mathematics

Article

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

Ala'a El-Nabawy [1,†], Nahla A. Belal [2,3,*,†] and Nashwa El-Bendary [2,3,†]

[1] Orange Labs., Smart Village 12511, Giza Governorate, Egypt; aelnabwy@gmail.com
[2] College of Computing and Information Technology, Arab Academy for Science, Technology, and Maritime Transport, Smart Village, Giza 12577, Egypt; nashwa.elbendary@aast.edu
[3] College of Computing and Information Technology, Arab Academy for Science, Technology, and Maritime Transport, Aswan 81531, Egypt
* Correspondence: nahlabelal@aast.edu
† These authors contributed equally to this work.

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.

Keywords: METABRIC dataset; breast cancer subtyping; deep forest; multi-omics data

check for updates

Citation: El-Nabawy, A.; Belal, N.A.; El-Bendary, N. A Cascade Deep Forest Model for Breast Cancer Subtype Classification Using Multi-Omics Data. *Mathematics* **2021**, *9*, 1574. https://doi.org/10.3390/math9131574

Academic Editor: Snezhana Gocheva-Ilieva

Received: 18 May 2021
Accepted: 27 June 2021
Published: 4 July 2021

1. Introduction

Breast cancer is one of the main causes of cancer death worldwide. Computer-aided diagnosis systems aim to reduce the cost of diagnosis while maintaining the same efficiency of the process. Conventional classification methods depend on feature extraction methods, and to overcome many difficulties of those feature-based methods, deep learning techniques are becoming important approaches to adopt.

Breast cancer classifiers use different methods and different data. Some methods use images [1–3], some use biological data [4,5], and some integrate many types of data [6,7].

Many recent studies have incorporated deep learning, and especially Deep Forest in their studies. Deep forest is still a young research area. However, a lot of work has shown promising results for employing this model in healthcare systems and bioinformatics. For example, reference [8] presents GcForest-PPI, which is a model that uses Deep Forest for the prediction of protein–protein interaction networks. Their model showed and enhanced prediction accuracy and a suggested improvement in drug discovery. The work in reference [9] combined Deep Forest and autoencoders, for the prediction of lncRNA-miRNA interaction, and their model showed improved results. Additionally, reference [10] uses deep learning with Random Forests on the METABRIC dataset, to make use of the different types of data. Their results enhanced the sensitivity values by 5.1%. Additionally, several studies have used deep learning and Deep Forest with the histopathological images data and mammography images [2,3,11–13].

IntClust is breast cancer subtyping technique into 10 subtypes. The IntClust subtyping is dependent on molecular drivers that are obtained using combined genomics and transcriptomic data. Pam50 is another breast subtyping method, and it consists of 5 subtypes [14].

For breast cancer subtyping, several data combined with several techniques have been employed. Since 2011, Mendes et al. [15] employed a clustering method with gene expression data, and showed that the subtyping obtained confirms with already established subtypes. Gene expression and methylation data have been used with different Random Forests models [16]. The study showed that gene expression data outperforms methylation data; however, some features are only discovered using methylation. The work done in [17] uses histopathological images, which was covered above; however, this work is specific to breast cancer. Histopathological images were used with a Stacked Sparse Autoencoder (SSAE), which is a deep learning strategy, and has shown improved performance, with an F-measure of 84.49% and an average area under Precision-Recall curve (AveP) 78.83%. Reference [18] proposes a method that uses histopathological images and extracts features using a convolutional neural network (CNN). The CNN designed obtained an enhanced performance, which was also slightly better when different CNNs were fused. In 2017, Bejnordi et al. [19] applied deep learning algorithms to detect lymph nodes for breast cancer in whole-slide pathology images cans and proposed an improved diagnosis.

Deep forest was used in [20] to classify cancer subtypes, and the model suffered from overfitting and ensemble diversity challenges because of small sample size and high dimensionality of biological data. This is overcome in the use of extensive biological data in this research by employing the METABRIC dataset.

A deep learning technique has also been proposed in [21], and it shows a higher performance than traditional machine learning methods for cancer subtype classification.

The small data size and imbalanced data problems have been addressed in [22], where an enhanced algorithm to handle the data was proposed, combining traditional techniques with deep learning methods. The results obtained confirmed that deep learning enhances performance; in addition to that, methylation data were suggested to be effectively used to improve diagnosis of cancer.

In reference [5], a deep neural networks model uses multi-omics data to classify breast cancer subtypes. The types of omics data used were mRNAdata, DNAmethylation data, and copy number variation (CNV), and the system achieved higher accuracy and area under curve.

Additionally, the authors in reference [6] confirm that deep neural networks perform better than traditional methods as it automatically extracts features from raw data. The data used is copy number alteration and gene expression data for breast cancer patients (METABRIC). The model presented integrates the datasets and the performance is superior to other models.

Moreover, the authors in reference [23] use a network propagation method with a deep embedded clustering (DEC) method to classify the breast tumors into four subtypes. Reference [24] employs deep learning techniques for feature extraction and classification to classify breast cancer lesions using mammograms. The system achieved high accuracy using fused deep features for two datasets compared to similar methods. Additionally, Zhang et al. [25] used a convolutional neural network (CNN) and a recurrent neural network (RNN) to classify three breast cancer subtypes using MRI data. The accuracy achieved was 91% and 83%, using CNN and CLSTM, respectively. Reference [26] combined graph convolutional network (GCN) and convolutional neural network (CNN) to analyze breast mammograms with an accuracy of 96.10 ± 1.60%. Reference [27] also used deep learning on histopathological images for breast cancer subtyping. Deep feature fusion and enhanced routing (FE-BkCapsNet) is used, and results achieved over 90% of accuracy.

Among many others, references [4,7] integrated multiple datasets to address the problem of cancer subtype classification. Xue et al. [4] used integrated omics data for cancer subtype classification using a deep neural forest model, HI-DFNForest, proposing an improved performance. It has been shown that integration of multi-omics data may enhance cancer subtype classification. However, not all types of data are available extensively, and not all types of data add to the classification process. In addition, employing all types of data imposes the restriction of very high time requirements. In the process of diagnosis of cancer, time becomes an important issue due to the critical cases of patients. In reference [7], it was shown that without using sampling techniques on the METABRIC dataset, the results obtained for classification of cancer subtypes are low. Additionally, there is a very big challenge regarding images, where the number of available samples is only 208, much lower than other omics data. Moreover, the techniques used to achieve the highest accuracy obtained were relatively high. In this paper, it was shown that gene expression alone can achieve comparable results on the Deep Forest configuration employed. In addition to achieving very fast performance.

The objective of this work is to extend the previous research [7] by employing a Deep Forest model for using feature combining and classifying the generated integrative data profiles, and enhancing the previously proposed framework through using the full dimension gene expression data and examining the computational performance.

In this paper, a cascade Deep Forest is employed to classify breast cancer subtypes for both subtyping, IntClust (10 subtypes) and Pam50 (5 subtypes), using the METABRIC datasets, namely clinical, gene expression, CNA, and CNV. The full dataset is used, without dimensionality reduction, and without sampling. Several configurations for the cascade Deep Forest are employed and the results obtained are an accuracy of 83.45% for 5 subtypes and 77.55% for 10 subtypes. Other obtained performance metrics also confirm the outperformance of employing gene expression solely, where the precision, recall, specificity, F1-measure, Jaccard, Hamming loss, and Dice are 0.822, 0.774, 0.961, 0.772, 0.640, 0.225, 0.709, respectively, for 10 subtypes. The measures are 0.8421, 0.833, 0.904, 0.820, 0.711, 0.166, and 0.852, for the 5 subtypes, respectively. The precision and Dice measures are slightly higher for the integrated profile gene expression, clinical, CNA, and CNV. 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.

The main contribution of this paper is to:

- Employ the omics METABRIC sub-datasets of gene expression, CNA, and CNV, in addition to the clinical dataset in full dimension without sampling.
- Develop a cascade Deep Forest-based model for breast cancer subtype classification using multi-omics data.
- Obtain comparable results using only omics data without using histopathological images.
- Improve the classification time for breast cancer subtyping through using the cascade Deep Forest classifier.

The rest of this paper is organized as follows. The methods used in evaluating the employed model for breast cancer subtyping are elaborated in Section 2. In Section 3, experimental results are presented, followed by a discussion in Section 4. Finally, a conclusion is presented in Section 5.

2. Materials and Methods

The proposed system in this manuscript uses integrative clinical data and genomics data generated from the extraction and combination of the gene expression, Copy Number Aberrations (CNA), and Copy Number Variations (CNV) feature sets from the genomics dataset.

As depicted in Figure 1, 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. The following subsections explain each phase in more details.

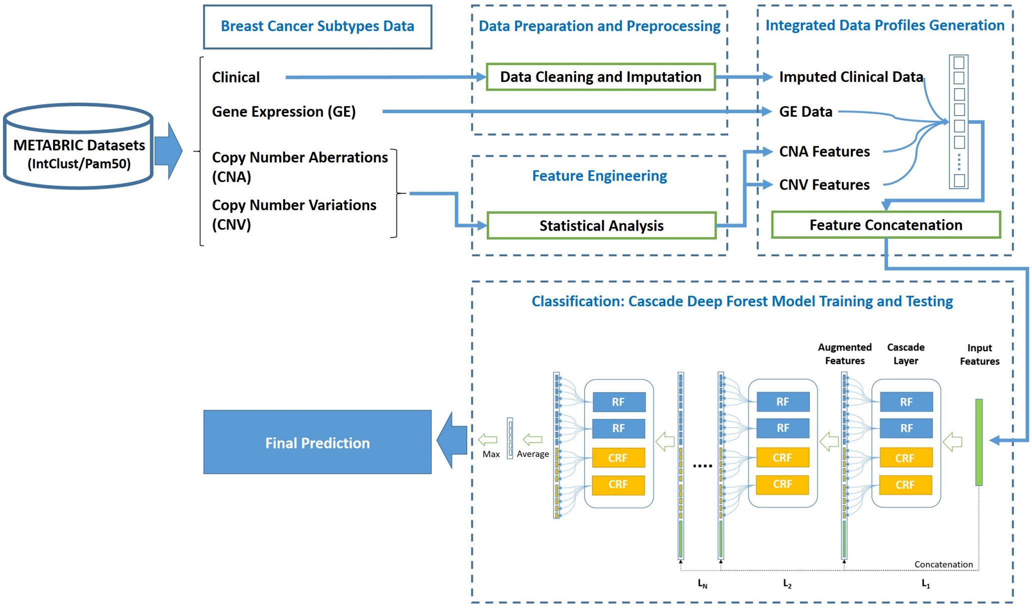
Breast Cancer Subtypes Data
Data Preparation and Preprocessing
Integrated Data Profiles Generation
METABRIC Datasets (IntClust/Pam50)
Clinical
Gene Expression (GE)
Copy Number Aberrations (CNA)
Copy Number Variations (CNV)
Data Cleaning and Imputation
Imputed Clinical Data
GE Data
CNA Features
CNV Features
Feature Engineering
Statistical Analysis
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

Figure 1. General structure of the proposed approach.

2.1. Data Acquisition

In this phase the breast cancer subtypes dataset used in the conducted experiments is the METABRIC dataset that contains several sub-datasets. The datasets considered for the research conducted in this manuscript are: clinical dataset, gene expression dataset, Copy Number Aberrations (CNA) dataset, and Copy Number Variations (CNV) dataset. The source of the gene expression, CNA and CNV data are the European Genome-phenome Archive (EGA) platform with the accession number, DAC ID, EGAC00001000484 [28]. However, the clinical data are obtained from the Synapse platform [29]. The datasets obtained contain datasets for validation and discovery.

The clinical data available is categorized into four main categories of 27 features. First, personal, which contains only the age at diagnosis. Second, the clinical pathology data, which is data about the tumor, including, size, lymph nodes data, grade, histological type, different hormonal levels, and other features. Third, the treatment category, which indicates the type of treatment received by the patient. Fourth, survival features, which are the status and time.

The Copy Number Aberration (CNA) dataset contains a total of 13 features describing chromosome regions, namely information in somatic tissues about the markers count and mutation type. The dataset also contains information about location, including five features.

In addition to, information about the number of genes in each segment and mutation type described in seven features. Similarly, the Copy Number Variation (CNV) dataset contains 13 features describing chromosome regions in germline tissues about the markers number, mutation type, location, and genes count.

The gene expression dataset contains 48,803 genes expressed using Illumina Sequenced HT 12 array v3.

At the end of breast cancer subtypes data acquisition phase, each obtained dataset is submitted to a data preparation and preprocessing module in phase 2.

2.2. Data Preparation and Preprocessing

This section presents a discussion for the datasets preparation and preprocessing.

2.2.1. Data Preparation

The METABRIC dataset was obtained as explained earlier, and the discovery part was extracted with its labels to include the clinical dataset, gene expression, CNA, and CNV sets. Each of the resulting datasets was prepared according to the following steps:

1. Submitting the CNA and CNV feature files to the statistical analysis feature engineering stage in the data preprocessing stage.
2. Submitting the clinical dataset to the preprocessing stage for data cleaning and imputation.
3. Transposing the gene expression data using Equation (1), then submitting it without any preprocessing to the third phase of the proposed system, where $(A)_{ji}$ represents the matrix of the original gene expression data and $(A^T)_{ij}$ represents the resulted transposed matrix.

$$(A^T)_{ij} = (A)_{ji} \quad \forall i, j. \tag{1}$$

2.2.2. Data Preprocessing

Following the preparation of data, the clinical dataset is submitted to the preprocessing stage for data cleaning and imputation. In this stage, a statistical analysis feature engineering scheme is applied to the CNA and CNV datasets. For that scheme, the frequency and the actual segment percentages Amplification (AMP), Insertions (GAIN), Homozygous Deletion (HOMD), Heterozygous Deletion (HETD) and Neutral (NEUT) were calculated for each chromosome. Figure 2 shows the detailed steps for statistical feature engineering of CNA and CNV data.

The clinical dataset features are encoded using textual categorical encoding. First, features with missing values more than 50% (like NOT_IN_OSLOVAL_P53_mutation_type), and features with 90% blank values (like NOT_IN_OSLOVAL_P53_mutation_details) are deleted. Data imputation [30] is performed on other features missing values with lower ratios. On the other hand, the gene expression dataset is submitted to the integrated data profiles generation phase as it is, without any preprocessing.

2.3. Integrated Data Profiles Generation

At the end of phase 2, the resulted CNA and CNV statistical feature sets, imputed clinical, and the features of gene expression datasets are submitted to phase 3 of integrated data profiles generation, to be concatenated and then generate the output set of integrated data profiles.

2.4. Cascade Deep Forest Based Classification

During phase 4 of the proposed system and after generating integrated data profiles through feature concatenation of imputed clinical data, GE data, and statistically engineered CNA and CNV data, a cascade Deep Forest model is applied for cancer subtype classification.As a preparatory step before the classification phase, the features obtained in phase 3 are split into subsets of 2/3 for training and 1/3 for testing.

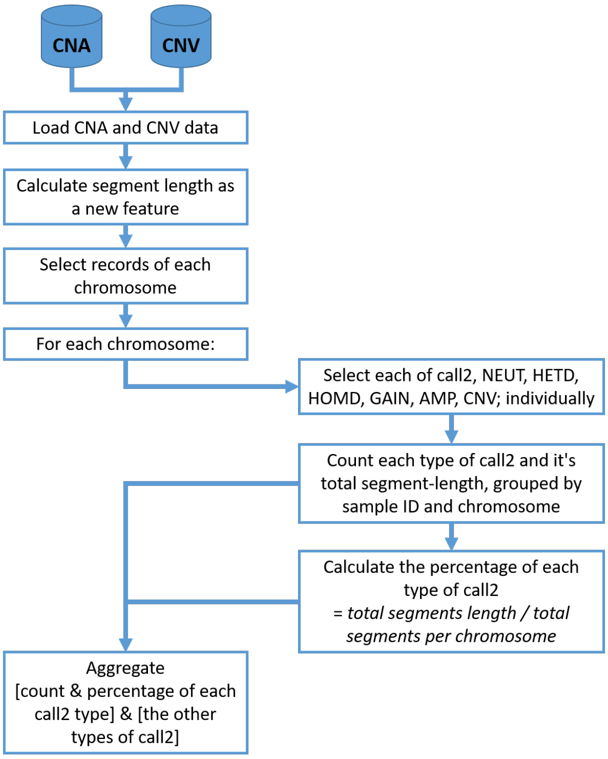
CNA
CNV
Load CNA and CNV data
Calculate segment length as a new feature
Select records of each chromosome
For each chromosome:
Select each of call2, NEUT, HETD, HOMD, GAIN, AMP, CNV; individually
Count each type of call2 and it's total segment-length, grouped by sample ID and chromosome
Calculate the percentage of each type of call2 = *total segments length / total segments per chromosome*
Aggregate [count & percentage of each call2 type] & [the other types of call2]

Figure 2. Statistical feature engineering of CNA and CNV data.

The motivation for considering cascade Deep Forest model in the proposed breast cancer subtype classification approach is that conventional supervised machine learning classifiers typically work with labeled data as well as neglecting a considerable amount of data with insufficient information. Consequently, small sample size of training data limits the progress in designing appropriate classifiers. Moreover, several challenges may limit the application of common conventional machine leaning models, such as Support Vector Machine (SVM) and Random Forests, to the task of cancer subtype classification. The sounding challenge is strengthening the risk of overfitting in training, which is characterized by using small sample size and high dimensionality of biology multi-omics data. Additionally, class-imbalance is a very common situation in multi-omics data, which augments the difficulties of model learning with the risk of weakening the ability of model estimation for large sequencing bias. Although several approaches have been recently developed to address the stated challenges [31,32], limited alternatives are proposed with validated methods for small-scale multi-omics data. Additionally, more accurate and robust methods still need further developments for achieving accurate of breast cancer subtype classification.

On the other hand, compared to the typical architecture of convolutional deep neural networks (DNNs) with several convolutional layers and fully connected layers, the DNNs are also highly prone to overfitting, with more chances for convergence to local optimums, when providing imbalanced or relatively small-size training data. However, dropout and regularization methods are widely applied to alleviate that problem, overfitting is still an inescapable problem for DNNs. Thus, the state-of-the-art recommended the cascade Deep Forest model as an efficient alternative to DNNs for learning hyper-level representations in more optimized way.

The cascade Deep Forest model fully uses the characteristics of both deep neural networks and ensemble models. The cascade Deep Forest learn features of class distribution by assembling decision tree-based forests while supervising the input, rather than the overhead of applying forward and backward propagation algorithms to learn hidden variables as in deep neural networks [33].

The cascade forest follows a supervised learning scheme based on layers, which employs ensemble Random Forests to obtain a class distribution of features that results in more precise classification [20,34]. The feature importance in the cascade Deep Forest model is not taken into account among multiple layers during the feature representation training. Accordingly, the prediction accuracy obtained is highly affected by the number of decision trees in each forest, especially with small-scale or imbalanced data, as it is critical in the construction of decision trees, where the discriminative features are used to decide splitting nodes. Figure 3 shows the architecture of the employed cascade forest.

As illustrated in Figure 3, considering the used cascade Deep Forest model, each level of the cascade consists of two Random Forests (RF) (the blue blocks) and two Completely Random Forests (CRF) (the yellow blocks). Therefore, suppose there are n subclasses to predict, then each forest should output an n-dimensional class vector, which is then concatenated for representing the original input.

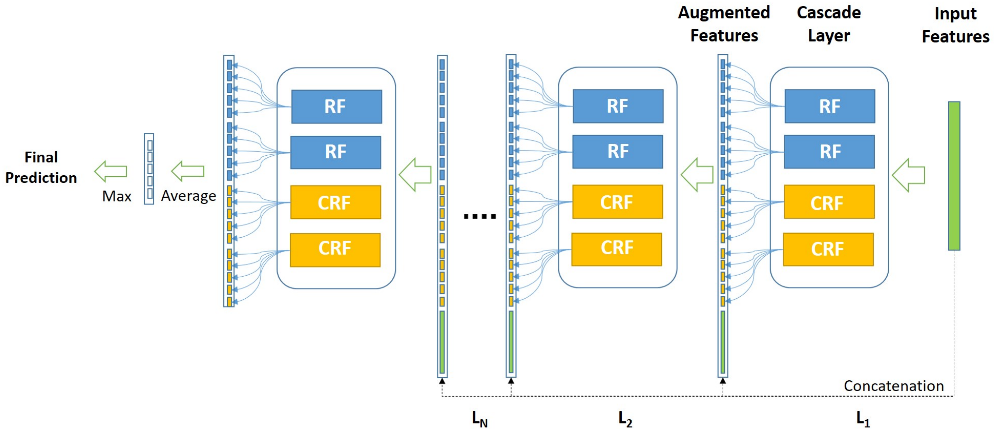

Figure 3. General structure of the cascade Deep Forest network.

3. Results

This section shows the results for different Deep Forest configurations against the 10 subtypes (IntClust) and the 5 subtypes (Pam50). The results in this paper are obtained using the full dimension dataset of the gene expression. The experiments use the whole 48,803 features set. Initially, the Deep Forest configuration was used with the dimensionally reduced gene expression data and the accuracy results were as low as 27%. This led to using the whole gene expression dataset of 48,803 features and the performance obtained showed promising results with high configurations. The performance is shown relative to the time taken per each run. Different Deep Forest configurations were used against the 10 and 5 subtypes. The number of estimators is increased to 900 to make sure that there is no increase in the accuracy. It is not easy to decide the most fitting configuration. For the 10 subtypes, the accuracy reached is 77.55% for the gene expression dataset using 100 trees in each forest, 100 estimators, 5 layers, and 10 k-folds. with time 5:08 s. For the 5 subtypes, the highest accuracy is achieved for the gene expression data and relatively for the CNV and CNA, using 300 trees in each forest, 300 estimators, 5 layers, and 5 k-folds. This accuracy is 83.45% for gene expression, with approximately 55% of accuracy for CNV and CNA, which is comparable to other configurations regarding time. However, performance was not the highest for the clinical data. The time achieved is 7:53 s.

Both experiments were performed using different number of trees per forest and different number of estimators. Specifically, 100, 300, 500, 700 for trees and estimators, using also 900 for the 5 subtypes. Those numbers were used once with 10 layers with 10 k-fold and once with 5 layers with 5 k-fold. To be more confident about the most fitting architecture, another run using 5 layers with 10 K-fold and 10 layers with 5 k-folds was performed. An extra experiment for gene expression data was performed using another combination of layers and k-fold to confirm the results, using 5 layers with 5 k-fold and 10 layers with 10 k-fold.

3.1. Results for Breast Cancer 10 Subtypes

Tables 1–4, show the results for 10 subtypes using all datasets, except images, since the number of samples in the images dataset is much lower than other datasets. The tables show that for gene expression, the highest accuracy achieved is 77.55%, with 100 trees, 100 estimators, 5 layers, and 10 k-folds, with time of 5:08 s. For clinical data, the highest accuracy achieved is 44.22%, with 100 trees, 100 estimators, 5 layers, and 5 k-folds, with time of 0:41 s. For CNV and CNA, the accuracy achieved is 55.78% and 53.40%, respectively, with 500 trees, 500 estimators, 10 layers, 10 k-folds, with a time of 23:01 and 22:40 s, respectively.

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

Tables 5 and 6, show results of different integrated data profiles for the IntClust (10 subtypes). The highest accuracy achieved for the 10 subtypes is 75.85%, for integrating clinical data with gene expression; however, gene expression alone still achieves a higher accuracy, as reported earlier. Please note that the accuracy reported in Table 6 is the overall accuracy, while the accuracy in the other tables is the reached max layer accuracy. This is the reason for the slight variation in the accuracy values. Moreover, the precision, recall, specificity, F1-measure, Jaccard, Hamming loss, and Dice are 0.822, 0.774, 0.961, 0.772, 0.640, 0.225, 0.709, respectively. The results of the obtained performance and statistical measures confirm the superiority of employing only gene expression.

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%

Table 6. Integrated Profiles Classification Performance Metrics for 10 Subtypes.

Data Profile	Accuracy	Precision	Recall (Sensitivity)	Specificity	F1-Measure	Jaccard	Hamming-Loss	Dice
GE	77.47%	0.822	0.774	0.961	0.772	0.640	0.225	0.709
Clinical	41.97%	0.369	0.4197	0.919	0.367	0.266	0.580	0.320
CNA	49.48%	0.445	0.494	0.926	0.450	0.309	0.505	0.297
CNV	53.24%	0.554	0.532	0.932	0.491	0.341	0.467	0.340
GE + Clinical	72.7%	0.789	0.726	0.956	0.708	0.573	0.273	0.561
GE + CNV	76.11%	0.812	0.761	0.961	0.749	0.613	0.239	0.644
GE + CNA	70.64%	0.749	0.706	0.951	0.683	0.546	0.293	0.555
Clinical + CNA	58.02%	0.5286	0.5802	0.938	0.541	0.402	0.419	0.522
Clinical + CNV	57.6%	0.521	0.576	0.938	0.537	0.3959	0.423	0.44
CNA + CNV	51.54%	0.472	0.515	0.9303	0.463	0.323	0.484	0.217
Clinical + CNV + GE	69.28%	0.753	0.693	0.950	0.676	0.532	0.307	0.611
Clinical + CNA + GE	70.9%	0.767	0.7098	0.953	0.6919	0.552	0.2901	0.623
CNA + CNV + Clinical	59.01%	0.557	0.6143	0.939	0.578	0.4387	0.3856	0.590
GE + CNA + CNV	74.74%	0.809	0.747	0.958	0.732	0.599	0.252	0.678
GE + CNV + CNA + Clinical	74.06%	0.809	0.7406	0.955	0.723	0.590	0.259	0.621

3.2. Results for Breast Cancer 5 Subtypes

For the 5 subtypes, Tables 7–10 show results of different datasets. For gene expression, the highest accuracy achieved is 83.45%, with 300 trees, 300 estimators, 5 layers, and 5 k-folds, with time of 7:53 s. For clinical data, the highest accuracy achieved is 76.35%, with 700 trees, 700 estimators, 10 layers, and 10 k-folds, with time of 16:28 s. For CNV, the accuracy achieved is 56.76%, with 500 trees, 500 estimators, 10 layers, 10 k-folds, with a time of 12:04 s. For CNA, the accuracy achieved is 56.42%, with 700 trees, 700 estimators, 5 layers, 5 k-folds, with a time of 3:54 s.

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

Tables 11 and 12 show results of different integrated data profiles for the Pam50 (5 subtypes) subtyping. The highest accuracy achieved for the 5 subtypes is 80.41%, for CNA, CNV, gene expression, and clinical data. Still, gene expression alone achieves a higher prediction accuracy. The slight variation in accuracy reported in Tables 6 and 12 is that it is the overall accuracy, while the accuracy in the other tables is the reached max layer accuracy. In addition to the obtained accuracy, precision, recall, specificity, F1-measure, Jaccard, Hamming loss, and Dice are 0.8421, 0.833, 0.904, 0.820, 0.711, 0.166, and 0.852, respectively. The precision and Dice measures are slightly higher for the integrated profile gene expression, clinical, CNA, and CNV. However, the remaining measures are in favor of the gene expression profile, which confirm the outperformance of the suggested data profile.

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%

Table 12. Integrated Profiles Classification Performance Metrics for 5 Subtypes.

Data Profile	Accuracy	Precision	Recall (Sensitivity)	Specificity	F1-Measure	Jaccard	Hamming-Loss	Dice
GE	83.38%	0.8421	0.833	0.904	0.820	0.711	0.166	0.852
Clinical	75.59%	0.761	0.755	0.892	0.748	0.611	0.244	0.796
CNA	55.59%	0.5474	0.555	0.705	0.500	0.360	0.444	0.432
CNV	55.93%	0.553	0.559	0.702	0.503	0.362	0.440	0.382
GE + Clinical	82.03%	0.848	0.8203	0.868	0.797	0.687	0.179	0.857
GE + CNV	79.05%	0.765	0.786	0.846	0.754	0.634	0.213	0.845
GE + CNA	79.39%	0.827	0.793	0.849	0.767	0.644	0.206	0.842
Clinical + CNA	67.7%	0.7009	0.677	0.807	0.654	0.504	0.322	0.644
Clinical + CNV	71.1%	0.725	0.711	0.818	0.694	0.546	0.288	0.783
CNA + CNV	56.76%	0.479	0.566	0.697	0.505	0.366	0.433	0.4
Clinical + CNV + GE	78.6%	0.769	0.786	0.841	0.756	0.635	0.213	0.842
Clinical + CNA + GE	78.9%	0.772	0.789	0.8447	0.759	0.639	0.210	0.842
Clinical + CNA + CNV	68.92%	0.644	0.688	0.809	0.659	0.514	0.311	0.727
GE + CNA + CNV	78.9%	0.768	0.789	0.846	0.759	0.64	0.210	0.842
GE + CNV + CNA + Clinical	82.37%	0.848	0.823	0.872	0.801	0.693	0.176	0.857

4. Discussion

In this work, the experiments first make use of different Deep Forest configurations on each dataset solely. Gene expression alone significantly gave the best performance, where the accuracy was 83.45% for 5 subtypes using 300 estimators, 5 layers, 5 k-folds, and the accuracy was 77.55% for 10 subtypes using 100 estimators, 5 layers, and 10 k-folds. This was concluded after experimenting 100,300,500,700 and 900 estimators across 5, 10 layers and 5, 10 k-folds. The integration of datasets was performed by concatenating the datasets and applying the best configuration of Deep Forest to classify it. The results reached did not give any improvement over the highest accuracy reached using gene expression. However, for the 5 subtypes, the integrated profile CNA + CNV achieved 56.7%, while CNA alone achieved 55.74%, and CNV alone achieved 55.41%. For the 10 subtypes, the clinical data alone 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%.

In the research [7], an accuracy of 88.36% was achieved for IntClust (10 subtypes) subtyping using Linear-SVM. The accuracy achieved was using the data profile of clinical, gene expression, CNA, and CNV datasets. For the Pam50 subtyping (5 subtypes), the accuracy was 97.1% using Linear-SVM and E-SVM classifiers, with all data including histopathological images features. However, the images data are not comprehensive, as they are only available for 208 samples, unlike other data, which are extensively available for all patients in the dataset. Moreover, the time taken to obtain the above-mentioned accuracy is extensive. hence, the Deep Forest used in this paper, makes use of the gene expression data alone to achieve comparable results without using any sampling techniques. In [7], the highest accuracy achieved used SMOTE sampling. The highest accuracy achieved among all different data profiles was 71.35% for IntClust, which was outperformed by the Deep Forest configuration in this paper, achieving 77.55%, and a running time of 5:08 s, which is extensively less than the model proposed in [7]. For the gene expression data alone in [7], the accuracy only reached 46.08%. Similarly, for Pam50 (5 subtypes), gene expression alone achieved 78.85%, while the highest recorded accuracy was 80.66%, without images, for the gene expression and clinical data profiles. However, the proposed Deep Forest configuration achieved up to 83.45% of accuracy and 7:53 s run time.

The current study could be further expanded by examining the technique on more datasets for breast cancer subtyping. Additionally, other deep learning methods could be employed to verify the robustness of using gene expression data.

5. Conclusions

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.

Author Contributions: Conceptualization, A.E.-N., N.A.B. and N.E.-B.; methodology, A.E.-N., N.A.B. and N.E.-B.; software, A.E.-N.; validation, A.E.-N., N.A.B. and N.E.-B.; formal analysis, A.E.-N., N.A.B. and N.E.-B.; investigation, A.E.-N., N.A.B. and N.E.-B.; resources, A.E.-N.; data curation, A.E.-N., N.A.B. and N.E.-B.; writing—original draft preparation, N.A.B. and N.E.-B.; writing—review and editing, N.A.B. and N.E.-B.; visualization, A.E.-N., N.A.B. and N.E.-B.; supervision, N.A.B. and N.E.-B.; project administration, N.A.B. and N.E.-B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The METABRIC dataset was obtained based on a formal access request from our institution to perform the study on the data. Due to the sensitive information in the dataset, Synapse offers access through a controlled use mechanism and the dataset could be requested through the following link: https://www.synapse.org/$#$!Synapse:syn1688369/wiki/27311 (accessed on 15 May 2021) .

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Araujo, T.; Aresta, G.; Castro, E.; Rouco, J.; Aguiar, P.; Eloy, C.; Polonia, A.; Campilho, A. Classification of breast cancer histology images using convolutional neural networks. *PLoS ONE* **2017**, *12*, e0177544. [CrossRef] [PubMed]
2. Pan, X.; Lu, Y.; Lan, R.; Liu, Z.; Qin, Z.; Wang, H.; Liu, Z. Mitosis detection techniques in H&E stained breast cancer pathological images: A comprehensive review. *Comput. Electr. Eng.* **2021**, *91*, 107038.
3. Chouhan, N.; Khan, A.; Shah, J.; Hussnain, M.; Khan, M. Deep convolutional neural network and emotional learning based breast cancer detection using digital mammography. *Comput. Biol. Med.* **2021**, *132*, 104318. [CrossRef] [PubMed]
4. Xu, J.; Wu, P.; Chen, Y.; Meng, Q.; Dawood, H.; Dawood, H. A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data. *BMC Bioinform.* **2019**, *20*, 527. [CrossRef] [PubMed]
5. Lin, Y.; Zhang, W.; Cao, H.; Li, G.; Du, W. Classifying Breast Cancer Subtypes Using Deep Neural Networks Based on Multi-Omics Data. *MDPI Genes* **2020**, *11*, 888. [CrossRef] [PubMed]
6. Mohaiminul Islam, M.; Huang, S.; Ajwad, R.; Chi, C.; Wang, Y.; Hu, P. An integrative deep learning framework for classifying molecular subtypes of breast cancer. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 2185–2199. [CrossRef] [PubMed]
7. El-Nabawy, A.; El-Bendary, N.; Belal, N.A. A feature-fusion framework of clinical, genomics, and histopathological data for METABRIC breast cancer subtype classification. *Appl. Soft Comput.* **2020**, *91*, 106238. [CrossRef]
8. Yu, B.; Chen, C.; Wang, X.; Yu, Z.; Ma, A.; Liu, B. Prediction of protein–protein interactions based on elastic net and deep forest. *Expert Syst. Appl.* **2021**, *176*, 114876. [CrossRef]
9. Wang, W.; Guan, X.; Khan, M.; Xiong, Y.; Wei, D.Q. LMI-DForest: A deep forest model towards the prediction of lncRNA-miRNA interactions. *Comput. Biol. Chem.* **2020**, *89*, 107406. [CrossRef]
10. Arya, N.; Saha, S. Multi-modal advanced deep learning architectures for breast cancer survival prediction[Formula presented. *Knowl. Based Syst.* **2021**, *221*, 106965. [CrossRef]
11. Sirinukunwattana, K.; Raza, S.; Tsang, Y.W.; Snead, D.; Cree, I.; Rajpoot, N. Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. *IEEE Trans. Med Imaging* **2016**, *35*, 1196–1206. [CrossRef] [PubMed]
12. Komura, D.; Ishikawa, S. Machine Learning Methods for Histopathological Image Analysis. *Comput. Struct. Biotechnol. J.* **2018**, *16*, 34–42. [CrossRef]
13. Sohail, A.; Khan, A.; Wahab, N.; Zameer, A.; Khan, S. A multi-phase deep CNN based mitosis detection framework for breast cancer histopathological images. *Sci. Rep.* **2021**, *11*, 1–18. [CrossRef]
14. Ali, H.R.; Rueda, O.M.; Chin, S.F.; Curtis, C.; Dunning, M.J.; Aparicio, S.A.; Caldas, C. Genome-driven integrated classification of breast cancer validated in over 7500 samples. *Genome Biol.* **2014**, *431*, 1–14.

15. Mendes, A. Identification of Breast Cancer Subtypes Using Multiple Gene Expression Microarray Datasets. In Proceedings of the Australasian Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, 16–22 July 2011; pp. 92–101.
16. List, M.; Hauschild, A.C.; Tan, Q.; Kruse, T.A.; Baumbach, J.; Batra, R. Classification of Breast Cancer Subtypes by combining Gene Expression and DNA Methylation Data. *J. Integr. Bioinform.* **2014**, *11*, 1–14. [CrossRef]
17. Xu, J.; Xiang, L.; Liu, Q.; Gilmore, H.; Wu, J.; Tang, J.; Madabhushi, A. Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images. *IEEE Trans. Med Imaging* **2016**, *35*, 119–130. [CrossRef]
18. Spanhol, F.; Oliveira, L.; Petitjean, C.; Heutte, L. Breast cancer histopathological image classification using Convolutional Neural Networks. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; Volume 2016, pp. 2560–2567.
19. Bejnordi, B.; Veta, M.; Van Diest, P.; Van Ginneken, B.; Karssemeijer, N.; Litjens, G.; Van Der Laak, J.; Hermsen, M.; Manson, Q.; Balkenhol, M.; et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA J. Am. Med. Assoc.* **2017**, *318*, 2199–2210. [CrossRef] [PubMed]
20. Guo, Y.; Liu, S.; Li, Z.; Shang, X. BCDForest: A boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data. *BMC Bioinform.* **2018**, *19*, 1–13. [CrossRef]
21. Gao, F.; Wang, W.; Tan, M.; Zhu, L.; Zhang, Y.; Fessler, E.; Vermeulen, L.; Wang, X. DeepCC: A novel deep learning-based framework for cancer molecular subtype classification. *Oncogenesis* **2019**, *8*, 527. [CrossRef]
22. Dong, Y.; Yang, W.; Wan, J.; Zhao, J.; Qiang, Y. MLW-gcForest: A Multi-Weighted gcForest Model for Cancer Subtype Classification by Methylation Data. *MDPI Appl. Sci.* **2019**, *9*, 3589. [CrossRef]
23. Rohani, N.; Eslahchi, C. Classifying Breast Cancer Molecular Subtypes by Using Deep Clustering Approach. *Front. Genet.* **2020**, *11*, 1108. [CrossRef] [PubMed]
24. Ragab, D.; Attallah, O.; Sharkas, M.; Ren, J.; Marshall, S. A framework for breast cancer classification using Multi-DCNNs. *Comput. Biol. Med.* **2021**, *131*, 104245. [CrossRef] [PubMed]
25. Zhang, Y.; Chen, J.H.; Lin, Y.; Chan, S.; Zhou, J.; Chow, D.; Chang, P.; Kwong, T.; Yeh, D.C.; Wang, X.; et al. Prediction of breast cancer molecular subtypes on DCE-MRI using convolutional neural network with transfer learning between two centers. *Eur. Radiol.* **2021**, *31*, 2559–2567. [CrossRef] [PubMed]
26. Zhang, Y.D.; Satapathy, S.; Guttery, D.; Górriz, J.; Wang, S.H. Improved Breast Cancer Classification Through Combining Graph Convolutional Network and Convolutional Neural Network. *Inf. Process. Manag.* **2021**, *58*, 102439. [CrossRef]
27. Wang, P.; Wang, J.; Li, Y.; Li, P.; Li, L.; Jiang, M. Automatic classification of breast cancer histopathological images based on deep feature fusion and enhanced routing. *Biomed. Signal Process. Control* **2021**, *65*, 102341. [CrossRef]
28. METABRIC Genomics Dataset, The European Genome-Phenome Archive (EGA). Available online: https://ega-archive.org/dacs/EGAC00001000484 (accessed on 15 April 2020).
29. METABRIC Clinical Dataset, Molecular Taxonomy of Breast Cancer International Consortium. Available online: https://www.synapse.org/#!Synapse:syn1688369/wiki/27311 (accessed on 15 April 2020).
30. Dziura, J.; Post, L.; Zhao, Q.; Fu, Z.; Peduzzi, P. Strategies for dealing with missing data in clinical trials: from design to analysis. *Yale J. Biol. Med.* **2013**, *86*, 343–358. [PubMed]
31. Jahid, M.; Huang, T.; Ruan, J. A personalized committee classification approach to improving prediction of breast cancer metastasis. *Bioinformatics* **2014**, *30*, 1858–1866. [CrossRef]
32. Saddiki, H.; McAuliffe, J.; Flaherty, P. GLAD: A mixed-membership model for heterogeneous tumor subtype classification. *Bioinformatics* **2015**, *30*, 225–232. [CrossRef]
33. Fan, Y.; Qi, L.; Tie, Y. The Cascade Improved Model Based Deep Forest for Small-scale Datasets Classification. In Proceedings of the 2019 8th International Symposium on Next Generation Electronics (ISNE), Zhengzhou, China, 9–10 October 2019; pp. 1–3.
34. Wang, H.; Tang, Y.; Jia, Z.; Ye, F. Dense Adaptive Cascade Forest: A Self Adaptive Deep Ensemble for Classification Problems. 2019. Available online: http://xxx.lanl.gov/abs/1804.10885 (accessed on 15 April 2020).