BMC Part of Springer Nature
BMC Bioinformatics

moBRCA-net: a breast cancer subtype classification framework based on multi-omics attention neural networks

Joung Min Choi[1] and Heejoon Chae[2*]

BIBS seminar

2023.09.21

YEONJU SEO

BIBS

Contents

1. Introduction

Background

- Breast cancer is a highly heterogeneous disease that comprises multiple biological components.
- Owing its diversity, patients have different prognostic outcomes

  → That's why early diagnosis and accurate subtype prediction are critical for treatment.

- Standardized breast cancer subtyping systems, mainly based on single-omics datasets, have been developed to ensure proper treatment in a systematic manner.

- Recently, multi-omics data integration has attracted attention to provide a comprehensive view of patients but poses a challenge due to the high dimensionality.
- Also, deep learning-based approaches have been proposed, but they still present several limitations

Background

- The present study introduces **moBRCA-net**, **a breast cancer subtype classification framework utilizing multi-omics data.**

- It integrates datasets through feature-selection modules considering biological relationships between **(1)DNA methylation, (2) gene expression and (3) microRNA expression**.

- Additionally, a **self-attention module** is applied to learn feature importance at the omics level, transforming each feature to a representation reflecting its significance for classification.
  → These representations are concatenated and fed into fully connected layers for predicting breast cancer subtypes in patients.

2. Materials & methods

Materials

< Datasets > : The breast cancer (BRCA) cohort datasets

- **From TCGA**
  1. **Gene expressions**
  2. **DNA methylation**
  3. **microRNA expression**

*PAM50 : a gene set used to classify different molecular subtypes of breast cancer based on the analysis of molecular characteristics in tumor samples.

* Patients who did not have all three omics data available were excluded.
*Breast cancer subtype information for each TCGA BRCA sample was retrieved from PAM50
*Total of 1059 samples were divided into 5 subtypes, as shown in Table 1.

Table 1 Number of samples for each breast cancer subtype

Breast cancer subtype	Number of samples
Luminal A	556
Luminal B	200
HER2-enriched	182
Basal-like	81
Normal-like	40

How to preprocessing?

< From the Cancer Cell Line Encyclopedia (CCLE)>

1. **Gene expression data:**
2. **microRNA data:**

step 1) remove which read counts were not available for all samples.

step 2) After calculating size factors, the read counts were normalized by library size and were log transformed using DESeq2.

How to preprocessing?

< From the Cancer Cell Line Encyclopedia (CCLE)>

3. **DNA methylation data :**

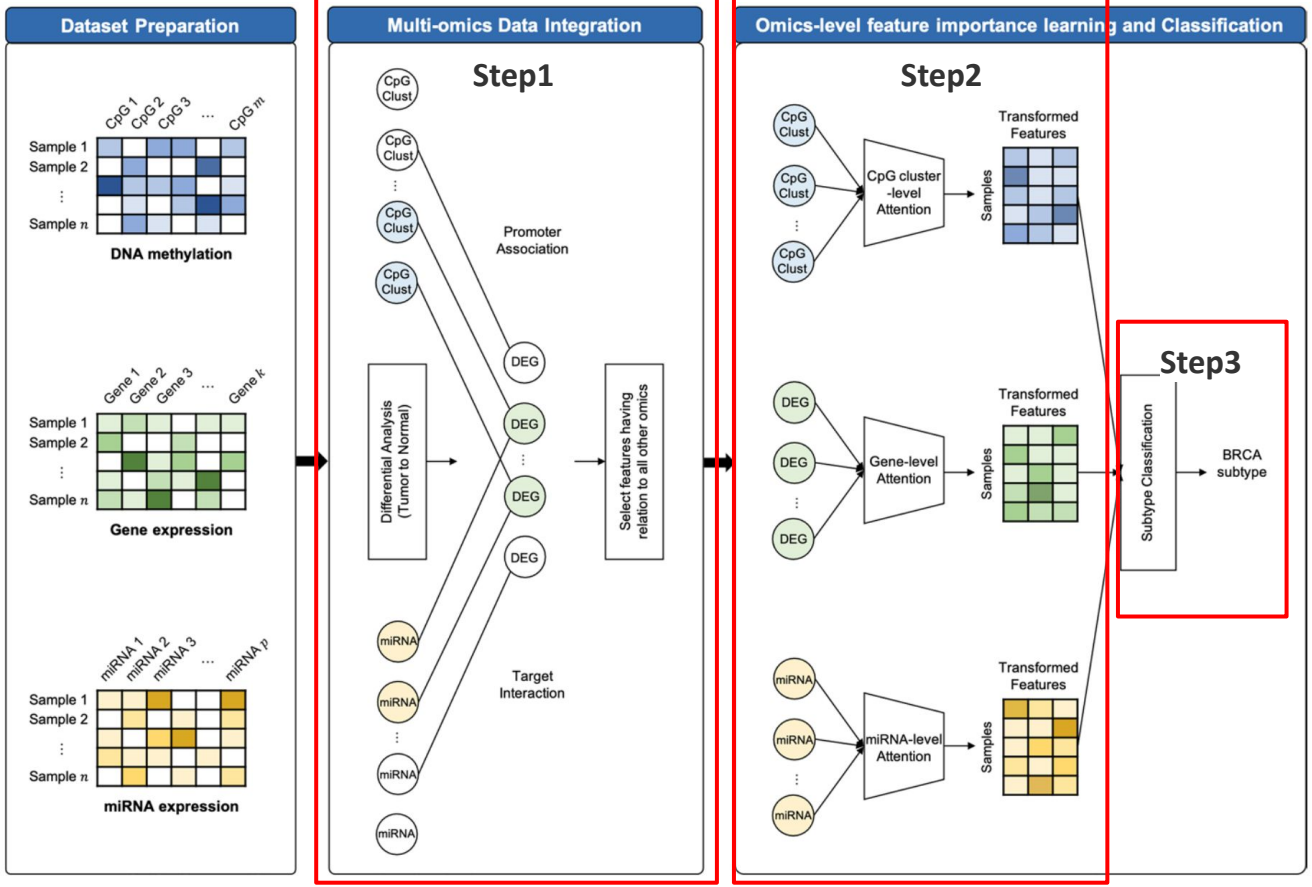
step 1) Both DNA methylome datasets measured by Illumina Human Infinium 450 K and 27 K platforms were used, with common features of both datasets being used for further analysis.

step 2) To eliminate the bias caused by a high frequency of missing values during model training, median imputation was performed, in which CpG sites with missing values for all samples were removed.

→ Therefore, 20,400 genes, 19,977 CpGs, and 1597 microRNAs were used!

About model architecture

Preprocessing



Step 1) Multi-omics data integration

- Feature selection module

✔ 1) Identifying informative genes as breast cancer signatures through differential analysis, yielding 1000 DEGs with high log-fold change and low adjusted p-value

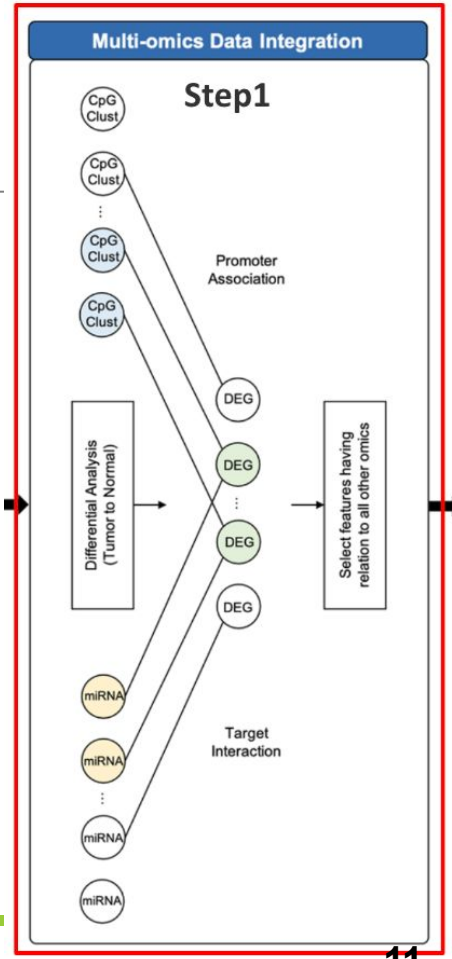
→ Genes with an absolute value of log (fold change) greater than 2 and an adjusted p-value less than 0.01 was considered as differentially expressed genes (DEGs).

✔ 2) Forming **CpG clusters** near DEG promoter regions to capture epigenetic relationships

*promoter-associated CpGs play important roles in gene silencing, genomic imprinting, and cancerogenesis.

→ After preprocessing, CpGs within 2 kb of the promoter regions of each DEG were grouped to form a cluster, where the average of the beta values were calculated.

→ DEGs without matched CpGs in the preprocessed dataset were filtered out to focus on features related to other omics.



Step 1) Multi-omics data integration

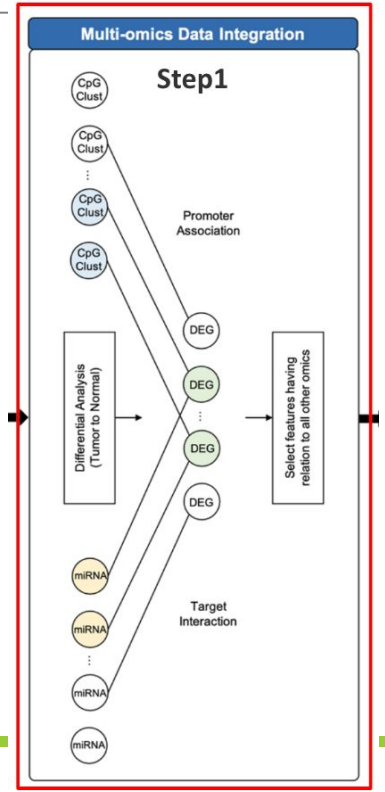
- Feature selection module

✔ 3) Selecting microRNAs interacting with the identified DEGs using the TargetScan database for comprehensive multi-omics integration.

→microRNAs control the function of their target mRNAs by downregulating the expression of their targets.

→ Thus, they have been recognized as drivers of diverse disease conditions including cancer .

→ microRNAs showing target interaction with the identified DEGs were selected based on the TargetScan database.



Step 2) Omics-level feature importance learning

- Omics-level self-attention module

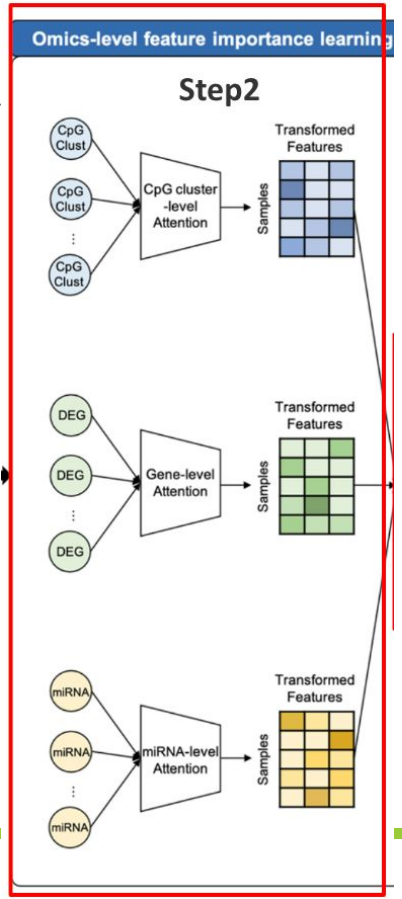
✔ To identify crucial features for classifying breast cancer subtypes and better understand the relative importance of those features

< self-attention >

step1) Utilized random vectors to create k-dimensional embedding vectors for each feature, transforming the original input data x into a new representation x̂.

$$\hat{x}_i = f_e(e_i, x_i) = e_i x_i$$

** Given a set of original input data x ∈ Rn, where n : dimension of the input data*
** They defIned the k-dimensional embedding vector ei for each feature i ∈ {1 : n} using random vectors and represented xi to xˆi via multiplication*



Step 2) Omics-level feature importance learning

- Omics-level self-attention module

< self-attention >

✔ To learn the level of importance for each feature to predict breast cancer subtype, each feature xˆi was assigned an attention score αi

step2) Assigned an attention score α to each feature x̂ to determine its importance in predicting breast cancer subtype.

→ Calculated attention scores using a series of mathematical operations involving weights (WFC, Wh1, Wh2) and a bias term

$$\bar{x}_i = tanh(W_{FC}\hat{x}_i + b)$$

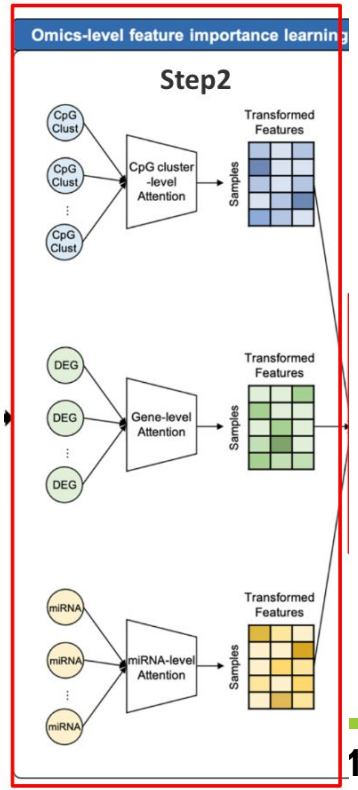
$$s_i = W_{h2}tanh(W_{h1}\hat{x}_i + b)$$

** WFC, Wh1, and Wh2: weights*

** b : bias term*

**** si i: attention score that represents the importance of each feature***

** xˆi : which was converted to a normalized weight αi by applying the softmax function.*



Step 2) Omics-level feature importance learning

- Omics-level self-attention module

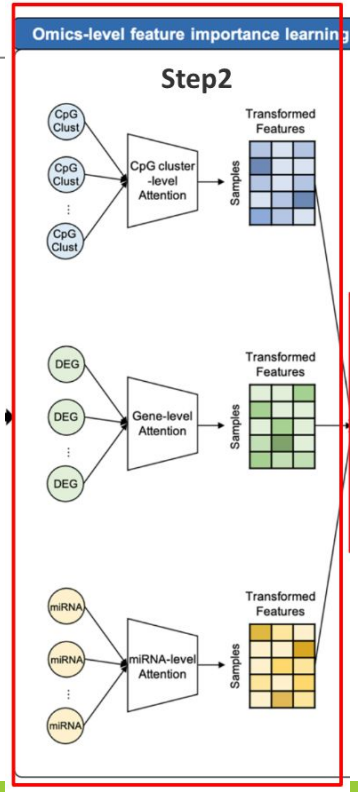
< self-attention >

step3) Normalized the attention scores α using the softmax function, ensuring they sum up to 1 for relative importance assessment.
→ Computed dense feature representations c by combining the encoded feature vectors x̄ with their normalized attention scores α.

$$\alpha_i = \frac{\exp(s_i)}{\sum_{j=1}^{n} \exp(s_j)}$$

$$c_i = \sum_{i=1}^{n} \alpha_i \bar{x}_i,$$

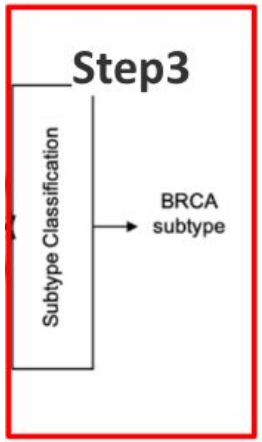
*Based on the calculated values, **xˆi was transformed to a dense feature representation ci** by the weighted sum of the encoded feature vectors x¯i and their normalized attention scores αi.



→ By self-attention module independently to each omics dataset, concatenated the transformed features, and fed them into the subtype classification module.

Step 3) Subtype classification

▪With two fully connected layers



✔ Classification module was constructed with two fully connected layers followed by the softmax function layer to achieve the final breast cancer subtype classification.

✔ moBRCA-net was trained to minimize the cross-entropy loss.

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \cdot \log(\hat{y}_i),$$

** C :# of breast cancer subtypes*

y ($\hat{y}$) : true (model predicted, respectively) subtype probability distribution.

✔ To prevent overfitting, dropout was applied, and L2 regularization was also added to the loss function.

Step 3) Subtype classification

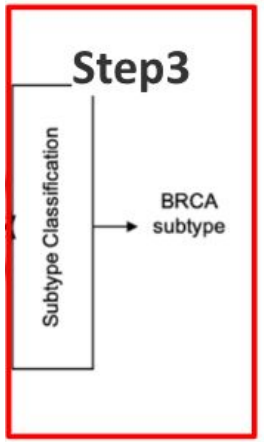
▪Hyper parameter & optimization

✔ Adam optimization algorithm for training.

✔ Split the dataset into 70% training and 30% test sets

→ repeated three times for each hyperparameter combination, and the architecture showing the best average accuracy.
→ best result was set as our moBRCA-net model!

✔ Set embedding vector dimension(k) to 128, encoding vector dimension ($\bar{x}$) to 64, and attention vector dimension (s) equal to the number of features.



***Embedding Vector Dimension (k):** used to represent categorical variables or discrete features in a continuous vector space.

***Encoding Vector Dimension ($\bar{x}$):** represents the internal state or a data point in a neural network.

→ It is a continuous vector that summarizes the information extracted from the input data.

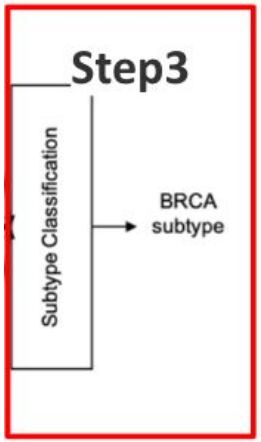
Step 3) Subtype classification

▪Hyper parameter & optimization

- ✔ Adam optimization algorithm for training.
- ✔ Split the dataset into 70% training and 30% test sets
  → repeated experiments three times to select the best model based on average accuracy.

- ✔ Set embedding vector dimension(k) to 128, encoding vector dimension (x^{-}) to 64, and attention vector dimension (s) equal to the number of features.

- ✔ **Hidden layer :** two fully connected layers with 200 and 5 hidden nodes
- ✔ **Activation function :**LU activation function
- ✔ **Dropout rate :** 0.7 for both omics-level attention and classification module training.
- ✔ **Learning rate :** 0.01 & **Epochs:**5000



Step 3) Subtype classification

- Hyper parameter

✔ Adam optimiz

✔ Split the datas
→ repeate
average ac

✔ Set embedding
64, and attenti

✔ **Hidden layer :**
✔ **Activation fun**
✔ **Dropout rate :**
training.
✔ **Learning rate :**

SupplementaryMaterial S1.

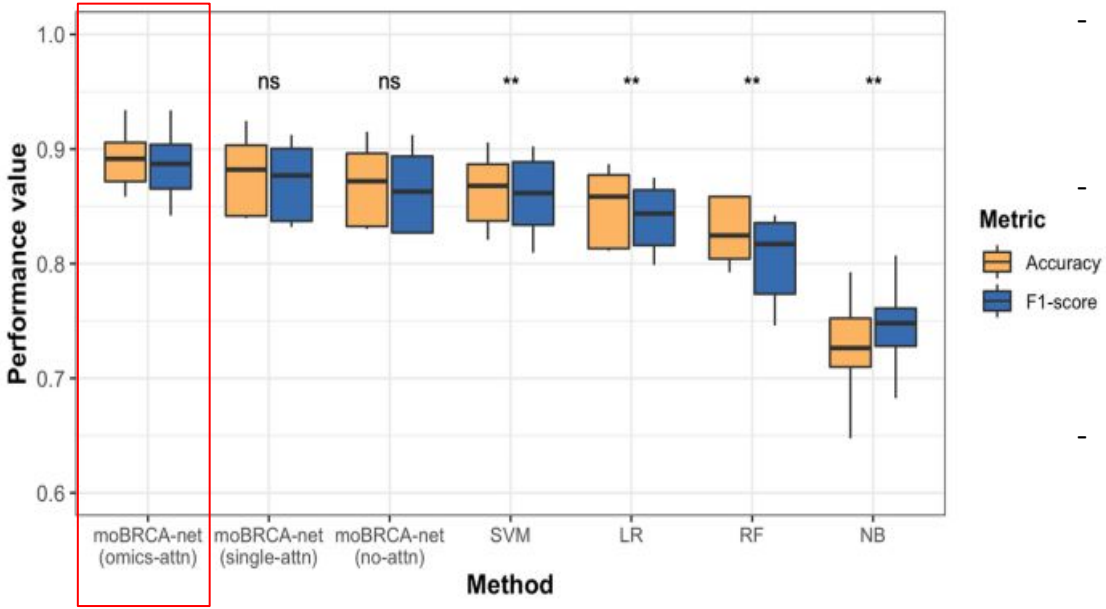
Table1. Hyperparmeter optimization results based on the average accuracy for classifying breast cancer subtypes.

Learning rate	Dimension of embedding vector	Dimension of encodig vector	# of hidden nodes in classification module	Dropout	Average accuracy
1e-3	64	32	50	0.7	0.8679
1e-3	64	32	100	0.6	0.8490
1e-3	64	32	100	0.7	0.8773
1e-2	64	32	100	0.7	0.8679
1e-2	64	32	200	0.8	0.8490
1e-3	64	32	300	0.6	0.8584
1e-2	128	32	200	0.7	0.8490
1e-2	128	64	50	0.7	0.8773
1e-3	128	64	100	0.7	0.8773
1e-2	128	64	100	0.7	0.9230
1e-3	128	64	200	0.6	0.8679
1e-3	128	64	200	0.7	0.9079
1e-2	**128**	**64**	**200**	**0.7**	**0.9433**
1e-2	128	64	200	0.8	0.8867
1e-2	128	64	300	0.7	0.8974
1e-3	256	128	100	0.7	0.8679
1e-3	256	128	200	0.7	0.8679

3. Results

1) Evaluation of moBRCA-net performance

✔ To evaluate the ability of moBRCA-net to classifying breast cancer subtypes
→ so let's compared its performance with that of widely-used machine learning(ML)-based classifiers!



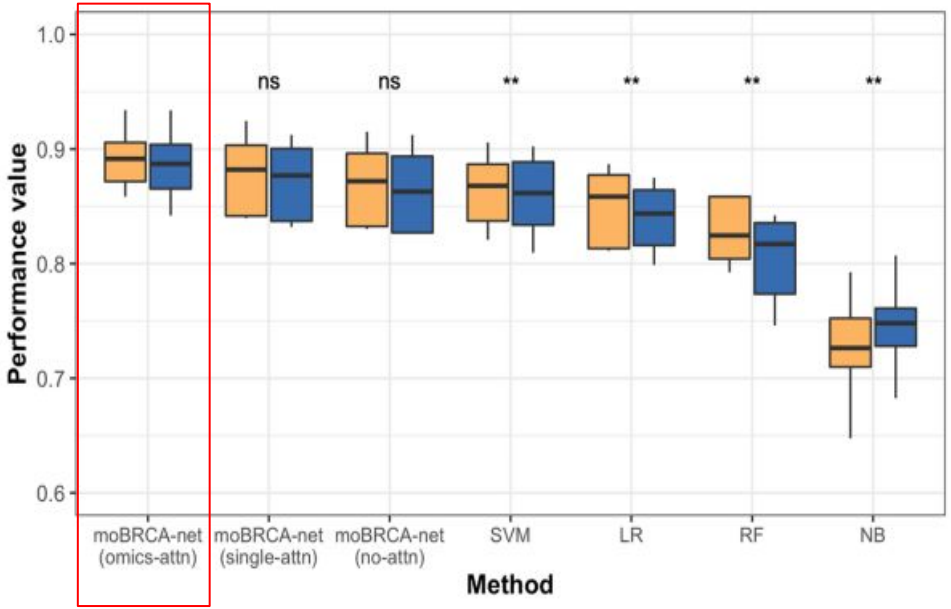
- The study optimized moBRCA-net and baseline methods using a 7:3 split of the TCGA-BRCA dataset for training and testing.
- Hyperparameters were fine-tuned through grid search, with each combination tested five times for accuracy.
  → Specific hyperparameters were selected for SVM, RF, and LR classifiers.
- Tenfold cross-validation was applied, using training data for integration and model training, and testing data solely for performance evaluation.

✔ moBRCA-net (omics-attn) outperformed other classifiers with an average **accuracy of 0.891**, **F1-score of 0.887,** and MCC of 0.831.

1) Evaluation of moBRCA-net performance

✔ To evaluate the ability of moBRCA-net to classifying breast cancer subtypes

→ so let's compared its performance with that of widely-used machine learning(ML)-based classifiers!



***previous other study's results**

Table 4

Classification results on the BRCA dataset.

Method	ACC	F1
KNN	74.22±2.63	72.56±2.90
SVM	74.68±1.14	73.78±0.87
LASSO	77.19±1.05	75.51±0.81
RF	71.48±1.68	69.55±1.74
FNN	73.61±1.49	72.56±1.51
MORONET	80.61±0.54	79.97±1.50
MOADLN	**82.97±1.35**	**83.06±1.58**

0.829 **0.830**

* The Matthews Correlation Coefficient (MCC) :

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

✔ moBRCA-net (omics-attn) outperformed other classifiers with an average **accuracy of 0.891**, **F1-score of 0.887,** and MCC of 0.831.

1) Evaluation of moBRCA-net performance

✔ To evaluate the ability of moBRCA-net to classifying breast cancer subtypes

→ so let's compared its performance with that of widely-used machine learning(ML)-based classifiers!

- **Subtype Wise performance result**

SupplementaryMaterial S4.

Average subtype-wise accuracy and weighted F1-score results of moBRCA-net (omics-attn) from the performance evaluation based on 10-fold cross validation in Fig 2.

Subtype	Accuracy	F1-score
Basal	0.9244	0.9289
Her2	0.9292	0.7942
Luminal A	0.9603	0.7254
Luminal B	0.9896	0.9737
Normal-like	0.5967	0.6099

1) Evaluation of moBRCA-net performance

✔ But still the data imbalance issue could impact the prediction performance in a subtype-specifIc fashion, where there is a large difference between the number of samples for each subtype.

→ so let's **adopted a data augmentation** based on the deep generative model to enlarge the training dataset size.

- **+) Conditional Variational Autoencoder (CVAE)** for data generation
  : a type of generative model used in machine learning.
  → It's particularly useful for generating new data samples that are similar to a given dataset.

  - It composed of a two-layered encoder and decoder, estimated the conditional distribution using latent variables and data, generating samples for specified breast cancer subtypes. * conditional information: the subtype of breast cancer
  - In each fold of tenfold cross-validation, the CVAE was optimized using the training dataset and generated samples to match the same number of samples for the "Luminal A" subtype, which had the largest number of samples.

  → These generated samples were incorporated during the training of moBRCA-net, and the performance was evaluated on the testing dataset.

1) Evaluation of moBRCA-net performance

✔ But still the data imbalance issue could impact the prediction performance in a subtype-specifIc fashion, where there is a large difference between the number of samples for each subtype.

→ so let's **adopted a data augmentation** based on the deep generative model to enlarge the training dataset size.

- **+) Conditional Variational Autoencoder (CVAE)** for data generation

SupplementaryMaterial S5.
Average overall and subtype-wise performance results of moBRCA-net with/without data augmentation based on 10-fold cross validation.

	Overall			Luminal A		Luminal B		Basal		Her2		Normal-like	
	ACC	F1	MCC	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
w/o aug	0.891	0.887	0.831	0.96	0.725	0.989	0.974	0.924	0.928	0.929	0.794	0.596	0.609
with aug	0.895	0.891	0.84	0.966	0.763	0.987	0.966	0.927	0.929	0.93	0.819	0.599	0.618

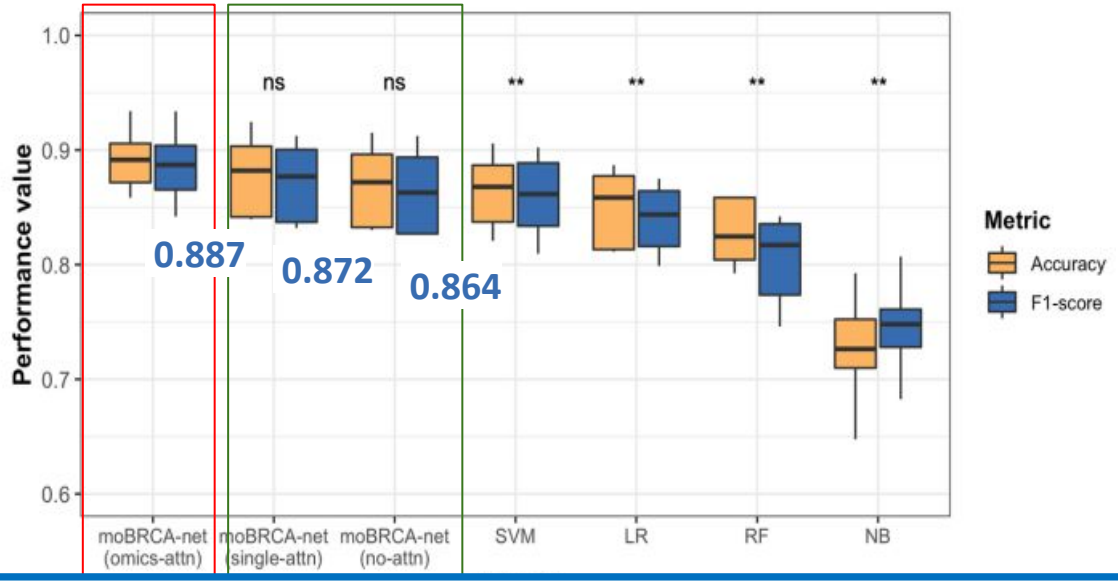
✔ The overall performance of moBRCA-net as well as the subtype-wise performance slightly improved compared to the model trained without the generated dataset.

✔ These results support that data augmentation strategy could help to alleviate the impact from the imbalanced dataset while training our model.

2) Effectiveness of each module in moBRCA-net

✔ To investigated the performance improvement of moBRCA-net by the introduction of the omics-level attention modules for feature importance learning

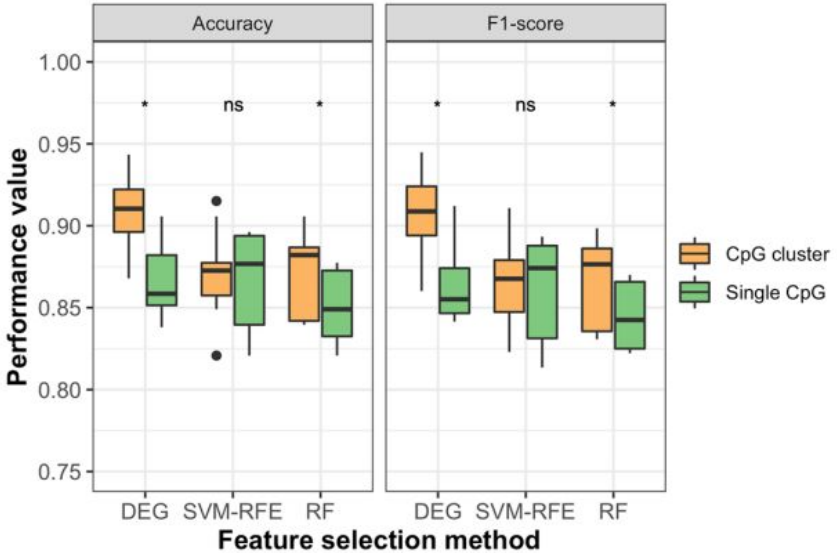
→ so let's implemented a single-attention module was applied to all features at once(single-attn), and in the other the attention module was removed to directly classify the breast cancer subtypes(no-attn).



✔ These experiments also suggest that applying attention at the proper level has an impact on learning the features and modeling the classification module.

3) Breast cancer subtype prediction improvement by multi-omics integration

✔ To validate whether utilizing multi-omics datasets could effectively improve the classification of breast cancer subtypes
→ let's compared the performance when using different combinations of multi-omics datasets and a single-omics dataset.



- investigated the impact of the CpG clusters by comparing the classification performance of moBRCA-net based on single CpG-based multi-omics integration using different feature selection method

* SVM-RFE (Support Vector Machine Recursive Feature Elimination) : aimed at selecting the most important features from a given dataset.

✔ When utilizing CpG clusters, the average classification performance significantly improved for DEGbased method (from F1-score of 0.864 to 0.908) and RF (from 0.845 to 0.866) SVM-RFE showed a slight performance increase (from 0.86 to 0.866).

✔ CpG cluster-based approaches achieved the best accuracy and F1-score compared with single-CpG approaches for all cases using different feature selection methods.

→ they assumed that CpGs located in regions relatively close to the promoter may share a similar methylation status, which could represent the methylation patterns related to breast cancers, consequently leading to performance improvement for subtype prediction.

3) Breast cancer subtype prediction improvement by multi-omics integration

✔ To validate whether utilizing multi-omics datasets could improve the classification of breast cancer subtypes
→ compared the performance when using different combinations of multi-omics datasets and a single-omics dataset.

Table 2 Average classification performance for breast cancer subtype classification using different omics datasets

	Multi-omics				Single omics		
	Gene + methyl + microRNA	Gene + methyl	Gene + microRNA	Methyl + microRNA	Gene	Methyl	MicroRNA
Accuracy	0.909	0.865	0.889	0.820	0.863	0.817	0.85
F1-score	0.908	0.857	0.886	0.807	0.852	0.806	0.842
p-value	–	< 0.05	< 0.05	< 0.01	< 0.05	< 0.01	< 0.01

Wilcoxon signed rank test was performed using the performance results using all three omics dataset and results from each combination of different omics dataset

✔ moBRCA-net showed a relatively higher performance when trained based on multi-omics compared with single-omics data

✔ These results indicate that multi-omics data provides more comprehensive information to distinguish breast cancer subtypes.

4) Interpretation of omics-level attention for breast cancer subtype classification

✔ To understand how omics-level attention helped improve the performance of the model
→ let's interpreted the attention scores of moBRCA-net.

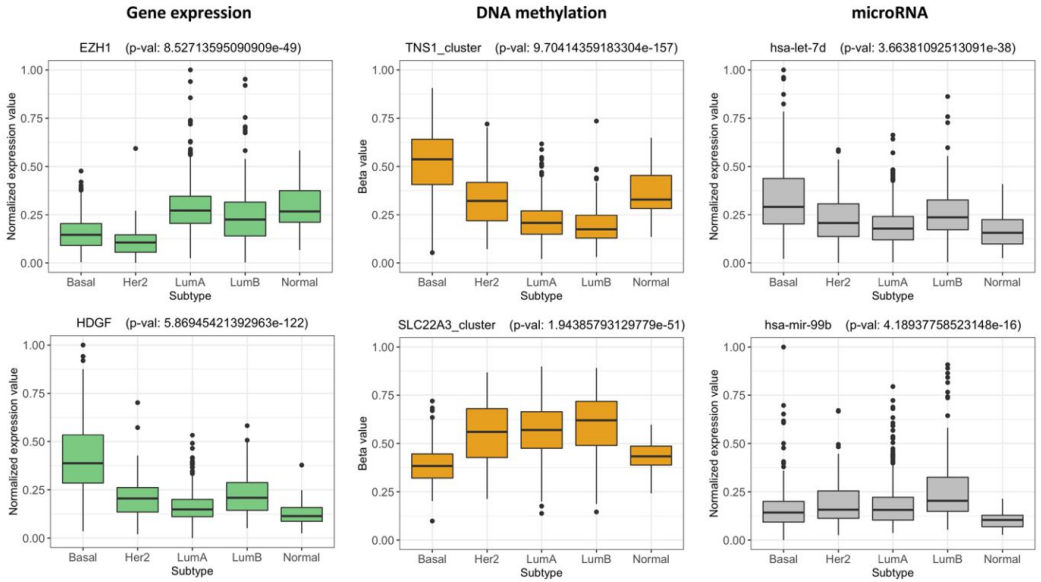


Fig. 4 Normalized values of the features from each omics dataset showing the top average attention scores

1) To directly compare the abundance difference between the feature subtypes with the highest attention scores

→ visualized the normalized gene expression values and beta values of those features obtained from samples of each breast cancer subtype

✔ Overall, they could conclude that the features with the highest attention scores showed significant differences across the five subtypes with p-value < 0.01, indicating that the attention module trained moBRCA-net to assign more weights for the features having discriminative power for classifying the subtypes.

4) Interpretation of omics-level attention for breast cancer subtype classification

+) DNA methylation shows distinct patterns for each breast cancer subtype

→ thus, it has the potential to be used as a subtype-specific marker.

---

2) Also, they hypothesized that the attention module would assign more weight to the biologically relevant features and identify the features showing a negative correlation (< −0.5) for each subtype.

→ so from each omics dataset, 200 features showing the highest average attention scores across patients were selected and the Pearson correlation between those features was analyzed.

- ✔ They identified feature pairs showing a negative correlation in different breast cancer subtypes, excluding the pairs of the normal-like subtype.
- ✔ These results were consistent with that of recent reports that showed that basal-like cancers more frequently present abundant NDRG2 expression in association with CpG-hypomethylation, with is associated with aggressiveness and unfavorable outcomes in the basal-like subtype

ex) NDRG2 showed a negative correlation with the CpG cluster composed of cg14030359 and cg18081258 in the basal subtype, STAT5 showed negative correlation with the CpG cluster composed of cg03001305 and cg16777510 in the luminal A, B, and basal-like subtypes.

4. Conclusion & Discussion

Conclusion

Novelties in their moBRCA-net

1) **Multi-Omics Integration:** gene expression, DNA methylation, and microRNA expression by maintaining their biological relationships.

   → This enables a comprehensive understanding of the molecular landscape of breast cancer.

2) **Self-Attention Mechanism:** allows the model to dynamically weigh the importance of different features within each omics data type.

   → This enhances the model's ability to discern critical features for accurate classification.

3) **Feature Importance Learning:** learns the importance of each feature at the omics level.

   → This means that it not only identifies important features, but also transforms them into new representations that reflect their significance in the classification task.

Conclusion

Novelties in their moBRCA-net

4) **Performance Superiority:** The study demonstrates that moBRCA-net outperforms established machine learning methods in predicting breast cancer subtypes.

→ This indicates its effectiveness in leveraging multi-omics data for accurate classification.

Results show that

- ✔ These experimental results confirmed that moBRCA-net has a significantly enhanced performance compared with other methods
- ✔ The effectiveness of multi-omics integration and omics-level attention were identified

Discussion

1. There will be potential for further improvement

   ✔ Due to the limitation of the computational resources, feature selection was performed to reduce the number of features for training our model.

   ✔ If the model could learn the dependency between the omics features directly via graph network, it could potentially be able to extract useful relations between the features of different omics datasets.

   → So, they plan to extend our moBRCA-net platform to utilize graph neural networks.

→ By doing this, we will be able to improve the moBRCA-net

Thank you for listening ☺️