Using Deep Learning for Survival Prediction of Breast Cancer using Microarray Data

by

Long Jingyu

Supervisor: Jackie Tan Yen

UpLevel

Submitted to the School of Biological Sciences in partial fulfillment of the requirements for Final Year Project (Practicum BS6211)

Nanyang Technological University

05, 2022

**DECLARATION**

I declare that,

in accordance with School requirements this report is under 10,000 words in length;

all presented work was performed within the official project time frame as stated below;

all presented work was performed by me, unless otherwise specified;

all relevant work experience gained before the Final Year Project is stated below;

this report is my own work, unless otherwise referenced, as defined by the [Academic Integrity Handbook](https://www.ntu.edu.sg/docs/default-source/tlpd-documents/academic-integrity-handbook_july-2017.pdf?sfvrsn=fc5a5b24_2) and I have read and understand the contents.

the included Abstract, Introduction, Results, Discussion and conclusions sections of the finalized report will be submitted to Turnitin. The similarity index meets the programme requirement. Both finalized Final Year report and similarity report are uploaded to NTULearn.

Final Year Project start date: 28/02/2022

Final Year Project submission date: 17/05/2022

Total number of weeks:

Pre-Final Year Project experience

*(List all relevant work experience gained before the Practicum, including, Immersion programme and volunteer work experience. Include type of work and duration.)*



Student’s signature: Date: 16/05/2022

Supervisor’s signature: Date: 16/05/2022

**Acknowledgements**

The days of studying in Singapore have come to an end. Although a year has passed quickly, the study and life are very fulfilling. At the moment of the epidemic, as a foreigner, the teachers and classmates of the school have given me warmth. I would like to express my special thanks to my supervisors, Jackie and Wilson, who have given their own suggestions and insights in this paper. They gave me great support and help. There are also eight other students in the class who are very friendly and take good care of me in group discussions and in daily life. And all the teachers at SBS Academy, they all made my time go smoothly this year, I am very grateful to them.In the days to come, I wish you all a bright future and live your life.

**Abbreviations**

|  |  |
| --- | --- |
| ADASYN | Adaptive Synthetic Sampling |
| BC | Breast Cancer |
| CI | Confidence Interval |
| CNN | Convolutional Neural Networ |
| DNN | Deep Neural Network |
| HR | Hazard Ratios |
| KM analysis | Kaplan-Meier analysis |
| Nadam | Nesterov-accelerated Adaptive Moment Estimation |
| RF | Random Forest |
| RUS | random sub-sampling |
| SMOTE | Synthetic Minority Over-sampling Technique |

**Content**

**[Abstract](#_Toc18966388_WPSOffice_Level1)** **[6](#_Toc18966388_WPSOffice_Level1)**

**[Introduction](#_Toc1462816476_WPSOffice_Level1)** **[7](#_Toc1462816476_WPSOffice_Level1)**

**[Materials and Methods](#_Toc1163721276_WPSOffice_Level1)** **[9](#_Toc1163721276_WPSOffice_Level1)**

**[Microarray data preprocessing](#_Toc1462816476_WPSOffice_Level2)** **[9](#_Toc1462816476_WPSOffice_Level2)**

**[Fixing Imbalanced Dataset with ADASYN](#_Toc1163721276_WPSOffice_Level2)** **[10](#_Toc1163721276_WPSOffice_Level2)**

**[Construction of the DNN prediction model](#_Toc1529912503_WPSOffice_Level2)** **[11](#_Toc1529912503_WPSOffice_Level2)**

**[Construction of the CNN prediction model](#_Toc1417732390_WPSOffice_Level2)** **[12](#_Toc1417732390_WPSOffice_Level2)**

**[Experimental details for benchmark models](#_Toc1497215265_WPSOffice_Level2)** **[13](#_Toc1497215265_WPSOffice_Level2)**

**[From AUC to reclassification](#_Toc1631066956_WPSOffice_Level2)** **[14](#_Toc1631066956_WPSOffice_Level2)**

**[Survival analysis](#_Toc713575537_WPSOffice_Level2)** **[14](#_Toc713575537_WPSOffice_Level2)**

**[Results](#_Toc1529912503_WPSOffice_Level1)** **[16](#_Toc1529912503_WPSOffice_Level1)**

**[Discussion](#_Toc1417732390_WPSOffice_Level1)** **[21](#_Toc1417732390_WPSOffice_Level1)**

**[Conclusion](#_Toc1497215265_WPSOffice_Level1)** **[24](#_Toc1497215265_WPSOffice_Level1)**

**[Reference](#_Toc1631066956_WPSOffice_Level1)** **[25](#_Toc1631066956_WPSOffice_Level1)**

**Abstract**

The disease with the highest mortality rate among women across the world has always been breast cancer. Hence, a reliable and accruable system needs to be established for the early diagnosis of patients. Based on the microarray data, a combination of rules and various techniques of machine learning has been proposed as a survival prediction model. First, this study applies the random over-sampling technique to resolve the problem of class imbalance within the microarray gene expression data. Second, this study applied the most powerful deep/machine learning models such as Convolutional Neural Networks, deep neural networks, XGBoost, and Random Forests for classification. Last but not least, each model’s performance is evaluated from different perspectives. The reclassification study is used to select the best threshold of output probability to further improve the performance of the training models. In survival analysis, KM analysis and the proportional-hazards model were conducted to evaluate the results for Improved DNN and CNN models. The result shows the proposed improved CNN outperformed all other models.

**Introduction**

Across the world, the most prevalent cancer with the highest death rate among female society is breast cancer. Based on research, breast cancer is the second most reason cause of death for women in the United States, as per the statistics of the American Cancer Society. For the prediction of survival of patients with breast cancer, independent of treatment, which is known as prognostication, several pathological and clinical indicators like lymph node involvement, tumour size, and histological grade have been used for the past two decades. However, the prediction of survival in patients with breast cancer is a complex task due to the clinically similar and molecularly heterogeneous nature of the breast tumours, which indicate varying outcomes clinically. For improving the quality of prognostication and for bringing added insights into breast cancer biology, researchers have begun to analyze to use an artificial intelligence approach to analyze gene expression profiles recently. Many studies intend to achieve an accuracy of prediction in a high standard manner through a robust approach with survival analysis for breast cancer microarray data. From merely 56% the survival could be increased to more than 86%, with early diagnosis of breast cancer. The Adjuvant Online (AOL, Olivotto et al., 2005), the NIH guidelines (Eifel et al., 2001), and the St Gallen consensus criteria (Goldhirsh et al., 2003) are some examples of the clinical guidelines for the selection of patients for adjuvant therapy. Identifying patients needing adjuvant systemic therapy continues to be a big challenge, despite breast cancer prognostication being an object of intense research. Ostensibly, an entirely correct prediction regarding the clinical manifestations of patients is not possible with prognostic factors for the recurrence of breast cancer like lymph node status and histology.

New insights were brought into breast cancer prognosis and biology by the sequencing of the human genome with the advent of array-based technology. Prognostic gene expression signatures were identified through the comprehensive genome-wide assessment of the gene expression profiling, by numerous research teams. The 76-gene (Wang et al., 2005) and the 70-gene (van’t Veer et al., 2002) signatures, were some of the examples of the gene signatures obtained through studies of the relationship between clinical outcomes and profiles of gene expression. These signatures could identify correctly a larger group of low-risk patients without needing treatment, with regard to the clinical guidelines. As the reduction of treatments could also reduce the costs and the inherent side effects, these were particularly relevant for the clinicians. Information regarding thousands of genes could be obtained simultaneously, through the use of microarray technology. However, due to the fact, that most of those data normally contain curvilinear effects with complex polynomial interaction among variables, and contain a small sample size which results in a lot of constraints, conventional statistical prediction models like regressions became significantly difficult for application in the models of survival prediction. Without taking the constraints from polynomial interaction terms and the statistical assumptions, the processing of thousands of independent variables could be possible with data mining techniques like decision trees and artificial neural networks. Being more advantageous for the development of models for survival prediction, these techniques tend to have better potential compared to logistic regression.

Probably, as smaller sample sizes were used, the analyses of the microarray data intended to make a prediction of the recurrence of breast cancer could rarely select the same group of genes, as reported earlier. One of the goals of this pertinent research was to enhance the sample size with the integration of the samples from multiple breast cancer microarray databases. In addition, discover the potential of artificial neural network models to do survival prediction for breast cancer by microarray data, with the goal of developing more robust survivability prediction models for breast cancer.

**Materials and Methods**

All the techniques were executed in Google Colab Python Development Environment. Figure. 1 illustrates the process of the data flow. The preprocessing technique being applied to the dataset was the initial step. Splitting the data into testing and training data was the second step. While the testing data was used for evaluating the performance of the model, the training data was utilized for training the model. After getting the prediction results, last step is doing survival analysis for different models’ results.

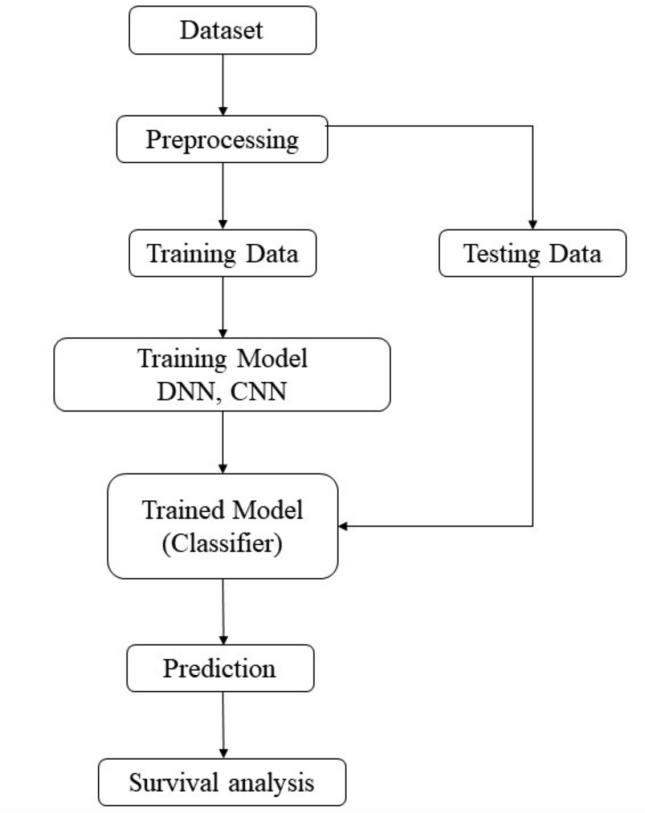


Figure 1. The Dataflow Diagram

**Microarray data preprocessing**

All the data were available from public sources:

1. Ge et al. data were downloaded from NCBI’s Gene expression Omnibus (www.ncbi.nlm.nih.gov/geo; accession, GDS1096). We renormalized the raw data (CEL files) using Bioconductor package gcrma.

2. Loi et al. data were downloaded from NCBI’s Gene expression Omnibus (accession GSE6532). We used the Rdata file.

3.The NKI, a.k.a. van de Vijver et al., data set was downloaded from the Rosetta Inpharmatics web site in 2007 (www.rii.com, this site is now defunct, the dataset is available in the supplementary code and data tar bundle).

Ignoring the flags, the normalization of the original authors was used. It is pertinent to note that due to the low dimensionality of the input space, no dimension reduction was conducted for BC microarray input data.

**Fixing Imbalanced Dataset with ADASYN**

Applied in numerous types of cancer research, the gene expression microarray are the profile data sets of the real world expression. Grouped into two classes (65% as 0 and 35% as 1), the breast cancer databases are highly dimensional (13110 dimensions), and have very few samples (295 patients). It needs to be noted that there was no balance between the number of samples of each class. It obvious that better results were obtained as against the original databases, using the over-sampling techniques. With the use of traditional oversampling techniques, like SMOTE and ROS, the classifier when working with high-dimension databases, high class imbalance, and few samples, could obtain better results as in case of the Random Sub-sampling (RUS) the performance of the classifier was found to be similar to that of the original database, although slightly less in certain cases. Moreover, an improved version of SMOTE was found to be Adaptive Synthetic Sampling ADASYN. It was made realistic as it added random small values to the points after creating the samples. Nevertheless, they were a bit scattered, since instead of the being linearly correlated to the parent, all the samples displayed slightly more variance.

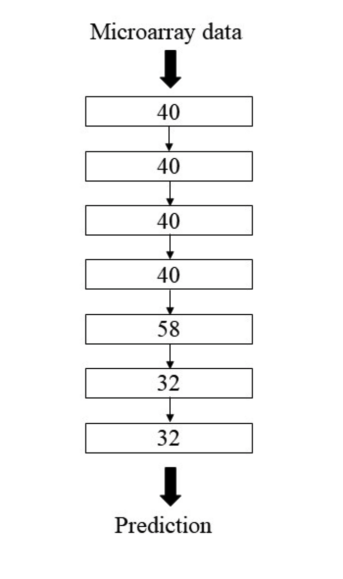
Studying the behavior of the deep learning classifiers applying high-class imbalance gene expression microarray databases, with low pattern, and high dimension, was the purpose of this research. So before separating the 295 patients, I used ADASYN algorithm to do random over-sampling. Finally, 394 samples were selected, with 79 patients as the test set, 63 as the validation set, and 315 as the training set. Besides, 79.32 months was considered as the overall median survival time.

**Construction of the DNN prediction model**

Representing multiple levels of abstraction, having several hidden layers, a DNN is composed of one output and one input layer. Several neurons constitute each of the hidden layers. Thus, in the supervised learning for combining different modalities, deep learning could be applied successfully. Varying the number of neurons their cross-entropy losses were observed in the validation set in case of DNN. The one with the least validation loss was selected as the best model. For identifying the best optimizer also, a similar search was conducted with the Nadam optimizer indicating the best performance.

Nesterov-accelerated Adaptive Moment Estimation, or the Nadam, is an extension of the Adam algorithm that incorporates Nesterov momentum and can result in better performance of the optimization algorithm. As an optimization algorithm, gradient descent follows the negative gradient of the objective function and can locate the minimum value of the function. But gradient descent has a limitation: if the gradient flattens out, the progress of the search may slow down. That's when momentum is added to gradient descent with inertia. To solve this problem, we further improve by incorporating the gradient of the projected new position instead of the current position, which is the Nesterov acceleration gradient (NAG) or Nesterov momentum. Another disadvantage of gradient descent is that all input variables use the same step size (learning rate). In the traditional adaptive motion estimation (Adam) algorithm, it uses a separate step size for each input variable, but may cause the step size to decrease rapidly to very small values.

With the ReLU activation function and the Nadam optimizer being adopted having L2-regularization, there were four layers with 40 neurons in each layer in the final model structure of our DNN. Following the default settings in Nadam for other parameters, the learning rate was set at 0.006.

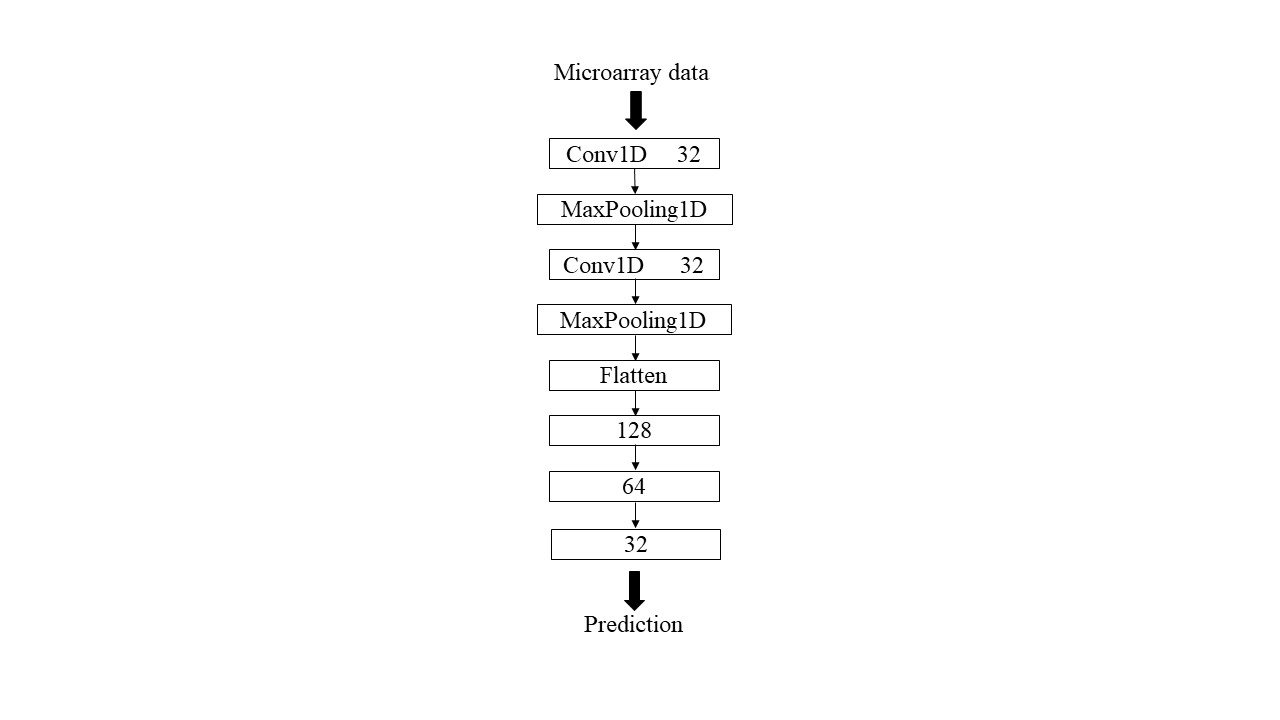
Figure 2. The DNN structure. The DNN architecture contains 7 layers, the 4th hidden layers with 40 neurons each, the fifth layer contained 58 neurons and were stacked with two hidden layers with

32 neurons each for the final prediction.

**Construction of the CNN prediction model**

To produce more complex patterns within higher layers, simple patterns in the data could be applied as the CNN was found to perform very well for classifying. The CNN as a form of multilayer neural networks, was found to be extraordinary. The backpropagation algorithm is used by ti for training as is one by almost all the other neural networks. It is its architecture that differentiates the CNN from the others. Along with an output layers, an input layer with several hidden layers is present in a typical CNN architecture. The input data with the kernel is multiplied basically for producing the modified output data, as the convolution layer implements a filter on it after receiving the input data. From the convolution layer, the pooling layer constitutes a subsampling process. The reduction of the dimensionality is the main purpose.

The first layer of the proposed CNN (Figure 3) algorithm in this research was in the form of an input layer. Applying the RELU activation function, a one-dimension convolution layer having 3 kernel sizes and 30 filters formed the second layer. With 2 pool sizes, the max polling layer formed the third layer. With the RELU activation function, the next layer was a fully connected layer. Flatten makes the input into a one-dimensional list for input to the dense layers. Dense activation shape state 128, 64, 32 hidden units individually are used in the next three dense layers. Finally, one neuron with sigmoid activation function constituted the output layer. For binary cross entropy used as a cost function and for learning, the Nadam optimizer was used.

Figure 3. The CNN structure. The CNN architecture contains

2 convolutional layers, 2 maxpooling layers, one flatten layer and 3 dense layers.

**Experimental details for benchmark models**

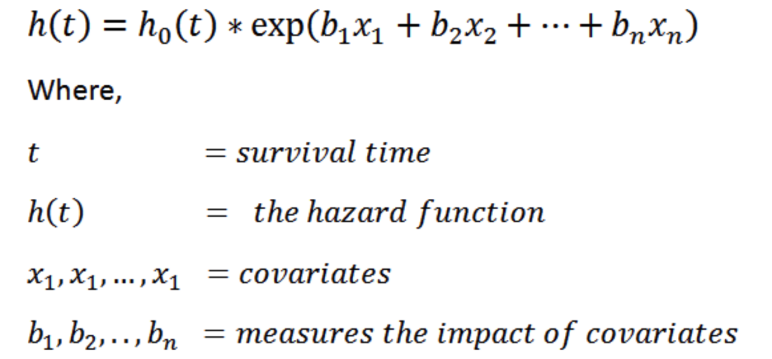
In machine learning, we mainly deal with classification and regression problems, both of which have many different types of algorithms. But we need to choose an algorithm that performs better on the corresponding data. This experiment is to complete a binary classification problem, and integration methods such as random forest, decision tree and XGBoost algorithm all show very good results. These algorithms provide high accuracy at fast speeds. To sum up, I choose two algorithms Random Forest and XGBoost as the benchmark models, both of which can achieve higher accuracy and are simple to use. And both RF and XGBoost, we need to pass few of the hyperparameters that effect the predictions of the model when training. Selecting a set of hyperparameters gives the best result, this can be done by RandomizedSearchCV. RandomizedSearchCV randomly passes the set of hyperparameters and calculate the score and gives the best set of hyperparameters which gives the best score as an output. We can define a grid of hyperparameter ranges, and randomly sample from the grid, performing 5-Fold CV with each combination of values. We restricted the maximum depth as one-third of the dimension of the input data dimension and varied the number of trees used.

**From AUC to reclassification**

The diagnostic ability of a binary classifier system generated through the plotting of the true positive rates against the false positive rates at different threshold settings, the receiver operating characteristic curve. How well the model performed was evaluated by using the area under the ROC curve. By adjusting the cut-off points, a performance showing better reclassification could be achieved under certain conditions. Frequently-used as summary measure of the ROC curve, the Youden index helped in determining the cut-off points. By comparing the probabilities for all classifiers having patient survival outcomes as 0.5, a patient was classified in the previous tasks of classification. New cut-off points obtained with the Youden index were utilized for reclassification for improving the predictive performance. The calculation of the Youden index is fairly simple. Just sumup the sensitivity of a diagnostic test and the specificity of the same diagnostic test, later substracting 100 from the value obtained from the addition. For being administered for the purpose of diagnostics, in case the Youden index is not more than 50%, then it would be considered as not meeting the empirical benchmarks. The measure of the ability of a diagnostic test to balance specificity (detecting health or no disease) and sensitivity (detecting disease) is known as the Youden index.

**Survival analysis**

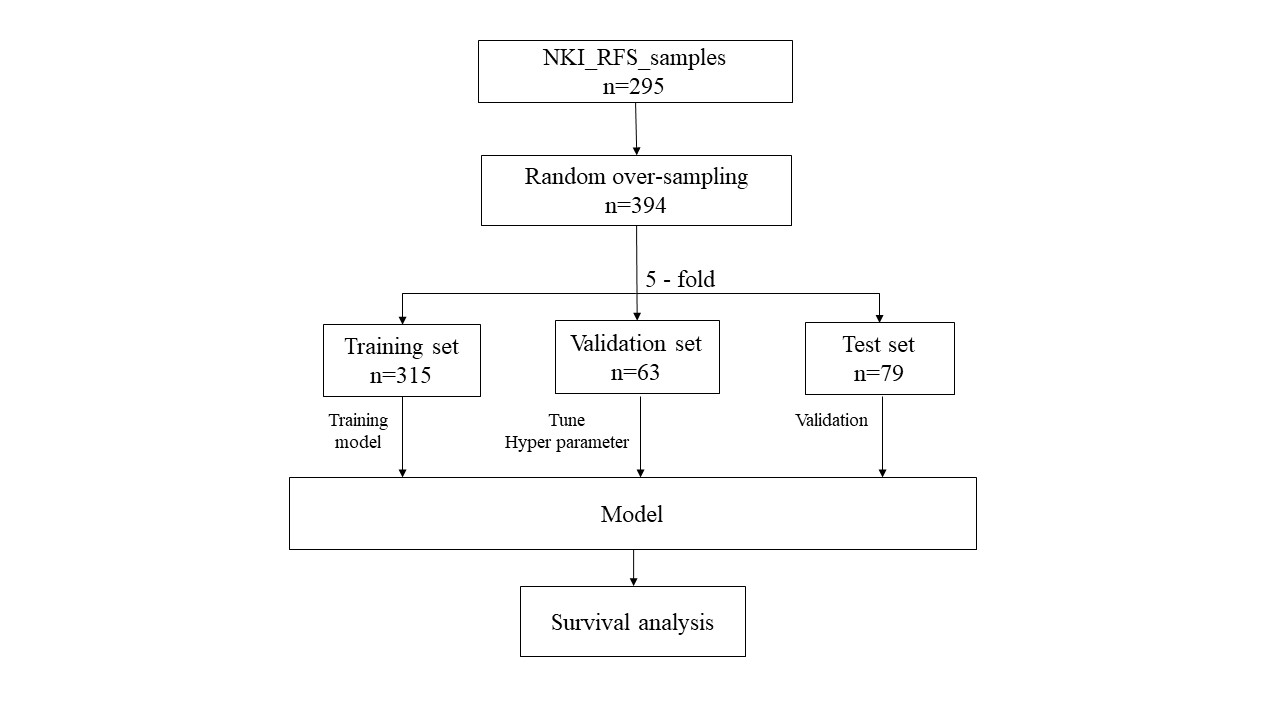
In our study, overall survival time was calculated from the date of surgery to the date of death. I divided the patients into high risk (which was predicted as dead) and low risk (which was predicted as survived). Survival curves were demonstrated based on a Kaplan-Meier estimation and compared using a log-rank test(KM analysis). We applied the cox proportional-hazards model to analyze the relationship between the prognostic genes for survival. The ultimate purpose of the Cox proportional hazard method is to notice how different factors in our dataset impact the event of interest. Hazard function is showed, the values exp(bi) is called the hazard ratio (HR).



As the value of the ith covariate increased, the event hazard increased, and thereby decreasing the duration of survival as indicated by the HR being greater than 1.To summarize, HR (Hazard Ratio) = exp(bi). Here notice the p-value of different parameters as we know that a p-value (<0.05) is considered significant. The hazard ratios (HR) and confidence intervals (CI) were reported.

**Results**

Several deep learning approaches like CNN, DNN, XGBoost, and random forest were applied to predict the survival status of patients who has breast cancer. Moreover, the measurement of the expression was substituted based on the sum of the residuals and the mean expression through the cohort, with the expression of each gene being fitted with R’s ‘lm’ function. The selected features (as the input profile for the DNN and CNN prediction based on the median 6.61-year (0.02~18.35 years) survival of breast cancer patients. Moreover, the study also adds the ADASYN algorithm to fix imbalances in real-life data sets through oversampling. The schematic of the entire framework is shown in Fig. 4.

Figure 4. Schematic of the study design. 295 samples were processed into 394 samples with random over-sampling (n=394) and divided into the training set (n=315) and testing set (n=79), and validation set (n=63). The DNN network and CNN network were trained and tuned based on the training set and the validation set, respectively. After training the DNN and CNN, we tested them on the test set and applied the survival analysis.

We used DNN and CNN to exploit the input features. Subsequent to the first therapeutic treatment. The results of the two models were the binary outcome of the survival probability of the patient. The optimized structure of the DNN consisted of seven hidden layers constituted by the Rectified Linear Unit (ReLU) with 40 neurons in each layer, based on Nadam as the optimizer. Besides, CNN also had Nadam as the optimizer. In order to prevent the problem of over-fitting, early stopping was used during the training process as if the validation loss does not improve for over 10 epochs. All of the models were trained under 100 epochs, with a batch size 32.

Under the same input features, XGBoost and Random Forest (RF) were selected for the purpose of comparing the performance and verifying the effectiveness of the CNN and DNN. Based on a 5-fold cross-validation of the training set, the parameters applied in the XGBoost and RF were optimized. There was a tendency to classify the patients as alive, as a result of the imbalance in the labels across the dataset (death = 194, survivals =101, and n= 295). Based on the observed results, the basic classifier without any reclassification is able to reach an accuracy of 0.84, and the AUC is able to achieve a higher score than accuracy taking the class imbalance. Overall, the performance of the DNN network exceeds the result of all other conventional machine learning methods, which has an accuracy of 91.96% and an AUC of 0.895. CNN achieved the best result in all classifications, with an accuracy of 94.78% and an AUC of 0.93, which is the only one with an AUC greater than 0.90.

Table 1. Performance comparison of the CNN, DNN, XGBoost and RF

|  |  |  |
| --- | --- | --- |
|  | **AUC** | **Accuracy** |
| RF | 0.8469 | 0.8513 |
| XGBoost | 0.8366 | 0.8424 |
| DNN | 0.8948 | 0.9196 |
| CNN | 0.9301 | 0.9478 |

The F1 score, recall, and precision were used to compare the overall performance of each approach. Finding the best threshold points for reclassification was necessary for the computation of these metrics. To select the threshold points, the Youden index was used for reclassification. Both on the DNN and the CNN results the reclassification was conducted. The cut-off points were determined respectively as 0.52 and 0.66 for the DNN and CNN. 50% was the cut-off point for having an acceptable Youden index. An overall failure of the diagnostic test in detecting either the disease or the health was considered if any value fell below 50%, which means that there is an increase after reclassification in the number of patients predicated as dead, the new cut-off point in DNN and CNN was more than 0.5 (the original cut-off). And the higher the index, the better the screening test is and the more real it is. CNN’s Youden index is 0.9045 and DNN’s is 0.8701, CNN wins this time.

Table 2. The performance of the DNN/CNN with/without reclassification and benchmark models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | | **Precision** | **Recall** | **F1-score** | **AUC** |
| XGBoost | | 0.84 | 0.84 | 0.84 | 0.84 | 0.85 |
| RF | | 0.85 | 0.85 | 0.85 | 0.85 | 0.84 |
| DNN before reclassification | | 0.91 | 0.90 | 0.89 | 0.90 | 0.89 |
| DNN after reclassification | | 0.93 | 0.92 | 0.92 | 0.92 | 0.92 |
| CNN before reclassification | | 0.94 | 0.94 | 0.93 | 0.94 | 0.93 |
| CNN after reclassification | | 0.95 | 0.94 | 0.95 | 0.95 | 0.94 |

In the case of the microarray DNN and CNN, their performances were evaluated with regard to the F1 score, recall, and precision, to confirm the effectiveness of reclassification. Pertinent to note that the F1 score and the recall of the original classifier were made low by the imbalanced class distribution of the data. In the case of both DNN and CNN, the evaluation criteria, such as F1 score and recall, are also improved when the Youden indices were introduced for reclassification. For the F1-score, which increased from 0.90 to 0.92, recall, which also increased from 0.89 to 0.92, was observed in the DNN results. On the other hand, the same trend can also be observed in the result of CNN, with the F1 score increasing from 0.94 to 0.95, and recall increasing from 0.93 to 0.95. Furthermore, there was an also increase in the precision of DNN, which raised from 0.90 to 0.92, but CNN did not improve in precision after reclassification. Overall, the CNN achieved a higher F1 score, precision, recall and accuracy than the DNN, which is a better classifier after the reclassification process.

图形用户界面, 图表, 折线图

描述已自动生成

Figure 5. KM analysis of each classifier

Table 3. Proportional-hazards model of each classifier.

|  |  |  |
| --- | --- | --- |
|  | **HR（95%CI）** | **p-value** |
| RF | 3.01 ( 0.75-12.88) | 0.12 |
| XGBoost | 2.62 (0.42- 16.36) | 0.30 |
| DNN | 1.02 ( 0.50-2.07) | 0.07 |
| CNN | 7.00 ( 2.19-22.19) | <0.005 |

Survival analysis for patients was conducted on the deep learning models to validate the result of each proposed approach. The patients were segregated as low risk (predicted as survived) and high risk (predicted as dead) as proposed by our DNN and CNN, respectively. To interpret the results of both CNN and DNN without and with reclassification, a proportional-hazard model and KM analysis were applied. On the KM analysis for CNN models, it was observed that the reclassification separated the two risk groups further apart. Nonetheless, in the proportional-hazards model, an improvement could also be observed. It is clear to observe that separation between the two status becomes more inapparent compare the DNN network (HR: 1.02, 95% CI: 0.50-2.07, p-value=0.07), the RF(HR: 3.01, 95% CI: 0.75-12.88, p-value=0.12) and XGBoost (HR: 2.62, 95% CI: 0.42- 16.36, p-value=0.30). Similar results can be seen for RF and XGBoost; however, the CNN network’s distances between the low-risk and high-risk patient groups were greater (HR: 7.00, 95% CI: 2.19-22.19, p-value<0.005) than for RF and XGBThe CNN achieves the highest hazard ratio compared to all other models, indicating that the CNN is best used for extracting information based on the predictive results.

**Discussion**

This study has a two-pronged purpose. First and foremost, to establish an independent and reliable framework of assessment for comparing the performances of the DNN and CNN methods of prediction. Second, comparing the accuracy of prediction of these approaches in the breast cancer microarrays, in order to identify the significant features of the most successful model of the risk prediction, and bring more insights into the breast cancer biology. Due to a plethora of issues related specifically to the survival microarray data, this task has been known to be a highly challenging one.

While the high-dimensionality of the microarray data is the first issue, which leads to the overfitting of the data mining methods being applied naively. We can see that the DNN model (Figure 6) has an overall validation accuracy of 90%, which is a slight improvement from our RF and XGBoost models, and also the train and validation accuracy are quite close to each other, indicating that the model is not overfitting. In figure 7, the preceding output that our model has obtained a validation accuracy of around 95%, which is a 5% improvement from the DNN model. Although the CNN model looks like slightly overfitting on the training data, we still get great validation accuracy.

|  |
| --- |
|  |
|  |  |

图表, 直方图

描述已自动生成

Figure 6. The accuracy and loss function graph of DNN model.

|  |
| --- |
|  |
|  |  |

Figure 7. The accuracy and loss function graph of the CNN model.

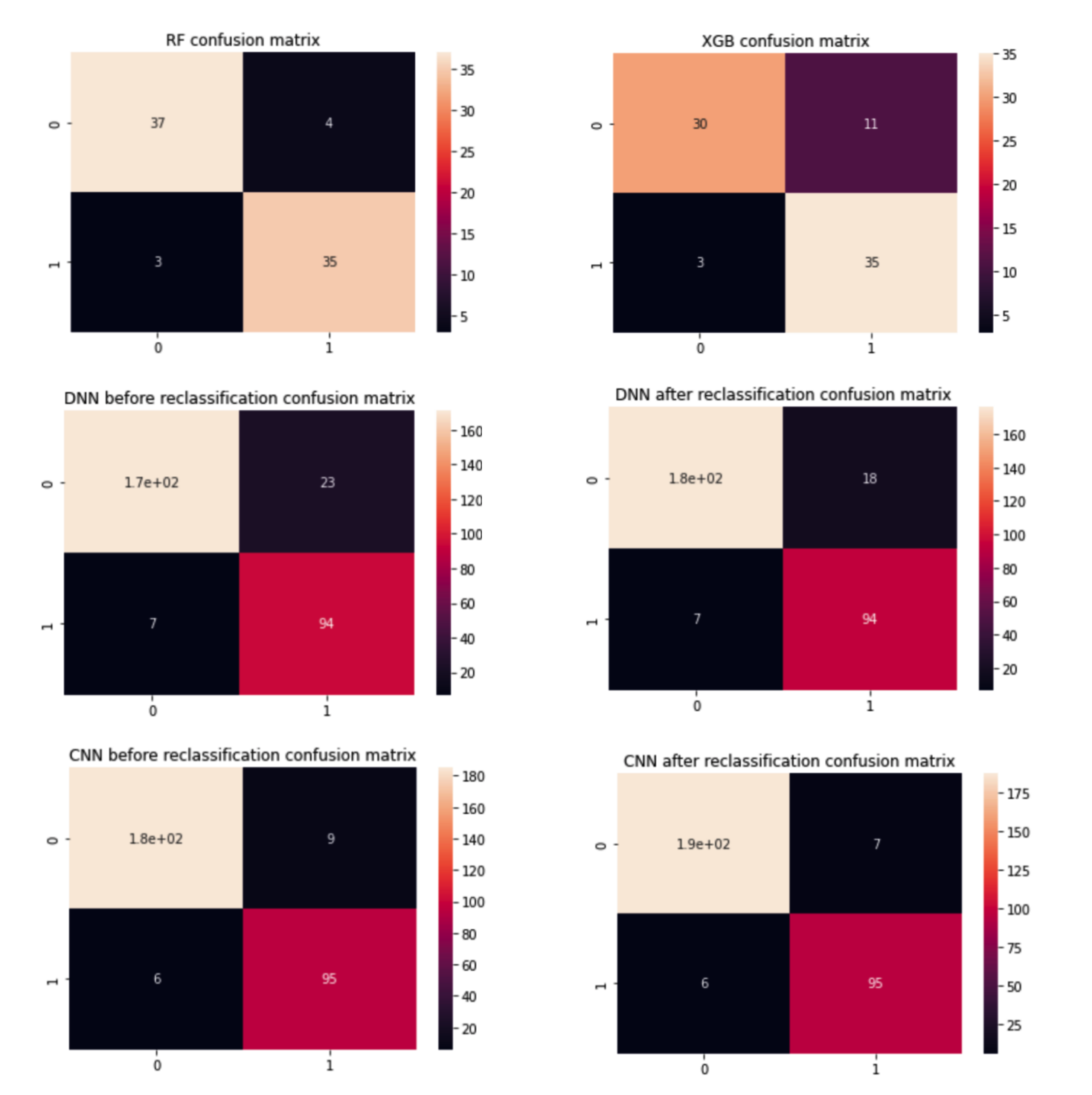
When the number of patients in a sample cohort exceeds the number of the explanatory variables (high feature to sample ratio), we would get the over-optimistic performance assessment. Observing the confusion matrix in different classifiers in figure 8, an overall failure of the diagnostic test in detecting either the disease or the health was considered if any value fell below 50%, which means that the after the process of reclassification, number of patients considered dead raised.

Figure 8. Confusion matrix of different models

The outperformance of the complex methods against the simplest breast cancer prognostication techniques lacking any statistical evidence despite a large number of samples has been a significant outcome of the analysis. Thus the results suggest that an improvement in the accuracy of prediction would not counterbalance the loss of interpretability derived from the application of the over-complex strategies of data analysis sufficiently. So I apply KM analysis and proportional-hazards model to CNN and DNN technology, which successfully increases the interpretability of the model.

One of the issues of survival analysis is few accuracy measurements for the models of risk prediction exist. Nonetheless, despite having certain measures for risk prediction accuracy, no studies have been conducted on their agreements on the same set of datasets and methods. It is almost impossible to obtain classified or censored information through the traditionally supervised methods of regression and classification, rather it requires specific survival techniques.

The performance of various methods of deep learning prediction was compared on 295 patients in this study. Nonetheless, to increase the universality and portability of the model, more studies on numerous datasets need to be conducted further. And the lack of independent data makes the comparison and validation of the BC microarray prognostication methods highly difficult.

The disadvantage of ADASYN is that it is vulnerable to the influence of outliers. Making the weight quite heavy, more samples would be generated, in case all majority class samples become the K neighbours of a minority class sample. A new state-of-the-art method helping to overcome the deficiencies of the existing methods is proposed and further research to study other classical algorithms for the treatment of class imbalances is suggested.

**Conclusion**

This study was intended at developing a prediction model that could be more reliable, for breast cancer survival, with the ability to predict the survival of the patients with breast cancer of the artificial neural network models being assessed. With regard to the precision and accuracy of each of these algorithms have been compared and measured. With the preprocessing technique for the classification of microarray gene expression data, this study presents the algorithms of deep learning such as CNN, improved CNN, DNN, and improved DNN. Initially, the preprocessing technique --- ADASYN algorithm was applied to alleviate the data imbalance(35% of patients were alive, while 65% patients were dead, of the total 295 patients), subsequent to the training of all models independently with good learning methods and structures, in case of all features of gene expression to handle the small sample data and high dimensions. For predictive analysis having an accuracy of around 84% and 85% XGB and RF Classifiers were respectively. Aiming to find the best results through the experiments, deep learning algorithms such as CNN and DNN (91% and 94%) were implemented to increase the accuracy of prediction. To predict the outcomes in terms of probabilities, activation functions like Sigmoid and Relu were applied. To select the cut-off points for reclassification on both DNN and CNN results, the Youden index was used subsequent to the application of our models for classification on testing data, 93% and 95% were the improved accuracies achieved on the DNN and CNN respectively. In order to evaluate the possible clinical utility of each classifier, the validation cohort (5-fold cross-validation) was analyzed on each model with respect to overall survival for evaluating their possible clinical utility. Most significantly, the distance between the high-risk and low-risk groups was significant for CNN in the survival analysis. The potential of CNN extracting useful information from predictive results was observed with the improved CNN (HR: 7.00, 95% CI: 2.19-22.19, p-value<0.005) attaining the highest hazard ratio. The improved CNN, which can be used as a rule-based classification model, achieves the best result with the highest accuracy and has been aptly confirmed by this study. As a result, for better medical decision making and as a useful tool for survival prediction of breast cancer patients this model is highly recommended.

**References**

Ahmed, O., & Brifcani, A. (2019, April). Gene expression classification based on deep learning. In 2019 4th Scientific International Conference Najaf (SICN) (pp. 145-149). IEEE.

Brueffer, C., Vallon-Christersson, J., Grabau, D., Ehinger, A., Häkkinen, J., Hegardt, C., ... & Saal, L. H. (2018). Clinical value of RNA sequencing–based classifiers for prediction of the five conventional breast cancer biomarkers: a report from the population-based multicenter Sweden Cancerome Analysis Network—Breast Initiative. JCO Precision Oncology, 2, 1-18.

Chou, H. L., Yao, C. T., Su, S. L., Lee, C. Y., Hu, K. Y., Terng, H. J., ... & Chu, C. M. (2013). Gene expression profiling of breast cancer survivability by pooled cDNA microarray analysis using logistic regression, artificial neural networks and decision trees. BMC bioinformatics, 14(1), 1-11.

Cheng, L. H., Hsu, T. C., & Lin, C. (2021). Integrating ensemble systems biology feature selection and bimodal deep neural network for breast cancer prognosis prediction. Scientific Reports, 11(1), 1-10.

Haibe-Kains, B., Desmedt, C., Sotiriou, C., & Bontempi, G. (2008). A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all?. Bioinformatics, 24(19), 2200-2208.

Loi, S., Haibe-Kains, B., Desmedt, C., Lallemand, F., Tutt, A. M., Gillet, C., ... & Sotiriou, C. (2007). Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. Journal of clinical oncology, 25(10), 1239.

Lai, Y. H., Chen, W. N., Hsu, T. C., Lin, C., Tsao, Y., & Wu, S. (2020). Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. Scientific reports, 10(1), 1-11.

Montazeri, M., Montazeri, M., Montazeri, M., & Beigzadeh, A. (2016). Machine learning models in breast cancer survival prediction. Technology and Health Care, 24(1), 31-42.

Ragab, A., Ouali, M. S., Yacout, S., & Osman, H. (2016). Remaining useful life prediction using prognostic methodology based on logical analysis of data and Kaplan–Meier estimation. Journal of Intelligent Manufacturing, 27(5), 943-958.

Schisterman, E. F., Faraggi, D., Reiser, B., & Hu, J. (2008). Youden Index and the optimal threshold for markers with mass at zero. Statistics in medicine, 27(2), 297–315.

Tiwari, M., Bharuka, R., Shah, P., & Lokare, R. (2020). Breast cancer prediction using deep learning and machine learning techniques. Available at SSRN 3558786.

Unal, I. (2017). Defining an optimal cut-point value in ROC analysis: an alternative approach. Computational and mathematical methods in medicine, 2017.

Venet, D., Dumont, J. E., & Detours, V. (2011). Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS computational biology, 7(10), e1002240.

Viloria, A., Lezama, O. B. P., & Mercado-Caruzo, N. (2020). Unbalanced data processing using oversampling: Machine Learning. Procedia Computer Science, 175, 108-113.