



Benchmarking Early, Latent and Decision Fusion of RNA Sequencing with Other Omics Data for Lymph Node Metastasis Prediction in Head and Neck Cancer

Aneela Nargis^{1*}, Raazia Sosan Waseem²

¹Department of Information Technology, DHA Suffa University, Karachi, Pakistan

²Department of Software Engineering, DHA Suffa University, Karachi, Pakistan

ARTICLE INFO

Article History:

Received: July 12, 2026
Revised: August 16, 2026
Accepted: September 15, 2026
Available Online: September 30, 2026

Keywords:

Multimomics Fusion
Head and Neck Cancer
Lymph Node Metastasis
Deep Learning
Benchmark Study

Classification Codes:

Funding:

This research received no external funding.

ABSTRACT

Lymph node spread is one of the strongest signs of poor outcome in head and neck squamous cell carcinoma, and small nodal deposits are often missed before surgery. This study is a benchmark that asks whether adding DNA methylation, copy number and protein data to RNA sequencing improves prediction of pathological node status from the primary tumour, and whether the way the layers are joined matters. We used 416 tumours from The Cancer Genome Atlas, of which 242 were node positive. Six data combinations were tested with three fusion strategies (early fusion of all features, latent fusion of autoencoder codes, and decision fusion of single layer models) and six model families: logistic regression, support vector machine, random forest, XGBoost, a multilayer perceptron and MO-LSFN, a deep model that places normalised autoencoder codes on a 32 by 32 map and classifies it with a wide residual network. All 76 models used the same fivefold cross validation, and every step that learns from data was fitted on training patients only. The main result is negative. Areas under the receiver operating characteristic curve ranged from 0.555 to 0.649. Logistic regression on RNA sequencing alone reached 0.609, and the best model, decision fusion with logistic regression on all four layers, reached 0.649 (95% confidence interval 0.591 to 0.705), a difference that was not significant. Adding layers changed the mean score by less than 0.01, and latent fusion was only slightly better than early fusion. A model using clinical variables alone reached 0.579. MO-LSFN gained the most from extra layers, from 0.556 to 0.628, but did not beat simpler models. A power analysis showed that this design could detect only differences of about 0.08 to 0.11, so smaller gains cannot be ruled out. Results were similar in patients with all four layers, and somewhat higher when N1 patients were excluded. The molecular profile of the primary tumour carries only a weak signal for nodal spread in this cohort, and fusion design matters less than is often assumed.

© 2026 The authors published by JCIS. This is an Open Access Article under the Creative Common Attribution Non-Commercial 4.0



Corresponding Author's Email: a.nargis@dsu.edu.pk

Citation: A. Nargis and R. S. Waseem, "Benchmarking Early, Latent and Decision Fusion of RNA Sequencing with Other Omics Data for Lymph Node Metastasis Prediction in Head and Neck Cancer," Journal of Computers and Intelligent Systems, vol. 4, no. 1, 2026.

1. Introduction

Head and neck squamous cell carcinoma (HNSCC) starts in the lining of the mouth, throat and larynx and is one of the most common cancers in the world [1], [2]. In South Asia, and in Pakistan in particular, it is very common because many people use smokeless tobacco and betel quid [3]. Spread of the tumour to the lymph nodes of the neck is one of the strongest signs of poor outcome in this disease [2], [4]. Some patients have small nodal deposits that imaging does not show, so surgeons often remove neck nodes as a precaution. A reliable way to estimate nodal spread from the primary tumour itself could help plan this treatment.

The molecular data of the primary tumour may carry part of this information. RNA sequencing (RNA-seq) measures the activity of about 20,000 genes and is the most widely used omics layer in cancer research. Other layers, such as

DNA methylation, copy number and protein levels, describe different parts of tumour biology [5], [6]. A common belief is that adding these layers to RNA-seq should improve prediction. This belief is rarely tested in a fair way. Each layer has thousands of features, cohorts have a few hundred patients, and models overfit easily [7]. The way the layers are joined, known as the fusion strategy, also matters [8], [9].

Three fusion strategies are common. Early fusion joins all features into one long vector. Latent fusion first compresses each layer with its own network and then joins the short codes. Decision fusion trains one model per layer and averages their predictions. Studies usually pick one strategy and one model, which makes it hard to know whether a gain comes from the extra data, the fusion strategy or the model [10].

We designed this study as a benchmark. Starting from RNA-seq alone, we add copy number (CNV), DNA methylation (MET) and reverse phase protein array (RPPA) data in six combinations. For each combination we test early, latent and decision fusion with six model families: logistic regression, support vector machine, random forest, XGBoost, a multilayer perceptron and MO-LSFN. MO-LSFN extends our earlier single layer model AIW-DynOpt [11] to several layers. The task is to predict pathological lymph node status from the primary tumour in TCGA-HNSC. All models use the same cross validation folds, and every step that learns from data is fitted on training patients only. We also report a clinical only baseline, sensitivity analyses and a power statement. The main result is negative: neither extra layers, nor the fusion strategy, nor the deep model clearly improved on logistic regression with RNA-seq alone. We report it because negative results of this kind help other groups choose simpler designs and avoid over interpreting small gains.

Section 1.1 reviews related work. Section 2 describes the data and methods, Section 3 gives the results, Section 4 discusses them together with the limits of the study, and Section 5 concludes.

1.1 Related Work

1.1.1 Fusion Strategies for Omics Data

Methods for combining omics layers are usually grouped by the stage at which the layers meet [8]. In early fusion, all features are joined before modelling. This is simple, but the largest layer tends to dominate. In joint or latent fusion, the model learns codes from the layers and joins them. Chaudhary et al. used an autoencoder on expression, microRNA and methylation data to find survival groups in liver cancer [12], and Tong et al. compared two autoencoder designs for breast cancer survival [13]. In decision fusion, one model is trained per layer and their outputs are combined. MOGONET follows a related idea, with one graph network per layer and a final network that combines their outputs [14]. Few studies compare all three strategies on the same data with the same folds.

1.1.2 Autoencoders and Convolutional Networks for Omics Data

An autoencoder learns to compress its input into a short code and then rebuild the input [15]. Variational autoencoders add a probabilistic code [16], and Way and Greene showed that such codes can match known biological programmes [17]. Franco et al. compared autoencoder types for multiomics subtype detection [18], and DeepProg combined several autoencoders with standard models to predict survival [19]. Convolutional and residual networks were designed for images [20]. Lyu and Haque placed expression values into 2D matrices for tumour type classification [21], Darendeli and Yilmaz turned expression values into images [22], and DeepInsight and IGTD place related features close together on a grid [23], [24]. Wide residual networks use fewer but wider layers and train well on small inputs [25].

1.1.3 Machine Learning Studies in HNSCC

Most computational studies of HNSCC use a single data type. Ju et al. searched TCGA data for diagnostic and prognostic genes [26], Li et al. studied prognostic genes in the tumour microenvironment [27], and Nunez Lopez et al. combined microRNA signatures with random survival forests [28]. Weusthof et al. built a model to predict perineural invasion [29], and Wang et al. derived a gene signature from cancer associated fibroblasts [30]. A review by Wang and Li found that deep learning for head and neck multiomics was still at an early stage [10]. We found no study that compares fusion strategies and model families side by side for nodal status in HNSCC.

2. Materials and Methods

2.1 Data and Cohort

We used the TCGA-HNSC data from the PanCancer Atlas release [31], as distributed by the cBioPortal data hub under the study name `hnscc_tcga_pan_can_atlas_2018` [32], [33]. The TCGA Research Network first described this

cohort in 2015 [5]. The label was the pathological N stage of the primary tumour. Patients with N0 were labelled node negative, and patients with N1, N2 (including N2a, N2b and N2c) or N3 were labelled node positive. Patients with NX or no N stage were excluded. We used one primary tumour sample per patient.

The final cohort had 416 patients with RNA-seq data, of whom 242 (58.2%) were node positive and 174 (41.8%) node negative. Among node positive patients, 66 were N1, 168 were N2 (12 N2, 8 N2a, 100 N2b and 48 N2c) and 8 were N3. TCGA does not record the extent of neck dissection, so the number of patients with limited dissections cannot be determined. RNA-seq was given as RSEM values for 20,511 genes. DNA methylation was given as gene level beta values for 13,063 genes, after removing genes with more than 20% missing values. Copy number was given as GISTIC2.0 calls for 25,087 genes [34], and RPPA covered 191 antibodies [35]. Methylation data were available for all 416 patients, copy number for 411 and RPPA for 176; 171 patients (100 node positive) had all four layers (Table 1). For the clinical analyses we also used age, sex, pathological T stage, tumour site from the ICD-O-3 topography code (oral cavity, oropharynx, larynx, or other) and the PanCancer Atlas HPV subtype.

Table 1. Data types in the study cohort (416 patients, 242 node positive)

Data type	Values	Genes or antibodies	Patients with data	Node positive with data	Features used in each fold
RNA-seq	$\log_2(\text{RSEM} + 1)$	20,511	416	242	5,000 most variable
DNA methylation (MET)	Gene level beta value	13,063	416	242	5,000 most variable
Copy number (CNV)	GISTIC2.0 calls, -2 to 2	25,087	411	238	5,000 most variable
Protein (RPPA)	RPPA level	191	176	104	All 191

2.2 Study Design

We compared six data combinations: RNA alone, RNA + CNV, RNA + MET, RNA + RPPA, RNA + CNV + MET, and all four layers. For each combination we tested three fusion strategies (Section 2.5) and up to six model families (Section 2.6). Decision fusion needs at least two layers, so it was not run for RNA alone, and it was run with two model families only. This gives 11 models for RNA alone and 13 for each of the other five combinations, 76 models in total. Throughout the paper, the reference model is logistic regression on RNA-seq alone with early fusion, which we call the RNA only LR reference. All models used the same stratified fivefold cross validation split. Inside each training fold, every step that learns from data was fitted on training patients only. This covered feature filtering, filling of missing values, scaling, autoencoder training, code normalisation and map ordering. We did not use oversampling such as SMOTE, because creating synthetic patients before splitting can leak test information into training [36], [37]. Class imbalance was handled with class weights where the model allowed it.

2.3 Preprocessing

In each training fold we kept the 5,000 most variable features for RNA-seq, methylation and copy number, and all 191 antibodies for RPPA. Missing values were filled with the training median. Each feature was then scaled with the training mean and standard deviation. The same steps, with the same values, were applied to the test fold. When a patient had no data for a layer, all features of that layer were set to zero, which equals the training mean after scaling.

2.4 Autoencoder for Each Data Type

Each data type m was compressed by its own deep undercomplete autoencoder (DUAЕ). For a scaled input vector x_m the encoder computes

$$z_m = W_2 \varphi(W_1 x_m + b_1) + b_2 \quad (1)$$

where φ is the exponential linear unit and W_1 , b_1 , W_2 and b_2 are learned weights and biases. The hidden layer had 1,024 units for RNA-seq, methylation and copy number and 191 units for RPPA. The decoder mirrors the encoder and ends in a linear layer. Each autoencoder was trained to reduce the mean squared reconstruction error

$$\mathcal{L}_m = (1/n) \sum_i \|x_{m,i} - \hat{x}_{m,i}\|^2 \quad (2)$$

with the Adam optimiser (learning rate 10^{-3} , weight decay 10^{-5} , batch size 64, dropout 0.1, 150 epochs). The code sizes were 320 for RNA-seq, methylation and copy number and 64 for RPPA. The autoencoders were trained once per fold, on the training patients who had that layer, and the same codes were used by every latent fusion model.

2.5 Fusion Strategies

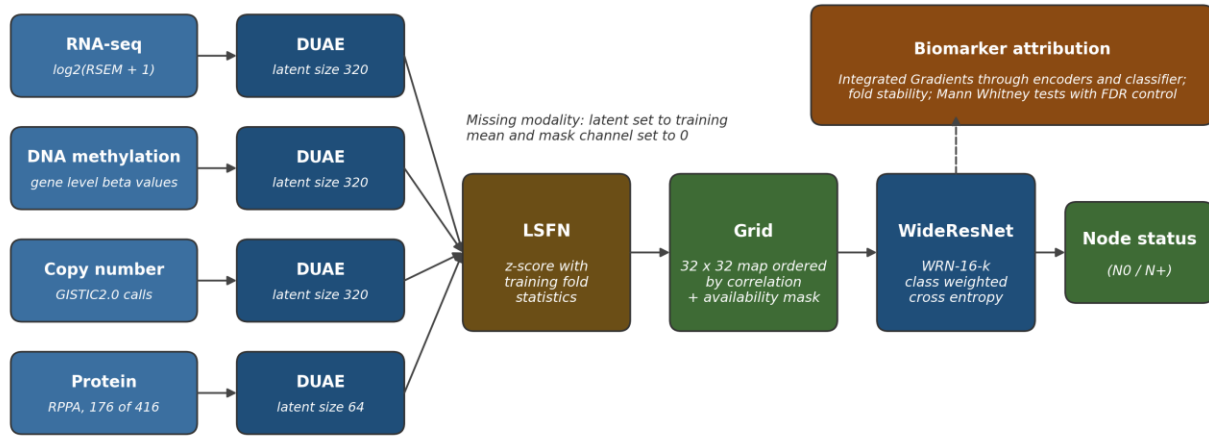
Early fusion joined the scaled features of all layers in a combination into one vector, giving up to 15,191 features. This is the simple concatenation baseline. Latent fusion joined the autoencoder codes of the layers in a combination. Each code value j was first standardised with the mean μ_j and standard deviation s_j of the training patients who had that layer,

$$\hat{z}_j = a_j (z_j - \mu_j) / s_j \quad (3)$$

where a_j is 1 if the patient has the layer of value j and 0 if not. We call this step latent space feature normalisation (LSFN). For the conventional models, one availability indicator per layer was added to the normalised codes. Decision fusion trained one model per layer on the patients who had that layer and averaged the predicted probabilities of the layers available for each test patient. We used logistic regression and XGBoost for decision fusion.

2.6 Models

Five conventional model families were used. Logistic regression (LR) had an L2 penalty and class weights, with the penalty chosen by inner fivefold cross validation. The support vector machine (SVM) used a radial basis kernel with $C = 1$ and class weights. The random forest (RF) had 500 trees and class weights. XGBoost used 300 trees of depth 3, a learning rate of 0.05, row sampling of 0.8, column sampling of 0.3 and a positive class weight equal to the class ratio in the training fold. The multilayer perceptron (MLP) had hidden layers of 256 and 64 units, L2 penalty 10^{-3} and early stopping; the scikit learn implementation does not support class weights. Models were implemented with scikit learn [38] and XGBoost [39].



Fixed settings: WRN-16-4, AdamW, 30 epochs, modality dropout 0.2

Figure. 1. The MO-LSFN model. Each data type is compressed by its own deep undercomplete autoencoder (DUAE). The codes are standardised with training statistics (LSFN), placed along a snake shaped path on a 32 by 32 map ordered by correlation, and joined with a mask channel. A wide residual network then classifies the map, and Integrated Gradients links each prediction back to the input features.

2.7 MO-LSFN

MO-LSFN is the deep model used for latent fusion (Figure 1). To give its convolutional layers meaningful neighbours, we measured the Pearson correlation r between all pairs of normalised code values on the training fold and ordered them by average linkage clustering with distance $1 - |r|$, following the idea behind DeepInsight and IGTD [23], [24]. The ordered values were placed on a 32 by 32 grid along a snake shaped path, so that values next to each other in the order are also next to each other on the grid. A second grid holds the availability mask, which gives an input of size 2 by 32 by 32. The map is classified by a wide residual network of depth 16 [25] with width factor $k = 4$. Each residual block computes

$$x_i = x_{i-1} + \mathcal{F}(x_{i-1}; \theta_i) \quad (4)$$

where $\mathcal{F}$ is the function learned by block l with parameters θ_l [20]. Global average pooling and a softmax layer give the class probabilities. The network was trained with class weighted cross entropy,

$$\mathcal{L}_{cls} = -(1/n) \sum_i \sum_c w_{c,i} y_{i,c} \log \hat{y}_{i,c} \quad (5)$$

where $w_{c,i} = n / (2 n_c)$ and n_c is the number of training patients in class c. We used AdamW with a learning rate of 10^{-3} , dropout 0.3, a cosine schedule, batch size 32 and 30 epochs. During training, modality dropout hid one available layer at random for each patient with probability 0.2, so that the network learned to work with incomplete inputs. For the four layer combination we also ran three ablation models: a random order on the map, no code normalisation, and no mask channel.

The design choices were made in advance and were not tuned. The code sizes were chosen so that the four codes fill a square map exactly: three layers of 320 values and one of 64 give 1,024 values, which is 32 by 32, so no padding is needed. RPPA received a smaller code because it has only 191 inputs, and an undercomplete code must be smaller than its input. The snake shaped path is the simplest layout in which every pair of consecutive values is also adjacent on the grid. Space filling curves such as the Hilbert or Z order curve keep more of this locality in two directions and are a natural alternative. WRN-16-4 was carried over from our earlier single layer work [11], where it trained stably on inputs of this size. We did not test other grid sizes, paths, network depths, network widths or code sizes, and we list this as a limitation.

2.8 Clinical Baseline and Sensitivity Analyses

We compared the distribution of T stage, tumour site, HPV subtype and age between node negative and node positive patients with the chi square test and the Mann Whitney U test. As a clinical only baseline, we trained LR with class weights on age, sex, T stage group (T1 or T2, T3 or T4), tumour site and HPV subtype, with missing categories coded as unknown, using the same folds and the same inner cross validation as the omics models. We then ran two sensitivity analyses on the existing out of fold predictions, without retraining any model. The first restricted the evaluation to the 171 patients with all four layers, to check that the results were not driven by missing RPPA data. The second excluded the 66 N1 patients and compared N0 with N2 or N3 (350 patients), to check how much the weak signal depends on the boundary between N0 and N1, where pathological staging is most likely to be uncertain.

2.9 Statistical Analysis

We pooled the test predictions from all five folds and computed each metric once. We report the area under the receiver operating characteristic curve (AUROC) with a 95% bootstrap confidence interval from 1,000 samples of patients, the area under the precision recall curve (AUPRC), balanced accuracy and the Matthews correlation coefficient (MCC), using a cut off of 0.5 for the last two. Because 58.2% of patients are node positive, a model with no skill has an AUPRC of 0.582, and this is the baseline for every AUPRC value in the paper. Fold level AUROC values of every model were compared with the RNA only LR reference using the corrected resampled t test of Nadeau and Bengio [40], which allows for the overlap between training sets.

To show what differences this design could detect, we computed the smallest fold level AUROC difference that the corrected t test would find at a two sided level of 0.05. We used the median standard deviation of the fold level differences between each model and the reference (0.044) and four degrees of freedom. The smallest detectable difference was 0.082 at 50% power and 0.110 at 80% power. Differences smaller than this, which include all the differences we observed, cannot be confirmed or ruled out with a single run of fivefold cross validation in a cohort of this size.

2.10 Feature Attribution

For MO-LSFN on all four layers, we computed Integrated Gradients [41] for the test patients in each fold, through the encoders, the normalisation, the map and the classifier, with the training mean as the reference input and 32 steps. This gives a score for each gene, methylation value, copy number call and protein. For each data type we computed its share of the total absolute score. We measured the stability of the rankings between folds in two ways: the Jaccard similarity of the top 50 feature sets, and the Spearman correlation of the scores of all features that were selected in both folds of a pair. For the ten top features of each data type we compared node positive and node negative tumours with the Mann Whitney U test and corrected the p values with the Benjamini and Hochberg method [42].

2.11 Implementation and Reproducibility

We wrote the code in Python 3.11 with PyTorch 2.14, scikit learn 1.8, XGBoost 3.2, pandas 3.0 and SciPy 1.17 [38], [39], [43]. All runs used a two core CPU without a GPU. All random seeds were fixed (seed 2026 for the fold split and for every model), so the fold assignments and results can be reproduced exactly. The code, the fold assignments, the result tables and the run logs are deposited in a public repository (see Data and Code Availability).

3. Results

3.1 Cohort, Clinical Factors and Autoencoder Reconstruction

The cohort had 416 patients, of whom 242 (58.2%) were node positive. Each training fold held about 333 patients and each test fold about 83. Table 2 shows the clinical factors by node status. Node positive patients were younger (median age 59 against 62 years, $p = 0.002$) and more often had T3 or T4 tumours (63.7% against 49.0% node positive, $p = 0.005$). Oropharyngeal tumours (73.3% node positive) and HPV positive tumours (66.7%) were also more often node positive, but these differences were not significant in this cohort ($p = 0.17$ and $p = 0.27$).

Table 2. Clinical factors by node status (416 patients; p values from the chi square test over known categories)

Factor	Category	Node negative	Node positive	Share node positive	p value
T stage	T1 or T2	79	76	49.0%	0.005
	T3 or T4	94	165	63.7%	
	Unknown	1	1	50.0%	
Tumour site	Oral cavity	97	120	55.3%	0.17
	Oropharynx	12	33	73.3%	
	Larynx	40	56	58.3%	
	Other or unknown	25	33	56.9%	
HPV subtype	Negative	151	199	56.9%	0.27
	Positive	15	30	66.7%	
	Unknown	8	13	61.9%	

Table 3 gives the reconstruction error of the autoencoders. For every layer the test error was much larger than the training error, with the widest gaps for RNA-seq (0.028 against 0.466) and methylation (0.046 against 0.476). We return to this gap in Section 4.

Table 3. Autoencoder reconstruction error (mean squared error on scaled features, mean $\pm$ SD over five folds)

Data type	Input features	Code size	Training error	Test error
RNA-seq	5,000	320	0.028 $\pm$ 0.005	0.466 $\pm$ 0.027
DNA methylation	5,000	320	0.046 $\pm$ 0.014	0.476 $\pm$ 0.045
Copy number	5,000	320	0.045 $\pm$ 0.009	0.230 $\pm$ 0.024
RPPA	191	64	0.060 $\pm$ 0.009	0.356 $\pm$ 0.021

3.2 Overall Comparison

Figure 2 shows the pooled AUROC of all 76 models. Every model found some signal, but none found much. AUROC values ranged from 0.555 to 0.649, and most lay between 0.59 and 0.64. The RNA only LR reference reached 0.609 (95% CI 0.554 to 0.666). Of the 75 other models, 46 had a higher pooled AUROC than the reference, but only one differed significantly at $p < 0.05$ (latent fusion with RF on RNA-seq, $p = 0.048$). With 75 comparisons, about four such results would be expected by chance alone, so we do not treat this one as a real difference. The clinical only baseline reached a pooled AUROC of 0.579 (95% CI 0.527 to 0.636) and an AUPRC of 0.638, lower than the RNA only LR reference in pooled terms, although its mean fold AUROC (0.608 $\pm$ 0.051) was almost the same ($p = 0.97$).

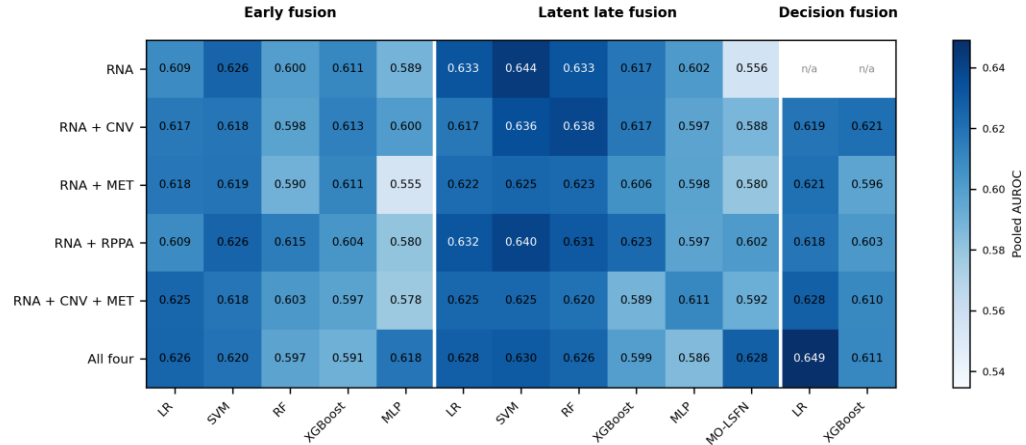


Figure 2. Pooled AUROC for every combination of data layers (rows), fusion strategy and model (columns). Decision fusion needs at least two layers, so it is not available for RNA alone (n/a). Darker cells mean higher AUROC.

Table 4 lists the main models. The best was decision fusion with LR on all four layers (AUROC 0.649, 95% CI 0.591 to 0.705; AUPRC 0.701; MCC 0.218). Latent fusion with SVM on RNA-seq alone came close (0.644), as did latent fusion with SVM on RNA + RPPA (0.640) and with RF on RNA + CNV (0.638). None of these was significantly better than the reference (p values from 0.08 to 0.36). The confidence intervals of all models in Table 4 overlapped widely. All observed differences from the reference (at most 0.040) were smaller than the smallest difference this design can detect (0.082 at 50% power; Section 2.9).

Table 4. Main models (pooled test predictions, 416 patients; Bal. acc., balanced accuracy; cut off 0.5 for balanced accuracy and MCC; the no skill AUPRC is 0.582; p value from the corrected t test against the RNA only LR reference)

Fusion	Layers	Model	AUROC (95% CI)	AUPRC	Bal. acc.	MCC	Fold AUROC, mean ± SD	p value
Early	RNA	LR	0.609 (0.554 to 0.666)	0.651	0.573	0.145	0.608 ± 0.030	Reference
None	Clinical only	LR	0.579 (0.527 to 0.636)	0.638	n/r	n/r	0.608 ± 0.051	0.97
Early	RNA	MLP	0.589 (0.534 to 0.643)	0.667	0.563	0.130	0.594 ± 0.078	0.80
Latent	RNA	SVM	0.644 (0.590 to 0.693)	0.700	0.623	0.246	0.657 ± 0.042	0.08
Latent	RNA	RF	0.633 (0.577 to 0.687)	0.697	0.590	0.194	0.635 ± 0.033	0.048
Latent	RNA + CNV	RF	0.638 (0.583 to 0.692)	0.688	0.588	0.206	0.649 ± 0.060	0.35
Latent	RNA + RPPA	SVM	0.640 (0.583 to 0.691)	0.695	0.634	0.270	0.652 ± 0.045	0.11
Early	All four	LR	0.626 (0.568 to 0.680)	0.683	0.575	0.148	0.633 ± 0.044	0.48
Early	All four	XGBoost	0.591 (0.537 to 0.645)	0.668	0.550	0.102	0.591 ± 0.052	0.64
Latent	All four	LR	0.628 (0.572 to 0.684)	0.687	0.580	0.158	0.629 ± 0.035	0.36
Latent	All four	SVM	0.630 (0.574 to 0.686)	0.687	0.595	0.188	0.639 ± 0.056	0.43
Latent	All four	MO-LSFN	0.628 (0.575 to 0.677)	0.664	0.608	0.217	0.626 ± 0.055	0.52
Decision	All four	LR	0.649 (0.591 to 0.705)	0.701	0.609	0.218	0.650 ± 0.049	0.20

The abbreviation n/r means not reported for the clinical only model, for which only ranking metrics were computed.

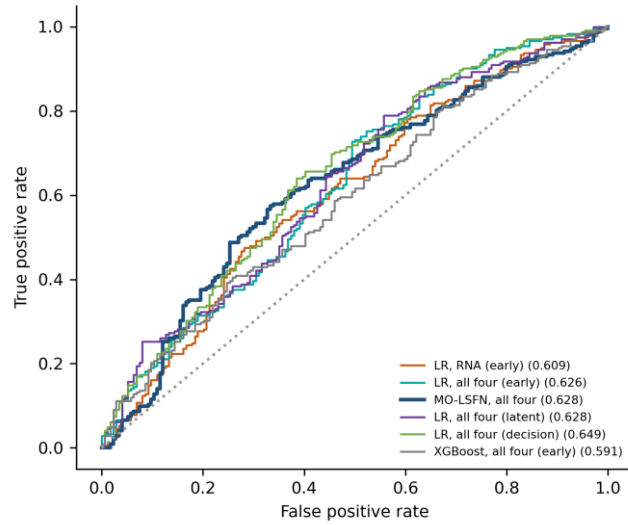


Figure. 3. Pooled ROC curves for selected models. Values in brackets are AUROC. The dotted line marks chance.

3.3 Effect of Data Layers, Fusion Strategy and Model Family

Table 5 summarises the results by data combination, fusion strategy and model family. Adding layers to RNA-seq made little difference. The mean AUROC over all models was 0.611 for RNA alone and between 0.605 and 0.616 for the five combinations with extra layers. The combination of all four layers had the highest mean (0.616) and the highest single value (0.649), but its advantage over RNA alone was 0.005 on average.

The fusion strategy mattered a little more. Over the five conventional model families and all six combinations, latent fusion gave a mean AUROC of 0.619 and early fusion 0.606. For the two models that we could test with all three strategies (LR and XGBoost) on the five combinations with more than one layer, the means were 0.611 for early, 0.616 for latent and 0.618 for decision fusion. Among model families, SVM (mean 0.627) and LR (0.623) were the most accurate and the most stable, followed by RF (0.614) and XGBoost (0.607). The MLP (0.593) and MO-LSFN (0.591) had the lowest means and the largest spread.

Table 5. Pooled AUROC summarised by data combination, fusion strategy and model family (mean $\pm$ SD over the models in each group, with the highest value and the number of models)

Group	Mean $\pm$ SD	Highest	Models
RNA	0.611 $\pm$ 0.025	0.644	11
RNA + CNV	0.614 $\pm$ 0.015	0.638	13
RNA + MET	0.605 $\pm$ 0.021	0.625	13
RNA + RPPA	0.614 $\pm$ 0.017	0.640	13
RNA + CNV + MET	0.609 $\pm$ 0.016	0.628	13
All four	0.616 $\pm$ 0.018	0.649	13
Early fusion, conventional models	0.606 $\pm$ 0.017	0.626	30
Latent fusion, conventional models	0.619 $\pm$ 0.016	0.644	30
Early fusion, LR and XGBoost, two or more layers	0.611 $\pm$ 0.011	0.626	10
Latent fusion, LR and XGBoost, two or more layers	0.616 $\pm$ 0.014	0.632	10
Decision fusion, LR and XGBoost, two or more layers	0.618 $\pm$ 0.014	0.649	10
LR	0.623 $\pm$ 0.010	0.649	17
SVM	0.627 $\pm$ 0.009	0.644	12
RF	0.614 $\pm$ 0.016	0.638	12
XGBoost	0.607 $\pm$ 0.010	0.623	17
MLP	0.593 $\pm$ 0.017	0.618	12
MO-LSFN	0.591 $\pm$ 0.024	0.628	6

The model counts in Table 5 follow from the design in Section 2.2. Each combination row counts five early fusion models and six latent fusion models, plus two decision fusion models for the five combinations with more than one layer (11 for RNA alone, 13 otherwise). Each fusion row counts one model per family and combination: five conventional families over six combinations gives 30, and LR and XGBoost over the five multi layer combinations gives 10. Each model family row counts its models over all combinations and strategies: LR and XGBoost appear in all three strategies (6 + 6 + 5 = 17), SVM, RF and the MLP appear in early and latent fusion (6 + 6 = 12), and MO-LSFN appears in latent fusion only (6).

3.4 MO-LSFN and Ablation

MO-LSFN behaved differently from the other models (Table 6, Figure 4a). On RNA-seq alone it was one of the weakest models in the study (AUROC 0.556). Its score rose as layers were added and reached 0.628 with all four layers, a gain of 0.072, which was the largest gain from extra data of any model family. With all four layers it matched LR with early fusion (0.626) and latent fusion (0.628), and it had one of the highest MCC values (0.217). It did not, however, outperform the RNA only LR reference by a significant margin or any of the strongest conventional models.

In the ablation study on all four layers (Figure 4b), every part of MO-LSFN helped. A random order on the map lowered the AUROC to 0.591, removing the code normalisation lowered it to 0.595, and removing the mask channel lowered it to 0.610. These differences were in the expected direction, but they were small compared with the width of the confidence intervals.

Table 6. MO-LSFN across data combinations and ablation models (pooled test predictions; Bal. acc., balanced accuracy; the no skill AUPRC is 0.582)

Model	AUROC (95% CI)	AUPRC	Bal. acc.	MCC
MO-LSFN, RNA	0.556 (0.493 to 0.613)	0.619	0.560	0.119
MO-LSFN, RNA + CNV	0.588 (0.532 to 0.640)	0.657	0.572	0.143
MO-LSFN, RNA + MET	0.580 (0.527 to 0.636)	0.664	0.547	0.093
MO-LSFN, RNA + RPPA	0.602 (0.543 to 0.654)	0.680	0.563	0.126
MO-LSFN, RNA + CNV + MET	0.592 (0.534 to 0.646)	0.642	0.554	0.111
MO-LSFN, all four	0.628 (0.575 to 0.677)	0.664	0.608	0.217
All four, random order on the map	0.591 (0.536 to 0.652)	0.630	0.575	0.150
All four, no code normalisation	0.595 (0.540 to 0.648)	0.642	0.532	0.063
All four, no mask channel	0.610 (0.551 to 0.665)	0.670	0.566	0.131

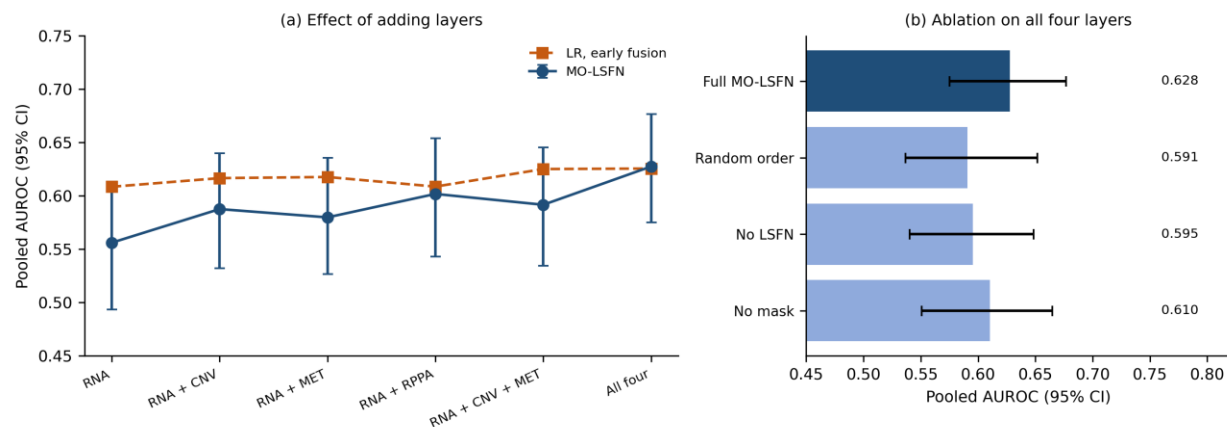


Figure 4. MO-LSFN results. (a) Pooled AUROC with 95% confidence intervals as layers are added, compared with the RNA only LR reference and LR with early fusion on each combination. (b) Ablation models on all four layers.

3.5 Sensitivity Analyses

Table 7 repeats the evaluation of seven key models in two subgroups, using the same out of fold predictions. In the 171 patients with all four layers, AUROC values ranged from 0.589 to 0.637 and stayed within the confidence intervals of the full cohort. The RNA only LR reference reached 0.623 and decision fusion with LR 0.637, so the main results were not driven by missing RPPA data. When the 66 N1 patients were excluded and N0 was compared with N2 or N3, every model scored higher, by 0.014 to 0.038. Decision fusion with LR reached 0.681 (95% CI 0.622 to 0.736) and the RNA only LR reference 0.638. The ranking of the models stayed about the same.

Table 7. Sensitivity analyses on the existing out of fold predictions (no retraining; AUROC with 95% bootstrap CI)

Model	All patients (n = 416, 242 N+)	All four layers (n = 171, 100 N+)	N0 vs N2 or N3 (n = 350, 176 N+)
RNA only LR reference	0.609 (0.554 to 0.666)	0.623 (0.536 to 0.702)	0.638 (0.581 to 0.696)
Early fusion, LR, RNA + RPPA	0.609 (0.553 to 0.665)	0.625 (0.542 to 0.705)	0.638 (0.581 to 0.697)
Latent fusion, SVM, RNA + RPPA	0.640 (0.583 to 0.691)	0.633 (0.546 to 0.712)	0.678 (0.624 to 0.732)
Early fusion, LR, all four	0.626 (0.568 to 0.680)	0.589 (0.503 to 0.675)	0.654 (0.593 to 0.713)
Latent fusion, LR, all four	0.628 (0.572 to 0.684)	0.594 (0.509 to 0.679)	0.659 (0.599 to 0.715)
MO-LSFN, all four	0.628 (0.575 to 0.677)	0.611 (0.524 to 0.699)	0.642 (0.582 to 0.697)
Decision fusion, LR, all four	0.649 (0.591 to 0.705)	0.637 (0.549 to 0.721)	0.681 (0.622 to 0.736)

3.6 Feature Attribution

For MO-LSFN on all four layers, methylation received the largest share of the total attribution (0.379 ± 0.049), followed by RNA-seq (0.348 ± 0.032), copy number (0.235 ± 0.049) and RPPA (0.037 ± 0.012). The small RPPA share partly reflects missing data, since RPPA was absent for 58% of patients and their RPPA scores are zero. Feature rankings were unstable between folds. The mean Jaccard similarity of the top 50 feature sets was 0.016 for RNA-seq, 0.028 for methylation, 0.055 for copy number and 0.188 for RPPA. The Spearman correlation over all shared features, which does not depend on a top 50 cut off, gave a fuller picture: 0.34 for RNA-seq, 0.36 for methylation, 0.41 for copy number and 0.14 for RPPA. The broad ordering of features was therefore partly preserved between folds, but the very top of each ranking was not. Table 8 lists the five top features of each layer. None of the 40 top features we tested differed significantly between node positive and node negative tumours after correction (all q values 0.21 or higher).

Table 8. Five top features per layer for MO-LSFN on all four layers (mean absolute Integrated Gradients score; Mann Whitney U test; Benjamini and Hochberg q value over the 40 features tested)

Layer	Feature	Mean absolute score ($\times 10^{-3}$)	Folds in top 50	Level in node positive tumours	q value
RNA-seq	PAX8	2.1	1	Lower	0.95
RNA-seq	TGM4	2.0	2	Higher	0.95
RNA-seq	CSAG2	2.0	2	Lower	0.48
RNA-seq	GLB1L3	1.9	1	Lower	0.34
RNA-seq	GSTT1	1.9	2	Higher	0.63
Methylation	BTG4, MIR34B, MIR34C	2.6	2	Lower	0.35
Methylation	ZNF57	2.5	3	Lower	0.34
Methylation	SRRT	2.4	2	Lower	0.21
Methylation	SPEF2	2.3	2	Lower	0.34
Methylation	DKK1	2.3	1	Lower	0.34
Copy number	PARD3	2.2	4	Lower	0.34
Copy number	CCNY	2.2	3	Lower	0.48
Copy number	SPIDR	2.1	1	Lower	0.95
Copy number	PARD3 AS1	2.1	3	Lower	0.48
Copy number	CEBPD	2.0	1	Lower	0.95
RPPA	p38 MAPK (pT180, pY182)	4.1	3	Lower	0.38
RPPA	Acetyl alpha tubulin (Lys40)	4.0	5	Higher	0.48
RPPA	A-Raf (pS299)	3.8	3	Lower	0.95
RPPA	HER3	3.7	3	Lower	0.34
RPPA	4E-BP1	3.6	3	Lower	0.34

4. Discussion

This benchmark gives a negative result. The molecular profile of the primary tumour carried only a weak signal for lymph node spread in TCGA-HNSC, and neither extra omics layers, nor the fusion strategy, nor our deep model changed this much. All 76 models scored between 0.555 and 0.649. The best model, decision fusion with LR on all four layers, gained 0.040 over the RNA only LR reference, but this gain was not significant, and averaged over all models, adding layers changed the AUROC by less than 0.01. MO-LSFN, the most complex model, did not outperform the simpler baselines. We think this result is useful. A study that tested one fusion design against one baseline could easily have reported a gain of this size as an improvement, and our grid shows how much such gains vary by chance.

The power analysis shows the limits of what this study can say. With a single run of fivefold cross validation and about 83 test patients per fold, only differences of about 0.08 to 0.11 in AUROC could be detected. All the differences we observed were smaller. Our data therefore do not show that fusion has no benefit; they show that any benefit is smaller than about 0.08 in this cohort. Repeated cross validation would narrow the confidence intervals of each model, but it would not raise the information in the 416 patients, and an external cohort would be more valuable.

Several factors could explain the weak signal. Nodal spread depends on features that a bulk tumour profile may not capture well, such as the invasive front, lymphatic invasion and the immune state of the draining nodes [2], [4]. Label

noise is a second factor. Pathological N stage depends on how many nodes are removed and examined, and the boundary between N0 and N1 is the most likely place for errors, for example when a small deposit is missed. When we removed N1 patients, every model improved by 0.014 to 0.038, which fits this explanation, although the gain also reflects the larger biological contrast between N0 and N2 or N3 disease. Clinical confounding is a third factor. T stage was strongly linked to node status in our cohort, and oropharyngeal and HPV positive tumours were more often node positive. A model built on clinical factors alone, however, reached a pooled AUROC of only 0.579, so the weak molecular signal is not simply a weaker copy of these clinical factors, and clinical factors were not strong predictors either.

The comparison of fusion strategies gives a practical message. Latent fusion was slightly better than early fusion for the conventional models, and decision fusion was about as good as latent fusion. Both approaches keep each layer separate for longer, which stops the largest layer from dominating a single long feature vector, in line with reviews that favour intermediate or late integration for heterogeneous omics data [8], [9]. The differences were small, though, and reporting a full grid of strategies and models, as in Figure 2, is a simple way to avoid reading too much into one combination. Early fusion also serves as the simple concatenation baseline that is often used to judge published fusion frameworks. We did not run MOGONET [14] or DeepProg [19]. MOGONET builds patient similarity graphs and was designed for larger multi class cohorts, and DeepProg targets survival rather than classification. Adding them is a natural next step, but given that the strongest models in our grid were regularised linear and kernel models, we would not expect a more complex framework to change the main conclusion at this sample size.

The autoencoders showed a large gap between training and test reconstruction error, for example 0.028 against 0.466 for RNA-seq. With about 333 training patients and 5,000 inputs, an autoencoder with a 1,024 unit hidden layer can memorise much of the training data. The codes still carried useful information for classification: latent fusion with conventional models matched or beat early fusion on the same data (mean AUROC 0.619 against 0.606), and LR, SVM and RF on the RNA-seq codes all scored above the reference. Good reconstruction and good discrimination are different goals. Even so, the gap means that the codes are not an efficient summary of the data. Stronger regularisation, smaller codes, denoising objectives or a linear method such as principal component analysis may give codes that generalise better, and comparing them is an obvious extension of this work.

MO-LSFN showed a different pattern from the other models. On RNA-seq alone it was one of the weakest models, but it improved steadily as layers were added and matched the best conventional latent models on all four layers. Its ablation results also pointed in the expected direction: the correlation order, the code normalisation and the mask channel each added to its score. This suggests that the design does what it was built to do, which is to combine several incomplete layers in one model. Its capacity, however, is too large for this cohort, and its fixed design choices (grid size, path, depth, width and code size) were not tuned. Other authors have also found that deep models rarely beat simpler models on small omics cohorts [8]. We therefore present MO-LSFN as one entry in the benchmark and not as a recommended model.

The attribution results support caution. Methylation and RNA-seq carried most of the attribution, but the top features changed from fold to fold and none of them differed significantly between node positive and node negative tumours. The Spearman correlations of 0.34 to 0.41 for the three large layers show that the broad ordering of features was partly stable, but the specific top features were not. A model with an AUROC near 0.63 cannot be expected to produce stable gene rankings, and we do not propose any of these features as biomarkers. Comparing the attributions with curated metastasis gene sets, and between HPV positive and HPV negative tumours, would be a useful next step in a larger cohort.

4.1 Limitations

This study has several limits. All data came from one public cohort, TCGA-HNSC, which is mostly North American, and we did not test the models on an external cohort. Pathological N stage depends on the extent of neck dissection and on pathology practice, which TCGA does not record, so the label carries some noise that we could only probe indirectly by excluding N1 patients. We did not adjust the molecular models for clinical factors; we reported their distribution, a clinical only baseline and their association with node status instead. Somatic mutation data were not included. We ran fivefold cross validation once, and the power analysis shows that differences below about 0.08 in AUROC cannot be detected. The settings of MO-LSFN, the autoencoders and the conventional models were fixed in advance and not tuned, and we did not compare other grid sizes, map paths, network sizes, code sizes, principal component analysis or published fusion frameworks such as MOGONET. RPPA was available for only 176 patients. Finally, the attribution results are computational and have not been tested in the laboratory.

5. Conclusions

This benchmark compared six combinations of RNA-seq with methylation, copy number and protein data, three fusion strategies and six model families, including our deep model MO-LSFN, for predicting lymph node status in 416 TCGA head and neck tumours. All models were tested with the same folds and with every step fitted on training patients only. The result is negative. Prediction was difficult for every model, with AUROC values between 0.555 and 0.649, and no model, fusion strategy or data combination clearly beat logistic regression on RNA-seq alone. Latent and decision fusion were slightly better than early fusion, SVM and LR were the most reliable models, and MO-LSFN gained the most from extra layers but did not beat simpler models. Gains smaller than about 0.08 cannot be ruled out with this design. Future studies should use larger and independent cohorts, including patients from South Asia, repeated cross validation, clinical and somatic mutation data, and tuned or alternative compression methods before claiming that multiomics fusion improves nodal prediction.

6. Declarations

Data and code availability: All data are public and can be downloaded from the cBioPortal data hub, study hnscc_tcga_pan_can_atlas_2018 (TCGA PanCancer Atlas).

Relationship to prior work: This study builds on our earlier single data type framework, AIW-DynOpt [11]. All experiments, results, tables and figures in this paper are new. The authors also use the public TCGA-HNSC cohort in other studies that address different clinical questions; no results are shared between those studies and this one. This paper is not under review at any other journal.

Conflict of interest: The authors declare no conflict of interest.

Funding: This research received no external funding.

Author contributions: Aneela Nargis: conceptualisation, methodology, software, formal analysis and writing of the first draft. Raazia Sosa Waseem: validation, review and editing. Both authors approved the final version.

7. Acknowledgements

The results shown here are based on data generated by the TCGA Research Network (<https://www.cancer.gov/tcga>).

8. References

1. F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J. Clin.*, vol. 74, no. 3, pp. 229-263, 2024, doi: 10.3322/caac.21834.
2. D. E. Johnson, B. Burtness, C. R. Leemans, V. W. Y. Lui, J. E. Bauman, and J. R. Grandis, "Head and neck squamous cell carcinoma," *Nat. Rev. Dis. Primers*, vol. 6, no. 1, p. 92, 2020, doi: 10.1038/s41572-020-00224-3.
3. U. Shamsi, "Cancer prevention and control in Pakistan: Review of cancer epidemiology and challenges," *Liaquat Natl. J. Prim. Care*, vol. 2, no. 1, pp. 34-38, 2020, doi: 10.37184/lnjpc.2707-3521.1.20.
4. C. R. Leemans, P. J. F. Snijders, and R. H. Brakenhoff, "The molecular landscape of head and neck cancer," *Nat. Rev. Cancer*, vol. 18, no. 5, pp. 269-282, 2018, doi: 10.1038/nrc.2018.11.
5. The Cancer Genome Atlas Network, "Comprehensive genomic characterization of head and neck squamous cell carcinomas," *Nature*, vol. 517, no. 7536, pp. 576-582, 2015, doi: 10.1038/nature14129.
6. S. Schmitz and J. P. Machiels, "Molecular biology of squamous cell carcinoma of the head and neck: Relevance and therapeutic implications," *Expert Rev. Anticancer Ther.*, vol. 10, no. 9, pp. 1471-1484, 2010, doi: 10.1586/era.10.128.
7. I. M. Johnstone and D. M. Titterton, "Statistical challenges of high-dimensional data," *Philos. Trans. R. Soc. A*, vol. 367, no. 1906, pp. 4237-4253, 2009, doi: 10.1098/rsta.2009.0159.
8. M. Picard, M.-P. Scott-Boyer, A. Bodein, O. Périn, and A. Droit, "Integration strategies of multi-omics data for machine learning analysis," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 3735-3746, 2021, doi: 10.1016/j.csbj.2021.06.030.

9. Y. J. Heo, C. Hwa, G.-H. Lee, J.-M. Park, and J.-Y. An, "Integrative multi-omics approaches in cancer research: From biological networks to clinical subtypes," *Mol. Cells*, vol. 44, no. 7, pp. 433-443, 2021, doi: 10.14348/molcells.2021.0042.
10. X. Wang and B. B. Li, "Deep learning in head and neck tumor multiomics diagnosis and analysis: Review of the literature," *Front. Genet.*, vol. 12, p. 624820, 2021, doi: 10.3389/fgene.2021.624820.
11. A. Nargis, M. M. Movania, and S. Siddiqui, "Autoencoder-integrated WideResNet with dynamic optimization (AIW-DynOpt): A novel hybrid deep learning approach for head and neck cancer gene expression analysis," *J. Univers. Comput. Sci.*, vol. 31, no. 2, pp. 159-188, 2025, doi: 10.3897/jucs.125224.
12. K. Chaudhary, O. B. Poirion, L. Lu, and L. X. Garmire, "Deep learning-based multi-omics integration robustly predicts survival in liver cancer," *Clin. Cancer Res.*, vol. 24, no. 6, pp. 1248-1259, 2018, doi: 10.1158/1078-0432.CCR-17-0853.
13. L. Tong, J. Mitchel, K. Chatlin, and M. D. Wang, "Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis," *BMC Med. Inform. Decis. Mak.*, vol. 20, p. 225, 2020, doi: 10.1186/s12911-020-01225-8.
14. T. Wang, W. Shao, Z. Huang, H. Tang, J. Zhang, Z. Ding, and K. Huang, "MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification," *Nat. Commun.*, vol. 12, p. 3445, 2021, doi: 10.1038/s41467-021-23774-w.
15. G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504-507, 2006, doi: 10.1126/science.1127647.
16. D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Banff, Canada, 2014.
17. G. P. Way and C. S. Greene, "Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders," in *Proc. Pac. Symp. Biocomput.*, vol. 23, 2018, pp. 80-91.
18. E. F. Franco, P. Rana, A. Cruz, V. V. Calderón, V. Azevedo, R. T. J. Ramos, and P. Ghosh, "Performance comparison of deep learning autoencoders for cancer subtype detection using multi-omics data," *Cancers*, vol. 13, no. 9, p. 2013, 2021, doi: 10.3390/cancers13092013.
19. O. B. Poirion, Z. Jing, K. Chaudhary, S. Huang, and L. X. Garmire, "DeepProg: An ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data," *Genome Med.*, vol. 13, p. 112, 2021, doi: 10.1186/s13073-021-00930-x.
20. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.
21. B. Lyu and A. Haque, "Deep learning based tumor type classification using gene expression data," in *Proc. ACM Int. Conf. Bioinformatics, Comput. Biol., Health Informatics (BCB)*, 2018, pp. 89-96, doi: 10.1145/3233547.3233588.
22. B. N. Darendeli and A. Yilmaz, "Convolutional neural network approach to predict tumor samples using gene expression data," *J. Intell. Syst. Theory Appl.*, vol. 4, no. 2, pp. 136-141, 2021.
23. A. Sharma, E. Vans, D. Shigemizu, K. A. Boroevich, and T. Tsunoda, "DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture," *Sci. Rep.*, vol. 9, p. 11399, 2019, doi: 10.1038/s41598-019-47765-6.
24. Y. Zhu et al., "Converting tabular data into images for deep learning with convolutional neural networks," *Sci. Rep.*, vol. 11, p. 11325, 2021, doi: 10.1038/s41598-021-90923-y.
25. S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2016, pp. 87.1-87.12, doi: 10.5244/C.30.87.
26. G. Ju, Z. Yao, Y. Zhao, X. Zhao, and F. Liu, "Data mining on identifying diagnosis and prognosis biomarkers in head and neck squamous carcinoma," *Sci. Rep.*, vol. 13, p. 10020, 2023, doi: 10.1038/s41598-023-37216-8.
27. Y. Li, J. Bi, G. Pi, H. He, Y. Li, and G. Han, "Exploration of prognostic biomarkers in head and neck squamous cell carcinoma microenvironment from TCGA database," *Ann. Transl. Med.*, vol. 11, no. 4, p. 163, 2023, doi: 10.21037/atm-22-6481.
28. Y. O. Nunez Lopez, B. Victoria, P. Golusinski, W. Golusinski, and M. M. Masternak, "Characteristic miRNA expression signature and random forest survival analysis identify potential cancer-driving miRNAs in a broad range of head and neck squamous cell carcinoma subtypes," *Rep. Pract. Oncol. Radiother.*, vol. 23, no. 1, pp. 6-20, 2018, doi: 10.1016/j.rpor.2017.12.004.

29. C. Weusthof, S. Burkart, K. Semmelmayr, F. Stögbauer, B. Feng, K. Khorani, et al., "Establishment of a machine learning model for the risk assessment of perineural invasion in head and neck squamous cell carcinoma," *Int. J. Mol. Sci.*, vol. 24, no. 10, p. 8938, 2023, doi: 10.3390/ijms24108938.
30. Q. Wang, Y. Zhao, F. Wang, and G. Tan, "Clustering and machine learning-based integration identify cancer associated fibroblasts genes' signature in head and neck squamous cell carcinoma," *Front. Genet.*, vol. 14, p. 1111816, 2023, doi: 10.3389/fgene.2023.1111816.
31. K. A. Hoadley et al., "Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer," *Cell*, vol. 173, no. 2, pp. 291-304, 2018, doi: 10.1016/j.cell.2018.03.022.
32. E. Cerami et al., "The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data," *Cancer Discov.*, vol. 2, no. 5, pp. 401-404, 2012, doi: 10.1158/2159-8290.CD-12-0095.
33. J. Gao et al., "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal," *Sci. Signal.*, vol. 6, no. 269, p. p11, 2013, doi: 10.1126/scisignal.2004088.
34. C. H. Mermel, S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhi, and G. Getz, "GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers," *Genome Biol.*, vol. 12, no. 4, p. R41, 2011, doi: 10.1186/gb-2011-12-4-r41.
35. R. Akbani et al., "A pan-cancer proteomic perspective on The Cancer Genome Atlas," *Nat. Commun.*, vol. 5, p. 3887, 2014, doi: 10.1038/ncomms4887.
36. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321-357, 2002, doi: 10.1613/jair.953.
37. S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," *ACM Trans. Knowl. Discov. Data*, vol. 6, no. 4, p. 15, 2012, doi: 10.1145/2382577.2382579.
38. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825-2830, 2011.
39. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD)*, 2016, pp. 785-794, doi: 10.1145/2939672.2939785.
40. C. Nadeau and Y. Bengio, "Inference for the generalization error," *Mach. Learn.*, vol. 52, no. 3, pp. 239-281, 2003, doi: 10.1023/A:1024068626366.
41. M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 70, 2017, pp. 3319-3328.
42. Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. R. Stat. Soc. Ser. B*, vol. 57, no. 1, pp. 289-300, 1995, doi: 10.1111/j.2517-6161.1995.tb02031.x.
43. A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019, pp. 8024-8035.