Benchmarking ensemble machine learning algorithms for multi-class, multi-omics data integration in clinical outcome prediction

A Comparative Study of Late Integration Methods

尹超

2025.08.19

中国科学院大学

Table of Contents

Research Overview

- Problem: Multi-omics data integration for clinical outcome prediction
- Challenge: High dimensionality, heterogeneous data, small sample sizes
- Solution: Ensemble machine learning with late integration strategies
- Application: Hepatocellular carcinoma, breast cancer, IBD

Motivation

- Multi-omics data provides complementary information
- Better understanding of disease mechanisms
- Improved clinical outcome prediction
- Challenges in multi-omics integration:
  - Curse of dimensionality
  - Heterogeneous data types
  - Missing values and batch effects

Benefits

- Enhanced accuracy
- Novel biomarker discovery
- Disease subtyping
- Therapeutic targets

Table of Contents

2 Background

Data Integration Strategies

2 Background

- Early Integration (Feature-level fusion)
  - Concatenate all omics data
  - Train single classifier
  - Problem: Exacerbates high dimensionality
- Intermediate Integration
  - Transform to common representation
  - Difficult clinical interpretation
- Late Integration (Decision-level fusion)
  - Train separate models on each modality
  - Aggregate results for final prediction
  - **Focus of this work**

Advantages of Late Integration

- Reduces dimensionality - separate models per modality

Advantages of Late Integration

- Reduces dimensionality - separate models per modality
- Flexible - different ML models for each modality

Advantages of Late Integration

- Reduces dimensionality - separate models per modality
- Flexible - different ML models for each modality
- Addresses heterogeneity - tailored preprocessing

Advantages of Late Integration

- Reduces dimensionality - separate models per modality
- Flexible - different ML models for each modality
- Addresses heterogeneity - tailored preprocessing
- Reduces overfitting - smaller feature spaces

Advantages of Late Integration

- Reduces dimensionality - separate models per modality
- Flexible - different ML models for each modality
- Addresses heterogeneity - tailored preprocessing
- Reduces overfitting - smaller feature spaces
- Lower computational complexity

Advantages of Late Integration

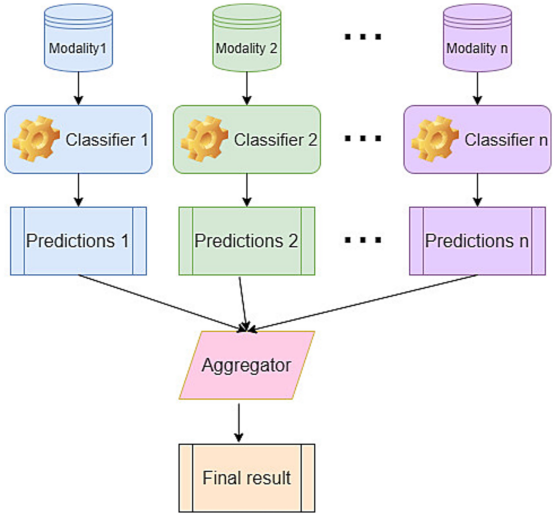
- Reduces dimensionality - separate models per modality
- Flexible - different ML models for each modality
- Addresses heterogeneity - tailored preprocessing
- Reduces overfitting - smaller feature spaces
- Lower computational complexity
- Modality-specific optimization

Table of Contents

3 Methods

Ensemble Methods Overview

3 Methods



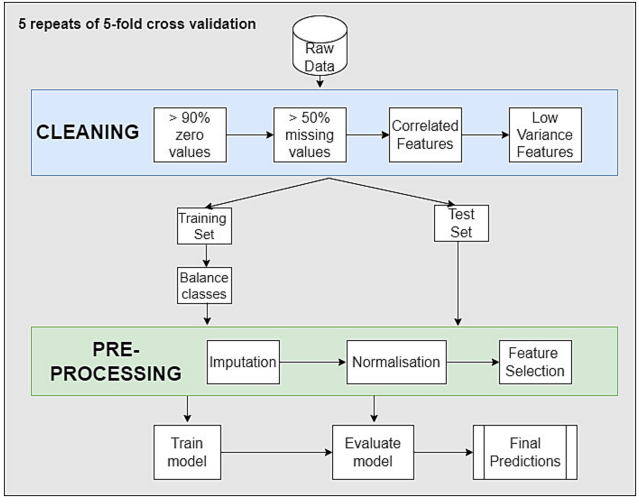
Ensemble Methods Evaluated

3 Methods

1. Voting Ensemble
   - Hard vote (majority)
   - Soft vote (probability averaging)
2. Meta Learner
   - Random forest as meta-classifier
3. Multi-modal AdaBoost
   - Hard vote, soft vote, meta learner variants
4. PB-MVBoost
   - Balances accuracy and diversity
5. Mixture of Experts
   - Novel application with gating function

Data Processing Pipeline

Comprehensive preprocessing approach



Data Processing Pipeline

Comprehensive preprocessing approach

Filtering Steps

- Remove features with >50% missing values
- Remove features with >90% zero values
- Eliminate correlated features
- Retain top 500 highest variance features if needed

Preprocessing Steps

- Balancing: SMOTE for imbalanced datasets
- Imputation: MICE or k-NN for missing values
- Normalization: CPM + log transformation for RNA/DNA
- Feature Selection: Boruta algorithm with GBM

Datasets

3 Methods

Table 1. Summary of the characteristics of the datasets used in the study, showing patient categories and number of samples in each, plus modalities and the number of features in each. The order of the number of samples and number of features is consistent with the list of patient categories and modalities, respectively

DB name	Reference	Patient categories (abbreviation)	No. samples	Omics modalities (abbreviation)	No. features[a]
HCC-Genus	*Private unpublished data*	Healthy Controls (CON)	28	Clinical (CLIN),	14
HCC-Species		MAFLD-cirrhosis (CIR)	28	Cytokine (CYT),	28
		MAFLD-related HCC (LN)	25	Pathology results (PATH), Metabolomic	48
		Viral HCC (LX)	25	(MET)	1046
				Lipoprotein (LIP),	112
				Oral Microbiome-Genus (OG),	243
				Oral Microbiome-Species (OS), Stool	583
				Microbiome-Genus (SG), and Stool	282
				Microbiome-Species (SS)	721
IBD-1	Mehta *et al.* (2023) [24]	Crohn's disease (CD)	50	Metagenomics (MTG)	934
	Nature medicine	Ulcerative Colitis (UC)	28	Metabolomics (MTB)	81,496
	https://doi.org/10.1038/	Non-IBD (non-IBD)	20	Metatranscriptomis (MTX)	83,227
	s41591-023-02217-7			Viromics (VIR)	262
IBD-2	Franzosa *et al.* (2019) [25]	Crohn's disease (CD)	88	Clinical (CLIN)	8
	Nature Microbiology	Ulcerative Colitis (UC)	76	Metabolites (METAB)	8850
	https://doi.org/10.1038/	Control	56	Microbiome (MICROB)	204
	s41564-018-0306-4				
Breast-1	Sammut *et al.* (2022) [13]	RCB-1	24	Clinicopathological (CLIN)	24
	Nature	RCB-II	59	Digital pathology (PATH)	8
	https://doi.org/10.1038/	RCB-III	27	RNA sequencing (RNA)	57,903
	s41586-021-04278-5	pCR	40	DNA sequencing (DNA)	31
Breast-2	Krug *et al.* (2020) [26]	Basal-like (Basal)	29	Clinical (CLIN)	28
	Cell	HER2-enriched (Her2)	14	mRNA (MRNA)	23,123
	https://doi.org/10.1016/j.	Luminal A (LumA)	57	Proteome (PROT)	9932
	cell.2020.10.036	Luminal B (LumB)	17		
		Normal-like (Normal)	5		

[a]The final coliumn (No. features) shows the total number of features, followed in brackets by the number of relevant features identified and used in the modelling.

Voting Ensemble

3 Methods

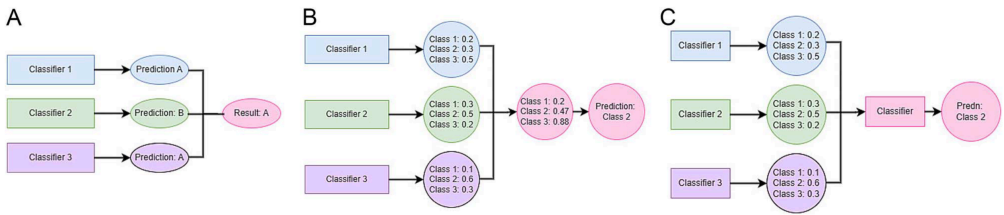


Figure 3. Techniques used to aggregate the results of the classifiers applied to each modality. (A) The hard vote is a simple majority vote. (B) The soft vote is an average of the probability scores. (C) The meta-learner is a classifier that learns from the results of the base-level classifiers.

Table of Contents

4 Results

Performance Results

4 Results

- Best performers: PB-MVBoost and AdaBoost with soft vote
- Peak performance: AUC up to 0.85 (HCC dataset)
- Multi-modal methods outperformed individual modalities in most cases
- Soft vote consistently better than hard vote

Key Finding

Boosting methods excel due to:

- Reduced overfitting
- Modality weighting
- Suited for high-dimensional, small-sample data

Evaluation Metrics

4 Results

Table 3. Evaluation metrics calculated on all models

Metric	Description	Formula
Accuracy (acc)	The ratio of correct predictions to the total number of samples.	$\frac{tp+tn}{tp+fp+tn+fn}$
Sensitivity (sens)	The fraction of positive samples that are correctly classified	$\frac{tp}{tp+fn}$
Specificity (spec)	The fraction of negative samples that are correctly classified	$\frac{tn}{tn+fp}$
Precision (p)	The fraction of samples identified as positive that were correctly classified	$\frac{tp}{tp+fp}$
Recall (r)	As for sensitivity	$\frac{tp}{tp+fn}$
F1 measure (f1)	The harmonic mean between precision and recall	$\frac{2*p*r}{p+r}$
AUC (auc)	Area under the Receiver Operating Curve: a measure of the classifier's ability to distinguish between classes	

Individual vs. Multi-modal Performance

4 Results

Table 4. Comparison of the best-performing multi-modal method with best-performing single modality for each dataset

Dataset		Modality/Method	AUROC	F1	Acc	Sens	Spec
HCC-Genus	Ind	CYT	0.77	0.64 (±0.15)	0.83 (±0.07)	0.65 (±0.14)	0.88 (±0.05)
	MM	PB-MVBoost	0.85	0.77 (±0.11)	0.89 (±0.06)	0.77 (±0.15)	0.93 (±0.06)
HCC-Species	Ind	CYT	0.77	0.64 (±0.15)	0.83 (±0.07)	0.65 (±0.14)	0.88 (±0.05)
	MM	PB-MVBoost	0.84	0.75 (±0.13)	0.88 (±0.06)	0.76 (±0.16)	0.92 (±0.06)
IBD1	Ind	MTB	0.56	0.39 (±0.17)	0.65 (±0.11)	0.41 (±0.22)	0.7 (±0.25)
	MM	Concatenation	0.61	0.46 (±0.19)	0.69 (±0.08)	0.48 (±0.2)	0.74 (±0.17)
IBD2	Ind	METAB	0.76	0.68 (±0.06)	0.79 (±0.07)	0.69 (±0.07)	0.83 (±0.07)
	MM	AdaBoost-Soft	0.8	0.74 (±0.05)	0.82 (±0.06)	0.74 (±0.05)	0.86 (±0.06)
		PB-MVBoost	0.8	0.73 (±0.07)	0.82 (±0.06)	0.73 (±0.09)	0.86 (±0.06)
Breast1	Ind	CLIN	0.8	0.65 (±0.2)	0.85 (±0.06)	0.71 (±0.18)	0.9 (±0.08)
	MM	Meta leaner	0.82	0.71 (±0.22)	0.89 (±0.05)	0.73 (±0.25)	0.92 (±0.04)
Breast2	Ind	PROT	0.74	0.57 (±0.36)	0.91 (±0.04)	0.58 (±0.38)	0.93 (±0.07)
	MM	Concatenation	0.74	0.58 (±0.37)	0.92 (±0.06)	0.59 (±0.4)	0.93 (±0.1)

Ind, individual modality. MM, multi-modal method.

Multi-modal integration provides consistent improvements

Feature Selection Stability

4 Results

- Concatenation: Least stable feature selection
- PB-MVBoost: Highest stability and accuracy
- AdaBoost (soft vote): Good balance of stability and signature length
- Integration methods overcome high-dimensional instability

Clinical Signature Characteristics

Desirable properties:

- Shorter signature length
- Fewer modalities required
- High stability and accuracy
- Clinical interpretability

Optimal Modality Subset

4 Results

Table 5. Performance of the incremental model in determining the best subset of modalities in each dataset, showing the degree to which performance improved as each modality was removed and the order in which the modalities were removed

Dataset	Best subset	Modality removed	F1 score after removal
HCC	CLIN, CYT, METAB	None	0.68
		OralSpecies	0.70
		OralGenus	0.71
		StoolSpecies	0.72
		StoolGenus	0.72
		Pathologic	0.73
IBD1	MTB, MTG	None	0.42
		VIR	0.41
		MTX	0.41
IBD2	METAB	None	0.67
		CLIN	0.69
		MICROB	0.69
Breast1	CLIN	None	0.6
		DNA	0.22
		RNA	0.25
		PATH	0.25
Breast2	PROT	None	0.44
		CLIN	0.57
		MRNA	0.58

Optimal Modality Subset

4 Results

Table 6. Results showing the performance of each data integration method on the optimal subset of modalities in the HCC dataset

Dataset	Method	All modalities		Optimal subset	
		AUC	F1	AUC	F1
HCC	Concatenation + RF	0.8	0.7 (±0.16)	0.69	0.53 (±0.14)
	Voting: hard vote	0.8	0.69 (±0.17)	0.81	0.68 (±0.13)
	Voting: soft vote	0.81	0.7 (±0.17)	0.79	0.7 (±0.14)
	Meta learner	0.77	0.65 (±0.19)	0.77	0.65 (±0.17)
	AdaBoost: hard vote	0.69	0.48 (±0.28)	0.71	0.5 (±0.28)
	AdaBoost: soft vote	0.84	0.76 (±0.15)	0.85	0.76 (±0.11)
	AdaBoost: meta learner	0.83	0.73 (±0.14)	0.81	0.71 (±0.13)
	Mixture of experts: soft vote	0.71	0.51 (±0.28)	0.74	0.6 (±0.19)
	PB-MVBoost	0.85	0.77 (±0.11)	0.85	0.77 (±0.1)
IBD1	Concatenation + RF	0.61	0.46 (±0.19)	0.62	0.48 (±0.13)
	Voting: hard vote	0.56	0.39 (±0.25)	0.55	0.37 (±0.24)
	Voting: soft vote	0.57	0.42 (±0.23)	0.59	0.45 (±0.17)
	Meta learner	0.5	0.3 (±0.27)	0.54	0.37 (±0.22)
	AdaBoost: hard vote	0.54	0.33 (±0.27)	0.51	0.18 (±0.2)
	AdaBoost: soft vote	0.56	0.4 (±0.22)	0.6	0.45 (±0.19)
	AdaBoost: meta learner	0.53	0.31 (±0.26)	0.51	0.08 (±0.1)
	Mixture of experts: soft vote	0.55	0.36 (±0.22)	0.58	0.37 (±0.16)
	PB-MVBoost	0.57	0.39 (±0.22)	0.56	0.39 (±0.19)

Optimal Modality Subset

4 Results

- Incremental model: Determines optimal modality subset
- Some datasets achieve equal performance with fewer modalities
- Clinical benefit: Fewer tests required
- Reduces cost and complexity

Example: HCC Dataset

- Full set: 7 modalities
- Optimal subset: 4 modalities
- Performance maintained with 43% fewer tests

Table of Contents

Key Findings

5 Discussion

- Boosting methods superior: PB-MVBoost and AdaBoost excel

Key Findings

5 Discussion

- Boosting methods superior: PB-MVBoost and AdaBoost excel
- Soft voting better: More nuanced than hard voting

Key Findings

5 Discussion

- Boosting methods superior: PB-MVBoost and AdaBoost excel
- Soft voting better: More nuanced than hard voting
- Multi-modal benefit: Complementary information utilization

Key Findings

5 Discussion

- Boosting methods superior: PB-MVBoost and AdaBoost excel
- Soft voting better: More nuanced than hard voting
- Multi-modal benefit: Complementary information utilization
- Stability matters: Feature selection consistency crucial

Key Findings

5 Discussion

- Boosting methods superior: PB-MVBoost and AdaBoost excel
- Soft voting better: More nuanced than hard voting
- Multi-modal benefit: Complementary information utilization
- Stability matters: Feature selection consistency crucial
- Modality optimization: Not all modalities equally valuable

Why Boosting Works

5 Discussion

- Reduces overfitting
  - Sequential weak learners
  - Focus on difficult samples
- Modality weighting
  - Prioritizes predictive modalities
  - Learned importance scores
- High-dimensional suitability
  - Designed for challenging datasets
  - Small sample, large feature spaces

Contrast

Other methods give equal weight to all modalities, potentially diluting predictive power

Clinical Implications

5 Discussion

- Personalized medicine: Better patient stratification
- Biomarker discovery: Stable feature signatures
- Cost reduction: Optimal modality subsets
- Diagnostic efficiency: Fewer required tests
- Mechanistic insights: Multi-modal feature importance

Future Applications

- Drug response prediction
- Disease progression monitoring
- Treatment selection guidance

Table of Contents

Best Practices for Multi-Modal Integration

6 Conclusion

1. Examine individual modalities first
   - Identify most predictive modalities
2. Apply incremental method
   - Determine optimal modality subset
3. Use boosting methods
   - PB-MVBoost or AdaBoost with soft vote
4. Consider stability
   - Balance accuracy and feature consistency
5. Validate across datasets
   - Ensure generalizability

Future Work

6 Conclusion

- Deep learning integration
  - Neural network ensemble methods
- Longitudinal data
  - Temporal multi-modal analysis
- Federated learning
  - Multi-center collaborative models
- Interpretability enhancement
  - Explainable AI for clinical decision support
- Real-world validation
  - Prospective clinical studies

Thank you for listening!