A Cascade Deep Forest Model for Breast Cancer Subtype Classification Using Multi-Omics Data

Ala’a El-Nabawy, Nahla Belal, Nashwa El-Bendary

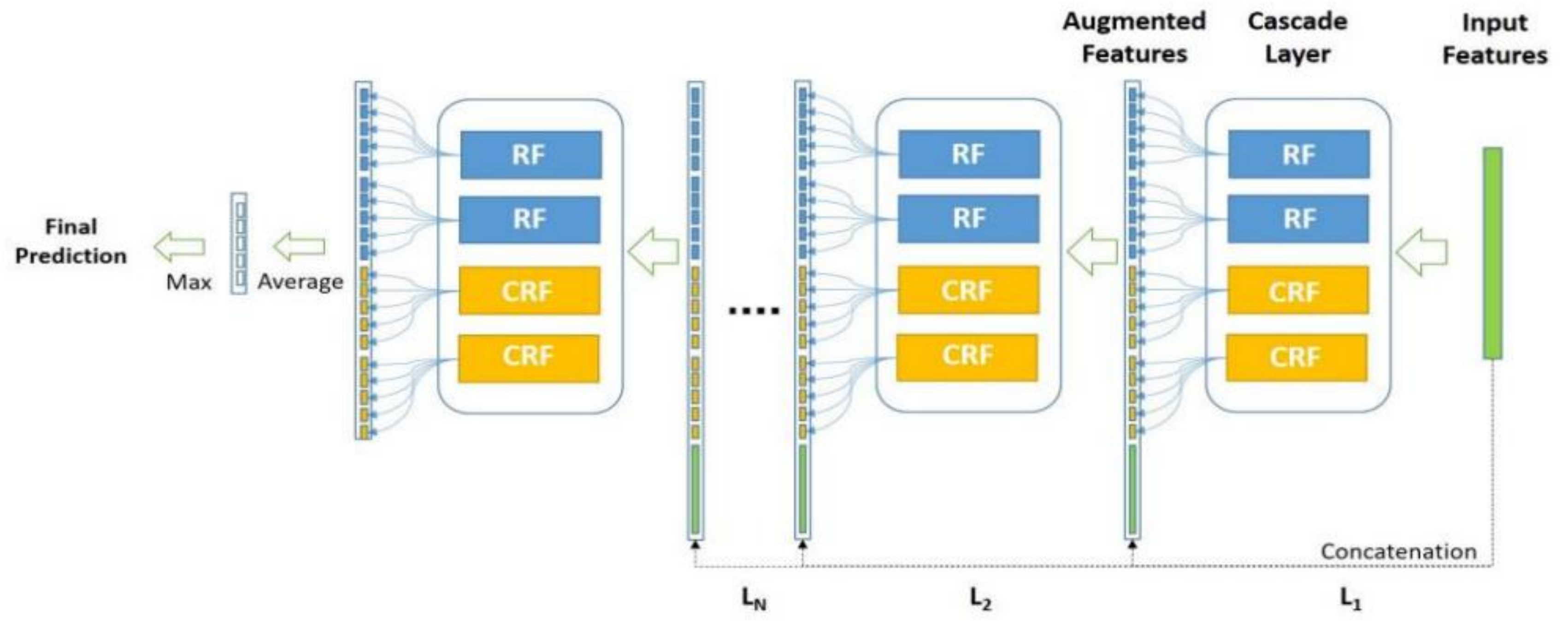
Arab Academy for Science and Technology and Maritime Transport

Abstract

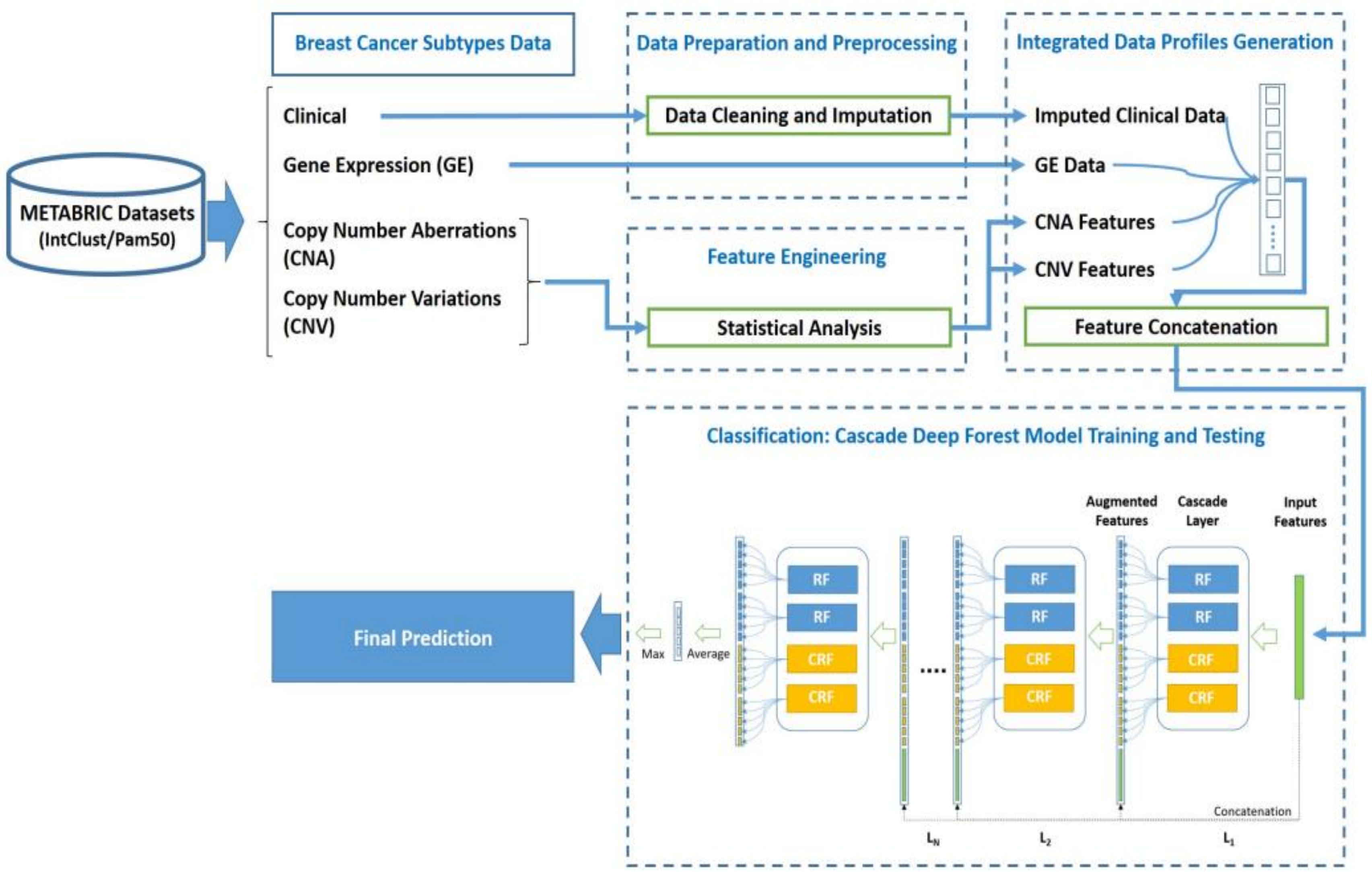
Automated diagnosis systems aim to reduce the cost of diagnosis while maintaining the same efficiency. Many methods have been used for breast cancer subtype classification. Some use single data source, while others integrate many data sources, the case that results in reduced computational performance as opposed to accuracy. Breast cancer data, especially biological data, is known for its imbalance, with lack of extensive amounts of histopathological images as biological data. Recent studies have shown that cascade Deep Forest ensemble model achieves a competitive classification accuracy compared with other alternatives, such as the general ensemble learning methods and the conventional deep neural networks (DNNs), especially for imbalanced training sets, through learning hyper-representations through using cascade ensemble decision trees. In this work, a cascade Deep Forest is employed to classify breast cancer subtypes, IntClust and Pam50,using multi-omics datasets and different configurations. The results obtained recorded an accuracy of 83.45% for 5 subtypes and 77.55% for 10 subtypes. The significance of this work is that it is shown that using gene expression data alone with the cascade Deep Forest classifier achieves comparable accuracy to other techniques with higher computational performance, where the time recorded is about 5 s for 10 subtypes, and 7 s for 5 subtypes.

Objective

Deep Forest ensemble model achieves a competitive classification accuracy compared with other alternatives through learning hyper-representations through using cascade ensemble decision trees. In this work, a cascade Deep Forest is employed to classify breast cancer subtypes, IntClust and Pam50, using multi-omics datasets and different configurations.



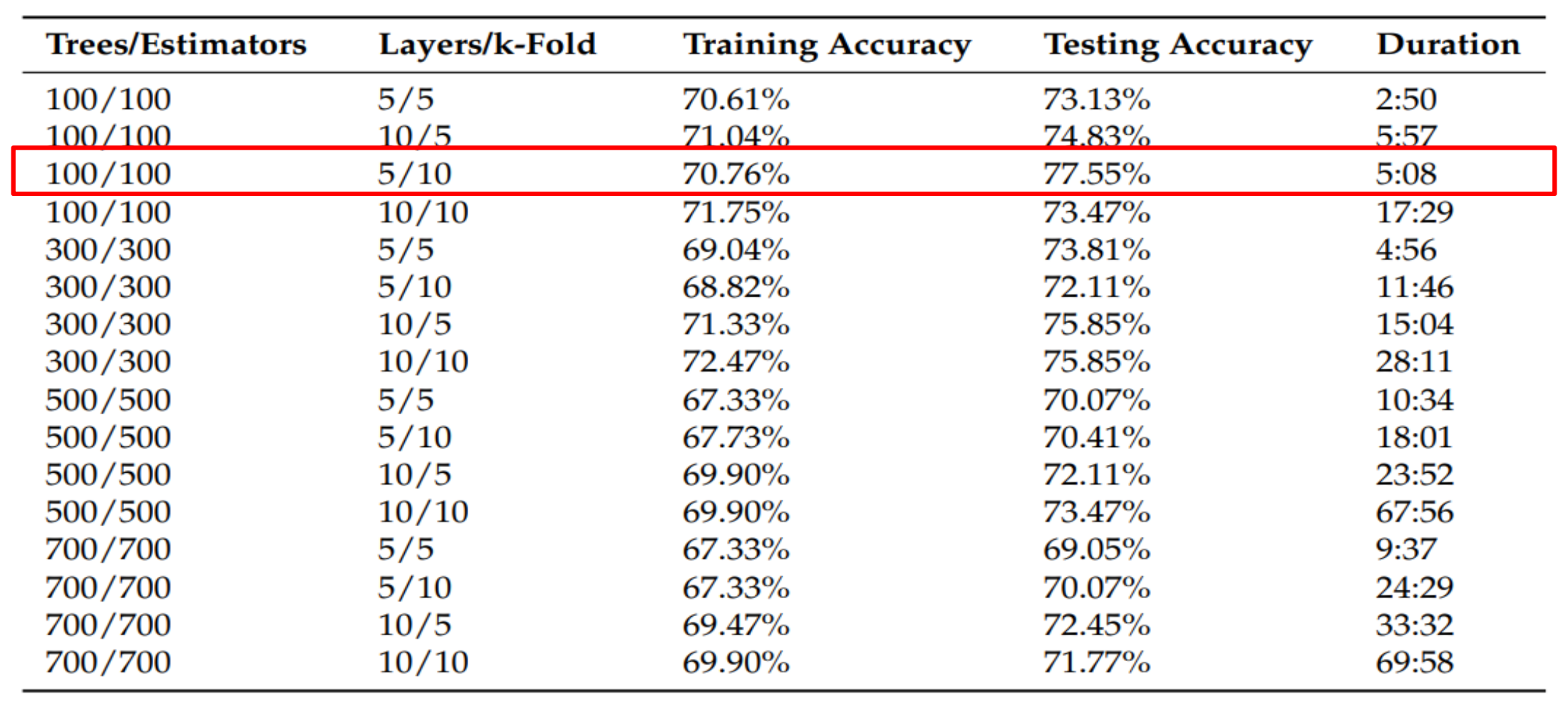
Methodology


Breast Cancer Subtypes Data
METABRIC Datasets (IntClust/Pam50)
Clinical
Gene Expression (GE)
Copy Number Aberrations (CNA)
Copy Number Variations (CNV)
Data Preparation and Preprocessing
Data Cleaning and Imputation
Feature Engineering
Statistical Analysis
Integrated Data Profiles Generation
Imputed Clinical Data
GE Data
CNA Features
CNV Features
Feature Concatenation
Classification: Cascade Deep Forest Model Training and Testing
Augmented Features
Cascade Layer
Input Features
RF
RF
CRF
CRF
....
Max
Average
Concatenation
L_N
L_2
L_1
Final Prediction

The proposed approach is composed of 4 phases; namely (1) Data acquisition of METABRIC breast cancer subtypes datasets, (2) Data preparation and preprocessing, (3) Integrated data profiles generation, and (4) Cascade Deep Forest-based classification. After the first phase of four breast cancer subtypes datasets acquisition, the proposed system moves to the second phase of data preparation and preprocessing with only three sub-datasets; namely the clinical data, the features of Copy Number Aberrations (CNA) and Copy Number Variations (CNV) data types, as the fourth sub-dataset of gene expression is submitted as it is without any preprocessing to the third phase of integrated data profiles generation. In the second phase, data cleaning and imputation preprocessing are applied to the clinical data, whereas statistical analysis is applied to the CNA and CNV features. Subsequently, in phase three, the data profiles are generated by concatenating the genomics and clinical features to obtain the integrated data profiles. Finally, the stages of classification process are employed in the fourth phase for training and teasing the proposed system through using the cascade Deep Forest model.

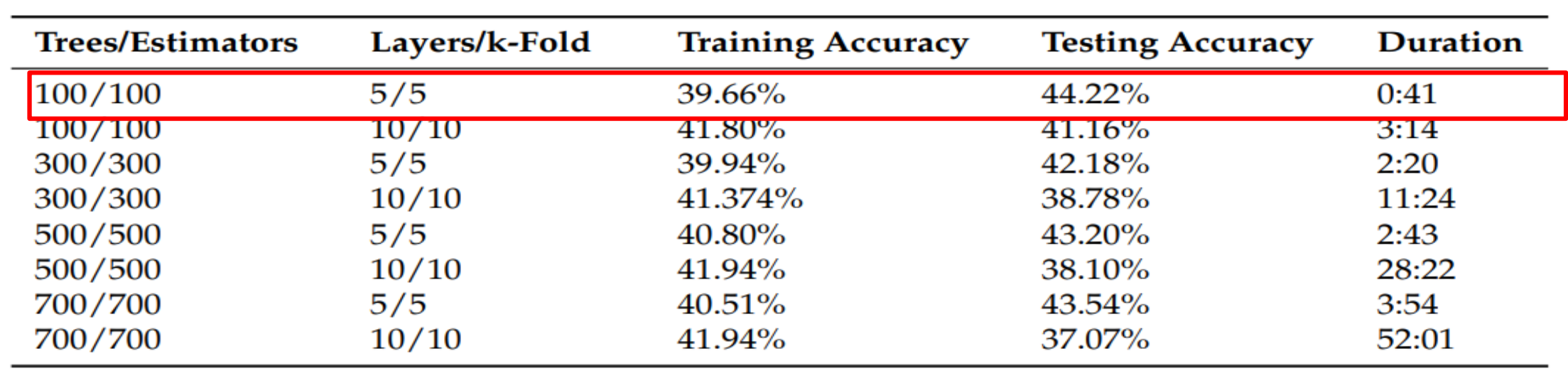
Results (IntClust)

Table 1. Gene Expression—10 Subtypes.



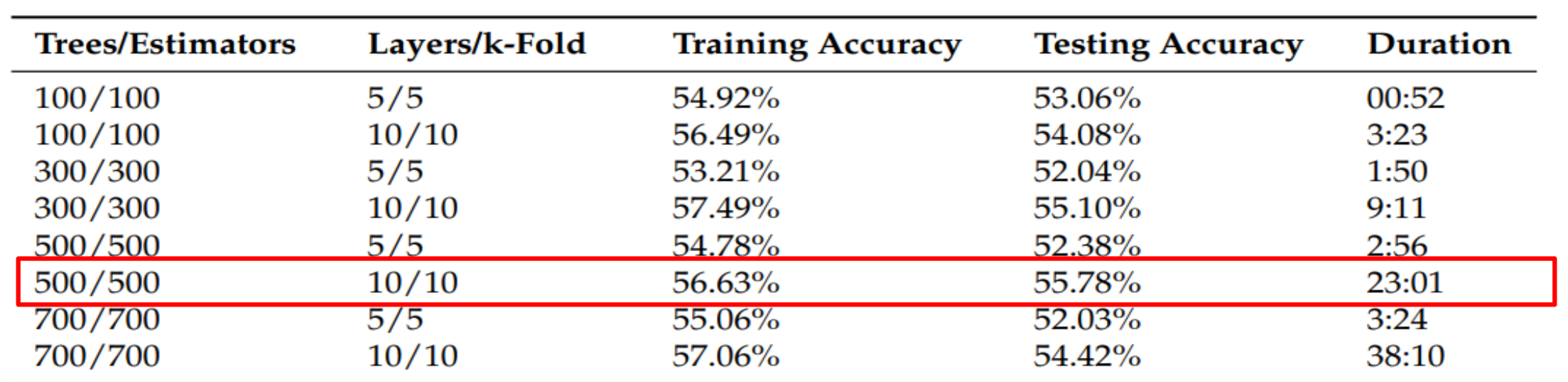
Trees/Estimators	Layers/k-Fold	Training Accuracy	Testing Accuracy	Duration
100/100	5/5	70.61%	73.13%	2:50
100/100	10/5	71.04%	74.83%	5:57
100/100	5/10	70.76%	77.55%	5:08
100/100	10/10	71.75%	73.47%	17:29
300/300	5/5	69.04%	73.81%	4:56
300/300	5/10	68.82%	72.11%	11:46
300/300	10/5	71.33%	75.85%	15:04
300/300	10/10	72.47%	75.85%	28:11
500/500	5/5	67.33%	70.07%	10:34
500/500	5/10	67.73%	70.41%	18:01
500/500	10/5	69.90%	72.11%	23:52
500/500	10/10	69.90%	73.47%	67:56
700/700	5/5	67.33%	69.05%	9:37
700/700	5/10	67.33%	70.07%	24:29
700/700	10/5	69.47%	72.45%	33:32
700/700	10/10	69.90%	71.77%	69:58

Table 2. Clinical Data—10 Subtypes.



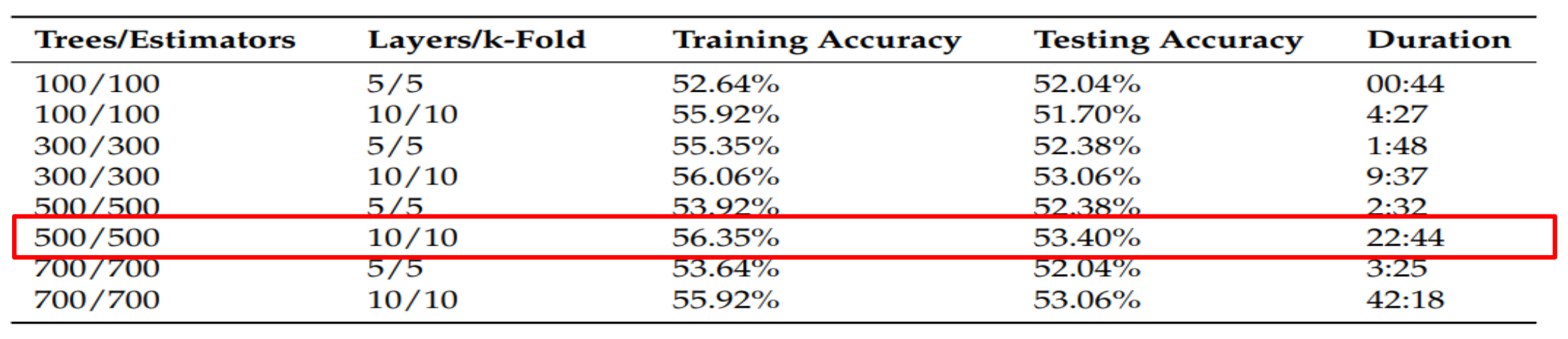
Trees/Estimators	Layers/k-Fold	Training Accuracy	Testing Accuracy	Duration
100/100	5/5	39.66%	44.22%	0:41
100/100	10/10	41.80%	41.16%	3:14
300/300	5/5	39.94%	42.18%	2:20
300/300	10/10	41.374%	38.78%	11:24
500/500	5/5	40.80%	43.20%	2:43
500/500	10/10	41.94%	38.10%	28:22
700/700	5/5	40.51%	43.54%	3:54
700/700	10/10	41.94%	37.07%	52:01

Table 3. CNV Data—10 Subtypes.



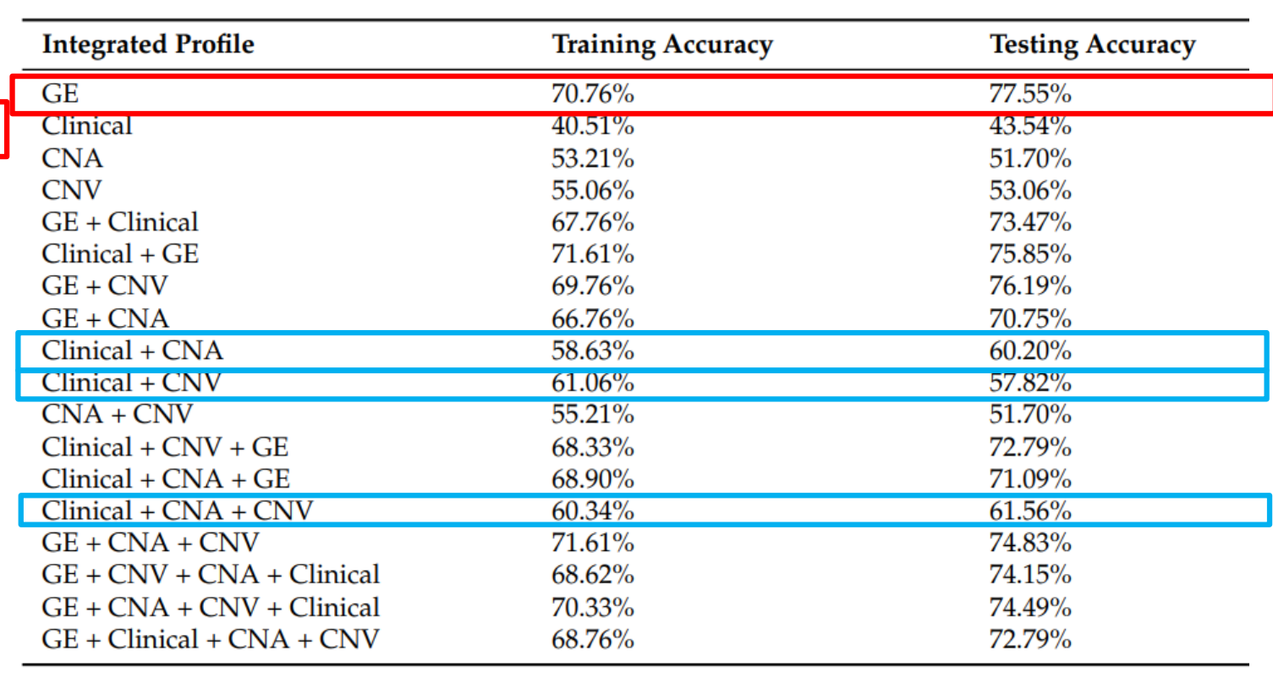
Trees/Estimators	Layers/k-Fold	Training Accuracy	Testing Accuracy	Duration
100/100	5/5	54.92%	53.06%	00:52
100/100	10/10	56.49%	54.08%	3:23
300/300	5/5	53.21%	52.04%	1:50
300/300	10/10	57.49%	55.10%	9:11
500/500	5/5	54.78%	52.38%	2:56
500/500	10/10	56.63%	55.78%	23:01
700/700	5/5	55.06%	52.03%	3:24
700/700	10/10	57.06%	54.42%	38:10

Table 4. CNA Data—10 Subtypes.



Trees/Estimators	Layers/k-Fold	Training Accuracy	Testing Accuracy	Duration
100/100	5/5	52.64%	52.04%	00:44
100/100	10/10	55.92%	51.70%	4:27
300/300	5/5	55.35%	52.38%	1:48
300/300	10/10	56.06%	53.06%	9:37
500/500	5/5	53.92%	52.38%	2:32
500/500	10/10	56.35%	53.40%	22:44
700/700	5/5	53.64%	52.04%	3:25
700/700	10/10	55.92%	53.06%	42:18

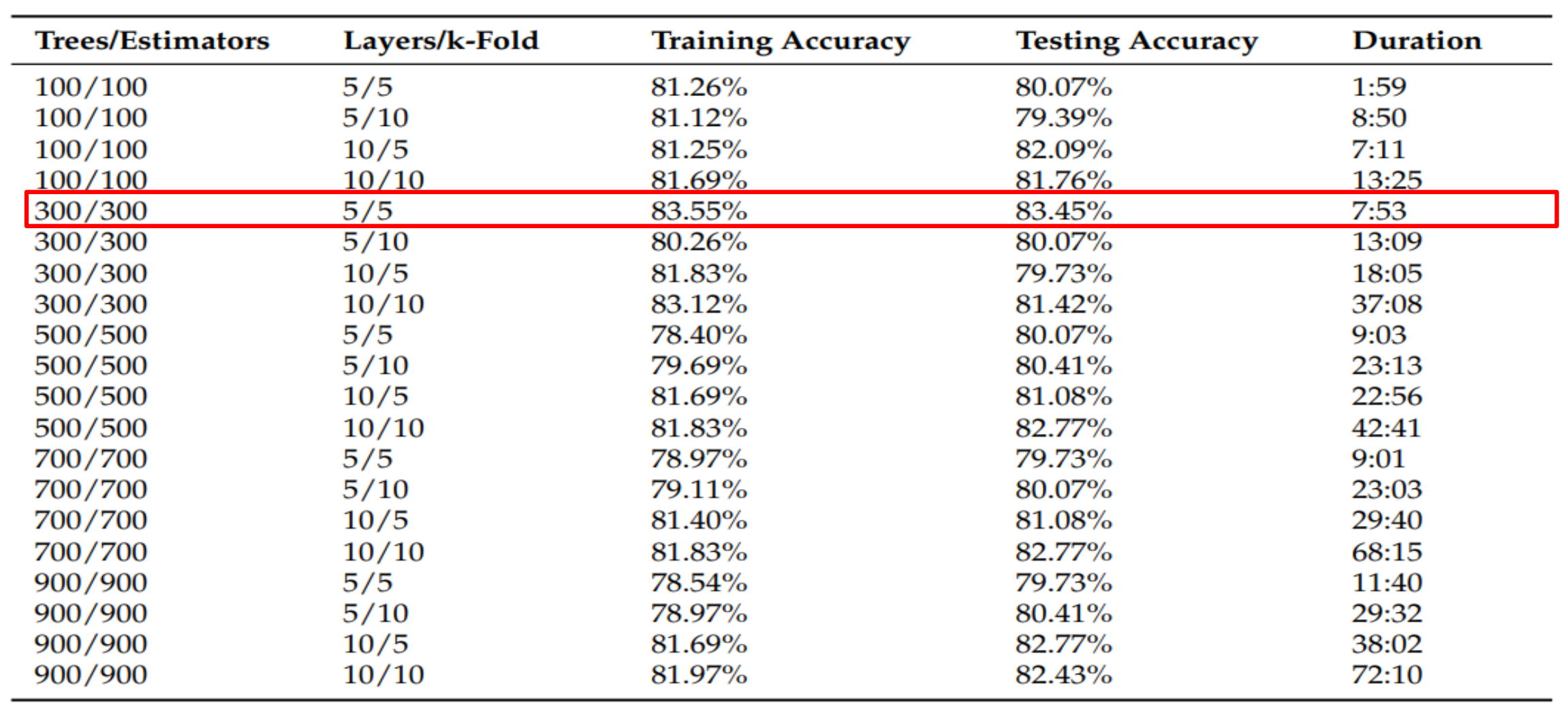
Table 5. Integrated Profiles Classification Accuracy for 10 Subtypes.



Integrated Profile	Training Accuracy	Testing Accuracy
GE	70.76%	77.55%
Clinical	40.51%	43.54%
CNA	53.21%	51.70%
CNV	55.06%	53.06%
GE + Clinical	67.76%	73.47%
Clinical + GE	71.61%	75.85%
GE + CNV	69.76%	76.19%
GE + CNA	66.76%	70.75%
Clinical + CNA	58.63%	60.20%
Clinical + CNV	61.06%	57.82%
CNA + CNV	55.21%	51.70%
Clinical + CNV + GE	68.33%	72.79%
Clinical + CNA + GE	68.90%	71.09%
Clinical + CNA + CNV	60.34%	61.56%
GE + CNA + CNV	71.61%	74.83%
GE + CNV + CNA + Clinical	68.62%	74.15%
GE + CNA + CNV + Clinical	70.33%	74.49%
GE + Clinical + CNA + CNV	68.76%	72.79%

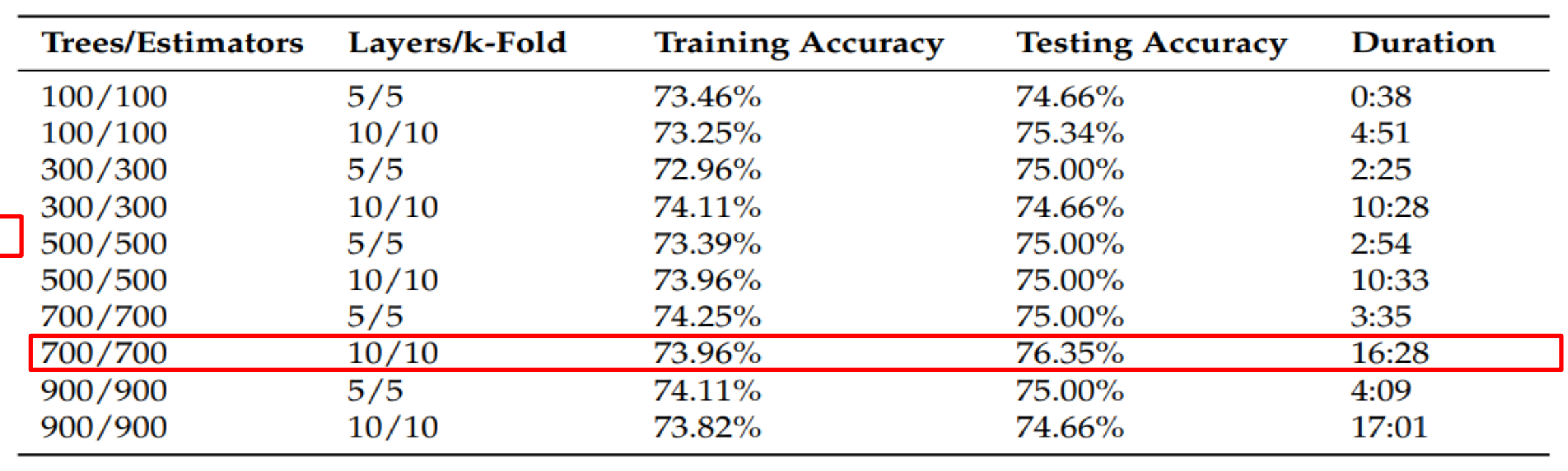
Results (Pam50)

Table 7. Gene Expression—5 Subtypes.



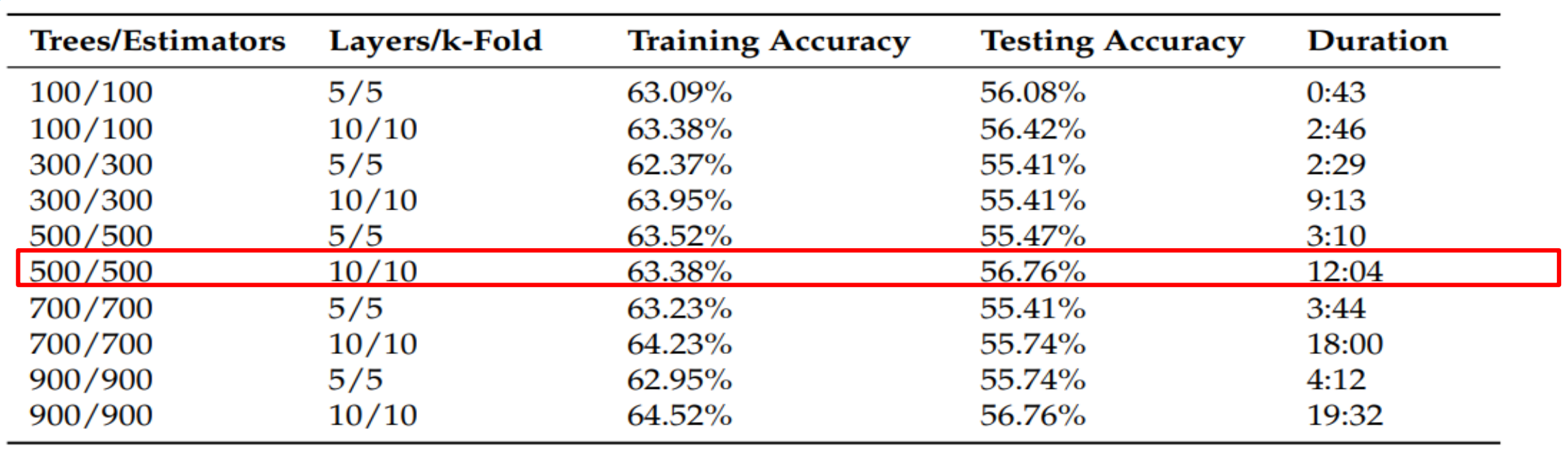
Trees/Estimators	Layers/k-Fold	Training Accuracy	Testing Accuracy	Duration
100/100	5/5	81.26%	80.07%	1:59
100/100	5/10	81.12%	79.39%	8:50
100/100	10/5	81.25%	82.09%	7:11
100/100	10/10	81.69%	81.76%	13:25
300/300	5/5	83.55%	83.45%	7:53
300/300	5/10	80.26%	80.07%	13:09
300/300	10/5	81.83%	79.73%	18:05
300/300	10/10	83.12%	81.42%	37:08
500/500	5/5	78.40%	80.07%	9:03
500/500	5/10	79.69%	80.41%	23:13
500/500	10/5	81.69%	81.08%	22:56
500/500	10/10	81.83%	82.77%	42:41
700/700	5/5	78.97%	79.73%	9:01
700/700	5/10	79.11%	80.07%	23:03
700/700	10/5	81.40%	81.08%	29:40
700/700	10/10	81.83%	82.77%	68:15
900/900	5/5	78.54%	79.73%	11:40
900/900	5/10	78.97%	80.41%	29:32
900/900	10/5	81.69%	82.77%	38:02
900/900	10/10	81.97%	82.43%	72:10

Table 8. Clinical Data—5 Subtypes.



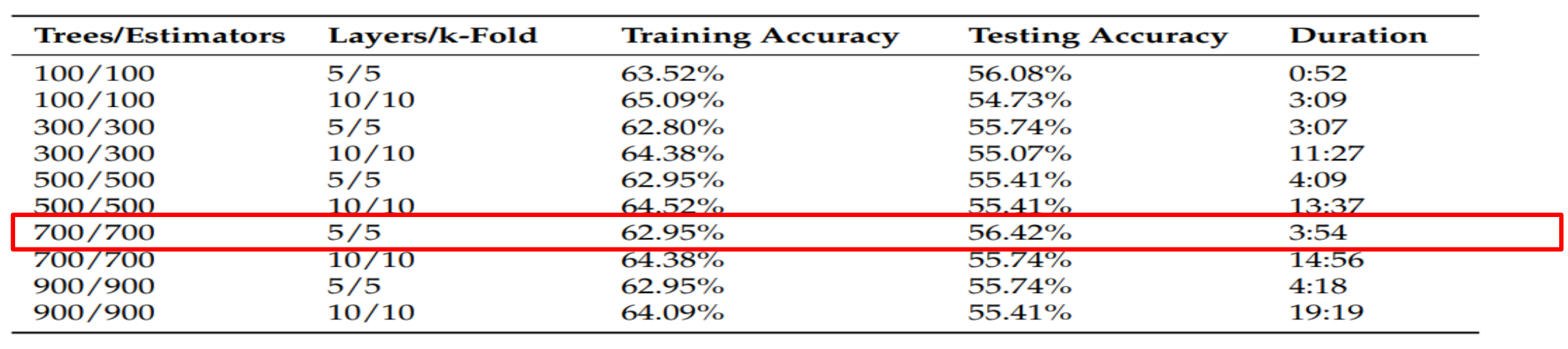
Trees/Estimators	Layers/k-Fold	Training Accuracy	Testing Accuracy	Duration
100/100	5/5	73.46%	74.66%	0:38
100/100	10/10	73.25%	75.34%	4:51
300/300	5/5	72.96%	75.00%	2:25
300/300	10/10	74.11%	74.66%	10:28
500/500	5/5	73.39%	75.00%	2:54
500/500	10/10	73.96%	75.00%	10:33
700/700	5/5	74.25%	75.00%	3:35
700/700	10/10	73.96%	76.35%	16:28
900/900	5/5	74.11%	75.00%	4:09
900/900	10/10	73.82%	74.66%	17:01

Table 9. CNV Data—5 Subtypes.



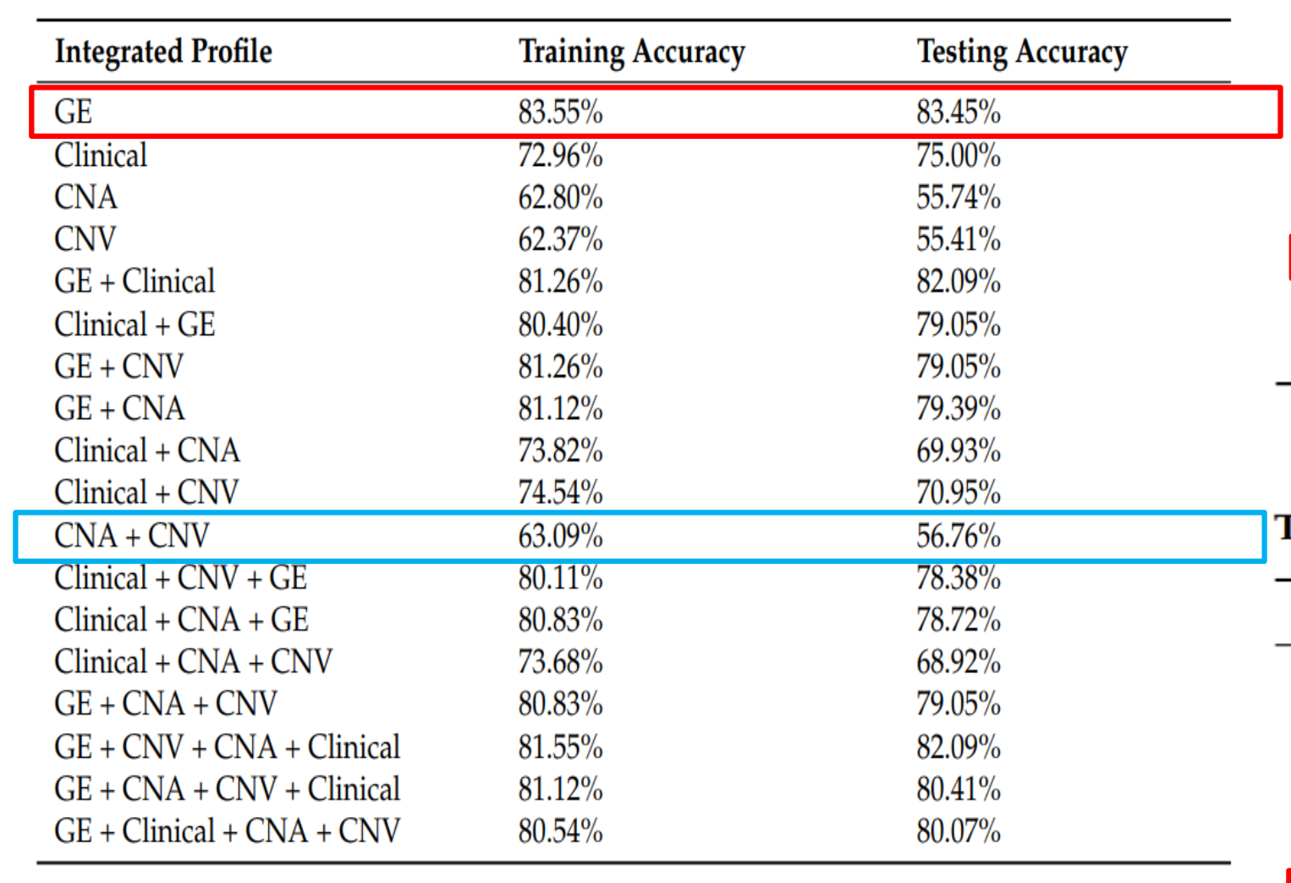
Trees/Estimators	Layers/k-Fold	Training Accuracy	Testing Accuracy	Duration
100/100	5/5	63.09%	56.08%	0:43
100/100	10/10	63.38%	56.42%	2:46
300/300	5/5	62.37%	55.41%	2:29
300/300	10/10	63.95%	55.41%	9:13
500/500	5/5	63.52%	55.47%	3:10
500/500	10/10	63.38%	56.76%	12:04
700/700	5/5	63.23%	55.41%	3:44
700/700	10/10	64.23%	55.74%	18:00
900/900	5/5	62.95%	55.74%	4:12
900/900	10/10	64.52%	56.76%	19:32

Table 10. CNA Data—5 Subtypes.



Trees/Estimators	Layers/k-Fold	Training Accuracy	Testing Accuracy	Duration
100/100	5/5	63.52%	56.08%	0:52
100/100	10/10	65.09%	54.73%	3:09
300/300	5/5	62.80%	55.74%	3:07
300/300	10/10	64.38%	55.07%	11:27
500/500	5/5	62.95%	55.41%	4:09
500/500	10/10	64.52%	55.41%	13:37
700/700	5/5	62.95%	56.42%	3:54
700/700	10/10	64.38%	55.74%	14:56
900/900	5/5	62.95%	55.74%	4:18
900/900	10/10	64.09%	55.41%	19:19

Table 11. Integrated Profiles Classification Accuracy for 5 Subtypes.



Integrated Profile	Training Accuracy	Testing Accuracy
GE	83.55%	83.45%
Clinical	72.96%	75.00%
CNA	62.80%	55.74%
CNV	62.37%	55.41%
GE + Clinical	81.26%	82.09%
Clinical + GE	80.40%	79.05%
GE + CNV	81.26%	79.05%
GE + CNA	81.12%	79.39%
Clinical + CNA	73.82%	69.93%
Clinical + CNV	74.54%	70.95%
CNA + CNV	63.09%	56.76%
Clinical + CNV + GE	80.11%	78.38%
Clinical + CNA + GE	80.83%	78.72%
Clinical + CNA + CNV	73.68%	68.92%
GE + CNA + CNV	80.83%	79.05%
GE + CNV + CNA + Clinical	81.55%	82.09%
GE + CNA + CNV + Clinical	81.12%	80.41%
GE + Clinical + CNA + CNV	80.54%	80.07%

Conclusion

This research proposes a Deep Forest classifier for the IntClust and Pam50 breast cancer subtypes. The experiments are carried out using different combinations of trees and estimators, specifically 100, 300, 500, 700, and 900, as well as layers and k-folds of 5 and 10. Gene expression alone significantly gave the best performance, with an accuracy of 83.45% for 5 subtypes and 77.55% for 10 subtypes, and time about 5 s for 10 subtypes, and 7 s for 5 subtypes. The integration of datasets did not give any improvement, where for the 5 subtypes, CNA and CNV data achieved 56.7%, while CNA alone achieved 55.74%, and CNV alone achieved 55.41%. For the 10 subtypes, the clinical data achieved 43.5%, CNV alone achieved 53.06%, and the CNA alone achieved 51.70%. The integrated clinical data with the CNA achieved 60.20%, the integrated clinical data with CNV achieved 57.82%, and the integrated clinical data with both CNA and CNV achieved 61.56%. It is concluded that using gene expression alone achieves comparable results.