Working with pathways and gene sets is a promising direction that could reduce multimodal model dimensionality by leveraging knowledge of the underlying biology and could allow more successful intermediate fusion approaches for more biologically interpretable insights to be drawn from the models.

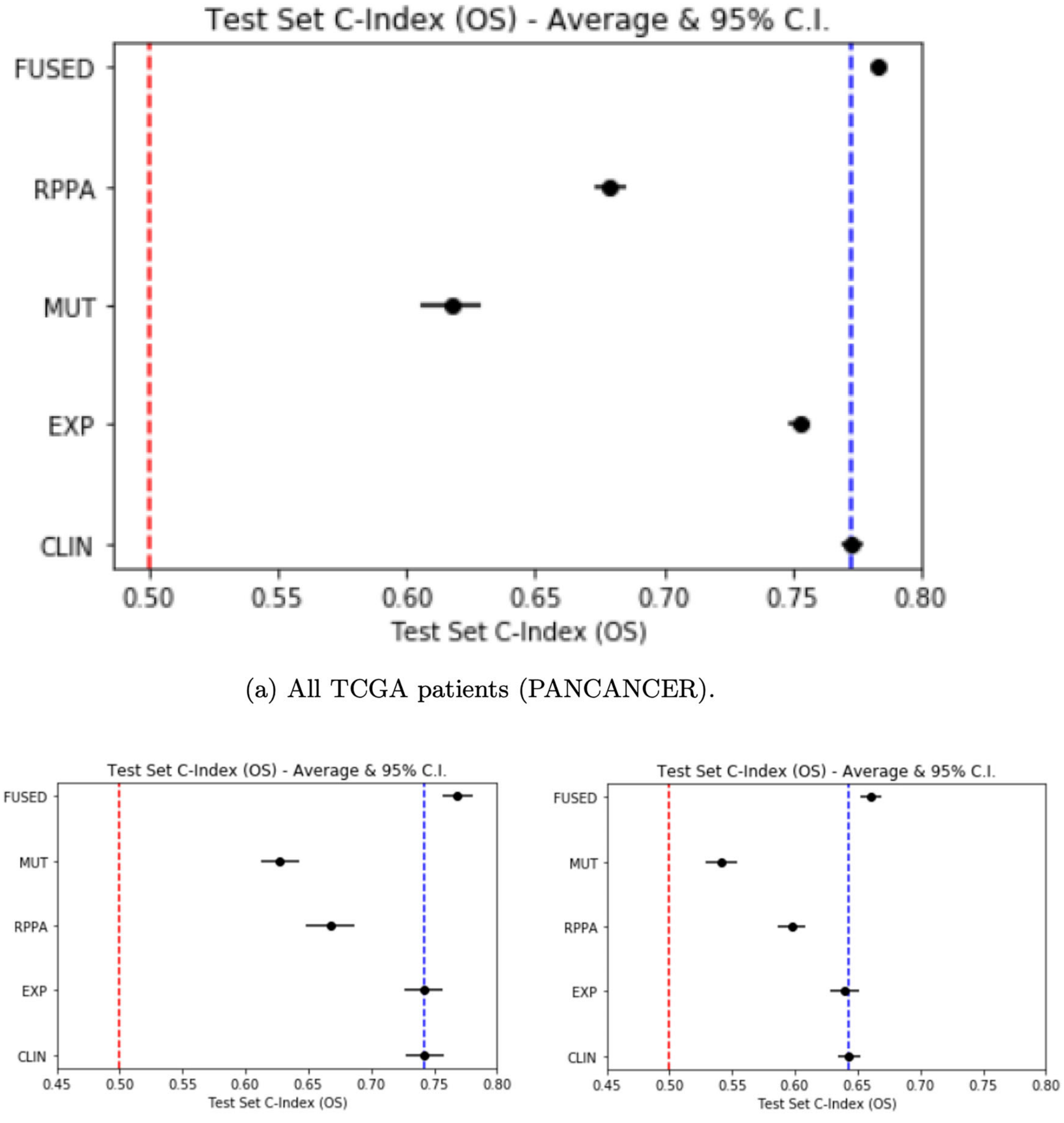
Methods

In this work, we address the gaps in the current literature on multimodal models by introducing the AZ-AI multimodal pipeline for the extensive exploration of multimodal fusion strategies that include several underexplored approaches. We developed a framework for the extensive exploration of multimodal fusion strategies that allows for rigorous evaluation of multimodal fusion approaches against one another and against unimodal models. We demonstrate the advantage of multimodal data fusion for the cancer patient OS endpoint and outline individual contributions of different modalities. Applying our framework to the TCGA data set, we identify a flexible, ensemble-based, late fusion strategy that consistently takes advantage of additional modalities across independent patient subpopulations for improving OS prediction.

OS prediction

We trained models on the task of predicting OS in patient data from TCGA. In this context, OS measures the length of time (in days) from the date of diagnosis to a patient's death by any cause. Monitoring OS requires longer follow-up times than other survival endpoints and is affected by deaths due to noncancer causes. Nevertheless, OS is one of the most common endpoints used in practice because there is minimal ambiguity in defining an OS event[67,68].

As with any survival modeling task, the influence of censoring[69,70] can complicate the analysis and introduce bias. Although most survival modeling methods listed in Table 3 account for right-censoring, most dimensionality reduction methods presented in Table 4 do not. We used the



(a) All TCGA patients (PANCANCER).

(b) Late stage patients (PANCANCER).

(c) Early stage patients (PANCANCER).

Fig. 7 | Performance of multimodal models (FUSED) versus unimodal models of each modality trained on TCGA patients of all cancer types (PANCANCER). Results shown for patients of **a** all stages, **b** late stages (III and V), and **c** early stages (I, II). Average test set C-index and 95% CIs across 10 runs are reported. The red dashed line denotes random prediction performance (C-index = 0.5). The blue dashed line denotes average C-index of best individual modality (here: clinical features). Multimodal models significantly outperformed all unimodal models and had lower variance.

concordance index (C-index)[71] as the most frequently used evaluation metric for the goodness-of-fit of survival models. The C-index is a measure of rank correlation between predicted risk scores and observed survival times. A C-index of 0.5 corresponds to a model whose ranking of predicted patient survival times is no better than random, whereas a C-index of 1 corresponds to a model whose ranking of predicted survival times perfectly matches that of true survival times.

Data

The TCGA project (2006–2018)[57] molecularly characterized more than 20,000 primary cancer samples and matched normal samples, spanning 33 cancer types. From these samples, the project produced multiple types of omics data (genomic, epigenomic, transcriptomic, and proteomic), which are listed along with any corresponding demographic, clinical, and imaging data for each sample. Table 3 lists the 33 cancer types in TCGA, along with the number of data points for each, after excluding patients with follow-up times of <1 day. Table 4 lists the modalities included in the experiments along with any preprocessing the samples underwent.

Pipeline overview

Figure 2 provides a summary of the AZ-AI multimodal pipeline. After specifying the subset of interest (cancer type[s] and stage[s], modalities to be included), the data are split into a training set and a test set, the former to be used for feature selection and training models, and the latter used exclusively for the final evaluation. The user specifies the desired train-test split ratio and (optionally) any variable(s) by which the split is to be stratified. The data then undergoes optional preprocessing, such as normalizing continuous features or imputing missing data. The transformations applied are learned on the training data and then applied to both training and test data. For

Table 3 | Cancer types included in TCGA

Abbreviation	Cancer type	No. of samples
BRCA	Breast invasive carcinoma	1017
UCEC	Uterine corpus endometrial carcinoma	513
LGG	Brain lower-grade glioma	509
HNSC	Head and neck squamous cell carcinoma	499
PRAD	Prostate adenocarcinoma	493
LUAD	Lung adenocarcinoma	491
THCA	Thyroid carcinoma	489
LUSC	Squamous cell lung cancer	470
SKCM	Skin cutaneous melanoma	450
BLCA	Bladder urothelial carcinoma	406
STAD	Stomach adenocarcinoma	402
COAD	Colon adenocarcinoma	398
KIRC	Kidney renal clear-cell carcinoma	367
LIHC	Liver hepatocellular carcinoma	357
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	286
KIRP	Kidney renal papillary cell carcinoma	277
SARC	Sarcoma	234
OV	Ovarian serous cystadenocarcinoma	207
ESCA	Esophageal carcinoma	183
PCPG	Pheochromocytoma and paraganglioma	178
PAAD	Pancreatic adenocarcinoma	168
GBM	Glioblastoma multiforme	155
READ	Rectal adenocarcinoma	141
TGCT	Testicular germ cell tumors	128
THYM	Thymoma	118
LAML	Acute myeloid leukemia	110
MESO	Mesothelioma	81
UVM	Uveal melanoma	80
ACC	Adrenocortical carcinoma	79
KICH	Kidney chromophobe	65
UCS	Uterine carcinosarcoma	57
DLBC	Lymphoid neoplasm, diffuse large B-cell lymphoma	37
CHOL	Cholangiocarcinoma	36
PANCANCER	All of the above cancer types	9481

example, we performed median imputation by using the training set's median on both the training and the test sets. Next, the (optional) dimensionality reduction step is employed (see "Dimensionality reduction"), again based on the training set. The chosen survival models are then trained on the training set. A user-specified fraction of the training data is (optionally) used as a validation set to determine the weights of the ensemble (see "Survival modeling") and of each modality in late fusion multimodal models (see Fig. 1, Late Fusion Strategy). During model training, fivefold cross-validation is applied to identify the optimal hyperparameter setup under a grid search. The final model is then trained on the full training set and evaluated on the held-out test set. We repeated the entire process multiple times ("runs") on different training-validation-test splits and computed the average C-index for the test set and the 95% CIs for each modality combination (see "Evaluation"). The different steps of the pipeline can be executed jointly or per individual modality, giving rise to different multimodal data fusion strategies (see "Fusion Strategies"). The user can select multiple cancer types and train a combined model or separate ones for each. Finally, there is the added option to report aggregate permutation feature importances[72,73] across all runs, affording some degree of model interpretability.

Dimensionality reduction. Table 1 lists the various dimensionality reduction methods explored. These include both feature selection methods (which return a subset of the original features) and feature extraction methods (which transform the original features), as well as both linear and nonlinear methods, univariate and multivariate methods, and others. Combinations are also possible. For instance, we can first select a subset of the original features by filtering out the least variant ones and then discard features that are highly correlated with other selected ones (both steps unsupervised). We can then train autoencoders to obtain a latent representation of the data (also unsupervised) and, finally, select features that are informative of OS by training univariate Cox PH models.

There is also the option to specify lists of genes of interest to include for certain modalities. An exhaustive exploration of these approaches and their combinations is beyond the scope of this work, but the provided code can be used to expand on what is presented here. We included the option to limit the risk of overfitting for feature selection approaches by selecting only the most stable features. This can be achieved by applying the feature selection method on k bootstraps of the training set and retaining only those features that are selected in at least $k_{\min}$ of the bootstraps by the feature selection method. A bootstrap is generated by uniformly sampling with replacement N_{train} data points from the original training set of size N_{train}.

Survival modeling. Table 2 shows the various survival model families included in the pipeline. "ENS" denotes a heterogeneous ensemble of models from all other families listed.

Let us denote with f_i the ith constituent survival model in the ensemble (trained on the training set). We denote with $f_i(\boldsymbol{x})$ the prediction of model f_i on a data point (patient) with feature vector $\boldsymbol{x}$. To obtain the ensemble's prediction $f_{ENS}(\boldsymbol{x})$ on the same data point, the predictions of each base model (survival risk scores) are first normalized to lie within [0, 1], and then the normalized predictions of all models on the same data point $\boldsymbol{x}$ are linearly combined via weights w_i by:

$$f_{ENS}(\boldsymbol{x}) = \frac{\sum_i w_i f_i(\boldsymbol{x})}{\sum_i w_i} \tag{1}$$

The weight w_i of each model is determined based on its performance (C-index, C_i) on either the training set or on a separate validation set by:

$$w_i = \begin{cases} C_i - 0.5, C_i \geq C_{\min} \\ 0, C_i < C_{\min} \end{cases}, C_{\min} > 0.5 \tag{2}$$

To obtain the results shown, we followed the latter approach. Once the weights w_i were determined on the validation set, retraining the base learners f_i on the combined training and validation set tended to further improve performance, as measured on the held-out test set. We suggest following this approach unless the sample size is too small, in which case we advise using the training set to determine the weights.

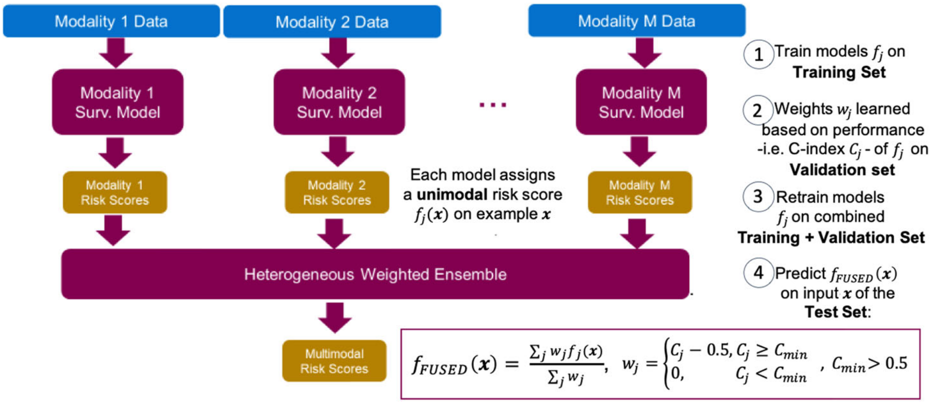
Equation 2 weighs the prediction of each model proportionally to its edge over the random survival model performance ($C = 0.5$). Weighted ensembles of this form are common in the literature[74,75]. We also added a minimal performance $C_{\min} > 0.5$ required to include the model in the ensemble[43]. This confers some robustness to the final prediction; poorly performing models are excluded because in small data sets, a perceived better-than-random performance can easily be due to chance. The optimal value of $C_{\min}$ is problem-specific, but we found that any $C_{\min} > 0.5$ can improve the ensemble's performance.

Fusion strategies. A common categorization of multimodal data integration methods is based on when the various modalities are combined and gives rise to three broad classes of fusion strategies: early, late, and intermediate[3,10,11,16,76]. The relative effectiveness of each depends heavily

Table 4 | Modalities included in this work and their preprocessing

Modality abbreviation	General description & preprocessing notes	No. of features
CLIN	• Clinical and demographic features • Included here: age, gender, race, cancer stage • Cancer stage is omitted in early- and late-stage models.	4
EXP	• RNA expression of protein-coding genes (gene expression) • Differential expression in normal tissues • Selected 1500 most variable across all indications • Selected 1000 most variable per indication	1000 (per individual indication) 1130 (NSCLC) 1500 (PANCANCER)
MUT	• DNA mutations (binarized; 0: wild type, 1: mutation) • Included only those with $$VEP_[impact] = 'high'$ • Selected 1500 most variable across all indications • Selected 1000 most variable per indication	1000 (per individual indication) 1292 (NSCLC)
METH	• DNA methylation probes (beta values) • Retained only those features with <80% missing values • Retained only those probes found in CpG islands, within 1500 base pairs upstream of the transcriptional start site (Chaudhury et al. 2018) • Included only differentially methylated regions for selected genes in EXP • Selected 25,000 most variable across all indications • Selected 1000 most variable per indication	1000 (per individual indication) 1660 (NSCLC)
MIRNA	• microRNA expression • Retained only those features with <80% missing values	867
RPPA	• Reverse-phase protein array (protein expression)	198
LNCRNA	• Long non-coding RNA expression • Extracted from mRNA expression, according to the list in Li et al.[81] • Retained only those features with <80% missing values	1952

mRNA messenger RNA.


Modality 1 Data
Modality 2 Data
Modality M Data
Modality 1 Surv. Model
Modality 2 Surv. Model
...
Modality M Surv. Model
Modality 1 Risk Scores
Modality 2 Risk Scores
Each model assigns a **unimodal** risk score $f_j(\boldsymbol{x})$ on example $\boldsymbol{x}$
Modality M Risk Scores
Heterogeneous Weighted Ensemble
Multimodal Risk Scores
1 Train models f_j on **Training Set**
2 Weights w_j learned based on performance -i.e. C-index C_j - of f_j on **Validation set**
3 Retrain models f_j on combined **Training + Validation Set**
4 Predict $f_{FUSED}(\boldsymbol{x})$ on input $\boldsymbol{x}$ of the **Test Set**:

$$f_{FUSED}(\boldsymbol{x}) = \frac{\sum_j w_j f_j(\boldsymbol{x})}{\sum_j w_j}, \quad w_j = \begin{cases} C_j - 0.5, & C_j \geq C_{min} \\ 0, & C_j < C_{min} \end{cases}, \; C_{min} > 0.5$$

Fig. 8 | Proposed heterogeneous weighted ensemble-based late fusion strategy. Schematic description of the strategy, with steps 1–4 detailing the implementation process, as well as the subset of the data used in each step. The equation to obtain the final multimodal (FUSED) prediction using the individual unimodal models is also provided.

on the prediction task and the amount of data available[3,16]. Fig. 1 describes each approach and provides a visual explanation along with its respective advantages and disadvantages, as well as alternative names used in the literature.

Proposed late-fusion method. Our proposed late fusion strategy is summarized in Fig. 8. It mimics the ensemble construction technique presented above in the section "Survival modeling." Each modality contributes a model f_j, which itself can be an ensemble of the form of Eq. 1. The individual modality predictions are then aggregated using Eq. 1. As before, the models f_j are trained on the training set and the weights w_j are determined based on the validation set performance by Eq. 2. The predictions $f_j(\boldsymbol{x})$ can be interpreted as the unimodal risk scores for a patient with feature vector $\boldsymbol{x}$ and the final ensemble prediction $f_{FUSED}(\boldsymbol{x})$ as the multimodal risk score. The normalized weight $w_j/\sum_j w_j$ of the jth modality can be interpreted as its relative importance in determining OS in the given setting.

Evaluation. With very few exceptions[32,33,62], studies in the existing literature use a simple evaluation scheme consisting of performing a single train-test split of the data. Assuming best practices are followed and no data leakage is taking place, this is a sensible practice for evaluating a single model on the given train-test split. Some studies[32,33] go a step further and repeat the model generation process multiple times, reporting average performance and, in the case of Chai et al.[33] a measure of variance of the performance of final models yielded on the same train-test split under their proposed methodologies. This is a sensible way of evaluating the modeling process on the given train-test split.

These evaluation practices, however, ignore the variability resulting from the data split itself, which can be very large, especially in small data sets. One implication is that any results reported (the produced models and their respective performance) are tied to a specific train-test split. This limits reproducibility and renders comparisons across studies meaningless. In the multimodal setting, this practice also hinders comparisons across modality combinations. Train-test split stratification is typically performed on one or more clinical features, but the resulting split can vary greatly in the distribution of nonclinical features. Different splits can therefore favor different modalities. This variability gives rise to the often-contradictory results reported in the literature.

Our goal was to answer the question, “Should we combine modalities, and if so, which ones and how?” To claim that one combination of modalities has an advantage over another, or even that there is indeed benefit in combining modalities, it is necessary to evaluate the modeling process across train-test splits under different modality combinations. We thus propose what is known in the machine learning literature as “repeated holdout validation”[77–80] as a more appropriate choice of validating the results in this setting. For each modality combination, we trained multiple models on different training sets and evaluated each on the corresponding held-out test set, reporting average performance and 95% CIs.

Experimental setup. For all results presented here, unless stated otherwise, we constructed multimodal models using the strategy described above in the section “Proposed late fusion method.” We first split the dataset into 80% training and 20% test data. For numerical features, we performed median imputation and then selected the top 25 features on each modality by Spearman correlation with OS time. For mutations, we provided a prespecified list of the top 25 genes associated with each indication as identified by AstraZeneca’s Biological Insights Knowledge Graph (BIKG) [78]. BIKG combines relevant data for drug development from public and internal data sources. On the selected feature subset, 80% of the available training data is then used to train survival models and perform hyperparameter optimization by using fivefold cross-validation. The remaining 20% of the training set was used as a validation set. We then evaluated the models by predicting the OS of the test data (20% of the total data). We repeated the process 10 times on different train-validation test splits and computed the average test set C-index and 95% CIs for each modality combination. The variance across runs can be very high, especially when modeling smaller subsets of TCGA (see Table 4 for sample sizes of each subpopulation). Therefore, we repeated the train-test split multiple times when comparing the relative importance of each individual modality or modality combination. We rounded results to the third decimal digit. For both unimodal and multimodal models, unless stated otherwise, we report the results of ensembles of survival models of all families shown in Table 2, except DS (excluded due to its relatively slow hyperparameter optimization), combined under Eq. 1. We set $C_{\min} = 0.53$ for Eq. 2.

Data availability

The data sets (TCGA) utilized in this study are available at https://portal.gdc.cancer.gov/.

Code availability

The pipeline and analysis code are available at https://github.com/kmhl548_azu/multimodal_pipeline.

Received: 12 November 2024; Accepted: 19 April 2025;
Published online: 06 May 2025

References

1. Picard, M., Scott-Boyer, M. P., Bodein, A., Perin, O. & Droit, A. Integration strategies of multi-omics data for machine learning analysis. *Comput. Struct. Biotechnol. J.* **19**, 3735–3746 (2021).
2. Rappoport, N. & Shamir, R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res.* **46**, 10546–10562 (2018).
3. Stahlschmidt, S. R., Ulfenborg, B. & Synnergren, J. Multimodal deep learning for biomedical data fusion: a review. *Brief. Bioinform.* **23**, bbab569 (2022).
4. Baltrusaitis, T., Ahuja, C. & Morency, L. P. Multimodal machine learning: a survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 423–443 (2019).
5. Huang, Y. et al. In *Adv Neural Inf Process Syst*. 34 (eds. M. Ranzato et al.) 1–13 (Neural Information Processing Systems Foundation, 2021).
6. Ramachandran, D. & Taylor, G. W. Deep multimodal learning: a survey on recent advances and trends. *IEEE Signal Process Mag.* **34**, 96–108 (2017).
7. Castanedo, F. A review of data fusion techniques. *Sci. World J.* **2013**, 704504 (2013).
8. Durrant-Whyte, H. F. In *Autonomous Robot Vehicles*. (eds. I. J. Cox & G. T. Wilfong) 73–89 (Springer, 1990).
9. Geng, J., Wang, H., Fan, J. & Ma, X. Deep supervised and contractive neural network for SAR image classification. *IEEE Trans. Geosci. Remote Sens* **55**, 2442–2459 (2017).
10. Khaleghi, B., Kkhamis, A., Karray, F. O. & Razavi, S. N. Multisensor data fusion: a review of the state of the art. *Inf. Fusion* **14**, 28–44 (2013).
11. Lahat, D., Adali, T. & Jutten, C. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proc. IEEE Inst. Electr. Electron Eng.* **103**, 1449–1477 (2015).
12. Turk, M. Multimodal interaction: a review. *Pattern Recognit. Lett.* **36**, 189–195 (2014).
13. Zhang, Y.-D. et al. Advances in multimodal data fusion in neuroimaging: overview, challenges, and novel orientation. *Inf. Fusion* **64**, 149–187 (2020).
14. Cantini, L. et al. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nat. Commun.* **12**, 124 (2021).
15. Huang, S., Chaudhary, K. & Garmire, L. X. More is better: recent progress in multi-omics data integration methods. *Front. Genet.* **8**, 84 (2017).
16. Lipkova, J. et al. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell* **40**, 1095–1110 (2022).
17. Chen, T. & Tyagi, S. Integrative computational epigenomics to build data-driven gene regulation hypotheses. *Gigascience* **9**, giaa064 (2020).
18. Schmidt, F., Kern, F. & Schulz, M. H. Integrative prediction of gene expression with chromatin accessibility and conformation data. *Epigenet. Chromatin* **13**, 4 (2020).
19. Silva, T. C. et al. ELMER v.2: an R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles. *Bioinformatics* **35**, 1974–1977 (2019).
20. Vaske, C. J. et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237–i245 (2010).
21. He, R. & Zuo, S. A robust 8-gene prognostic signature for early-stage non-small cell lung cancer. *Front. Oncol.* **9**, 693 (2019).
22. Zhou, Q., Chen, Q., Chen, X. & Hao, L. Bioinformatics analysis to screen DNA methylation-driven genes for prognosis of patients with bladder cancer. *Transl. Androl. Urol.* **10**, 3604–3619 (2021).
23. Liou, C.-Y., Cheng, W.-C., Liou, J.-W. & Liou, D.-R. Authoencoder for words. *Neurocomputing* **139**, 84–96 (2014).
24. Li, Y. et al. A large cohort study identifying a novel prognosis prediction model for lung adenocarcinoma through machine learning strategies. *BMC Cancer* **19**, 886 (2019).
25. Xie, G. et al. Group lasso regularized deep learning for cancer prognosis from multi-omics and clinical features. *Genes (Basel)* **10**, 240 (2019).
26. Myers, J. L., Well, A. D. & Lorch, R. F. J. *Research Design and Statistical Analysis*. 3rd edn, (Routledge, 2010).
27. Bennasar, M., Hicks, Y. & Setchi, R. Feature selection using Joint Mutual Information Maximisation. *Expert Syst. Appl* **42**, 8520–8532 (2015).
28. Brown, G., Pocock, A., Zhao, M.-J. & Luján, M. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *J. Mach. Learn Res.* **14**, 27–66 (2012).
29. Fleuret, F. Fast binary feature selection with conditional mutual information. *J. Mach. Learn Res.* **5**, 1531–1555 (2004).

30. Macedo, F., Oliveira, M. R., Pacheco, A. & Valadas, R. Theoretical foundations of forward feature selection methods based on mutual information. *Neurocomputing* **325**, 67–89 (2019).
31. Vergara, J. R. & Estévez, P. A. A review of feature selection methods based on mutual information. *Neural Comput. Appl.* **24**, 175–186 (2014).
32. Neums, L., Meier, R., Koestler, D. C. & Thompson, J. A. Improving survival prediction using a novel feature selection and feature reduction framework based on the integration of clinical and molecular data. *Pac. Symp. Biocomput.* **25**, 415–426 (2020).
33. Chai, H. et al. Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. *Comput. Biol. Med.* **134**, 104481 (2021).
34. Cheerla, A. & Gevaert, O. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics* **35**, i446–i454 (2019).
35. Lee, T. Y., Huang, K. Y., Chuang, C. H., Lee, C. Y. & Chang, T. H. Incorporating deep learning and multi-omics autoencoding for analysis of lung adenocarcinoma prognostication. *Comput Biol. Chem.* **87**, 107277 (2020).
36. Wang, R., Huang, Z., Wang, H. & Wu, H. AMMASurv: asymmetrical multi-modal attention for accurate survival analysis with whole slide images and gene expression data. *2021 IEEE Int. Conf. Bioinformatics Biomed.* **2021**, 757–760 (2021).
37. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).
38. Ishwaran, H. & Lu, M. *Random Survival Forests*. (Wiley, 2019).
39. Akai, H. et al. Predicting prognosis of resected hepatocellular carcinoma by radiomics analysis with random survival forest. *Diagn. Interv. Imaging* **99**, 643–651 (2018).
40. Jiang, D. et al. A machine learning-based prognostic predictor for stage III colon cancer. *Sci. Rep.* **10**, 10333 (2020).
41. Miao, F., Cai, Y., Zhang, Y.-X., Fan, X. & Li, Y. Predictive modeling of hospital mortality for patients with heart failure by using an improved random survival forest. *IEEE Access* **6**, 7244–7253 (2018).
42. Moncada-Torres, A., van Maaren, M. C., Hendriks, M. P., Siesling, S. & Geleijnse, G. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Sci. Rep.* **11**, 6968 (2021).
43. Alshdaifat, E., Al-Hassan, M. & Aloqaily, A. Effective heterogeneous ensemble classification: an alternative approach for selecting base classifiers. *ICT Express* **7**, 342–349 (2021).
44. Large, J., Lines, J. & Bagnall, A. The heterogeneous ensembles of standard classification algorithms (HESCA): the whole is greater than the sum of its parts. *arXiv* https://doi.org/10.48550/arXiv.1710.09220. https://arxiv.org/abs/1710.09220 (2017).
45. Sabzevari, M., Martínez-Muñoz, G. & Suárez, A. Building heterogeneous ensembles by pooling homogeneous ensembles. *Int. J. Mach. Learn Cyber* **13**, 551–558 (2021).
46. Borisov, V. et al. Deep neural networks and tabular data: a survey. *IEEE Trans. Neural Netw. Learn Syst.* **35**, 7499–7519 (2022).
47. Grinsztajn, L., Oyallon, E. & Varoquaux, G. Why do tree-based models still outperform deep learning tabular data? *arXiv* https://doi.org/10.48550/arXiv.2207.08815. (2022).
48. Shwartz-Ziv, R. & Armon, A. Tabular data: deep learning is not all you need. *Inf. Fusion* **81**, 84–90 (2022).
49. Li, Y. C., Wang, L., Law, J. N., Murali, T. M. & Pandey, G. Integrating multimodal data through interpretable heterogeneous ensembles. *Bioinform. Adv.* **2**, vbac065 (2022).
50. Pölsterl, S. et al. Heterogeneous ensembles for predicting survival of metastatic, castrate-resistant prostate cancer patients. *F1000Res* **5**, 2676 (2016).
51. Khairalla, M., Ning, X., Al-Jallad, N. T. & El-Faroug, M. O. Short-term forecasting for energy consumption through stacking heterogeneous ensemble learning model. *Energies* **11**, 1605 (2018).
52. Xia, Y., Liu, C., Da, B. & Xie, F. A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Syst. App.l* **81**, 182–199 (2018).
53. Zhao, C., Xin, Y., Li, X., Yang, Y. & Chen, Y. A heterogeneous ensemble learning framework for spam detection in social networks with imbalanced data. *Appl Sci.* **10**, 936 (2020).
54. Bayoudh, K., Knani, R., Hamdaoui, F. & Mtibaa, A. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *Vis. Comput.* **38**, 2939–2970 (2022).
55. Driess, D. et al. PaLM-E: an embodied multimodal language model. *arXiv* https://arxiv.org/abs/2303.03378 (2023).
56. Jain, A. et al. MURAL: multimodal, multitask retrieval across languages. *arXiv*. https://arxiv.org/abs/2109.05125 (2021).
57. National Cancer Institute. The Cancer Genome Atlas Program. https://www.cancer.gov/ccg/research/genome-sequencing/tcga (2023).
58. Zhu, W., Xie, L., Han, J. & Guo, X. The application of deep learning in cancer prognosis prediction. *Cancers***12**, 603 (2020).
59. Chen, L. et al. Histopathological images and multi-omics integration predict molecular characteristics and survival in lung adenocarcinoma. *Front. Cell Dev. Biol.* **9**, 720110 (2021).
60. Feng, G. et al. Predicting the survival period of non-small cell lung cancer based on deep learning. 11th International Conference on Data Science and Advanced Analytics (DSAA), San Diego, CA, 1–7 (2024).
61. Ye, Q. et al. Multi-omics immune interaction networks in lung cancer tumorigenesis, proliferation, and survival. *Int. J. Mol. Sci.* **23**, 14978 (2022).
62. Wulczyn, E. et al. Deep learning-based survival prediction for multiple cancer types using histopathology images. *PLoS ONE* **15**, e0233678 (2020).
63. Patwardhan, K. A. et al. Towards a survival risk prediction model for metastatic NSCLC patients on durvalumab using whole-lung CT radiomics. *Front. Immunol.* **15**, 1383644 (2024).
64. Takahashi, S. et al. Predicting deep learning based multi-omics parallel integration survival subtypes in lung cancer using reverse phase protein array data. *Biomolecules* **10**, 1460 (2020).
65. Vale-Silva, L. A. & Rohr, K. Long-term cancer survival prediction using multimodal deep learning. *Sci. Rep.* **11**, 13505 (2021).
66. Lundberg, S. M. & Lee, S.-I. in *NIPs'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*. (eds U. von Luxburg & I. Guyon) 4768–4777 (Curran Associates, 2017).
67. Liu, J. et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400–416 e411 (2018).
68. Pazdur, R. Endpoints for assessing drug activity in clinical trials. *Oncologist* **13**, 19–21 (2008).
69. Klein, J. P. & Moeschberger, M. L. *Survival Analysis: Techniques For Censored And Truncated Data*. 2nd edn, (Springer, 2003).
70. Leung, K. M., Elashoff, R. M. & Afifi, A. A. Censoring issues in survival analysis. *Annu. Rev. Public Health* **18**, 83–104 (1997).
71. Harrell, F. E. Jr., Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15**, 361–387 (1996).
72. Breiman, L. Random forests. *Mach. Learn* **45**, 5–32 (2001).
73. Molnar, C. *Interpretable Machine Learning*. (2020).
74. Fang, Z., Wang, Y., Peng, L. & Hong, H. A comparative study of heterogeneous ensemble-learning techniques for landslide susceptibility mapping. *Int. J. Geogr. Inf. Sci.* **35**, 321–347 (2020).
75. Tsoumakas, G., Katakis, I. & Vlahavas, I. In *Maching Learning: ECML 2004*. 3201 Lecture Notes in Computer Science (eds J. F. Boulicaut, F. Esposito, F. Giannotti, & D. Pedreschi) 465–476 (Springer, 2004).
76. Vieira, S., Lopez Piñaya, W. H., Garcia-Dias, R. & Mechelli, A. In *Machine Learning: Methods and Applications to Brain Disorders*. (eds. A. Mechelli & S. Vieira) Ch. 16, 283–305 (Academic Press, 2019).
77. Kim, J.-H. Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap. *Comput. Stat. Data Anal.* **53**, 3735–3745 (2009).

78. Raschka, S. Model evaluation, model selection, and algorithm selection in maching learning. *arXiv*. https://arxiv.org/abs/1811.12808 (2018).
79. Tanner, E. M., Bornehag, C. G. & Gennings, C. Repeated holdout validation for weighted quantile sum regression. *MethodsX* **6**, 2855–2860 (2019).
80. Tantithamthavorn, C., McIntosh, S., Hassan, A. E. & Matsumoto, K. An empirical comparison of model validation techniques for defect prediction models. *IEEE Trans. Softw. Eng.* **43**, 1–18 (2017).
81. Li, Y. et al. Pan-cancer characterization of immune-related lncRNAs identifies potential oncogenic biomarkers. *Nat. Commun.* **11**, 1000 (2020).

Acknowledgements

The authors thank Jacob Gould Ellen, currently of Harvard Medical School, and Arthur Liberzon, and Gustavo Alonso Arango-Argoty of AstraZeneca for their helpful suggestions, which improved this manuscript, and Deborah Shuman and Elizabeth Choe of AstraZeneca for editorial support.

Author contributions

N.N., N.M., R.M., D.S., and E.J. have made contributions to the conception of this study. N.N., E.J., D.S., and N.M. contributed to the design of this study. N.N., R.M., D.S., and M.G. contributed to the analysis of the data. N.N., E.J., D.S., N.B., and N.M. contributed to the interpretation of the data. N.N., D.S., and H.R. contributed to the creation of new software used in this work. N.N., N.M., H.R., and E.J. have drafted and revised the manuscript.

Competing interests

All authors are or were employees of AstraZeneca at the time this work was performed and may have stock ownership, options, or interests in the company.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41698-025-00917-6.

Correspondence and requests for materials should be addressed to Natasha Markuzon or Etai Jacob.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.