Briefings in Bioinformatics, 2025, **26(2)**, bbaf116
https://doi.org/10.1093/bib/bbaf116
Review

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

Annette Spooner[1,*], Mohammad Karimi Moridani[2], Barbra Toplis[3], Jason Behary[3,4], Azadeh Safarchi[5], Salim Maher[3,4], Fatemeh Vafaee[2,6], Amany Zekry[3,4], Arcot Sowmya[1]

[1]School of Computer Science and Engineering, University of New South Wales, High St, Kensington, NSW 2052, Australia
[2]School of Biotechnology and Biomolecular Sciences, University of New South Wales, NSW 2052, Australia
[3]St George and Sutherland Clinical Campuses, University of New South Wales, Short St, Kogarah, NSW 2217, Australia
[4]Department of Gastroenterology and Hepatology, St George Hospital, Gray St, Kogarah, NSW 2217, Australia
[5]Health and Biosecurity, Microbiome for One System Health, Commonwealth Scientific and Industrial Research Organisation, 160 Hawkesbury Rd, Westmead, NSW 2145, Australia
[6]UNSW Data Science Hub, University of New South Wales, High St, Kensington, NSW 2052, Australia

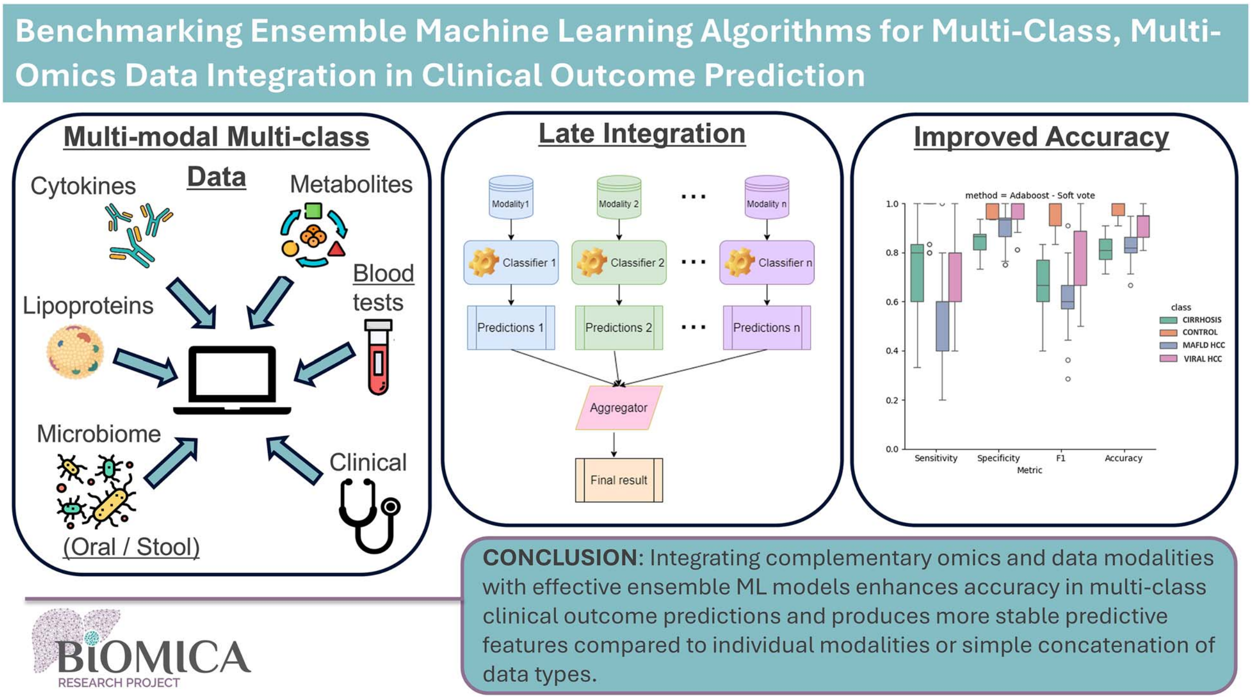
[*]Corresponding author. School of Computer Science and Engineering, University of NSW, Anzac Parade, Kensington, NSW 2052, Australia.
E-mail: a.spooner@unsw.edu.au

Abstract

The complementary information found in different modalities of patient data can aid in more accurate modelling of a patient's disease state and a better understanding of the underlying biological processes of a disease. However, the analysis of multi-modal, multi-omics data presents many challenges. In this work, we compare the performance of a variety of ensemble machine learning (ML) algorithms that are capable of late integration of multi-class data from different modalities. The ensemble methods and their variations tested were (i) a voting ensemble, with hard and soft vote, (ii) a meta learner, and (iii) a multi-modal AdaBoost model using hard vote, soft vote, and meta learner to integrate the modalities on each boosting round, the PB-MVBoost model and a novel application of a mixture of expert's model. These were compared to simple concatenation. We examine these methods using data from an in-house study on hepatocellular carcinoma, plus validation datasets on studies from breast cancer and irritable bowel disease. We develop models that achieve an area under the receiver operating curve of up to 0.85 and find that two boosted methods, PB-MVBoost and AdaBoost with soft vote were the best performing models. We also examine the stability of features selected and the size of the clinical signature. Our work shows that integrating complementary omics and data modalities with effective ensemble ML models enhances accuracy in multi-class clinical outcome predictions and produces more stable predictive features than individual modalities or simple concatenation. We provide recommendations for the integration of multi-modal multi-class data.

Received: November 19, 2024. **Revised:** February 6, 2025. **Accepted:** February 21, 2025

Graphical Abstract



Benchmarking Ensemble Machine Learning Algorithms for Multi-Class, Multi-Omics Data Integration in Clinical Outcome Prediction

CONCLUSION: Integrating complementary omics and data modalities with effective ensemble ML models enhances accuracy in multi-class clinical outcome predictions and produces more stable predictive features compared to individual modalities or simple concatenation of data types.

Keywords: multi-omics; multi-modal; multi-class; machine learning; late integration; clinical outcome prediction; cancer; hepatocellular carcinoma

Introduction

Modelling complex biological systems has the potential to expand our understanding of many challenging diseases. Data encompassing entire biological systems, known as ‘omics data, can provide potentially complementary information, giving a more accurate picture of a patient’s disease state. Multi-omics data integration is the process of combining two or more ‘omics datasets in order to gain a better understanding of the mechanisms of biological processes and the complex interactions between biological systems [1, 2]. Multi-omics data integration can improve the accuracy of clinical outcome prediction, provide novel insights into mechanisms underlying complex diseases, aid in subtyping diseases and stratifying patient cohorts, discover new biomarkers, or identify potential therapeutic targets [2–5].

However, the analysis of multi-omics data presents many challenges. ‘Omics datasets typically contain a large number of features and only a small number of samples because of the high cost of clinical data collection and the limited number of study participants [6]. If not addressed, this leads to overfitting and, therefore, biased results in most machine learning (ML) algorithms. In addition, these datasets often contain many irrelevant and redundant features, misleading the algorithm and increasing computational complexity. The different ‘omics datasets are often heterogeneous, having been collected using different technologies, and can vary greatly in size, statistical distribution, scale, and signal strength [7]. They may be imbalanced in terms of the number of features or the composition of the classes. In addition, they may have missing values or suffer from batch effects. Finally, multi-omics data integration must provide stable and interpretable results if it is to be trusted by clinicians.

Data integration strategies are often categorized as early, intermediate, or late [8]. Early integration, also known as feature-level fusion, simply concatenates all of the ‘omics data’ into a single dataset and trains a classifier on it directly. This exacerbates the problems of high dimensionality, noise, and highly correlated features [2]. Intermediate integration transforms the omics datasets into a common representation space prior to modelling [9]. It can capture the relationships between the various modalities, but a clinical interpretation of the results is difficult. Late integration, also known as decision-level fusion, trains a model on each ‘omics dataset independently and the results are aggregated to give a final prediction. Techniques for aggregating the results may include a majority vote, weighted majority vote, a sum of the probabilities in each class, and a meta-learning [10]. Late integration is illustrated in Fig. 1.

Although late integration or decision-level fusion does not take into account the relationships between modalities, it is well-placed to overcome many of the challenges of multi-omics data integration. By training a separate model on each omics modality, the curse of dimensionality is reduced, meaning overfitting is less likely to occur, and computational complexity is also reduced. The approach is flexible as different ML models can be trained on each modality, taking advantage of the best-performing model in each case and also addressing the problems of heterogeneity and feature imbalance between modalities [6]. Pre-processing steps such as filtering, imputation of missing values, normalization, and feature selection can also be tailored to individual modalities.

Despite the many advantages of late integration strategies in overcoming the challenges of multi-omics data integration, there is a paucity of literature examining these methods.

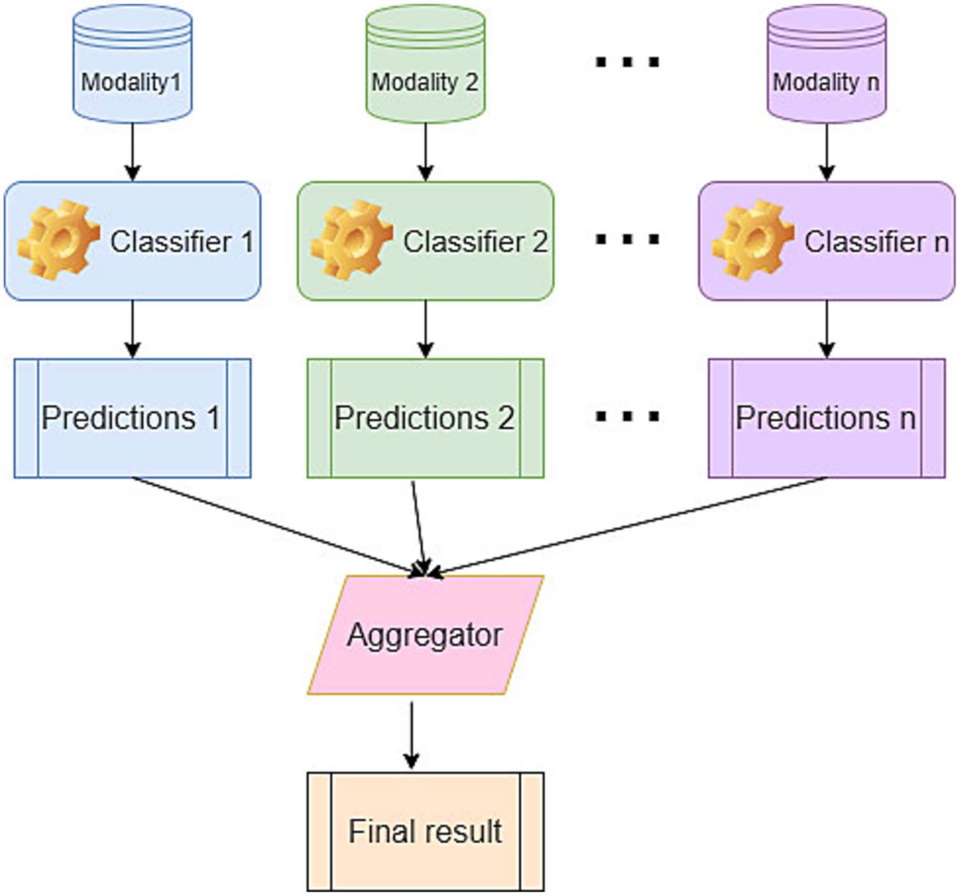


Figure 1. The structure of a model using late integration. Classifiers are applied independently to each modality and the predictions from those classifiers are aggregated to give a final result.

Previous studies have examined the use of ML for multi-modal data integration in various clinical settings, including oncology [11, 12], breast cancer [13], autoimmune disease [14], and inflammatory bowel disease [15] and in studies of type 2 diabetes, osteoarthritis, Alzheimer's disease, and systemic lupus erythematosus [5]. However, none of these studies focused on the advantages of late integration, and many surveys only briefly touch on the subject [3]. In contrast, Malik *et al.* [16] developed a framework for the quantification of survival and drug response for breast cancer patients using late integration of multi-omics data. Sharifi-Noghabi *et al.* [17] also predicted drug response using late integration of deep neural networks. Carillo-Perez *et al.* [18] used a late integration strategy to differentiate between non-small-cell lung cancer subtypes.

Our work aims to fill this gap by benchmarking late integration strategies using multi-modal ensemble ML and proposing improvements to existing late integration methods. Multi-modal ML is the task of learning from multiple modalities of the data to exploit complementary information and improve learning performance [19]. Ensemble ML is a technique that combines multiple weak learners to form a strong learner, i.e. more accurate than its base learners [20]. Multi-modal ensemble learning, therefore, combines these two techniques and can integrate and learn from multiple sources of high-dimensional 'omics data.

In this work, we benchmarked five ensemble ML methods, including multiple variations, for the late integration of multi-modal, multi-class data and compared these methods to a simple concatenation of the data. We applied these methods to an in-house dataset containing data collected from patients with liver disease [21] and validated them on data from four publicly available multi-omics datasets. We used existing methods and improved some methods to enhance their predictive accuracy. We examined the predictive accuracy of these methods as well as the stability of the features they selected and compared the results of the multi-modal methods with those of the individual modalities. For each dataset, we also determined an optimal subset of modalities that performed as well as the full set, thereby permitting patient diagnosis using fewer tests.

Materials and methods

To demonstrate the benefits of ensemble ML for multi-omics data integration, we benchmarked five late-integration methods that are capable of multi-modal, multi-class learning and tested their predictive accuracy as well as their suitability for knowledge discovery, measured by their ability to select a stable and interpretable set of predictive features. The ML methods tested include: (i) a simple voting ensemble, that aggregates the results from the different modalities using either a hard or soft vote; (ii) meta-learning, an algorithm that learns from the outputs of other ML algorithms; (iii) an enhanced multi-modal version of the well-known boosting algorithm AdaBoost [22]; (iv) a multi-view boosting method known as PB-MVBoost [23], which takes into account both the accuracy and the diversity of the classifiers trained on each view; and (v) a mixture of experts model that trains a different model for each class and integrates these with a gating function. We compared these methods to a single classifier trained on the concatenated data as a baseline.

Data

The proposed multi-modal, multi-class ML methods were applied to an in-house dataset containing data collected from patients with liver disease [21], and validated on data from four publicly available multi-omics datasets, two containing data from patients with inflammatory bowel disease (IBD) and two from patients with breast cancer, to differentiate between different stages and types of disease. The validation datasets were selected for their multi-modal, multi-class structure, demonstrating that the proposed methods can be applied to any clinical data. The characteristics of these datasets are summarized in Table 1, where references are given.

The in-house dataset [21] consists of data collected from prospectively enrolled patients to investigate the effect of the gut microbiota's composition and function on the development of liver disease and primary hepatocellular carcinoma (HCC). Adult study subjects with liver cirrhosis and/or HCC were enrolled from two liver centres in Sydney, Australia, as were healthy adult volunteers to function as controls. Baseline demographic and clinical data were collected at time of enrolment, with blood samples collected for multi-omics data and oral/faecal samples collected for metagenomic sequencing of oral/gut microbiota.

Study subjects were subsequently divided into one of four classes: healthy controls (CON); individuals with liver cirrhosis secondary to metabolic-associated fatty liver disease (MAFLD-cirrhosis) (CIR); individuals with HCC secondary to MAFLD-CIR (LN); individuals with HCC secondary to viral hepatitis (LX). Of the 122 study subjects, the number of samples common to all modalities was 106, and samples were approximately evenly distributed across classes, as shown in Table 1. Seven modalities of data were available, and these are also listed in Table 1. For the stool and oral microbiome, two microbial levels of genera and species were used separately in the models.

The study was approved by the Sydney Local Health District Human Research Ethics Committee (HREC), New South Wales Health: approval number HREC/16/ RPAH/701; SSA18/G/058.

Data pre-processing

The data processing pipeline is shown in Fig. 2. Special consideration was given to memory management because of the large data file sizes, resulting in a first-step process to read the files into memory. The first step was to read the raw files one at a time, apply filtering to identify potentially relevant features, as opposed to those with no predictive ability, record their names, and then release the memory used. In the second step, only the relevant features of all files were read in, reducing the amount of memory required.

Prior to modelling, filtering was performed to identify potentially relevant features and reduce the dimensionality of the data. This was also a multi-step process that could be tailored to each dataset and depended on the characteristics of the dataset. The following command line options could be chosen:

(1) Features with >50% missing values and >90% zero values were eliminated.
(2) Of the remaining features, one feature from each pair of correlated features was eliminated.
(3) If the dimensionality of the dataset was still very large (i.e. if the ratio of the number of features to the number of samples was greater than 10), then only the 500 features with the highest variance were retained.

The following pre-processing steps were applied to the data during model building:

(1) Balancing: for imbalanced datasets, classes were balanced using the Synthetic Minority Oversampling Technique (SMOTE) [27]. SMOTE was applied to the training set only during model building, following the method of [28], balancing one class at a time against the majority class.
(2) Imputation: missing values were imputed using Multiple Imputation by Chained Equations (MICE) [29]. However, if the size of the dataset caused the computation time of MICE to increase unacceptably, then k-Nearest Neighbours (KNN), a single imputation method, was employed instead.
(3) Normalization: Counts per million normalization, followed by a log transformation, was applied to the RNA sequence and DNA data. Standardization was applied to all other data.
(4) *Feature selection:* Feature selection was performed using the Boruta feature selection algorithm [30], with the gradient boosting machine (GBM) [22] as the underlying learner.

Boruta [30] is a model-agnostic algorithm which can wrap around any base learner that provides feature importance scores. It creates a set of shadow features, which are randomly shuffled copies of the original features, and then repeatedly trains a random forest on the combined set of features. Any features that achieve a lower importance score than the highest-scoring shadow feature are considered irrelevant. In this way, Boruta naturally generates a feature selection threshold, which is based on P-values.

Boruta is known as an 'all-relevant' method, as it identifies all features relevant to the target variable, including those that are correlated. This is in contrast to the more common 'minimal-optimal' feature selection methods, which identify a small subset of features that maximize predictive accuracy. The 'all-relevant' method is advantageous when the aim of model development is knowledge discovery.

Individual modalities

In order to see the benefits of data integration, a classifier was first trained on each of the individual modalities without performing data integration. The GBM classifier was chosen because of its superior performance in previous experiments. The same pre-processing pipeline, shown in Fig. 2 was applied to the individual modalities. The aim of these experiments was not only to set a baseline for comparison with the integration techniques but also to observe which modalities gave the best predictive performance.

Data integration techniques

ML methods capable of integrating multiple modalities of data using the late integration strategy were examined and compared in this study. To ensure a fair comparison, all methods used the GBM [22], a multi-class classifier, as their underlying classifier. However, it should be noted that the methods investigated allow different underlying classifiers to be trained for each modality. The methods, which were developed using custom code, are summarized in Table 2 and the integration methods are described graphically in Fig. 3.

Concatenation

This method, also known as feature-level fusion [6], combines the data from all modalities into a single vector and trains a single

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.

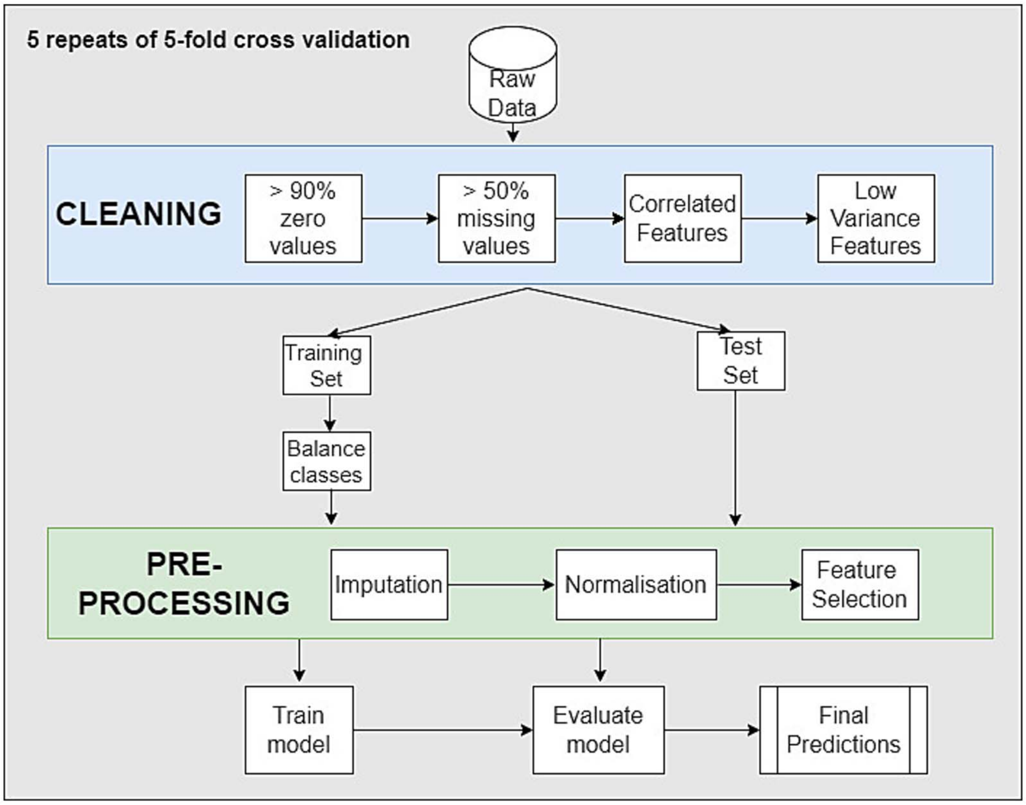


Figure 2. The data processing pipeline used in the ML experiments. The data were first filtered, removing features with >90% zero values and >50% missing values. Correlated and low variance features were also removed in the larger datasets. The data were repeatedly split into training and test sets using five repeats of cross-fold validation. Imbalanced classes were balanced during training. Multiple imputation was used to impute missing values, values were normalized and feature selection was applied prior to training. The test set was used to evaluate the model performance.

Table 2. Data integration methods benchmarked in this study

Integration method	Results aggregator	Abbreviation	Brief description
Concatenation		CONCAT	All modalities are joined into a single dataset and a classification model is trained on this dataset.
Voting ensemble	Hard vote	ENS-H	A classification model is trained on each modality and the predictions are combined using majority vote.
Voting ensemble	Soft vote	ENS-S	A classification model is trained on each modality and the predictions are combined by adding the class probabilities.
Meta learner	Meta learner	ML	A classification model is trained on each modality and the predictions are combined by training a learner (in this case a random forest) on the results of the individual models.
AdaBoost	Hard vote	ADA-H	A multi-modal AdaBoost model is trained using all modalities. In each boosting round the predictions are combined using majority vote.
AdaBoost	Soft vote	ADA-S	A multi-modal AdaBoost model is trained using all modalities. In each boosting round the predictions are combined using the sum of probabilities.
AdaBoost	Meta learner	ADA-M	A multi-modal AdaBoost model is trained using all modalities. In each boosting round the predictions are combined using a meta-learner.
PB-MVBoost	Weighted vote	PBMV	A classification model is trained on each modality and weights are learned for each modality and each classifier. Results are combined using a weighted vote.
Mixture of experts using voting ensemble	Gating function	MOE-COMBN	A multi-modal voting ensemble classifier is trained for each class and a gating function selects the most confident result for each sample.

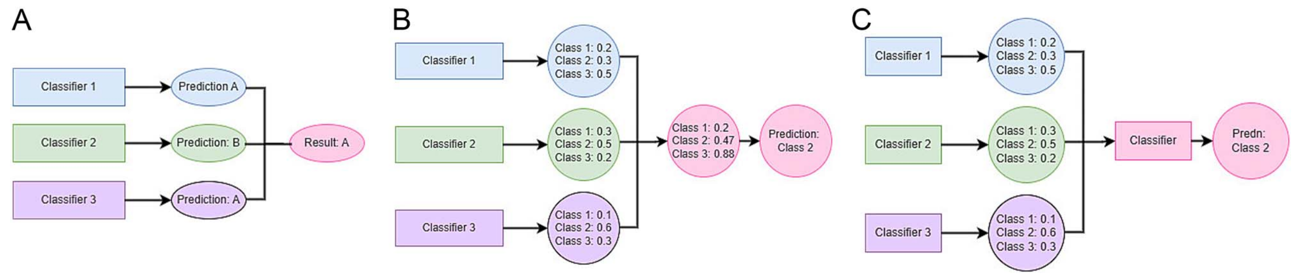


Figure 3. Techniques used to aggregate the results of the classifiers applied to each modality. (A) The hard vote is a simple majority vote. (B) The soft vote is an average of the probability scores. (C) The meta-learner is a classifier that learns from the results of the base-level classifiers.

multi-class classifier on the concatenated data. This method was included as a baseline comparison for the other, more advanced methods.

Voting ensemble

A voting ensemble trains a classifier on each modality and aggregates the outputs of the individual classifiers using a hard vote or a soft vote, illustrated in Fig. 3. In a hard vote, or majority vote, the class, i.e. predicted by the greatest number of modalities, is the final prediction. In the case of a tie, the first predicted class is selected. In a soft vote or weighted vote, the probability scores from each modality for each class are averaged. The class with the highest probability score is the predicted class.

Meta learner

Meta-learning is a general term for ML algorithms that learn how to combine the outputs of other algorithms to maximize predictive accuracy. In the case of multi-modal data, a different base learner can be trained on each modality in parallel, and the meta-learner is trained on the outputs of those base learners. The advantage of meta-learning is that it can produce reliable results with only a relatively small number of examples [31]. A recent survey gives many examples of the use of meta-learning in healthcare in areas such as clinical risk prediction, disease diagnosis, and drug interaction detection [31].

In this study, a meta-learner, illustrated in Fig. 3C was constructed using the GBM as the base classifier and a random forest as the meta-learner.

Multi-modal AdaBoost

Boosting is an ensemble technique that trains a series of weak classifiers sequentially, such that each classifier learns from the mistakes of its predecessors [32]. Ultimately, these weak classifiers are combined to form a strong classifier.

AdaBoost or Adaptive Boosting [22] was one of the first boosting algorithms to be proposed. On each iteration of the AdaBoost algorithm, the samples that were misclassified in the last iteration are given increased weight, forcing the algorithm to focus on the more difficult-to-classify samples, with the aim of correcting the errors made in the last iteration. The final model is a weighted linear sum of all the models in the ensemble.

AdaBoost was initially developed for binary classification, but a multi-class version of AdaBoost was introduced by Zhu *et al.* [33]. More recently, multi-view versions of AdaBoost have been proposed, which can also be used to model multi-modal data. Xu and Sun developed a multi-view AdaBoost algorithm, but it was limited to two views [34]. Xiao and Guo extended the multi-view AdaBoost framework to an arbitrary number of views in the context of multilingual subjectivity analysis [35]. They developed two approaches, each of which trains a learner separately on each view and then combines the results of the single-view classifiers on each iteration, either using a hard majority vote or a linear weighted combination of the outputs of the single-view classifiers.

Here, a multi-modal version of AdaBoost was implemented following the method of Xiao and Guo [35], with some novel modifications. On each round of the boosting process, a classifier was trained independently on each modality. The results from these classifiers were then aggregated to give a final decision using one of three methods—a hard vote, a soft vote, or a meta-learner trained on the results of the independent classifiers.

Whilst in the original AdaBoost, it is a simple matter to identify the misclassified samples to calculate the classification error rate, in multi-modal AdaBoost, all modalities must be taken into account when calculating the error rate. Here, a sample is considered to be correctly classified only if it is classified with high confidence. If the final decision is made by a hard vote or meta-learner, a sample is classified with high confidence if at least half of the modalities agree on its classification. The soft vote gives a final probability for each class, and a sample is classified with high confidence if the highest probability is at least double that of the next highest probability. If these conditions are not met, then the sample is considered to be misclassified.

The optimal number of boosting steps within each iteration was chosen empirically, based on tests run using different numbers of boosting steps, and was determined to be 20. Beyond this value, no added benefit was achieved.

PB-MVBoost

PB-MVBoost [23] is a multi-view ensemble method, based on AdaBoost, that aims to balance the accuracy of the classifiers trained on each view and the diversity of their outputs by learning two sets of weights—weights over the classifiers and weights over the views. It then combines the results of each classifier using a weighted vote, learning the weights by minimizing an upper bound on the error of the majority vote. Fadnavis *et al.* [36] used PB-MVBoost in a novel framework to distinguish healthy controls from those in the early stages of Huntington's disease.

PB-MVBoost was implemented in R, directly following the author's Python implementation, which is available on GitHub.

Mixture of experts

A mixture of experts model [37] divides a complex ML task into multiple sub-tasks, based on domain knowledge, and trains a model on each sub-task. Each of these models focuses solely on its specific sub-task, becoming an expert in that sub-space. A gating function learns which expert to trust and selects the best expert to predict each sample. Minoura *et al.* [38] developed a model for integrated analysis of single-cell multi-omics data using a mixture of experts model.

Here, a novel adaptation of this method was developed: a separate model was trained for each class in the multi-class setting, using a one-vs-rest approach, i.e. each expert was a binary classifier, distinguishing its own class from all other classes combined. The experts were trained in parallel and each one independently balanced the data for its own class during training. Once training was complete on all folds of cross-validation, the gating function was applied to determine the best response for each sample. The gating function operated using the following rules:

- Each expert can only predict the class it has been trained to predict or 'REST'.
- If an expert predicts its own class, and it is the only one to do so, then its prediction is accepted as correct.
- If more than one expert predicts its own class, then the prediction of the expert that predicts with the highest confidence (probability) is accepted as correct.
- If no experts predict their own class (i.e. all predict 'REST'), then the sample is classified as unknown. This indicates which samples are difficult to predict.

Incremental method

From a clinical perspective, the ability to make an accurate prediction from a smaller subset of modalities means that fewer tests are required, making diagnosis simpler, less expensive, and possibly less invasive for the patient. With this in mind, we designed an incremental model that determines the subset of the modalities that gives maximum predictive performance.

The incremental model adds (or removes) one modality at a time to (or from) the model. Here, we included all modalities at the start and eliminated one modality at a time, comparing the performance of the models trained on the remaining subsets of modalities after each elimination. The modality missing from the model that gave the best performance score was the next modality to be removed. So the order of removal was determined by the model itself and subsequent tests were carried out on the reduced modality set. The modalities were integrated using a soft voting ensemble and the metric used for comparison was the F1 score. A small margin of error was allowed in the performance comparison.

Feature selection and calculation of feature importance scores

Feature selection identifies the features that are most relevant to the model outcome [39]. These features can then be examined as potential biomarkers or may provide useful insights into the underlying mechanisms of disease.

When multiple classifiers are combined to give a final result, the final feature importance scores must be calculated in a meaningful way. For most models, the final set of selected features consists of those selected by each individual classifier, normalized, and scaled to a range of [0,1] for comparison.

Some models required additional consideration. The features provided as input to the meta-learner are the results of the base classifiers applied to each modality. Therefore, the feature importance scores generated by the meta-learner give an indication of the relative importance of each modality.

The calculation of feature importance scores for AdaBoost and PB-MVBoost was more complex. For each fold of data to which the model was applied, multiple boosting iterations were run. Each iteration produced a model weight and a set of feature importance scores for each modality. The final feature importance scores for each modality in each fold were calculated as the sum of the raw scores multiplied by the model weight, divided by the sum of the weights.

The Mixture of Experts model selects a set of features for each expert i.e. for each class. The fact that these feature sets can differ shows clearly that different features can influence the model for different classes.

In addition, for all models, only those features that were selected in 75% of cross-validation iterations and that had a normalized score of 0.5 or higher were included in the final set of selected features.

Feature selection stability

Stability of feature selection can be defined as the reproducibility of the features selected when the method is applied to different samples of the data [40]. Various measures of stability have been proposed, each of which has its own strengths and limitations. The relative weighted consistency index, proposed by Somol and Novovičová [41], has been chosen here because of its ability to measure the stability of sets of features of different lengths, such as those selected from different data samples when running repeated experiments, and because it does not over-emphasize low-frequency features.

Experimental framework and model evaluation

The methods being assessed were evaluated for their predictive performance as well as the stability of the set of features they selected. All models were evaluated using five repeats of five-fold cross-validation. The evaluation metrics are shown in Table 3 were calculated for each fold and averaged over folds and classes to give a final result [42].

All experiments were run on the Katana high-performance computing cluster at the University of NSW [43]. All code was written in R [44]. The R package *mlr* [45] was used as a framework to implement the ML experiments and the R package *Future* was used to implement parallel processing [46].

Tests of statistical significance were applied to each group of experiments, using the corrected resampled paired *t*-test, proposed by Nadeau and Bengio [47]. This test takes into account the fact that the Type I error is inflated when applying a standard *t*-test on results from a repeated k-fold cross-validation because the results are not independent and correct for this.

Results

Individual modalities

The performance of the individual modalities in each dataset is shown in the boxplots in Supplementary Figs S1–S5 and in Supplementary Table S1. The values shown represent the average across all iterations of cross-validation.

In the HCC dataset, the best-performing modality was the Cytokine data, with an AUC of 0.77 and an F1 score of 0.64 (±0.15). The metabolomic data also performed well, with an AUC of 0.74 and an F1 score of 0.6 (±0.15). In two other datasets, IBD1 and IBD2, the metabolomic data also proved to be the most predictive with AUC scores of 0.56 and 0.76 respectively and F1 scores of 0.39 (±0.17) and 0.68 (±0.06), respectively. In the breast cancer datasets, the clinical data were the most predictive in the Breast1 dataset, with an AUC of 0.8 and an F1 score of 0.65 (±0.2), and the proteomic data were the most predictive in the Breast2 dataset, with an AUC of 0.74 and an F1 score of 0.57 (±0.36).

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	

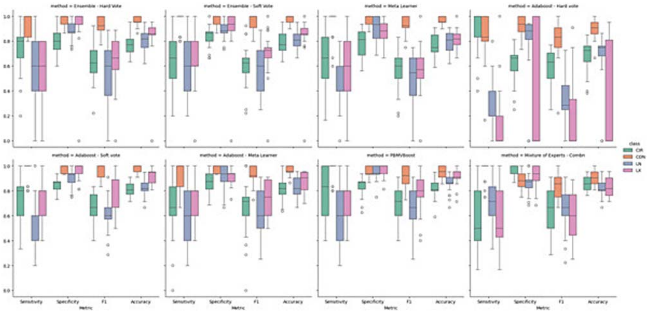


Figure 4. Performance of the multi-modal models on the HCC (Genus) dataset. The colours of the bars represent the patient classes. Each group of boxplots in each subplot represents one metric, from left sensitivity, specificity, F1 score, and accuracy.

Multi-modal data integration

Each of the multi-modal data integration strategies was applied to each dataset and the results of these experiments are shown in the boxplots in Fig. 4 for the HCC (Genus) dataset, and in Supplementary Figs S6–S11 and Supplementary Table S2 for the remaining datasets. The PB-MVBoost method was either the best or second-best performing model in every dataset, achieving an AUROC score of 0.85 in the HCC-Genus dataset. The AdaBoost method with a soft vote aggregator also performed well, achieving AUROC scores of up to 0.84 and performing equally as well as the PB-MVBoost method in some cases. In the IBD1 and Breast2 datasets, a simple concatenation of modalities performed better than the multi-modal methods. However, further examination of the feature selection results in Section 3.4 will reveal that the features selected by this method are less stable than those selected by the multi-modal methods.

A comparison of the performance of the best multi-modal method in each dataset to the best individual modality from the same dataset is shown in Table 4. In almost all cases, the best multi-modal method outperforms the best individual modality. This is evidence of the benefit of multi-modal data integration. The exception is the Breast2 dataset, where the performance is equal. As seen in Table 1, the classes in Breast2 are very imbalanced and there are very few normal samples with which to compare the cancerous samples. This is likely to affect performance of the models trained on the Breast2 data.

Optimal subset of modalities

The incremental model, described in Section 2.5, was used to determine the most predictive subset of modalities in each dataset. On each iteration of this model, modalities were removed one at a time and the performance of the model trained on the remaining subsets of modalities was examined. The modality missing from the model that gave the best performance score was the next modality to be removed.

For some datasets, namely IBD2, Breast1, and Breast2, a single modality gave the best predictive performance, whilst for others, namely HCC and IBD1, a small number of modalities gave equal or better performance than the full set of modalities. The results of this analysis are given in Table 5, which shows the degree to which performance, measured by the F1 score, improved as each modality was removed and the order in which the modalities were removed. The subset of modalities selected as optimal corresponds to the top-performing individual modalities in the two datasets, confirming the validity of the incremental model.

To further validate this method, Table 6 compares the AUC and F1 scores for each of the multi-modal methods on all modalities versus the optimal subset for HCC and IBD1, as the optimal subset size for these two datasets was greater than one. It can be seen that each method performs as well on the optimal subset of modalities as it does on the full set of modalities.

A limitation of the incremental method is that it aggregates the predictions from the subsets of modalities using a soft voting ensemble only, for reasons of efficiency, but this ensemble may not be as accurate as some of the more advanced methods. Therefore, the method provides a guide only, indicating which modalities may be important, but it may be possible to achieve better results than it indicates.

Feature selection

As the primary purpose of developing these multi-modal models is to identify a clinical signature that can predict the development of the disease in question, the methods must also be judged on the clinical signature they identify. The desirable characteristics of a clinical signature include its length, its stability, the number of modalities from which it draws features, and its accuracy in prediction. Ideally, a signature consisting of a smaller number of features and modalities will mean fewer tests for the patient and be more economical. The more stable, or reproducible, and the more accurate the feature selection results, the more confidence clinicians can have in their reliability in predicting disease.

The results of the feature selection for each multi-modal method applied to each dataset are summarized in Supplementary Table S3, which lists the number of features selected, the number of modalities those features are drawn from, the stability of

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.

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

the feature selection, measured using the relative weighted consistency index [41], the predictive accuracy of the selected features and the mean of these two measurements. Note that the Meta Learner uses a random forest as its meta classifier, and a random forest applies a non-zero feature importance score to every feature, rather than selecting a subset of features. Since every feature achieves a score on every iteration this method artificially achieves a perfect stability score of 1.

In every dataset, simple concatenation produced the least stable feature selection results. Concatenation of the modalities greatly increases the dimensionality of the data and feature selection is less stable in high dimensions. The data integration methods overcome this problem by training a separate classifier on each modality, illustrating another benefit of these methods.

In most datasets, the PB-MVBoost method achieved the highest mean of stability and accuracy, but its signature length tended to be among the longest of all the methods and was selected from the largest number of modalities. By contrast, in most cases, the AdaBoost model with a soft vote produced a slightly shorter signature with little loss in predictive accuracy or stability.

Discussion

In this work, we have presented a benchmarking study of multimodal multi-class ML techniques for late integration of multiomics data applied to different datasets. We examined their predictive accuracy as well as the stability of the features they selected and compared the results of the multi-modal methods with those of the individual modalities. For each dataset, we also determined an optimal subset of modalities that performed as well as the full set, thus permitting patient diagnosis using fewer tests.

We employed existing and enhanced methods for late integration. Existing methods included a simple voting ensemble using hard and soft voting, a meta learner, and the PB-MVBoost algorithm. Enhancements were made to the multi-modal AdaBoost algorithm to improve its predictive accuracy and a novel application of the mixture of experts model was developed, which builds an expert for each class and combines the results of these experts using a novel gating function.

Overall, the multi-modal methods showed superior performance to the individual modalities, with the PB-MVBoost and the

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)

AdaBoost models being the most predictive. This shows that different modalities can provide complementary information about a patient's disease state, and integrating those modalities in a single model can improve predictive performance. Therefore, the use of these data integration techniques is recommended.

Boosting is an ensemble technique that trains a series of weak learners and combines them to form a stronger learner, improving predictive accuracy by reducing overfitting. Because of its ability to reduce overfitting, boosting is particularly suited to datasets that have a large number of features and a small number of samples, which is a characteristic of the datasets examined in this work. In addition, the two boosted methods, PB-MVBoost and AdaBoost, calculate a weight for each modality, thereby prioritizing the more predictive modalities. Combining boosting, modality weighting, and data integration gives these methods an advantage, so it is not surprising that they performed well.

In contrast, the meta learner, a mixture of experts model and voting ensembles gives equal weight to each modality, and our results on the individual modalities show that some modalities are more predictive than others. Therefore, the less predictive modalities are likely to be detrimental to the overall predictive power of the model.

Further, our results from the incremental model show that it is not always necessary to incorporate all modalities of data in a model to achieve the maximum predictive power. Finding the right subset of modalities to maximize predictive performance is crucial, not only to optimize the model outcome, but also for future clinical use, where a smaller set of modalities can simplify and reduce the cost of screening patients for disease. Therefore, the use of our incremental model to determine the most predictive subset of modalities is recommended.

In the IBD1 and Breast2 datasets, a simple concatenation was the best-performing model. The proteomic modality of the Breast2 dataset significantly outperforms the other two modalities and is identified by the incremental model as being the best subset of modalities on its own. Therefore, it seems likely that the other two modalities do not contribute significantly to the model and that the proteomic modality dominates in the concatenation model.

In methods where the final decision could be made via a hard vote or soft vote, such as the voting ensembles and AdaBoost, the soft vote was the better performer each time. Averaging the probability scores for each class in each modality, as in the soft vote, gives a finer tuning of the results than a simple hard cut-off, as in the case of the hard vote. In a hard vote ensemble, the more modalities in the model, the greater the chance of errors compounding and confusing the final decision.

The ability to identify a stable set of predictive features is an essential, yet often neglected, aspect of modelling any biological system with the aim of knowledge discovery, as these features are likely to provide valuable insights into the underlying biological processes. However, stability must be considered in conjunction with predictive performance. A method that selects the same set of features on each iteration will be very stable but may have poor predictive accuracy, whilst another method that selects a different set of features on each iteration may be quite accurate but highly unstable and unsuitable for knowledge discovery. Our results show that the PB-MVBoost model and the AdaBoost model produced the most accurate and stable feature selections, with the AdaBoost model generally producing a slightly shorter clinical signature.

Based on our results, we can provide the following recommendations for integrating multi-modal, multi-class data. We recommend first examining the individual modalities to identify which are the most predictive. Following this, the incremental method should be applied to determine whether an optimal subset of modalities can be found, and this optimal subset should align with the most predictive individual modalities. Finally, we recommend the training of a PB-MVBoost or AdaBoost model with a soft vote for integrating the data modalities.

Conclusion

The ability to integrate data from multiple modalities can allow more accurate modelling of a patient's clinical outcome, as these modalities can provide complementary information. Such modelling may also assist in understanding the mechanisms underlying complex diseases and help identify novel biomarkers to aid in diagnosis.

The aim of this paper was to compare the performance of ensemble ML algorithms capable of late integration of multi-class data from different modalities. The ensemble methods and their variations tested were (i) a voting ensemble, with hard and soft vote, (ii) a meta learner, and (iii) a multi-modal AdaBoost model using a hard vote, a soft vote and a meta learner to integrate the modalities on each boosting round, the PB-MVBoost model and a novel application of a mixture of experts model. These were compared to simple concatenation as a baseline and to the individual modalities.

We examined the predictive accuracy of these methods, as well as the stability of the features they selected and the size of the clinical signature determined by applying them to an in-house dataset containing data collected from patients with liver disease [21] and validated them on data from four publicly available multi-omics datasets. Our results demonstrated that on the whole the multi-modal methods outperformed the individual modalities and that two boosted methods, PB-MVBoost and AdaBoost with a soft vote were the overall best-performing models, both in terms of predictive accuracy and stability of feature selection. We also provided a means of determining an optimal subset of modalities that could lead to a smaller clinical signature without loss of predictive accuracy. Finally, we have provided recommendations for the integration of multi-modal multi-class data.

Key Points

- Modelling multiple modalities of data can improve the accuracy of clinical outcome prediction.
- Multiple ensemble machine learning methods using late integration of multi-omics data were benchmarked across diverse clinical datasets.
- The multi-modal methods outperformed the individual modalities in almost all cases.
- Two boosted methods, PB-MVBoost and AdaBoost, were the best-performing models.
- Stability of feature selection and size of clinical signature were also considered.

Author contributions

Annette Spooner (Methodology, Software, Validation, Writing—original draft), Mohammad Karimi (Writing—review & editing), Barbra Toplis (Data Curation, Resources), Jason Behary (Data Curation, Resources), Azadeh Safarchi (Writing—review & editing), Salim Maher (Writing—review & editing), Fatemeh Vafaee (Supervision, Writing—review & editing), Amany Zekry (Conceptualization, Funding acquisition, Supervision, Writing—review & editing), and Arcot Sowmya (Supervision, Writing—review & editing).

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: None to declare.

Funding

This work was supported by a grant from the Medical Research Future Fund (MRFF), grant id 2008996.

Data availability

The HCC dataset is available from the authors upon reasonable request. Data availability for the other datasets can be found in the relevant papers—references are given in Table 1.

Code availability

The code used in this work can be downloaded from https://github.com/annette987/MOMENT

References

1. Krassowski M, Das V, Sahu SK. *et al.* State of the field in multi-omics research: from computational needs to data mining and sharing. *Front Genet* 2020;**11**:1–17. https://doi.org/10.3389/fgene.2020.610798
2. Chicco D, Cumbo F, Angione C. Ten quick tips for avoiding pitfalls in multi-omics data integration analyses. *PLoS Comput Biol* 2023;**19**:e1011224. https://doi.org/10.1371/journal.pcbi.1011224
3. Picard M, Scott-Boyer MP, Bodein A. *et al.* Integration strategies of multi-omics data for machine learning analysis. *Comput Struct Biotechnol J* 2021;**19**:3735–46. https://doi.org/10.1016/j.csbj.2021.06.030
4. Dickinson Q, Aufschnaiter A, Ott M. *et al.* Multi-omic integration by machine learning (MIMaL). *Bioinformatics* 2022;**38**:4908–18. https://doi.org/10.1093/bioinformatics/btac631
5. Kreitmaier P, Katsoula G, Zeggini E. Insights from multi-omics integration in complex disease primary tissues. *Trends Genet* 2023;**39**:46–58. https://doi.org/10.1016/j.tig.2022.08.005
6. Tabakhi S, Suvon MNI, Ahadian P. *et al.* Multimodal learning for multi-omics: a survey. *World Sci Annu Rev Artif Intell* 2023;**01**:1–39. https://doi.org/10.1142/s2811032322500047
7. Santiago-Rodriguez TM, Hollister EB. Multi ‘omic data integration: a review of concepts, considerations, and approaches. *Semin Perinatol* 2021;**45**:151456. https://doi.org/10.1016/j.semperi.2021.151456
8. A Serra, P Galdi, and R Tagliaferri, *Multiview Learning in Biomedical Applications*. Elsevier Inc., Amsterdam, Netherlands. 2018. https://doi.org/10.1016/B978-0-12-815480-9.00013-X.
9. Wang Y, Tang S, Ma R. *et al.* Multi-modal intermediate integrative methods in neuropsychiatric disorders: a review. *Comput Struct Biotechnol J* 2022;**20**:6149–62. https://doi.org/10.1016/j.csbj.2022.11.008
10. Galar M, Fernández A, Barrenechea E. *et al.* An overview of ensemble methods for binary classifiers in multi-class problems: experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognit* 2011;**44**:1761–76. https://doi.org/10.1016/j.patcog.2011.01.017
11. Raufaste-Cazavieille V, Santiago R, Droit A. Multi-omics analysis: paving the path toward achieving precision medicine in cancer treatment and immuno-oncology. *Front Mol Biosci* 2022;**9**:1–18. https://doi.org/10.3389/fmolb.2022.962743
12. Acharya D, Mukhopadhyay A. A comprehensive review of machine learning techniques for multi-omics data integration: challenges and applications in precision oncology. *Brief Funct Genomics* 2024;**23**:1–12. https://doi.org/10.36001/ijphm.2024.v15i2.3850

13. Sammut SJ, Crispin-Ortuzar M, Suet-Feung C. *et al.* Multi-omic machine learning predictor of breast cancer therapy response. *Nature* 2022;**601**:623–9. https://doi.org/10.1038/s41586-021-04278-5
14. Martorell-Marugán J, Chierici M, Jurman G. *et al.* Differential diagnosis of systemic lupus erythematosus and Sjögren's syndrome using machine learning and multi-omics data. *Comput Biol Med* 2022;**152**:2023. https://doi.org/10.1016/j.compbiomed.2022.106373
15. Gardiner LJ, Carrieri AP, Bingham K. *et al.* Combining explainable machine learning, demographic and multi-omic data to inform precision medicine strategies for inflammatory bowel disease. *PLoS One* 2022;**17**:1–23. https://doi.org/10.1371/journal.pone.0263248
16. Malik V, Kalakoti Y, Sundar D. Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer. *BMC Genomics* 2021;**22**:1–11. https://doi.org/10.1186/s12864-021-07524-2
17. Sharifi-Noghabi H, Zolotareva O, Collins CC. *et al.* MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics* 2019;**35**:i501–9. https://doi.org/10.1093/bioinformatics/btz318
18. Carrillo-Perez F, Morales JC, Castillo-Secilla D. *et al.* Machine-learning-based late fusion on multi-omics and multi-scale data for non-small-cell lung cancer diagnosis. *J Pers Med* 2022;**12**:601. https://doi.org/10.3390/jpm12040601
19. Kline A, Wang H, Li Y. *et al.* Multimodal machine learning in precision health: a scoping review. *npj Digit Med* 2022;**5**:1–14. https://doi.org/10.1038/s41746-022-00712-8
20. Zhao J, Xie X, Xu X. *et al.* Multi-view learning overview: recent progress and new challenges. *Inf Fusion* 2017;**38**:43–54. https://doi.org/10.1016/j.inffus.2017.02.007
21. Behary J, Amorim N, Jiang X-T. *et al.* Gut microbiota impact on the peripheral immune response in non-alcoholic fatty liver disease related hepatocellular carcinoma. *Nat Commun* 2021;**12**:1–14. https://doi.org/10.1038/s41467-020-20422-7
22. Freund Y, Schapire R. A decision-theoretic generalization of on-line learning and an application to boosting. *Comput Learn theory* 1995;**55**:119–39. https://doi.org/10.1006/jcss.1997.1504
23. Goyal A, Morvant E, Germain P. *et al.* Multiview boosting by controlling the diversity and the accuracy of view-specific voters. *Neurocomputing* 2019;**358**:81–92. https://doi.org/10.1016/j.neucom.2019.04.072
24. Mehta RS, Mayers JR, Zhang Y. *et al.* Gut microbial metabolism of 5-ASA diminishes its clinical efficacy in inflammatory bowel disease. *Nat Med* 2023;**29**:700–9. https://doi.org/10.1038/s41591-023-02217-7
25. Franzosa EA, Sirota-Madi A, Avila-Pacheco J. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat Microbiol* 2019;**4**:293–305. https://doi.org/10.1038/s41564-018-0306-4
26. Krug K, Jaehnig EJ, Satpathy S. *et al.* Proteogenomic landscape of breast cancer tumorigenesis and targeted therapy. *Cell* 2020;**183**:1436–1456.e31. https://doi.org/10.1016/j.cell.2020.10.036
27. Chawla KWP, N V, Bowyer KW. *et al.* SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002;**16**:321–57. https://doi.org/10.1613/jair.953
28. Rachmatullah MIC. The application of repeated SMOTE for multi class classification on imbalanced data. *Tek Inform dan Rekayasa Komput* 2022;**22**:13–24. https://doi.org/10.30812/matrik.v22i1.1803
29. van Buuren, Oudshoorn CGM. MICE: multivariate imputation by chained equations. *R Packag version* 2007;**1**:2007.
30. Kursa MB, Jankowski A, Rudnicki WR. Boruta—a system for feature selection. *Fundam Informaticae* 2010;**101**:271–85. https://doi.org/10.3233/FI-2010-288
31. Rafiei A, Moore R, Jahromi S. *et al.* Meta-learning in healthcare: a survey. *SN Computer Science* 2024;**5**:792, 1–25.
32. Schapire RE. The strength of weak learnability (extended abstract). *Mach Learn* 1990;**5**:197–227. https://doi.org/10.1002/nbm.1810
33. Zhu J, Rosset S, Zou H. *et al.* Multi-class AdaBoost. *Stat Interface* 2009;**2**:349–60. https://doi.org/10.4310/sii.2009.v2.n3.a8
34. Xu Z, Sun S. An algorithm on multi-view AdaBoost. *Neural Inf Process—Theory Algorithms* 2010;**6443**:3236–6. https://doi.org/10.1007/978-3-319-24612-3_301724
35. Xiao M, Guo Y. Multi-view AdaBoost for multilingual subjectivity analysis. *24th International Conference on Computing Linguistics—Proceedings of theCOLING 2012 Technical Papers* 2012; 2851–66.
36. Fadnavis S, Polosecki P, Garyfallidis E. MVD-fuse: detection of white matter degeneration via multi-view learning of diffusion microstructure. bioRxiv 2021.
37. Yuksel SE, Wilson JN, Gader PD. Twenty years of mixture of experts. *IEEE Trans Neural Networks Learn Syst* 2012;**23**:1177–93. https://doi.org/10.1109/TNNLS.2012.2200299
38. Minoura K, Abe K, Nam H. *et al.* A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. *Cell Reports Methods* 2021;**1**:100071. https://doi.org/10.1016/j.crmeth.2021.100071
39. Li J, Wang S. Feature selection: a data perspective. *ACM Comput Surv* 2018;**50**:1–45. https://doi.org/10.1145/3136625
40. Turney P. Technical note: bias and the quantification of stability. *Mach Learn* 1995;**20**:23–33. https://doi.org/10.1007/bf00993473
41. Somol P, Novovičová J. Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. *IEEE Trans Pattern Anal Mach Intell* 2010;**32**:1921–39. https://doi.org/10.1109/TPAMI.2010.34
42. Hossin M, Sulaiman MN. A review on evaluation metrics for data classification evaluations. *Int J Data Min Knowl Manag Process* 2015;**5**:01–11. https://doi.org/10.5121/ijdkp.2015.5201
43. "Katana," https://doi.org/10.26190/669X-A286, 2010. https://doi.org/10.26190/669X-A286
44. R Core Team, "*R: A language and environment for statistical computing*," Vienna, Austria: R Foundation for Statistcal Computin. https//www.R-project.org. 2019.
45. B. Bischl, Lang M, Kotthoff L. *et al.* "Mlr: machine learning in R." *J Mach Learn Res* 2016;**17**:5938–42, [Online]. Available: https://dl.acm.org/citation.cfm?id=3053452
46. Bengtsson H. A unifying framework for parallel and distributed processing in R using futures. *R J* 2021;**13**:273–91. https://doi.org/10.32614/RJ-2021-048
47. Nadeau C, Bengio Y. Inference for the generalization error. *Mach Learn* 2003;**52**:239–81. https://doi.org/10.1023/A:1024068626366

Briefings in Bioinformatics, 2025, **26(2)**, bbaf116
https://doi.org/10.1093/bib/bbaf116
Review