Investigating and Assessing Diverse Strategies and Classification Techniques Applied in the Integration of Multi-Omics Data

Funda İPEKTEN[1,2], Necla KOÇHAN[3], Gözde Ertürk ZARARSIZ[1,4], Halef Okan DOĞAN[5], Vahap ELDEM[6], Gökmen ZARARSIZ[1,4]

[1]Department of Biostatistics, Faculty of Medicine, Erciyes University, Kayseri, Türkiye
[2]Department of Biostatistics, Faculty of Medicine, Adiyaman University, Adiyaman, Türkiye
[3]Department of Mathematics, Izmir University of Economics, Izmir, Türkiye
[4]Drug Application and Research Center (ERFARMA), Erciyes University, Kayseri, Türkiye
[5]Department of Biochemistry, Cumhuriyet University, Sivas, Türkiye
[6]Department of Biology, Faculty of Sciences, Istanbul University, Istanbul, Türkiye

High-throughput technologies have recently attracted much attention in academia. These technologies enable gathering information from various biological datasets, including genomics, transcriptomics, proteomics, and metabolomics data. Integrating and analyzing various datasets will allow us to accurately diagnose the disease, deepen our understanding of biological processes, and develop advanced treatment approaches. Numerous integration techniques have been developed for this aim, enabling researchers to unravel the complex patterns and underlying mechanisms behind biological occurrences, including those associated with illnesses. This study aims to explore and employ various data integration methodologies and machine-learning techniques on multi-omics data to predict disease classes using real datasets.

Keywords: Multi-omics data integration, machine-learning, system biology, precision medicine

1. INTRODUCTION

The emergence of advanced high-throughput technologies has enabled the acquisition of diverse omics data, including genomics, transcriptomics, proteomics, metabolomics, and epigenomics. Integrating these multiple layers of omics data has become a crucial focus in biomedical research, particularly in diagnostics, drug discovery, and precision medicine, because it plays a pivotal role in comprehending the underlying mechanisms of complex diseases.

Numerous studies highlight the significant impact of combining omics data in biomedical research (1–3). Integrating these omics data is essential to improving the accuracy of predictions, diagnosis, and prognosis within medical and clinical settings. For instance, Tao et al. showed that a multi-omics approach rather than single-omics data yields better results while classifying breast cancer patients (4). The integration also facilitates the analysis of complex diseases, such as cancer (5,6). Furthermore, it can help identify disease-associated biomarkers, facilitating the development of more targeted and personalized diagnostic and therapeutic strategies that can be found across multiple layers of omics data. Many articles cover the importance of integrating different omics datasets from different perspectives (7–9).

Data integration aims to create a meaningful representation of the biological information contained in each omics data preserving the information content in each data. However, it is not easy to integrate these diverse omics datasets as their high dimensionality and complexity lead to significant computational and statistical difficulties for the downstream analysis. Numerous algorithms have been proposed and developed to integrate diverse omics datasets effectively to address these challenges. These algorithms can be categorized as unsupervised and supervised. In the unsupervised approach, there are no class labels in the data. In multi-omics data integration using the unsupervised approach, one tries to discover hidden structures in the observed data. However, in the supervised approach, the class labels are known. Here, the aim is to develop learning rules by integrating different omics data and thus predict the class labels of new observations. The algorithms discussed in this paper are limited to supervised approaches only.

'Model-based integration', one of the methods used in integrating multi-omics datasets, involves compiling a training set from various data types, including genomics, proteomics, transcriptomics, and more. This method generates multiple models during the training phase and subsequently combines these models into a unified model incorporating interactions between different omics structures (10). One of the advantages of this method is that considerate considers the interaction between different omics structures associated with a specific disease or genotype. Additionally, it offers increased predictive power compared to the models built only on single omics data (8). Also, it is reliable and effective in capturing the complexity of the underlying biological system. On the other hand, the 'model-based integration' methods depend on the assumption about data distribution and violation of this assumption can lead to inaccurate results (11). Additionally, one cannot observe the variables

contributing to the final model. Therefore, it becomes challenging to identify the specific genes associated with the disease. The model-based integration strategy can amalgamate prediction models derived from various data kinds. This may facilitate the integration of data sets wherein each data type is sourced from distinct patient cohorts, although all patients share the same disease (10). Therefore, the training phase is performed independently for each -omics data. In this process, cross validation method can be used to reduce overfitting (12).

Another approach employed for integrating multi-omics datasets is 'concatenate-based integration'. This method combines measurements of individual omics structures into a multidimensional matrix before modeling, facilitating downstream analysis. An advantage of this approach is its simplicity once the determination of how variables will be combined in the matrix is made, allowing the use of various machine-learning methods for analyzing quantitative or categorical data. However, a challenge in data integration arises from variations in the number of biomarkers across omics datasets, given their diverse sizes. To address this issue, it is recommended that block scaling factors be applied to the omics datasets (11). This helps eliminate problems arising due to differences in the number of biomarkers in the model, ensuring that each biomarker carries equal weight in the analysis. Moreover, omics datasets having varying dimensions and are obtained from different technologies, exhibiting diverse structures, unexpected values, distinct noise distributions, and disparate variances (13). Managing these differences is crucial for successful concatenate-based integration and subsequent multi-omics data analysis.

The 'transformation-based integration' method is introduced to transform individual omics data into a common shared format or representation such as a graph or kernel matrix (14). Graph-based algorithms consider subjects as nodes and their relationships as edges whereas kernel-based algorithms generate a classifier in feature space. Recently, Yang et al. (15) used kernel fusion, which leverages the similarities between subjects, to integrate multilevel heterogeneous omics datasets. While the transformation aids in integrating datasets with varying scales and measurement units, facilitating their utilization in machine-learning algorithms for subsequent analysis, this approach has several drawbacks. One limitation is the possibility of information loss from the original datasets due to the employed transformation technique. In this method, the original data is converted into a new data matrix format where the diagonal elements are zero, and the off-diagonal elements display the correlations

between variables. However, issues such as singularity or a non-invertible variance-covariance matrix may lead to inaccuracies and/or challenges (16). Therefore, selecting an appropriate transformation method becomes pivotal, as it can significantly influence the outcomes of integrated analysis.

While data integration methodology plays a pivotal role in integrating diverse datasets and creating common models for comprehensive analyses, it is essential to highlight that integration itself may not be sufficient for robust disease diagnosis and treatment strategies. Beyond integration, the analysis of integrated omics data using powerful machine learning algorithms, such as Random Forest (RF), Nearest Shrunken Centroids (NSC), and Support Vector Machines (SVMs), is crucial for understanding the underlying mechanisms of diseases and for accurate disease classification and diagnosis. In 2015, Taskesen et al. showed that the results derived from integrated data surpass those obtained from single omics datasets (17). Specifically, integrating gene expression and DNA methylation datasets significantly enhanced the prediction accuracy of the known molecular subtype of acute myeloid leukemia.

However, to the best of our knowledge, there has not been a comprehensive study that systematically compares the performance of classification algorithms along with data integration methods. This gap in the existing literature highlights the need for further research to thoroughly evaluate and compare the effectiveness of various classification algorithms along with different data integration strategies. Therefore, in this study, we aimed to investigate and implement the three data integration approaches described earlier, along with machine-learning algorithms, on multi-omics data to forecast disease subtypes using real datasets. We first integrated multi-omics data employing various strategies. We then applied diverse machine-learning algorithms for further analysis. Specifically, Multiple Kernel Learning (MKL), NSC, RF, and SVM algorithms were selected for application to the integrated data, obtained through concatenate-based and model-based integration techniques. Furthermore, we explored the performance of Composite Association Network (CANetwork), Relevance Vector Machine (RVM) and Ada-boost RVM algorithms for the data integrated using a transformation-based integration approach.

2. MATERIALS and METHODS

2.1. Datasets

This study used three different real cancer data sets were used: colon, kidney, and thyroid. Colon cancer datasets: The LinkedOmics platform provided the colon cancer miRNA, proteome, and RNA-Seq datasets. The dataset has two distinct classes including colon cancer and non-colon with 20 and 75 class sizes, respectively. From 95 samples, miRNA 989, the proteome 8058, and RNA-Seq 13482 have features.

Kidney cancer datasets: The LinkedOmics platform provided the kidney cancer methylation, miRNA, and RNA-Seq datasets. The dataset has two distinct classes including kidney papillary renal cell carcinoma and kidney clear cell renal carcinoma with 73 and 16 class sizes, respectively. From 89 samples, methylation 13744, the miRNA 799, and RNA-Seq 20190 has features.

Thyroid cancer datasets: The LinkedOmics platform provided the kidney cancer methylation, miRNA, and RNA-Seq datasets. The dataset has two distinct classes including thyroid papillary carcinoma and not thyroid papillary carcinoma with 142 and 354 class sizes, respectively. From 496 samples, methylation 20118, the miRNA 808, and RNA-Seq 19927 have features.

Also information about these datasets is provided in **Fig. 2.1**. These datasets were downloaded from the web application resource “LinkedOmics”, developed by Vasaikar et al. in 2018 (18). LinkedOmics encompasses omics data and clinical information from 11,158 patients across 32 cancer types within the The Cancer Genome Atlas (TCGA) program/project. We used the 'reduce()' and 'lapply()' functions in the R programming language to organize data and extract matching samples for the same patient across individual omics datasets (e.g., transcriptomics, proteomics), ensuring consistency for further analysis. The data was split into 80% training set and 20% test set in order to be standard for the integration strategies and classification models to be described in the following subsections. All models were trained, and parameters were optimized on the training set and performance evaluation was performed on the test set.

2.2. Feature selection

In the classification problems, certain variables may significantly influence the outcome, while others may have negligible effects. Including variables without impact in the model can lead to unnecessary complexity and issues such as variable crowding. On the other hand, if feature selection is not done, processing such large data will require significant computational resources and time for accurate modelling and analysis.

Feature selection helps mitigate unnecessary complexity by identifying and including only the most relevant variables in the model. In our study, we employed the minimum Redundancy Maximum Relevance (mRMR) feature selection method (19). This method was chosen based on a literature review, and it is widely recognized for its applicability in pooled data scenarios (20,21).

The mRMR method is a feature selection method used to measure a variable's relevance to the corresponding class while considering redundancy with other variables in the dataset (19). Assume that the dataset has two class labels such as -1 and 1. Also, assume that S^* is the final feature set. Then, the mRMR method aims to find the best S^* with the minimum redundancy maximum correlation level solving the following formula:

$$\max_{S^*} \frac{\sum_{i \in S^*, h \in \{+1,-1\}} I(h,i)}{\frac{1}{|S^*|}\sum_{i,j \in S^*} I(j,i)} \qquad (1)$$

where $I(j,i)$ and $I(h,i)$ shows the mutual information between the j^{th} and i^{th} variables and h^{th} and i^{th} variables, respectively. To enhance computational performance, the mRMR method employs a recursive and additive approach to construct the final S^* by transforming Eq. (1) into the following equation:

$$\max_{i \in \Omega_{S^*}} \frac{I(+1,i) + I(-1,i)}{\frac{1}{|S^*|}\sum_{i,j \in S^*} I(i,j)} \qquad (2)$$

where Ω is the whole variable set and $\Omega_S = \Omega - S$. See (16) for more information on the mRMR method.

2.3. Dimension reduction

Despite the utilization of feature selection methods, there is still a possibility of having a high number of variables, potentially leading to increased complexity and challenges in interpretation. One approach to address these challenges is implementing Principal Component Analysis (PCA) after performing feature selection. PCA reduces the number of variables in a dataset while retaining as much of the variability in the data as possible. The first principal component accounts for the maximum variance in the data. Subsequent components capture the remaining variance in decreasing order. PCA was performed using the "*prcomp()*" function in R.

2.4. Integration methods

In this study, we focused on three distinct integration methods: (i) 'model-based integration' methods, (ii) 'concatenate-based integration' methods, and (iii) 'transformation-based integration' methods. Each approach offers a unique way to combine and analyze multi-omics datasets.

The 'model-based integration' method generated individual prediction models using each dataset. Subsequently, a final forecasting model was constructed by combining and utilizing the outcomes derived from these individual estimation models (8). We employed the **caret** package in R for the 'model-based integration' method.

The 'concatenate-based' method involves merging three different datasets into a single matrix. However, using this combined matrix directly for classification could lead to inaccuracies. This is because the data, when directly merged into a single matrix, carries the risk of unequal weighting of variables due to varying variable sizes in each dataset. As Spicker et al. proposed, block scaling was implemented according to each dataset to address this issue (13). The block scaling was performed using the "*blockscal()*" function in the **rnirs** package. In this process, the test matrix was derived by scaling the block scaling through the square root of the sum of variances for each column within the block of the reference matrix. This approach ensured that the variables used in the model carried equal weight, overcoming the challenge posed by different variable sizes in the original datasets. The subsequent classification analyses were conducted using this appropriately scaled test matrix.

The 'transformation-based integration' is an approach that combines multiple datasets after transforming the dataset into a shared format, such as a graph or kernel matrix, before creating the model. This approach involves a mapping or data transformation tailored to the

data type, aiming to retain the data type-specific features (10). In the ‘transformation-based integration’ process, the distance between datasets was calculated using the "*gausskernel()*" function in the **KRLS** package.

Various strategies were explored in the downstream analysis for the TCGA datasets obtained from “LinkedOmics”. These strategies were investigated under two main categories: strategies tailored for single omics data and strategies falling under integration methods. Before applying the strategies mentioned in the next step, the data set should be pre-processed with methods specific to the relevant -omics data. These data preprocessing steps may include steps such as normalization, data transformation, etc. Therefore, before applying the following models, the data should be normalized and transformed into a data set which is ready for classification modelling. For instance, in RNA-sequencing data, systematic variations in the data can be removed with methods such as DESeq2, TMM (Trimmed Mean of M-values) normalization. In addition, data can be transformed by methods such as voom transformation, vst (Variance Stabilizing Transformation) or rlog transformation. Details on preprocessing steps in omics data can be found in (23–26).

The strategies for single omics data **(Fig. 2.2)** can be further investigated in three scenarios:

i. *raw data:* This involves considering raw omics datasets separately. Subsequently, machine-learning algorithms were applied to these datasets. What is meant by raw data here is that no data integration has been made. We would like to state that this should be a data set that has passed preprocessing steps such as data normalization and data transformation and is ready for modelling.*raw data + feature selection:* In this scenario, the mRMR approach, the feature selection method, was applied to raw omics data prior to implementing machine-learning algorithms.
ii. *raw data + feature selection + dimension reduction:* Here, dimension reduction is employed following mRMR feature selection. Machine-learning algorithms were then implemented to compare the classification performances.

The strategies falling under integration methods investigated in the following scenarios:

i. *integrated raw data:* This scenario entails integrating raw omics datasets and implementing machine-learning algorithms **(Fig. 2.3)**.

ii. *integrated raw data + feature selection:* Raw omics datasets were initially integrated, followed by the application of the mRMR feature selection and machine-learning algorithms, respectively **(Fig. 2.3)**.

iii. *integrated raw data + feature selection + dimension reduction:* In this scenario, dimension reduction is applied in addition to the mRMR feature selection method. Machine-learning algorithms are then applied **(Fig. 2.3)**.

iv. *integrated data with applied feature selection method:* Here, feature selection, mRMR, is initially applied to raw single omics data separately. Then, these datasets are integrated and used in the classification process **(Fig. 2.4)**. We note here that only the 'concatenate-based integration' method is employed within this strategy.

v. *integrated data with applied feature selection method + dimension reduction:* In this scenario, after integrating the data that have undergone the feature selection method, precisely the mRMR method, PCA is performed. Machine-learning is then implemented as the ultimatestep **(Fig. 2.4)**. Similar to the fourth strategy, only the 'concatenate-based integration' method is employed within this scenario.

2.5. Classification Methods

Following integrating multi-omics datasets, we proceeded the downstream analysis using machine-learning algorithms. Our approach involved implementing several machine-learning algorithms, including MKL, SVM, NSC, RF, CANetwork, RVM, and Ada-boost RVM. We note here that MKL, SVM, NSC and RF algorithms were applied to model-based and concatenate-based integrated data whereas CANetwork, RVM, and Ada-boost RVM algorithms were used for the data combined via transformation-based integration technique (14,19,27). SVM is a supervised machine-learning algorithm widely used for classification and regression tasks (28). Classification aims to find a hyperplane (a.k.a decision boundary) that best separates different classes (groups) of data points. We employed the "*gausskernel()"* function in our analyses.

MKL is a technique used to enhance SVM by incorporating multiple kernels (29). Kernels define the similarity between data points and can be linear or nonlinear. MKL allows for selecting and combining different kernels to capture complex relationships within the data. By learning the weights or coefficients associated with each kernel, MKL enables the algorithm to adaptively emphasize or de-emphasize the contribution of each kernel for

different regions of the input space. The advantages of MKL include the ability to handle heterogeneous data sources, capture diverse patterns, and improve the generalization performance of machine-learning models. Cross-validation is implemented to optimize the parameters of the MKL method.

NSC is a method that calculates class centroids (i.e., the center of gravity of each class), shrinks them towards zero (i.e., Δ as described in the original article by Tibshirani et al. (30)), and automatically performs feature selection without requiring further adjustments. This makes NSC a useful tool for both dimensionality reduction and classification tasks (30). We used the **pamr** package in R for our analysis.

RF, known as ensemble learning, is one of the frequently used machine-learning algorithms as it is to implement. Moreover, it is a powerful and robust classification algorithm with two parameters to be optimized: the number of variables used in each node and the number of trees to be generated (32). To implement RF, we used the **caret** package.

RVM and Ada-boost RVM are kernel-based machine-learning algorithms. These approaches use kernel functions to map input data into a higher-dimensional space, capturing complex relationships within the data (14). The parameters, resampling size and the maximum number of iterations, were fine-tuned using cross-validation. The training of the models was executed using the "*ParaRVM()"* function, and the integration of omics datasets was achieved through the utilization of the *"RVMInt()"* function from the **MDIntegration** package.

The CANetwork graph-based method is another machine-learning approach used for high-dimensional data. This approach initially estimates the correlation matrices. Then, it uses the function *AFcorMI()* from the **WGCNA** package, which computes a predicted weighted mutual information adjacency matrix from a given correlation matrix (14). Ultimately, the function *CANetwork()* from **MDIntegration** package integrates weighted mutual information adjacency correlation matrices obtained for each dataset. In the classification step, the classification models were constructed independently, employing various machine-learning algorithms on the training data. In the training step, specific parameters for each classification method were fine-tuned to optimize their performance. Furthermore, to ensure robustness and validity, the models underwent a resampling process five times, and the data were cross-validated five times within each of the ten resampled datasets. The performance of

the models was ultimately assessed using the test data based on the metrics elaborated in Subsection 2.6.

All analyses were conducted using the R programming language (version 3.6.2) on the Centos Enterprise 7.3 Linux operating system. The computational resources utilized for these analyses were part of the TRUBA resources available at the TÜBİTAK ULAKBİM High Performance and Grid Computing Center.

2.6. Model Performance Evaluation

The Area Under Curve (AUC), accuracy, F1-score, and Matthew Correlation Coefficient (MCC) were used to evaluate the performance of classification methods. These metrics are calculated by comparing the predicted disease status with the actual disease status in the dataset (**Table 2.1**).

Table 2.1. Confusion Matrix

		TRUE CLASS	
PREDICTED CLASS		**Positive**	**Negative**
	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

AUC (Area Under the Curve) is commonly used performance metric in classification problems to measure the performance of a classification model. It is the area under the ROC (Receiver Operating Characteristic) curve illustrating the trade-off between sensitivity (true positive rate) and 1-specificity (false positive rate). Additionally, accuracy and F1-score offer quantitative measures that capture various aspects of model performance, ensuring a comprehensive evaluation. Also, MCC (Matthews Correlation Coefficient) is a powerful performance measure, especially for imbalanced data sets in binary classification problems. MCC is calculated by taking into account the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values. MCC takes a value between -1 and +1. +1 indicates perfect classification, -1 indicates completely incorrect classification, and 0 indicates random guess. The calculation for the MCC measure is as follows (33) (Table 2.2). (**Table 2.2**).

Table 2.2. Model performance metrics

Metrics	Formula
Sensitivity	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$
Specificity	$\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$
Accuracy	$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$
F1-score	$2*\left(\frac{\mathrm{TP}}{2}*\mathrm{TP}+\mathrm{FP}+\mathrm{FN}\right)$
MCC	$(\mathrm{TP}*\mathrm{TN})-\frac{\mathrm{FP}*\mathrm{FN}}{\mathrm{TP}+\mathrm{FP}}*(\mathrm{TP}+\mathrm{FN})*(\mathrm{FP}+\mathrm{TN})*(\mathrm{FN}+\mathrm{TN})$
AUC	$\int_0^1 TPR(x)d(FPR(x))$

MCC: Matthew Correlation Coefficients, AUC: Area Under the Curve, TPR: True Positive Rate (sensitivity), FPR: False Positive Rate (1-specificity)

3. Results for real datasets

The performance of the models under different scenarios is detailed separately, along with the corresponding processing times for each model. Comprehensive results for all real datasets are available in the **Supplementary File.**

3.1. Results for kidney cancer dataset

The results for the kidney cancer dataset revealed that, when considering the raw data, the performance metrics of SVM methods applied to methylation and miRNA raw data were found to surpass those of NSC and RF methods (**Table 3.1**). However, in RNA-Seq raw data, the RF method's performance was superior to that of the NSC and SVM methods. The SVM method exhibited a higher classification performance when examining concatenated-based integrated data than the NSC and RF methods. In model-based integrated data, the NSC method's classification performance was higher than that of the RF and SVM methods. Notably, the NSC method applied to model-based integrated data achieved the highest performance among the integration methods. When we use the 'transformation-based integration' method, the highest performance was obtained with RVM (**Table 3.2**).

Table 3.1. The results of NSC, RF, and SVM methods on raw kidney cancer dataset

Performance metrics	Raw data			Concatenate-based integrated raw data	Model-based integrated raw data
	Methylation	miRNA	RNA-Seq	Methylation + miRNA+ RNA-Seq	Methylation + miRNA+ RNA-Seq
NSC					
AUC	0.919(0.007)	0.941(0.006)	0.910(0.007)	0.915(0.011)	**0.987(0.009)**
Accuracy	0.955(0.004)	0.969(0.003)	0.941(0.004)	0.951(0.006)	**0.995(0.003)**
F1-score	0.972(0.002)	0.981(0.002)	0.962(0.003)	0.969(0.004)	**0.997(0.002)**
MCC score	0.857(0.012)	0.903(0.010)	0.823(0.012)	0.845(0.019)	**0.983(0.012)**
RF					
AUC	0.883(0.008)	0.909(0.008)	**0.929(0.008)**	0.890(0.013)	0.947(0.016)
Accuracy	0.948(0.003)	0.958(0.004)	**0.972(0.003)**	0.958(0.005)	0.981(0.006)
F1-score	0.969(0.002)	0.974(0.002)	**0.983(0.002)**	0.974(0.003)	0.989(0.003)
MCC score	0.817(0.014)	0.863(0.012)	**0.906(0.010)**	0.857(0.017)	0.932(0.020)
SVM					
AUC	**0.922(0.003)**	**0.971(0.002)**	0.926(0.003)	**0.927(0.006)**	0.960(0.015)
Accuracy	**0.960(0.001)**	**0.988(0.001)**	0.967(0.001)	**0.969(0.003)**	0.986(0.005)
F1-score	**0.975(0.001)**	**0.993(0.001)**	0.980(0.001)	**0.981(0.002)**	0.992(0.003)
MCC score	**0.873(0.005)**	**0.962(0.003)**	0.888(0.005)	**0.900(0.009)**	0.949(0.018)

The number of features is given as mean(standard error). AUC: Area Under the Curve, MCC: Matthew Correlation Coefficient, NSC: Nearest Shrunken Centroids, RF: Random Forest, SVM: Support Vector Machine. Bold values indicate the best performances.

Table 3.1. The results of RVM, Ada-boost, and CANetwork methods on transformation-based integrated kidney cancer dataset

Performance metrics	**Kidney Cancer Dataset**
	Methylation + miRNA+ RNA-Seq
RVM	
AUC	**0.927(0.022)**
Accuracy	**0.974(0.008)**
F1-score	**0.985(0.004)**
MCC score	**0.904(0.029)**
Ada-boost RVM	
AUC	0.885(0.023)
Accuracy	0.958(0.008)
F1-score	0.975(0.005)
MCC score	0.845(0.030)
CANetwork	
AUC	0.746(0.030)
Accuracy	0.875(0.015)
F1-score	0.926(0.009)
MCC score	0.537(0.060)

The number of features is given as mean(standard error). AUC: Area Under the Curve, MCC: Matthew Correlation Coefficient, RVM: Relevance Vector Machine, Ada-boost RVM: Adaptive-boosting Relevance Vector Machine, CANetwork: Composite Association Network. Bold value indicates the best performance.

'**Table 3.3** presents the results obtained from applying mRMR feature selection to the kidney cancer dataset. It can be seen from **Table 3.3** that the best performance was obtained when MKL classification was applied to the concatenated-based integrated data, and mRMR feature selection was implemented after the integration. The performance of the SVM method was observed to be higher than the performance of MKL and RF methods when single miRNA and RNA-Seq data were utilized. On the other hand, the performance of the MKL method was observed to be higher than the performance of RF and SVM methods when methylation data was used. For concatenated-based integrated data, the classification performance of the MKL method was superior to the classification performance of RF and SVM methods. In the model-based integrated data, the SVM performed better than the RF. For the data where first mRMR and then the integration method is applied, the performance of the MKL was higher than that of RF and SVM methods.

Table 3.4 provides the results after applying mRMR and PCA to the kidney. The highest performance was achieved by applying SVM to the model-based integrated data with mRMR and PCA.

Table 3.2. The results of MKL, RF, and SVM methods on FS applied kidney cancer dataset across different integration methods

Performance metrics	**FS applied single data**			**FS applied concatenate-based integrated data**	**FS applied model-based integrated data**	**Integrated data after applying FS to each single omics data**
	Methylation	miRNA	RNA-Seq	Methylation + miRNA+ RNA-Seq	Methylation + miRNA+ RNA-Seq	Methylation + miRNA+ RNA-Seq
MKL						
AUC	**0.916(0.019)**	0.913(0.004)	0.802(0.006)	**1.000(0.000)**	**-**	**0.940(0.011)**
Accuracy	**0.965(0.007)**	0.965(0.002)	0.911(0.003)	**1.000(0.000)**	**-**	**0.974(0.005)**
F1-score	**0.978(0.004)**	0.978(0.001)	0.947(0.002)	**1.000(0.000)**	**-**	**0.984(0.026)**
MCC score	**0.884(0.024)**	0.885(0.005)	0.703(0.010)	**1.000(0.000)**	**-**	**0.921(0.012)**
RF						
AUC	0.853(0.005)	0.897(0.005)	0.720(0.006)	0.841(0.010)	0.953(0.018)	0.893(0.009)
Accuracy	0.930(0.002)	0.958(0.002)	0.876(0.003)	0.919(0.005)	0.984(0.006)	0.953(0.004)
F1-score	0.957(0.001)	0.975(0.001)	0.927(0.002)	0.950(0.003)	0.990(0.004)	0.971(0.003)
MCC score	0.757(0.008)	0.859(0.006)	0.547(0.012)	0.725(0.017)	0.939(0.024)	0.848(0.013)
SVM						
AUC	0.867(0.003)	**0.981(0.001)**	**0.946(0.003)**	0.901(0.006)	**0.967(0.014)**	0.927(0.006)
Accuracy	0.939(0.001)	**0.986(0.001)**	**0.976(0.001)**	0.932(0.004)	**0.988(0.005)**	0.969(0.003)
F1-score	0.963(0.001)	**0.991(0.001)**	**0.985(0.001)**	0.957(0.002)	**0.993(0.003)**	0.981(0.002)
MCC score	0.785(0.005)	**0.961(0.002)**	**0.925(0.004)**	0.779(0.012)	**0.958(0.017)**	0.900(0.009)

The number of features is given as mean(standard error). FS: Feature Selection, PCA: Principal Component Analysis, AUC: Area Under the Curve, MCC: Matthew Correlation Coefficient, MKL: Multiple Kernel Learning, RF: Random Forest, SVM: Support Vector Machine. Bold values indicate the best performances. Since the MKL method was not included in the available packages we implemented in this study, we excluded it for model-based integrated datasets.

Table 3.3. The results of MKL, RF, and SVM methods on FS and PCA applied kidney cancer dataset across different integration methods

Performance metrics	**FS+PCA applied single omics data**			**FS+PCA applied concatenate-based integrated data**	**FS+PCA applied model-based integrated data**	**PCA applied concatenate-based integrated data, applying FS to each single omics data before the integration**
	Methylation	miRNA	RNA-Seq	Methylation + miRNA+ RNA-Seq	Methylation + miRNA+ RNA-Seq	Methylation + miRNA+ RNA-Seq
MKL						
AUC	**0.916(0.004)**	0.909(0.004)	0.752(0.006)	**0.926(0.004)**	**-**	**0.919(0.004)**
Accuracy	**0.965(0.001)**	0.960(0.002)	0.892(0.003)	**0.967(0.002)**	**-**	**0.965(0.002)**
F1-score	**0.978(0.001)**	0.975(0.001)	0.937(0.002)	**0.980(0.001)**	**-**	**0.978(0.001)**
MCC score	**0.884(0.005**	0.868(0.005)	0.627(0.010)	**0.888(0.006)**	**-**	**0.886(0.005)**
RF						
AUC	0.845(0.006)	0.898(0.006)	0.734(0.005)	0.853(0.006)	0.933(0.019)	0.807(0.007)
Accuracy	0.940(0.002)	0.965(0.002)	0.884(0.003)	0.940(0.002)	0.976(0.007)	0.912(0.003)
F1-score	0.965(0.001)	0.980(0.001)	0.931(0.002)	0.964(0.001)	0.986(0.004)	0.948(0.002)
MCC score	0.775(0.010)	0.857(0.008)	0.584(0.009)	0.788(0.009)	0.914(0.025)	0.683(0.012)
SVM						
AUC	0.856(0.003)	**0.952(0.002)**	**0.834(0.005)**	0.829(0.003)	**0.950(0.015)**	0.911(0.004)
Accuracy	0.894(0.002)	**0.965(0.001)**	**0.930(0.002)**	0.896(0.002)	**0.979(0.006)**	0.947(0.002)
F1-score	0.933(0.001)	**0.978(0.001)**	**0.958(0.001)**	0.936(0.001)	**0.987(0.003)**	0.966(0.001)
MCC score	0.680(0.007)	**0.887(0.005)**	**0.760(0.007)**	0.662(0.007)	**0.928(0.020)**	0.831(0.007)

The number of features is given as mean(standard error). FS: Feature Selection, PCA: Principal Component Analysis, AUC: Area Under the Curve, MCC: Matthew Correlation Coefficient, MKL: Multiple Kernel Learning, RF: Random Forest, SVM: Support Vector Machine. Bold values indicate the best performance. Since the MKL method was not included in the available packages we implement in this study, we exclude it for model-based integrated datasets.

3.2. Results for thyroid cancer dataset

According to the results of thyroid cancer dataset, the performance of NSC method used in miRNA and RNA-Seq raw data was higher than that of RF and SVM methods. On the other hand, the performance of the SVM method was observed to be higher than that of NSC and RF methods in methylation raw data (**Supplementary File, Table S1)**. For concatenated-based integrated data, the classification performance of the NSC method was higher than the classification performance of RF and SVM methods. Similarly, when the model-based integrated data was used in the classification, the performance of the NSC method was better than that of RF and SVM methods. If omics data were integrated using a 'transformation-based integration' method, CANetwork was identified as the most effective classification algorithm, demonstrating the highest classification performance (**Supplementary File, Table S2)**. The results after applying mRMR feature selection for the thyroid cancer dataset are given in **Supplementary File, Table S3**. Like in the kidney cancer dataset, the best performance was obtained when MKL classification was applied to the concatenate-based integrated data where mRMR feature selection was implemented after the integration. The performance of the MKL method was observed to be higher than the performance of the RF and SVM methods when single miRNA and methylation data were utilized.

When concatenated-based integrated data was used, the classification performance of the SVM method was superior to the classification performance of MKL and RF methods. Similar results were obtained in the kidney cancer dataset when the data were integrated using 'model-based integration', where SVM outperformed RF. Additionally, for the data where mRMR was applied first and then the integration method, the performance of MKL was found to be higher than that of the RF and SVM methods.

After applying mRMR and PCA for kidney cancer, the results are presented in **Supplementary File, Table S4**. When each omics data is first processed through mRMR feature selection, followed by PCA, and finally integrated, the MKL classification algorithm achieves the highest performance with these integrated data.

3.3. Results for colon cancer dataset

Table S5 illustrates results for the colon cancer dataset when analyzing raw data across different strategies. The NSC method exhibited the lowest performance when raw miRNA was employed, whereas the SVM demonstrated the highest performance when utilizing either raw proteome or raw RNA-Seq data. However, the SVM method performed best in concatenate-based integrated data. In contrast, the RF method was identified as the most efficient approach in the model-based integrated data.

In the case of transformation-based integrated data, the CANetwork method yielded the lowest performance, while the AdaBoost RVM method demonstrated a slightly higher performance than RVM (**Supplementary File, Table S6)**.

The results for various scenarios after applying mRMR feature selection are outlined in **Supplementary File, Table S7**. The findings indicated that, following the application of mRMR feature selection, the RF approach exhibited superior performance compared to the MKL and SVM approaches in miRNA data. In contrast, for proteome and RNA-Seq data, the MKL and SVM methods were identified as the most effective approaches, respectively. When integrating omics datasets with the mRMR feature selection method implemented prior to integration, the performance of RF was superior to other approaches.

Table S8 summarizes the results for various scenarios after applying mRMR feature selection and PCA. SVM outperformed other methods when the data was integrated with a model-based approach followed by mRMR feature selection and PCA processes.

3.4. Processing times and the results for the feature selection method

We investigated processing (or running) times using various methods across different scenarios. Detailed results are available in **Supplementary File, Table S9-S12**.

In the single omics data of colon, kidney, and thyroid cancers, the NSC method demonstrated superior computational efficiency, making it the most efficient method compared to other methods (**Supplementary File, Table S9**). Conversely, after implementing the feature selection and PCA, SVM emerged as the top-performing method in terms of computational cost. SVM was the most efficient algorithm after integrating the datasets using concatenate-

based and model-based integration techniques following feature selection and PCA (**Supplementary File, Table S9**).

For the concatenate-based integrated datasets after applying feature selection to each single omics cancer dataset, the RF method was found to be the fastest algorithm for colon and kidney cancer datasets. In contrast, for the thyroid cancer dataset, SVM was observed to be the most efficient algorithm in terms of computational cost (**Supplementary File, Table S10**). After applying mRMR feature selection to single omics datasets followed by concatenate-based integration, the least costly method for colon, kidney, and thyroid cancer datasets after applying PCA was observed to be the SVM method. See **Supplementary File, Table S10**). For the transformation-based integrated datasets, the least computational cost for colon, kidney and thyroid cancer datasets was achieved by RVM, Ada-boost RVM, and CANetwork, respectively (**Supplementary File, Table S11**).

In the model-based integrated datasets of colon, kidney, and thyroid cancer, NSC surpassed other methods in computational cost efficiency. Nevertheless, after applying feature selection and PCA to the model-based integrated data, SVM emerged as the most efficient algorithm for all cancer datasets. See **Supplementary File, Table S12** for details.

Additionally, we investigated the number of variables determined in the feature selection methods across various strategies, and these findings are documented in **Supplementary File, Table S13-S14**. In RF and SVM algorithms, we did not perform any feature selection process; hence, the number of variables used in these models was the same for all cancer datasets. However, feature selection process is conducted in the NSC algorithm, determining the most appropriate number of variables for each dataset and incorporating them into the model. The number of features selected across different strategies are presented in **Supplementary File, Table S13**. Due to the feature selection step in the NSC algorithm, the data subjected to mRMR feature selection and PCA are not included in the analyses conducted by the NSC method. In concatenate-based integrated data, the number of features used in the NSC method was found to be lower than that of the RF and SVM methods.

4. DISCUSSION

Integrating multi-omics data and applying machine-learning techniques are crucial in advancing our understanding of complex biological systems, particularly in cancer diagnosis and precision medicine. In this study, we conducted a comprehensive comparison study of three integration methods in conjunction with powerful machine-learning algorithms on real datasets. Thus, we aimed to provide insights into how omics datasets should be assessed, highlighting the significance of evaluating them separately and in an integrated fashion. We also compared each strategy in terms of computational time. Feldner-Busztin et al. proposed an end-to-end pipeline for machine learning-based subgroup identification in non-small cell lung cancer (NSCLC) in their study (34). They also proposed and validated fusion-based classification models to identify subgroups in new samples. Unlike this study, we also demonstrated model-based and graph-based methods. We aimed to demonstrate the effectiveness of various integration methods in subgroup classifications. By comparing the integration and classification methods used in our study and the information obtained from related studies in the literature, each researcher will be able to easily choose the appropriate scenario for their own dataset (35).

Recently, Chierici et al. (36) investigated the performance of RF and SVM methods using concatenate-based integrated data instead of single omics data. Their results indicated similar outcomes for both RF and SVM methods across each strategy. As per their findings, the classification methods employed to predict estrogen receptor status, invasive subtypes of breast carcinoma, and survival in renal clear cell carcinomas demonstrated robust performance when applied to integrated data. However, for myeloid leukemia, they observed that the performance of the classification methods was superior when utilizing single omics data (36). In our study, we reached similar outcomes. For instance, we observed that the performance of the concatenate-based integrated colon cancer dataset surpassed that of single miRNA and RNA-Seq colon cancer data in the SVM method **(Supplementary File, Table S5)**. Similarly, the performance of the NSC method, when omics data were integrated using a concatenate-based approach, outperformed the same method when only the RNA-Seq kidney cancer dataset was preferred **(Table 3.2).** This aligns with the observations made by Chierici et al. study (36) and supports the notion that the choice of integration method can impact the performance of classification methods in omics data analysis.

In another study, Ma et al. (37) implemented the ‘concatenate-based integration’ method to integrate omics datasets to predict breast cancer diagnosis across different strategies. They emphasized that the analyses on integrated data were more informative. They followed different strategies: (i) trained the model by integrating the data in its raw form, (ii) trained the model after integrating the data in its raw form and applying variable selection methods, (iii) applied variable selection methods before integrating the data, and then integrated data followed by training the model. They concluded that there was no superiority among these strategies. In our study, we observed that the performances of RF and SVM methods for concatenate-based integrated colon, kidney, and thyroid cancer datasets were close, indicating no significant difference in terms of superiority in applying variable selection before or after integration. The classification performances of concatenate-based data and data with variable selection applied in concatenate-based could be either higher or lower than the classification performances of single data.

Multi-omics integration methods typically rely on statistical and conventional machine-learning techniques, such as iClusterBayes (38), NEMO (39), Multiple Similarity Network Embedding (MSNE) (40), and weight-boosted Multi-Kernel Learning (wMKL) (41). However, dealing with the high dimensionality of omics datasets and substantial variations in the number of genomic features among different modalities can pose challenges. The use of machine-learning models may encounter the "curse of dimensionality", complicating the analysis of these datasets. Therefore, employing feature selection and/or dimension reduction techniques may be beneficial before the classification process. For instance, Zhang et al. (19) focused on the mRMR feature selection method to predict the diagnosis of glioblastoma multiforme. The authors obtained a dataset comprising a total of 128 variables, including 71 gene expression, 50 gene methylation, 3 miRNAs, and 4 copy number variations. Classification algorithms were applied to the integrated dataset using the MKL method. Their results demonstrated that, compared to other statistical techniques, MKL in the integrated data could enhance the accuracy of glioblastoma multiforme diagnosis. Like Zhang et al. study, we employed the mRMR feature selection method together with integrating the methylation, miRNA, and RNA-Seq data separately for the colon and kidney cancer datasets, and then implemented the MKL classification algorithm. The results showed that the MKL method in integrated data could enhance the accuracy of each colon and kidney cancer diagnosis compared to other methods. However, while the results for miRNA data were consistent with the literature in the thyroid cancer dataset, the results for integrated

methylation data were differed. Based on our findings, we conclude that the MKL method for integrated data improves disease classification performance overall. However, it is important to note that the performance of this method may not constantly improve due to potential issues arising from preprocessing steps, as data obtained from innovative technologies can be noisy and require preprocessing stages such as normalization and filtering. Thus, conducting assorted studies with diverse data and evaluating preliminary results while considering these preprocessing steps is critical for future research.

Others have developed 'transformation-based integration' techniques to integrate the omics datasets (14). The kernel-based classification approach was found to be efficient in detecting the nonlinear relationships between samples (14). For colon and kidney cancer datasets, we found that the kernel-based integration method's performance yielded better results than the graphical-based integration method, which is consistent with results in the literature. However, we observed a slightly higher performance with the graphical-based method for thyroid cancer dataset.

Despite the availability of powerful and sophisticated methodologies, there is still an urgent need for appropriate techniques to harness the potential of large high-throughput datasets (42). For instance, Cao et al. (41) have introduced a new perspective with their developed method, weight-boosted Multi-Kernel Learning (wMKL). Recently, there has been a shift towards addressing this issue by incorporating deep learning approaches into multi-omics integration methods (5,43). Deep learning techniques may offer a promising solution to overcome the challenges posed by the complex nature of multi-omics data, providing a more effective means of extracting meaningful patterns and insights from such high-dimensional and diverse datasets.

In addition, although many powerful and advanced statistical learning methods have been developed for multi-omics data, class imbalance, commonly seen in these methods, is an inevitable problem. There are resampling approaches such as up-down-smote-rose as a solution to this problem. However, few studies consider the class imbalance situation in multi-omics data. In the study of Yang et al., the class imbalance problem was solved using methods such as synthetic minority oversampling technique and random under-sampling (44) . In the study of Novoloaca et al., the class imbalance problem was addressed using the MCC measure. We also considered this problem by using the MCC measure similarly (45).

In this study, we considered cases where there are only two categories of classes/groups. Nevertheless, this can extend to situations involving multiple classes, increasing the research's applicability and impact. The study included algorithms related to the classification problem and conducted a comprehensive comparison. We believe that a study comparing unsupervised algorithms such as iClusterBayes, NEMO, and MSNE in a similar scope will contribute to the literature.

5. CONCLUSION

The performances of integration methods along with machine-learning algorithms can exhibit variability across datasets under different scenarios (i.e., raw data, raw data with feature selection applied, etc.). The results showed that the classification models perform relatively better when 'concatenate-based' and 'model-based integration' methods are used for data integration than when 'transformation-based integration' methods are. Furthermore, these two methods yield better results after applying feature selection techniques. On the other hand, considering computational efficiency, the MKL method, which outperforms other classification methods and is relatively faster, can be leveraged after applying feature selection to concatenate-based integrated data.

In situations where the importance of variables is not considered, a 'model-based integration' method can be preferred due to its superior classification performance. Nonetheless, it is recommended that more expensive experimentation be conducted to yield findings of broader applicability.

Funding: This study was supported by the Research Fund of Erciyes University [TYL-2019-9600]. The funders had no role in study design, data collection and analysis, for decision to publish or prepare the manuscript.

6. REFERENCES

1. Shruthi BS, Vinodhkumar P. Proteomics: A new perspective for cancer. Adv Biomed Res. 2016;5(1):67.
2. Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Brief Bioinform. 2018;19(2):286–302.
3. Wörheide MA, Krumsiek J, Kastenmüller G, Arnold M. Multi-omics integration in biomedical research–A metabolomics-centric review. Anal Chim Acta. 2021;1141:144–62.
4. Tao M, Song T, Du W, Han S, Zuo C, Li Y, et al. Classifying breast cancer subtypes using multiple kernel learning based on omics data. Genes (Basel). 2019;10(3):200.
5. Gong P, Cheng L, Zhang Z, Meng A, Li E, Chen J, et al. Multi-omics integration method based on attention deep learning network for biomedical data classification. Comput Methods Programs Biomed. 2023 Apr 1;231:107377.
6. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights. 2020;14:1177932219899051.
7. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biology. 2017;18:1-15.
8. Lin E, Lane HY. Machine learning and systems genomics approaches for multi-omics data. Biomark Res. 2017;5:1–6.
9. Cantafio MEG, Grillone K, Caracciolo D, Scionti F, Arbitrio M, Barbieri V, et al. From single level analysis to multi-omics integrative approaches: a powerful strategy towards the precision oncology. Road from Nanomedicine to Precis Med. 2019;82969.
10. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet. 2015;16(2):85–97.
11. Cavill R, Jennen D, Kleinjans J, Briedé JJ. Transcriptomic and metabolomic data integration. Brief Bioinform. 2016;17(5):891–901.
12. Kim D, Li R, Dudek SM, Ritchie MD. ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. BioData Min. 2013;6(1):1–14.
13. Spicker JS, Brunak S, Frederiksen KS, Toft H. Integration of clinical chemistry, expression, and metabolite data leads to better toxicological class separation. Toxicol Sci. 2008;102(2):444–54.
14. Yan KK, Zhao H, Pang H. A comparison of graph- and kernel-based -omics data integration algorithms for classifying complex traits. BMC Bioinformatics. 2017 Dec;18(1):1-13.
15. Yang H, Cao H, He T, Wang T, Cui Y. Multilevel heterogeneous omics data integration with kernel fusion. Brief Bioinform. 2020 Jan 17;21(1):156–70.
16. Alpar R. Uygulamalı Çok Değişkenli İstatistiksel Yöntemler [Internet]. Detay; 2017.
17. Taskesen E, Babaei S, Reinders MMJ, de Ridder J. Integration of gene expression and DNA-methylation profiles improves molecular subtype classification in acute myeloid leukemia. BMC Bioinformatics. 2015;16(4):1–8.
18. Vasaikar S V, Straub P, Wang J, Zhang B. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Res. 2018;46(D1):D956–D63.
19. Zhang Y, Li A, Peng C, Wang M. Improve Glioblastoma Multiforme Prognosis Prediction by Using Feature Selection and Multiple Kernel Learning. IEEE/ACM Trans Comput Biol Bioinforma. 2016 Sep 1;13(5):825–35.
20. El-Manzalawy Y, Hsieh TY, Shivakumar M, Kim D, Honavar V. Min-redundancy and

max-relevance multi-view feature selection for predicting ovarian cancer survival using multi-omics data 06 Biological Sciences 0604 Genetics. BMC Med Genomics. 2018;11:19-31.
21. Mallik S, Bhadra T, Maulik U. Identifying Epigenetic Biomarkers using Maximal Relevance and Minimal Redundancy Based Feature Selection for Multi-Omics Data. IEEE Trans Nanobioscience. 2017;16(1):3-10.
23. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):1–21.
24. Robinson MD, Oshlack A. Deseq2论文附录. Genome Biol. 2010;11(3):1–9.
25. Law CW, Chen Y, Shi W, Smyth GK. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):1–17.
26. Nadler SG, Tritschler D, Haffar OK, Blake J, Bruce AG, Cleaveland JS. Differential expression and sequence-specific interaction of karyopherin α with nuclear localization sequences. J Biol Chem. 1997;272(7):4310–5.
27. Kim M, Tagkopoulos I. Data integration and predictive modeling methods for multi-omics datasets. Molecular Omics. 2018; 14(1):8-25.
28. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97.
29. Gönen M, Alpaydın E. Multiple Kernel Learning Algorithms. The Journal of Machine Learning Research. 2011;12:2211-68.
30. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS. 2002; 99(10):6567:72.
31.
32. Breiman L. Random Forests. Mach Learn.2001;45:5-32.
33. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21:1-13.
34. Feldner-Busztin D, Nisantzis PF, Edmunds SJ, Boza G, Racimo F, Gopalakrishnan S, et al. Dealing with dimensionality: the application of machine learning to multi-omics data. Bioinformatics. 2023;39(2):btad021.
35. Khadirnaikar S, Shukla S, Prasanna SRM. Machine learning based combination of multi-omics data for subgroup identification in non-small cell lung cancer. Sci Rep. 2023;13(1):4636.
36. Chierici M, Bussola N, Marcolini A, Francescatto M, Zandonà A, Trastulla L, et al. Integrative Network Fusion: A Multi-Omics Approach in Molecular Profiling. Front Oncol. 2020;10:1065.
37. Ma C, Zhang HH, Wang X. Machine learning for Big Data analytics in plants. Trends Plant Sci. 2014;19(12):798–808.
38. Mo Q, Shen R, Guo C, Vannucci M, Chan KS, Hilsenbeck SG. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics. 2018;19(1):71–86.
39. Rappoport N, Shamir R. NEMO: cancer subtyping by integration of partial multi-omic data. Bioinformatics. 2019;35(18):3348–56.
40. Xu H, Gao L, Huang M, Duan R. A network embedding based method for partial multi-omics integration in cancer subtyping. Methods. 2021;192:67–76.
41. Cao H, Jia C, Li Z, Yang H, Fang R, Zhang Y, et al. wMKL: multi-omics data integration enables novel cancer subtype identification via weight-boosted multi-kernel learning. Br J Cancer. 2024;130(6):1001-1012.
42. Manochkumar J, Cherukuri AK, Kumar RS, Almansour AI, Ramamoorthy S, Efferth T. A critical review of machine-learning for “multi-omics” marine metabolite datasets. Computers in Biology and Medicine. 2023;165:107425.

43. Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. Briefings in Bioinformatics. 2022;23(1):bbab454.
44. Yang Y, Mirzaei G. Performance analysis of data resampling on class imbalance and classification techniques on multi-omics data for cancer classification. PLoS One. 2024 Feb 1;19(2):e0293607.
45. Novoloaca A, Broc C, Beloeil L, Yu WH, Becker J. Comparative analysis of integrative classification methods for multi-omics data. Brief Bioinform. 2024 Jul 1;25(4): bbae331.

FIGURES

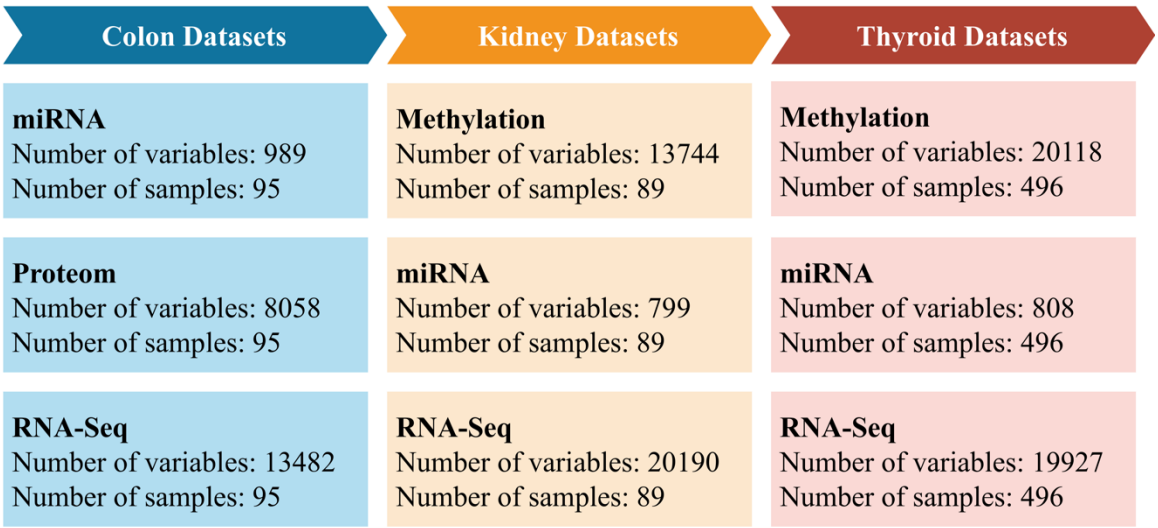

Fig. 2.1. Datasets used in this study.

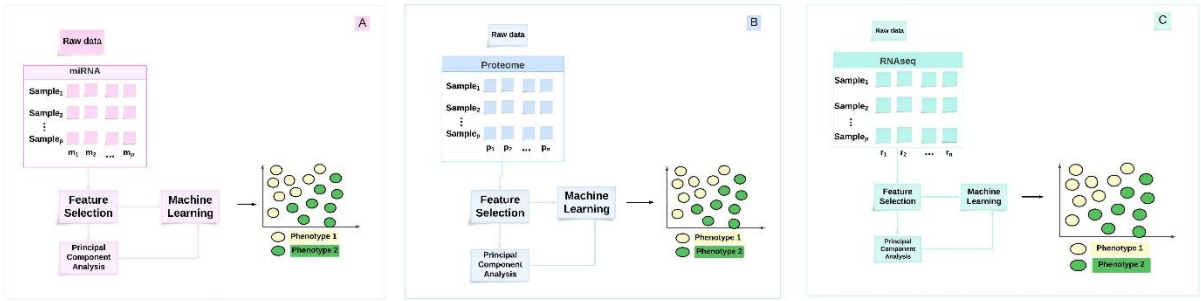

Fig. 2.2. The workflow of disease diagnosis using single omics data separately. We note here that the labels depicted with yellow and green may vary for each method.

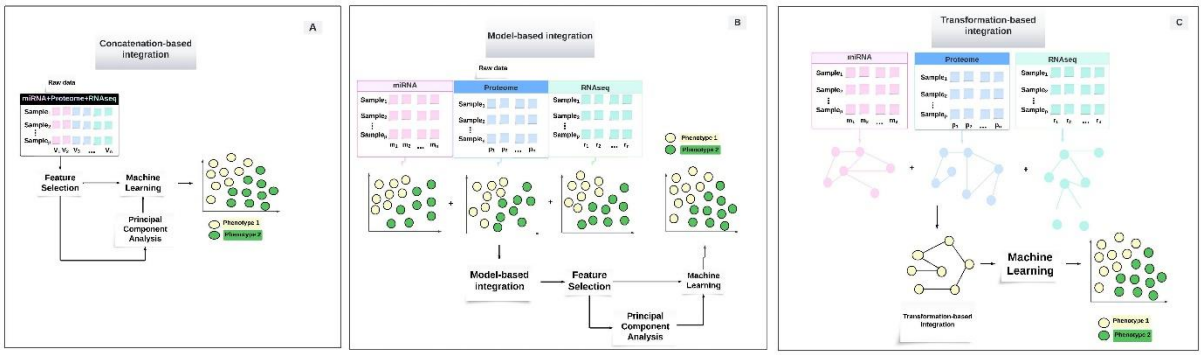

Fig. 2.3. The workflow of disease diagnosis with different data integration approaches. This figure outlines the main steps involved in disease diagnosis, highlighting the various approaches used for integrating different omics datasets.

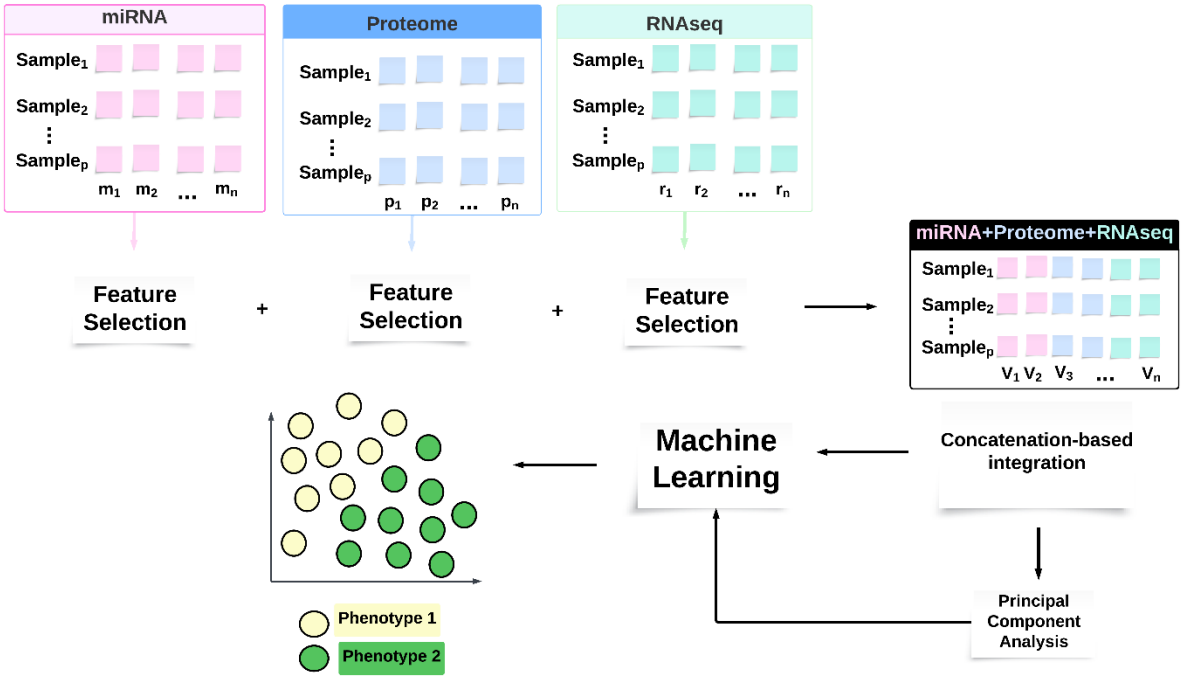

Fig. 2.4. The workflow of disease diagnosis with concatenate-based integration method. The workflow illustrates disease diagnosis, with concatenate-based integration employed after applying feature selection to each omics dataset.