



Article

Application of Feature Selection and Deep Learning for Cancer Prediction Using DNA Methylation Markers

Rahul Gomes ^{1,*} , Nijhum Paul ², Nichol He ¹, Aaron Francis Huber ¹ and Rick J. Jansen ^{2,3,4,5,*} 

¹ Department of Computer Science, University of Wisconsin-Eau Claire, 133 Phillips Science Hall, 101 Roosevelt Ave, Eau Claire, WI 54701, USA

² Department of Public Health, North Dakota State University, 640S Aldevron Tower, 1455 14th Ave N, Fargo, ND 58102, USA

³ Genomics, Phenomics, and Bioinformatics Program, North Dakota State University, 640S Aldevron Tower, 1455 14th Ave N, Fargo, ND 58102, USA

⁴ Center for Immunization Research and Education (CIRE), North Dakota State University, 640S Aldevron Tower, 1455 14th Ave N, Fargo, ND 58102, USA

⁵ Center for Diagnostic and Therapeutic Strategies in Pancreatic Cancer, North Dakota State University, 640S Aldevron Tower, 1455 14th Ave N, Fargo, ND 58102, USA

* Correspondence: gomesr@uwec.edu (R.G.); rick.jansen@ndsu.edu (R.J.J.)

Abstract: DNA methylation is a process that can affect gene accessibility and therefore gene expression. In this study, a machine learning pipeline is proposed for the prediction of breast cancer and the identification of significant genes that contribute to the prediction. The current study utilized breast cancer methylation data from The Cancer Genome Atlas (TCGA), specifically the TCGA-BRCA dataset. Feature engineering techniques have been utilized to reduce data volume and make deep learning scalable. A comparative analysis of the proposed approach on Illumina 27K and 450K methylation data reveals that deep learning methodologies for cancer prediction can be coupled with feature selection models to enhance prediction accuracy. Prediction using 450K methylation markers can be accomplished in less than 13 s with an accuracy of 98.75%. Of the list of 685 genes in the feature selected 27K dataset, 578 were mapped to Ensemble Gene IDs. This reduced set was significantly ($FDR < 0.05$) enriched in five biological processes and one molecular function. Of the list of 1572 genes in the feature selected 450K data set, 1290 were mapped to Ensemble Gene IDs. This reduced set was significantly ($FDR < 0.05$) enriched in 95 biological processes and 17 molecular functions. Seven oncogene/tumor suppressor genes were common between the 27K and 450K feature selected gene sets. These genes were RTN4IP1, MYO18B, ANP32A, BRF1, SETBP1, NTRK1, and IGF2R. Our bioinformatics deep learning workflow, incorporating imputation and data balancing methods, is able to identify important methylation markers related to functionally important genes in breast cancer with high accuracy compared to deep learning or statistical models alone.

Keywords: DNA methylation; deep learning; breast cancer; TCGA



Citation: Gomes, R.; Paul, N.; He, N.; Huber, A.F.; Jansen, R.J. Application of Feature Selection and Deep Learning for Cancer Prediction Using DNA Methylation Markers. *Genes* **2022**, *13*, 1557. <https://doi.org/10.3390/genes13091557>

Academic Editor: Quan Zou

Received: 10 July 2022

Accepted: 25 August 2022

Published: 29 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

DNA methylation is important in cancer development and progression due to its role in silencing tumor suppressor genes or enhancing oncogene expression [1]. It involves adding a methyl group to the cytosine base pair position in the DNA of a living organism. This epigenetic modification has been demonstrated to directly influence gene expression [2]. Methylation data can be generated using high throughput sequencing techniques [3] where a single donor can have over 850,000 detectable methylation markers (CpGs) [4] across the human genome. Newer sequencing technologies now allow the evaluation of methylation at each genome location with whole genome sequence data. These datasets tend to be too large to reasonably parse through manually. Given the high ratio of markers to samples in these cancer datasets, it is necessary to establish a standardized automated framework

that is capable of processing such a massive amount of information, reducing prediction bias, and providing researchers a pipeline with access to pre-trained prediction knowledge. Predictions need to have high accuracy, sensitivity, and specificity. This will save training time and facilitate better knowledge discovery across research groups and cancer types.

Deep learning [5,6] (a subset of machine learning) has been one solution to this big data issue [7] that has gained significant momentum due to its ability to extract a meaningful subset of features from these different datasets [8,9] without any preprocessing or feature transformation. While deep learning algorithms excel in prediction, they can also be computationally expensive. That coupled with high dimensional datasets [10] can make accurate training and prediction very challenging. To overcome these obstacles feature selection prior to application of machine learning algorithms can be used [11]. Feature selection algorithms remove redundant and correlated information from big data thereby scaling down memory constraints. Additionally, data imbalance [12] also poses a challenge to machine learning and is a common issue with most datasets. An imbalanced dataset has more records belonging to a particular class than another. For example, a cancer dataset with 90% normal patients and 10% cancer patients can bias the model to predict a patient as normal. All of these are significant problems that affect neural network models applied to genomic datasets including DNA methylation data generated by high throughput sequencing.

There has been a focus in research on investigating methylation information to predict the relationship between specific gene methylation or expression and cancer. The earliest identified research article surveyed to use deep learning on methylation datasets was published in 2016 [13] and shows a deep learning model capable of predicting DNA methylation state from CpG markers using immortalized myelogenous (K562) cells. Since that time, research in this domain has been restricted to small datasets or to certain types of cancers. In a study by Angermuelle et al. [14], DNA and CpG modules from Single Cell Bisulfite Sequencing (scBS-seq) data and Single Cell Reduced Bisulfite Sequencing (scRRBS-seq) data for *Mus musculus* (house mouse) were trained using deep learning for prediction of methylated states in a cell. The DNA prediction module utilized a CNN with 2 hidden layers to extract features from DNA sequences while the CpG module utilized a Bidirectional Gated Recurrent Unit to extract features from CpG neighborhoods. This study was limited to only 6 human liver cancer cells (HepG2) [15] and mouse embryonic stem cells (ESCs) [16]. The dataset was small, and the CNN only implemented two convolution layers. Researchers in [17] utilized deep learning to extract DNA methylation states from Nanopore sequencing reads and found the prediction accuracy to be better than traditional techniques such as Hidden Markov Models (HMM). Liu et al. [18] utilized machine learning to extract CpG methylation markers for 27 cancer types from a total of 13,526 samples, where 10,140 samples were cancerous and 3386 were normal. The authors utilized t-statistics to extract the top 2000 CpG markers from 485,000 original CpG sites. The chosen markers were further filtered based on Random Forest and only 12 markers were used to train a deep learning model. While this research utilized a much larger dataset, the manual feature extraction process for selecting the top 2000 markers eliminated more than 99% of the CpG sites. Tian et al. [19], used whole genome bisulfite sequencing (WGBS) data of Human ESCs to predict if the input samples were hypo, hyper or mid-methylated. DNA sequences selected for this analysis were fixed at 400 bps and the input data was fed in CNN as a (400×4) feature matrix where 4 stood for bases A, T, C, and G. Authors also noted a data imbalance issue where more data was available on hyper-methylation which led to smaller prediction errors compared to hypo and mid-methylation sites.

Based on the previous work in this domain showing limited use of deep learning and feature selection, the objectives of this research was to develop a bioinformatics workflow incorporating both these aspects to select the most important methylation features associated with breast cancer thereby enabling high predictive accuracy but being scalable at the same time. Thereby our workflow maximizes deep learning predictive accuracy while maintaining scalability.

2. Materials and Methods

2.1. Dataset

To verify the proposed approach, Illumina 27K and 450K datasets were obtained from the Breast Invasive Carcinoma project from GDC Data Portal [20]. The disease types chosen were Ductal and Lobular Neoplasms. A total of 1188 samples were retrieved. Table 1 shows the distribution of methylation data from the Illumina platform.

Table 1. Characterization of CpG markers in the breast cancer datasets.

Dataset	Total Samples	Tumor Samples	Normal Samples	# CpG Markers
27K	337	309	28	27,578
450K	851	750	101	485,577

Our dataset had CpG markers corresponding to samples with no data. Further analysis showed that the percentage of these markers with null values was independent of cancer tissue type (normal or tumor). As machine learning algorithms do not tend to work well with ‘no data’, these markers were either removed or imputed before proceeding further. We selected either removal or imputation based on research in Lena et al. [21] where authors performed age correlations with methylation beta values before and after imputation. found that imputation for missing 20% information would not introduce a significant margin of error and that statistical tests could validate up to 30% of markers with no data imputed.

Analysis of the 27K dataset revealed 2597 CpG markers with null values across all 337 samples. Hence, a cut-off of 80% for a missing CpG marker across all samples was used to remove a specific CpG marker. The remaining 24,981 CpG markers across the 337 samples had a total of 4911 missing values. Four different imputation techniques were used to fill in these missing values namely zero, k-nearest neighbor (KNN), mean, and iterative imputation. During zero imputation, the missing values are replaced with zero. For KNN-imputation, the missing CpG markers are imputed using the average values or weighted by Euclidean distance from CpG markers distributed across ‘k’ neighboring samples [22]. Mean imputation calculates a simple average of CpG markers and assigns that value to the missing CpG marker. Finally, with iterative imputation, the samples with missing CpG markers are imputed by modeling each marker with missing value as a function of other markers in an iterative and round-robin manner. The imputer used Bayesian Ridge regression [23] to draw a probabilistic model for estimating missing values.

A similar path was followed for imputation for the 450K dataset. Initially, there were 485,577 CpG markers with almost 89,671 CpG markers having null values across all 851 samples. However, unlike in the 27K dataset, there were several samples with a significant percentage of null values. Hence, a cut-off of 30% for a missing CpG marker across samples was used to remove them. After removal, the dataset had 395,722 CpG markers across the 851 samples with a total of 332,330 missing values that required imputation. A similar process of mean, zero, and KNN imputation techniques were followed with the exclusion of iterative imputation due to memory constraints. Table 2 shows the error associated with these imputation approaches while Figure 1 shows the missing percentage of methylation values in the samples across datasets. These imputation techniques were verified using the Random Forest [24] regression technique. Based on the lowest MSE results, the mean imputation dataset was used for 27K, and zero imputation was used for 450K.

Table 2. Mean Squared Error (MSE) and Standard Deviation (STD) for the imputed results using the 27K and 450K datasets. A smaller value of MSE and STD signifies better model imputation.

	Metric	Zero Impute	KNN Impute	Mean Impute	Iterative Impute
27K	MSE	0.016648	0.016755	0.016749	0.016777
	STD	0.007299	0.007253	0.007245	0.007340
450K	MSE	0.017244	0.017253	0.017251	—
	STD	0.005273	0.005286	0.005307	—

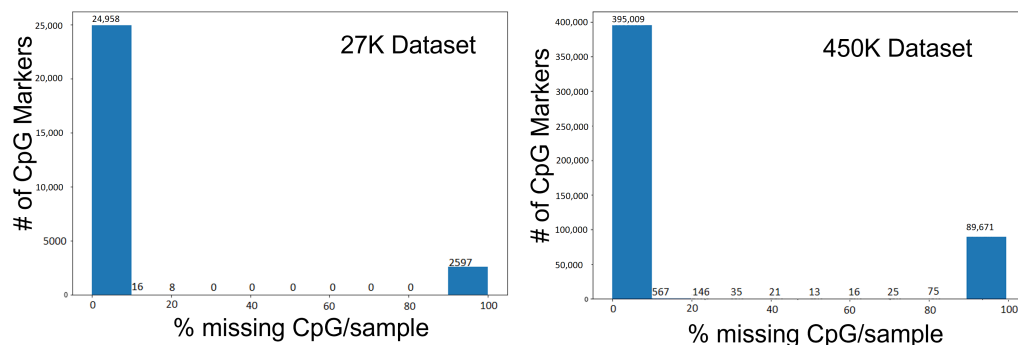


Figure 1. Missing percentage of CpG markers per sample among the 27K and 450K datasets.

2.2. Feature Selection to Reduce Dimensionality

Our developed workflow is visualized in Figure 2. It begins with the pre-processing steps explained in the methods, the feature selection as discussed in this section above, and the implementation of the deep learning model for prediction. In the dimension reduction step, two processes were used, Analysis of variance (ANOVA) and Random Forest.

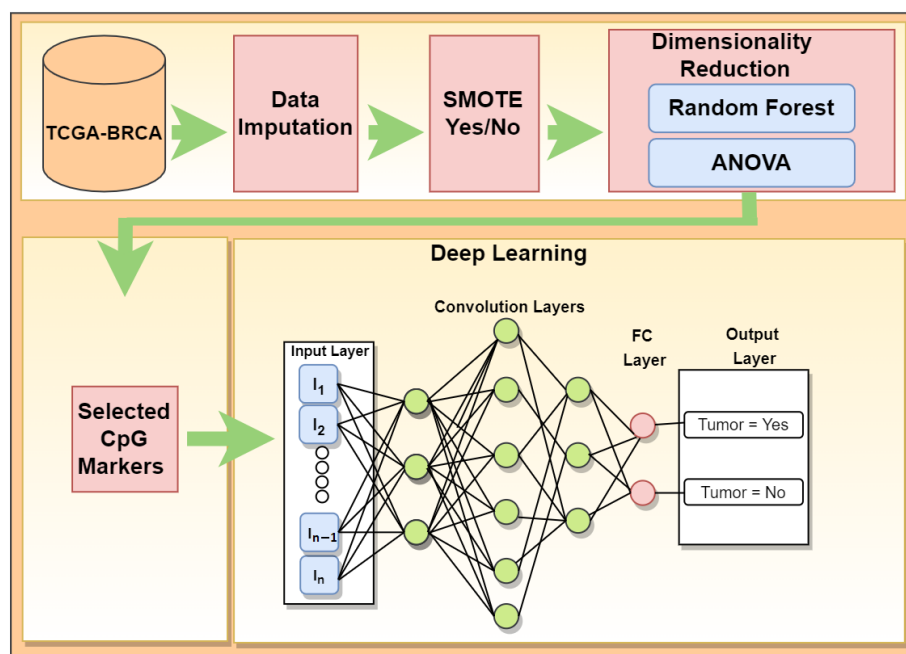


Figure 2. Developed workflow to classify samples as tumor or normal.

ANOVA is a technique that can compare the means of different groups. ANOVA uses F-tests (ratio of variances) to statistically test the equality of means. Larger values represent greater dispersion. Therefore if a specific feature, results in more separation in the means or classes, it will have a higher score. ANOVA offers an advantage over the common T-test

which is known to use a repeating set of comparisons among two attributes at a time [25]. ANOVA F-test model was trained separately on 24,981 markers from the 27K dataset and 395,722 markers from the 450K dataset.

Random Forests [24,26] are an ensemble learning method that constructs a multitude of decision trees for classification and regression at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees are [26]. By contrast, variables with low importance might be omitted from a model, making it simpler and faster to fit and predict.

2.3. Handling Data Imbalance

The class imbalance of tumor and normal samples is very high in the datasets as observed in Table 1. Such a high imbalance often results in biased prediction and misleading accuracy. One approach to address this challenge is increasing the observations of the minority class, also known as oversampling. For the model, Synthetic Minority Oversampling Technique (SMOTE) [27] was used for oversampling the minority (normal) samples. SMOTE creates new synthetic samples rather than just duplicating examples from minority classes, as duplicating the examples does not add any new information to the model. This technique works by selecting training data that are close in the feature space (nearest neighbors) and generating a new sample in that feature space like the neighbors. SMOTE was applied after oversampling, to ensure that the percentage of tumor and normal samples in our model are equal and to reduce the bias and misinterpretation.

2.4. Deep Learning Application for Cancer Prediction

The proposed deep learning sequential model is built using TensorFlow [28]. Two variants of a sequential deep neural network were implemented based on the size of the dataset. The 27K dataset was classified using a neural network with four hidden layers and an output layer. These hidden layers have 10, 20, 30, and 20 neurons respectively. These neurons are passed through a non-linear ReLU [29] activation function. To prevent overfitting the model, a dropout of 0.25 was used after each hidden layer. Dropouts prevent overfitting by turning off a few neurons at random. Since it is a binary classification, the binary cross-entropy loss function was used to evaluate the performance of the model. This loss function was optimized using Adam [30] optimizer. Since the 450K dataset is significantly larger, two different models were tested. The difference between these models can be found in the number of neurons in the model. The standard version had 10, 20, 30, and 20 neurons respectively in the hidden layers while the extended version had 100, 200, 300, and 200 filters thereby allowing the model to pass more information from the dataset. For future reference, the smaller deep sequential network will be referred to as the base model while the extended network will be referred to as the larger model. The model architecture was chosen for its simplicity, which allows for quick compilation and low computation cost. The datasets are simplistic, which means the problem does not require a complex architecture to produce good results. Each variant of the model is trained for 30 epochs. While training, model weights were updated using loss obtained from a validation dataset. The purpose to monitor validation loss is again related to overfitting. Training on a fixed sample size and tuning the model by monitoring its performance on untrained validation data ensures that the model does not overfit the training set. In addition, the learning rate reduction is applied with a factor of 0.5 if validation accuracy does not improve after 5 epochs, with a minimum learning rate of 0.0001.

Due to the imbalance of outcomes in the datasets, preprocessing is aimed toward generating a random, balanced training dataset where there is an equal number of tumor and normal samples. The following steps were used to ensure this selection:

- Dataset is separated into positive and negative tumor outcomes.
- The limiting outcome is randomly separated into two sets containing 70% (for training) and 30% (for testing) of the data.

- A subset of the non-limiting outcome, equal to 70% of the limiting outcome, is randomly chosen.
- The two subsets of the two outcomes, equivalent in number, is combined to form the training data.
- All remaining samples are combined to form the testing data set.
- Both data sets are randomly shuffled internally.

2.5. Gene Set Enrichment Analysis (GSEA)

Gene Set Enrichment Analysis (GSEA) was performed on the genes associated with the reduced set of CpG markers. GSEA identifies sets of genes that are enriched in a particular dataset when compared to a control. GSEA considers all genes in the dataset instead of considering only the subsets of genes with significant changes in gene expression. TCGAbiolinks package and ShinyGo [31] were used for performing GSEA.

2.6. Survival Analysis

Cox Proportional Hazards modeling was used to determine significant survival differences based on the 7-gene set expression score. These scores were then summed across each of the seven genes split into tertiles for each sample, so each sample would have one total score. This total expression score was categorized into five groups (<2, 2–3, 3–4, 4–5, 5+) because sample sizes were small for more extreme scores. A survival analysis was performed using this total expression score. We used a log-rank p -value < 0.05 to indicate a significant difference likely exists between score categories. We feel that further combining categories would obscure any ability to investigate a potential trend or pattern in the data. In this survival analysis, the categories created were combined already to alleviate any sample size or power issues. We propose that presenting the data with more categories will be more informative and helpful to the reader when considering this data and considering similar analyses with their own datasets.

3. Results

3.1. Feature Selection to Reduce Dimensionality

Figure 3 highlights the feature selection methods implemented in our analysis of four individual datasets. In ANOVA analysis, the markers with a p -value greater than a threshold value (0.05/total features) were removed from the total features. The number of features is reduced to 3704 for 27K and 125,949 for 450K datasets. Similar analysis of datasets after application of SMOTE, creating a balanced dataset, saw 15,483 features being selected for 27K and 260,159 features for the 450K. Results from Random Forests saw a lesser number of markers being selected from both datasets compared to ANOVA. Finally, to incorporate results from both these algorithms, the reduced features obtained from the ANOVA F-test are applied to the random forest model. This approach reduced the number of important markers from 24,981 to 336 for the 27K imbalanced dataset, while the final list of markers for the balanced dataset was 475. For the 450K dataset, it was observed that the number of features reduced from 395,722 to 1044 for imbalanced data and 1445 for the balanced dataset.

3.2. Deep Learning Application for Cancer Prediction

To verify the efficacy of the proposed feature selection technique, the baseline deep learning model was applied to three variants of the 27K dataset as highlighted in Table 3. The training sample for this dataset without SMOTE application is heavily biased towards tumor samples. To prevent the bias of the result of the sequential model, a 5-fold cross-validation technique was used each time comparing a subset of tumor samples with normal samples, while maintaining an equal distribution of records. For the datasets without SMOTE application there were 309 tumor samples and 28 normal samples. However, the application of SMOTE resulted in 618 samples with 309 samples each for normal and tumor. This process ensures all models are trained on an equal number of positive and

negative outcomes (e.g., normal and tumor samples), preventing bias. Figure 4 and Table A1 highlight the results from the proposed deep learning model on 27K dataset.

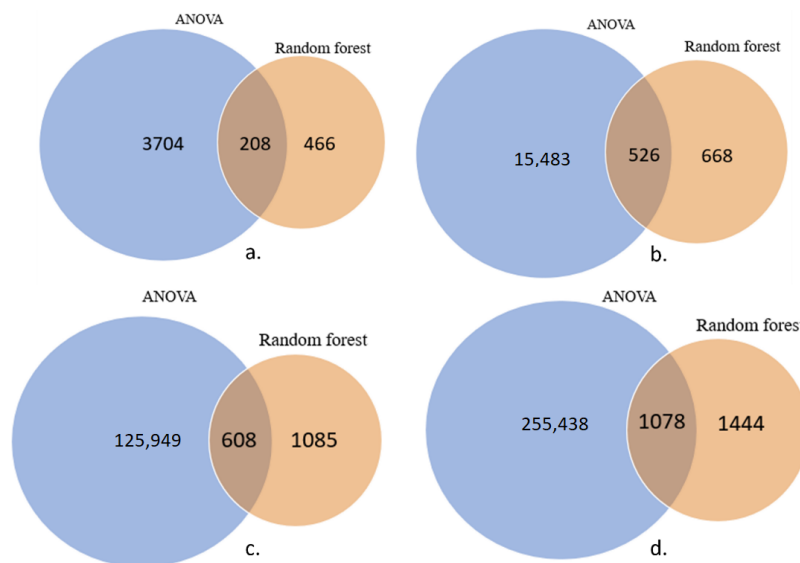


Figure 3. Number of CpG markers selected from (a) 27K before SMOTE, (b) 27K after SMOTE, (c) 450K datasets before SMOTE and (d) 450K after SMOTE application.

Table 3. Model training set-up for 27K and 450K dataset.

	Dataset	# Features	Sample Size	Tumor Samples	Normal Samples	Runtime
27K	All markers	24,981	337	309	28	21 s
	Anova_RF	336	337	309	28	12 s
	Anova_RF (with Smote)	475	618	309	309	13 s
450K	450K All (base + large)	395,722	851	750	101	1:44:10 s
	Anova_RF (base + large)	1044	851	750	101	38:41 s
	Anova_RF with SMOTE (base + large)	1445	1500	525	525	13 s

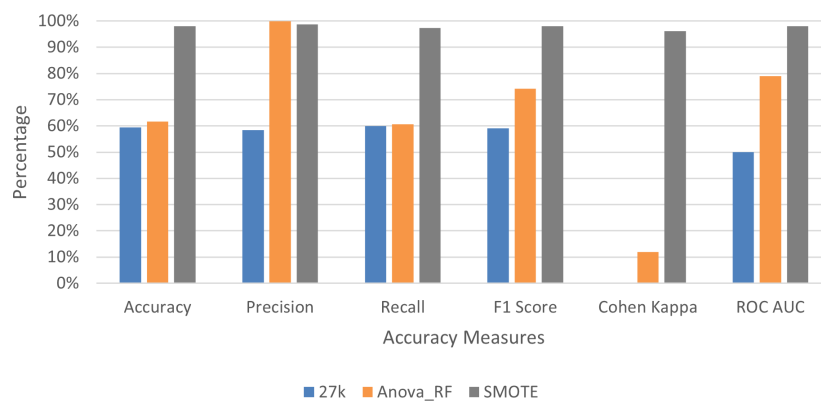


Figure 4. Accuracy metrics from deep learning models on 27K datasets before and after SMOTE application on training data using validation split of 30%. These graphs have been derived from average accuracy values of five trials shown in Table A2.

Two different models were tested on the original 450K datasets. As mentioned earlier, the difference between the two models is the number of filters per layer of the model. The base model is denoted with filters [10, 20, 30, 20] while the larger model has filters [100, 200, 300, 200]. A total of 851 samples were used for both models where 750 samples were tumor and 101 normal as shown in Table 1. Table 3 summarizes the model training set-up for the 450K dataset. Figure 5 summarizes the performance of different approaches for 450K. These results are a summary of accuracy metrics derived in our analysis from Table A2 for the base model and Table A3 for the larger model. Confusion matrices for both 27K and 450K datasets are also available from a separate testing sample that was not used during the training process and are shown in Table A4. Based on these findings, we can conclude that the original 450K and 27K datasets performed poorly in tumor prediction. In Table A4, it can be observed that without SMOTE application and deep learning, the majority of samples have been predicted as tumors even though they were normal. The filtered dataset performs much more reasonably with accuracy values significantly higher than the original dataset and comparable to the SMOTE application dataset. For example, in Table A3, the average accuracy for the filtered dataset is 91.28% while after SMOTE application it is 98.75%. Results after the application of SMOTE on this dataset were very promising. It is conclusive from these graphs that the majority of models trained on data without feature selection and with data imbalance are heavily biased towards predicting only one outcome. This large variability is shown through the excessive standard deviation bars. It is observed that the model performance was significantly better and consistent across different trials using selected features and more so with the balanced dataset produced by SMOTE.

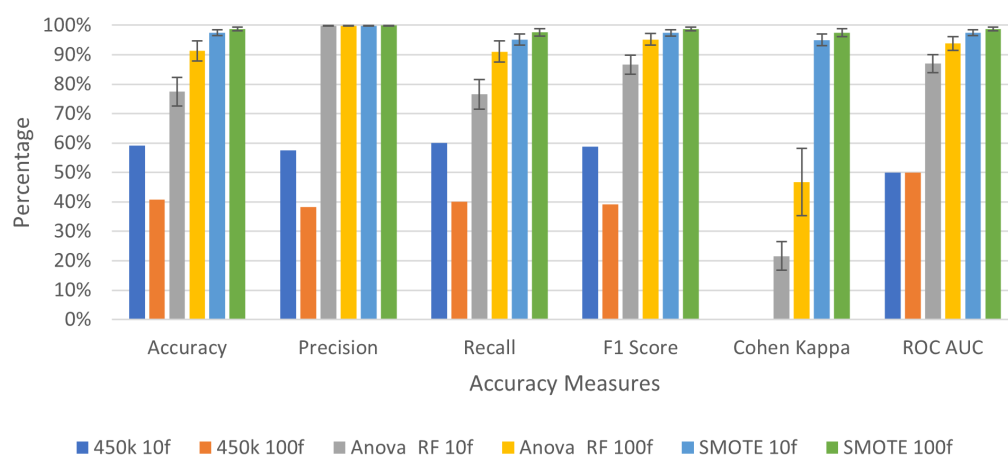


Figure 5. Accuracy metrics from deep learning models on 450K datasets before and after SMOTE application using validation split of 30%. These graphs have been derived from average accuracy values of five trials shown in Table A3.

3.3. Gene Set Enrichment Analysis (GSEA)

It is important to know the metastatic potential of primary malignant tissue as it is related to the choice of therapy. Previous studies indicate that sets of gene expression profiles can successfully predict survival [32]. After feature selection, four sets of important CpG markers were obtained. The CpGs with the lowest p -values (those associated with breast cancer), were annotated to identify which genes were associated with those CpGs. This gene list was then used to perform enrichment analysis and understand how these genes interact, and to infer the functional impact of the CpGs.

Gene Set Enrichment Analysis (GSEA) was performed on six different gene sets associated with the markers identified in Table 3. These were overall 27K and 450K datasets, 27K and 450K with feature selection but no SMOTE, and 27K and 450K with both feature selection and SMOTE. The EAcocomplete tool in TCGA biolinks package in R [33–35] was used on these sets to identify classes of genes or proteins that are over-

represented using annotations for that gene set. The barplot in Figures 6 and 7 shows canonical pathways significantly overrepresented (enriched) by the DEGs (differentially expressed genes) identified from reduced marker datasets after the application of SMOTE. The most statistically significant canonical pathways identified in the DEGs list are listed according to their p-value corrected FDR ($-\text{Log}$) (colored bars) and the ratio of list genes found in each pathway over the total number of genes in that pathway (red line). Plots corresponding to other sets are shown in Figures A1–A4.

Additionally, GSEA analysis was performed on the 27K and 450K with SMOTE gene lists using ShinyGo [31]. Of list of 685 genes in the 27K SMOTE set, 578 were mapped to Ensemble Gene IDs as shown in Table 4. This reduced set was significantly (FDR < 0.05) enriched in five biological processes, one molecular function, a P53 signaling pathway, and a network visualization of functional associations. (Figure 6b,c) Of the list of 1572 genes in the 450K SMOTE set, 1290 were mapped to Ensemble Gene IDs as shown in Table 4. This reduced set was significantly (FDR < 0.05) enriched in 95 biological processes and 17 molecular function, cellular senescence pathway, and a network visualization of functional associations (Figure 7c).

To evaluate the enriched gene sets identified above and understand their association with cancer, a comparative analysis was performed between these genes with cancer-related genes from other investigated references, one being the combined gene sets from COSMIC [36] and TSGene [37,38]. As shown in Table 4, the TSGene database contains 1217 tumor suppressor genes and the COSMIC has 2172 oncogenes. After combining these two databases (TSGene + COSMIC), a total of 3326 unique genes were identified. Results identified 55 genes that were common between the 27K ANOVA-RF with SMOTE and the list of combined 3326 genes in the COSMIC and TSGene database. Another 136 genes were common between the 450K ANOVA-RF with SMOTE and the 3326 genes. Of the 55 genes in the 27K SMOTE set the top 10 most significantly (FDR < 0.05) enriched biological processes, molecular function, and pathways, and network are visualized. (Figure 8b,c) Of the 136 genes in the 450K SMOTE set the top 10 most significantly (FDR < 0.05) enriched biological processes, molecular function, and pathways, and network are visualized (Figure 9b,c). Of note is the P53 signaling pathway was identified in both gene sets and the gene network is dominated by signaling and cancer-related processes. A list of seven genes were identified as common between them as shown in Figure 10.

Table 4. Evaluation of selected oncogenic and tumor suppressor associated gene sets identified to be associated with breast cancer.

Dataset	CpG Markers	Total Genes	COSMIC + TSGene Overlap (3326 Genes)	Sample Genes Overlap (100 Genes)
27K all	24,981	18,166	1214	98
27K ANOVA-RF	336	470	36	2
27K ANOVA-RF SMOTE	475	685	55	6
450K all	395,722	35,555	1455	100
450K ANOVA-RF	1044	1208	88	7
450K ANOVA-RF SMOTE	1445	1572	136	9

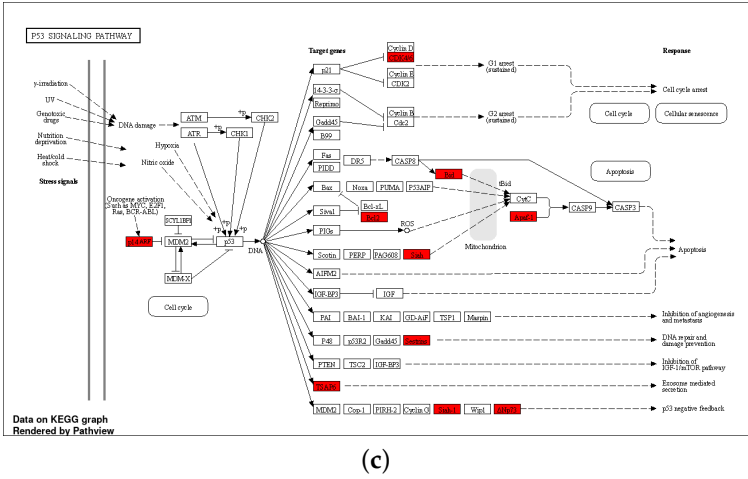
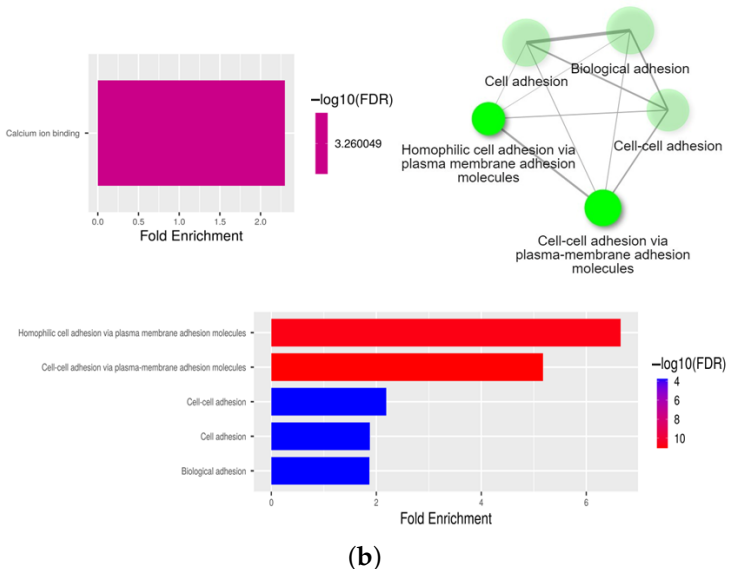
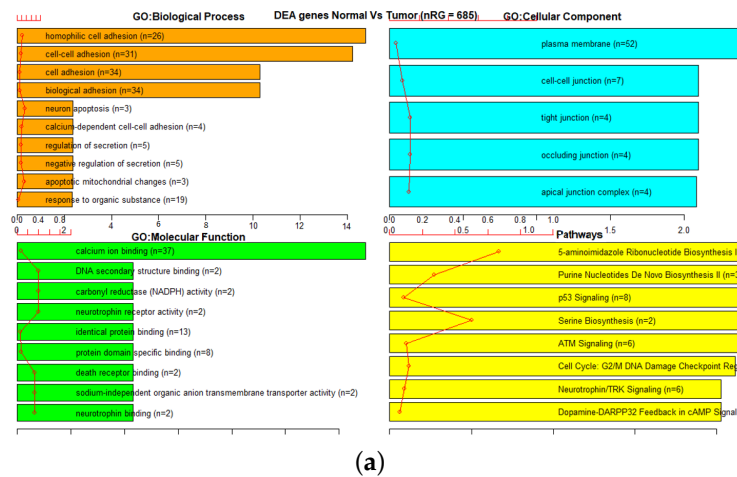
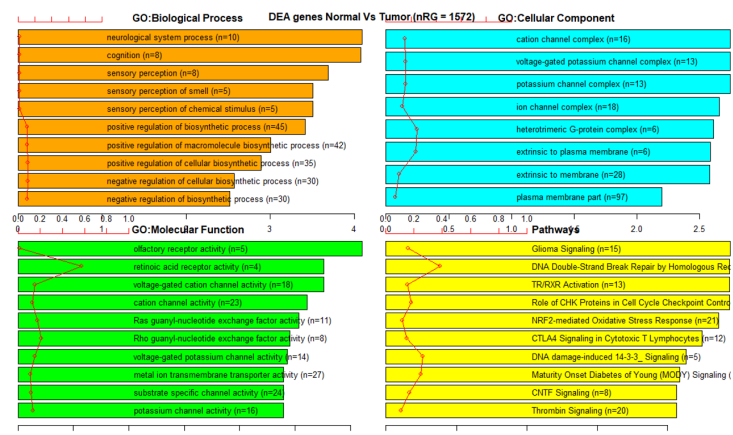
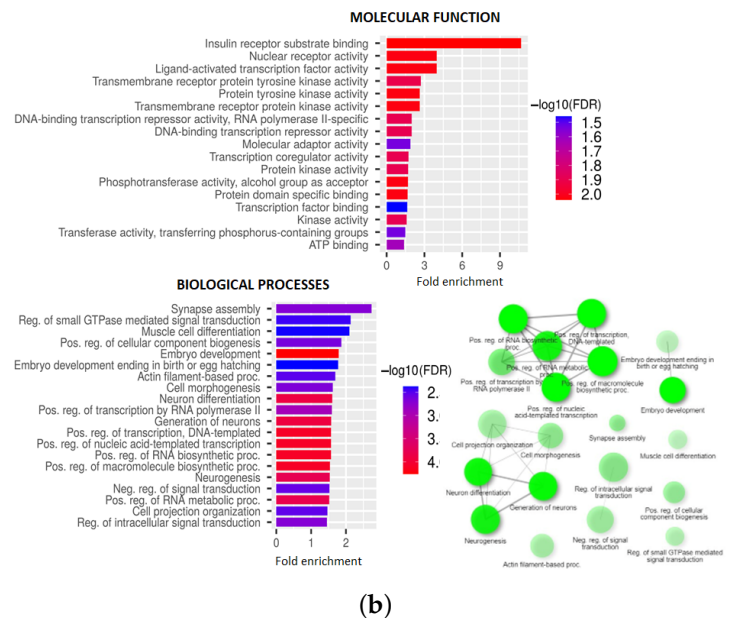


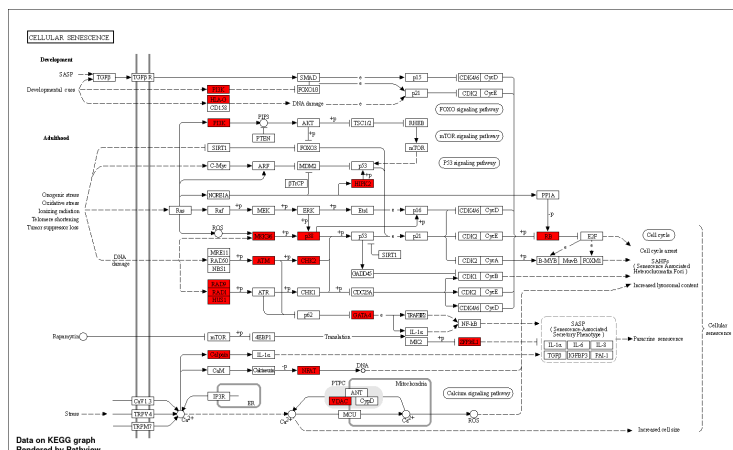
Figure 6. GSEA on reduced markers of 27K SMOTE using (a) TCGA EAanalysis where orange represents biological processes, cyan is a cellular component, green is molecular function, and yellow represents pathways, (b) ShinyGO with network visualization of functional associations where maroon color represents a molecular function, green represents network-based interaction of biological processes, while the graph in red and blue color represents the strength of the molecular functions identified and (c) ShinyGO enriched pathway visualization where red genes are in the gene set [39,40].



(a)

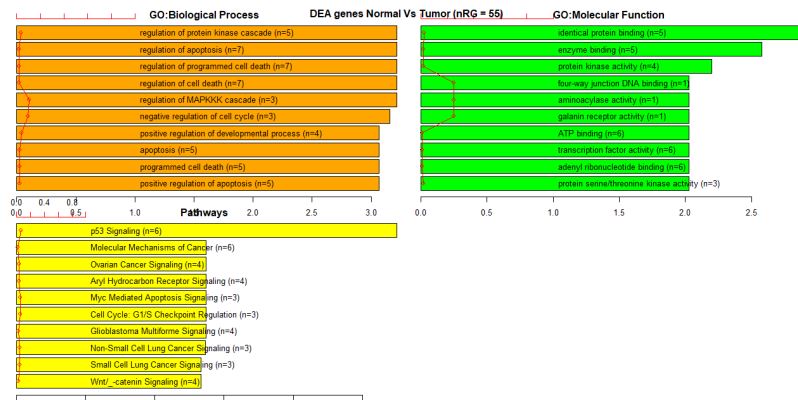


(b)

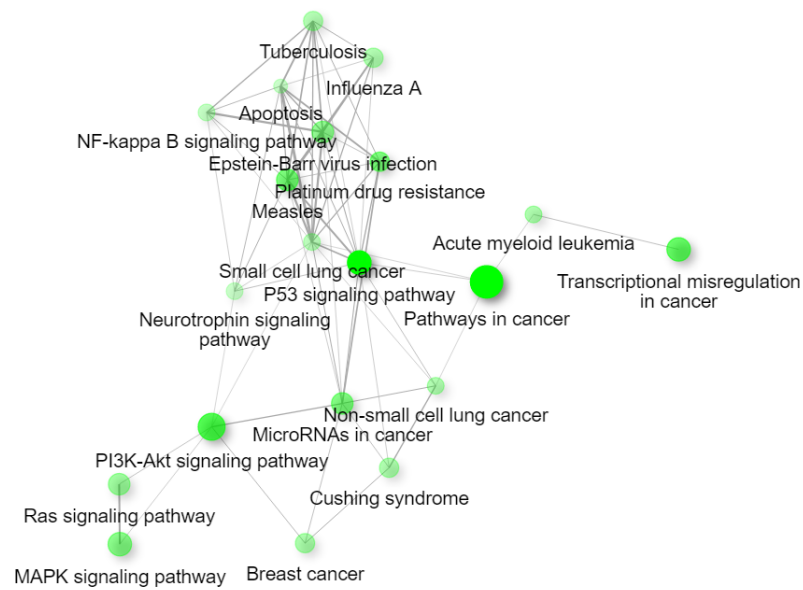


(c)

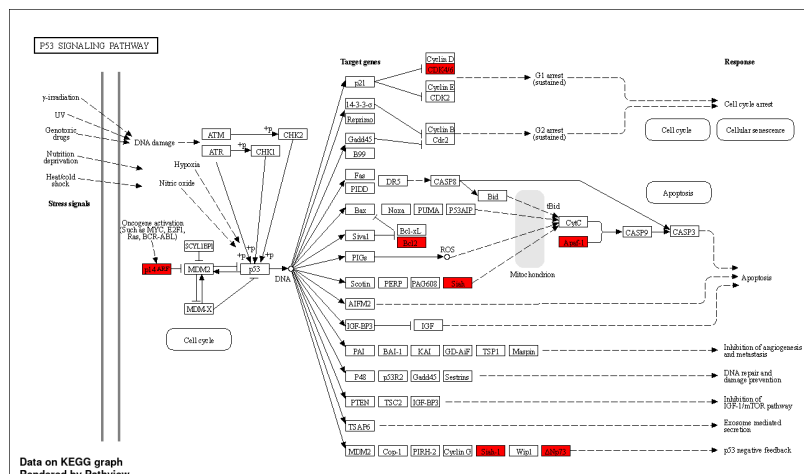
Figure 7. GSEA on reduced markers of 450K SMOTE using (a) TCGA EAanalysis where orange represents biological processes, cyan is a cellular component, green is molecular functions, and yellow represents pathways, (b) ShinyGO derived biological processes and molecular functions and network-based interaction of biological processes in green nodes, and (c) ShinyGO enriched pathway visualization where red genes are in the gene set [39,40].



(a)

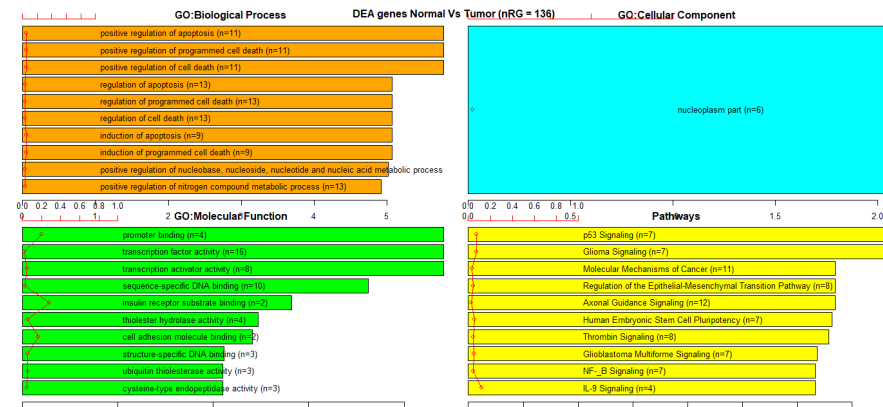


(b)

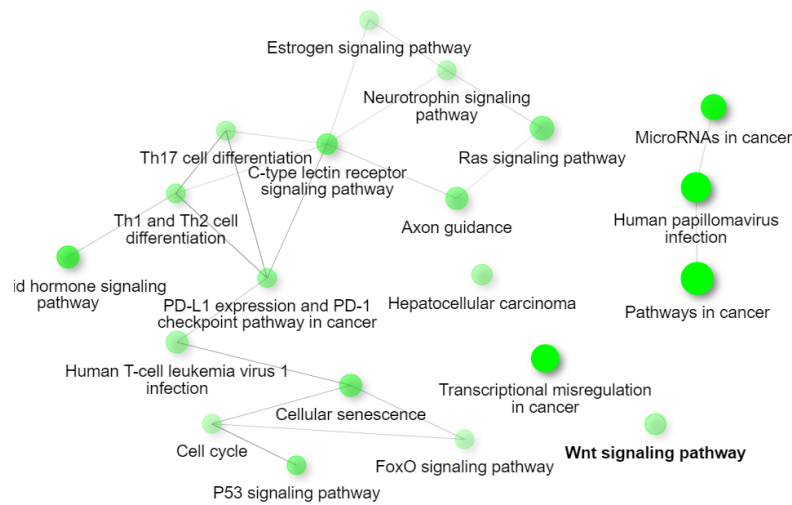


(c)

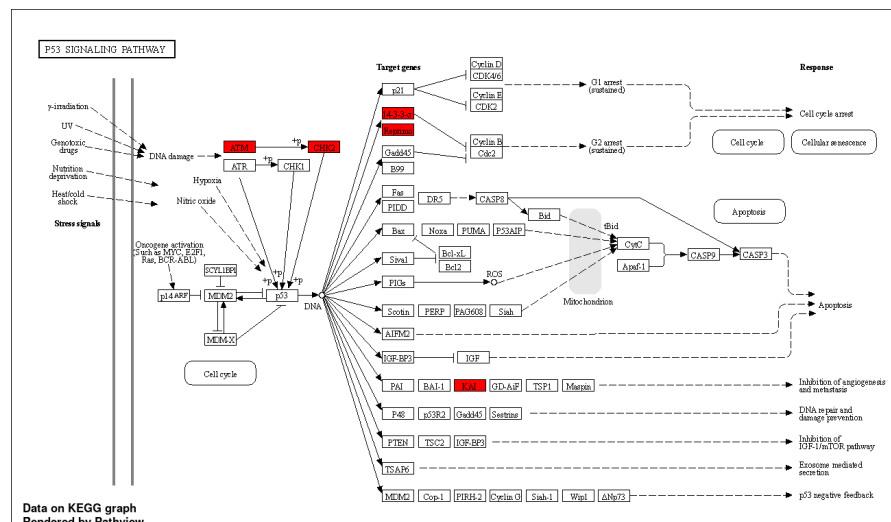
Figure 8. GSEA on Tumor suppressor/oncogene overlap subset of 27K with SMOTE, using (a) TCGA EAanalysis where orange represents biological processes, green is molecular functions, and yellow represents pathways, (b) ShinyGO with network visualization of functional associations, and (c) ShinyGO enriched pathway visualization where red genes are in the gene set [39,40].



(a)



(b)



(c)

Figure 9. Gene Set Enrichment Analysis on Tumor suppressor/oncogene overlap subset of 450K with SMOTE, (a) TCGA EAanalysis where orange represents biological processes, cyan is a cellular component, green is molecular functions, and yellow represents pathways, (b) ShinyGO with network visualization of functional associations, and (c) ShinyGO enriched pathway visualization where red genes are in the gene set [39,40].

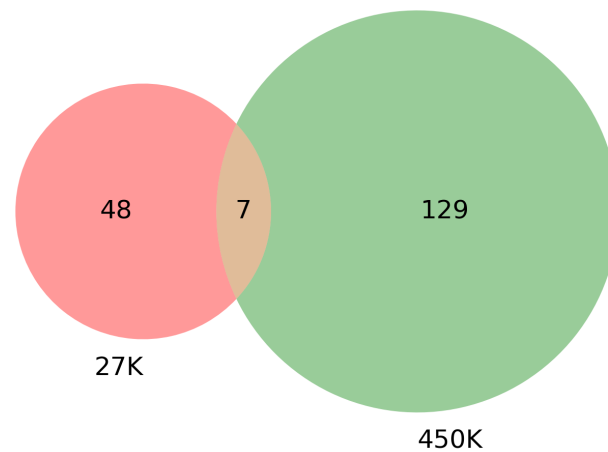


Figure 10. Venn diagram showing the overlap between genes that were found in TSGene + COSMIC [36–38] and also found in the 585 out of 685 mapped genes in 27K ANOVA-RF with SMOTE and 1290 out of 1572 mapped genes in 450K ANOVA-RF with SMOTE. A summary of these results can be seen in Table 4. Results indicate that 55 genes were common between TS + COSMIC and 27K while 136 between TS + COSMIC and 450K. There were 7 genes common between all of them.

3.4. Survival Analysis Using Seven Overlapping Genes

Using the seven genes *RTN4IP1*, *MYO18B*, *ANP32A*, *BRF1*, *SETBP1*, *NTRK1*, *IGF2R* identified as overlapping between the 27K with SMOTE, 450K with SMOTE, and the tumor suppressor/oncogene list, an expression score was calculated. For each of the seven genes, expression values were broken into tertiles for all samples. The lowest tertile received a score of 0, middle tertile score of 0.5, and highest tertile score of 1. A survival analysis was performed using this total expression score. The overall log-rank *p*-value (=0.0027) from the survival analysis indicates a significant difference in survival is observed based on the total expression score (Figure 11). Visually, the largest difference appears between the highest expression score group (5+) and the lowest expression score group (<2) with some irregularity in the middle categories when looking at the trend between increasing score and decreasing survival.

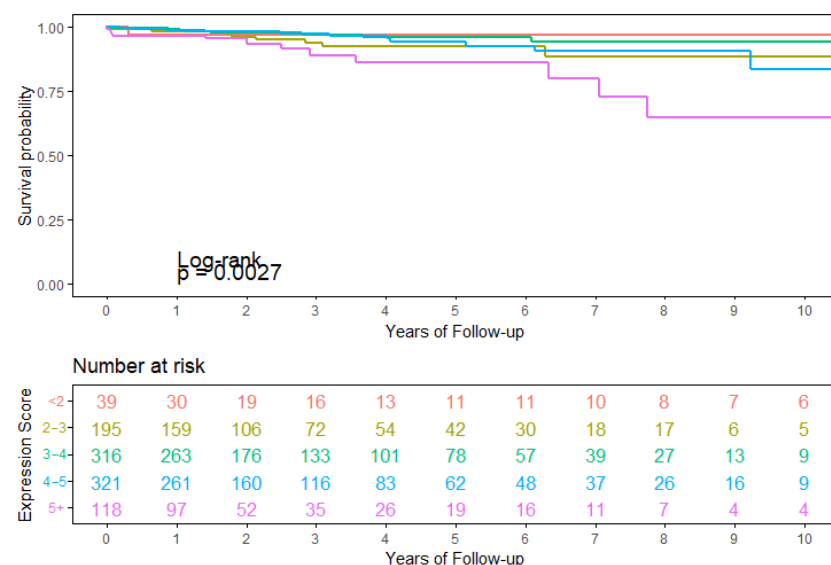


Figure 11. 10-year survival using TCGA-BRCA data and an expression score calculated across the seven genes which overlapped between the 27K with SMOTE, 450K with SMOTE, and the tumor suppressor/oncogene lists.

4. Discussion

In this paper, we demonstrated that imputing missing data and balancing methylation datasets is an important pre-analysis step in the bioinformatics workflow. Our workflow improves the accuracy of breast cancer case prediction using either 27K or 450K methylation datasets. An important note is that the imputation method selected was different depending on the size of the dataset, however, this difference was minimal. Our workflow consistently identified cell signaling and cancer-related processes as important features in predicting breast cancer cases.

Important pathways identified for genes associated with significant CpG markers were the P53 signaling pathway for 27K and cellular senescence for 450K datasets. There is strong evidence in the literature with over 68 studies linking altered P53 signaling with breast cancer, however, none of these studies have demonstrated altered methylation as a potential reason for P53 signaling disruption as we do with this study. Likewise, strong evidence also exists for the association between cellular senescence and breast cancer with 48 study results returned in PubMed. Although again, the evidence linking methylation alterations with these pathways in breast cancer are lacking.

Biological functions identified for genes associated with significant CpG markers were focused around adhesion for 27K and around post transcription and differentiation of cells for 450K datasets. There are 18 studies related to adhesion biological function and breast cancer in PubMed with two studies providing evidence of a potential role of methylation in this relationship. The study by [41] observes that methylation can alter focal adhesion pathways when MCF-7 cells are exposed to cadmium and then selenium. Kominsky et al. [42] reported greater discohesion with hypermethylation of CLDN-7 in breast cancer cell lines but not invasive ductal carcinomas.

When focusing on the oncogene/tumor suppressor gene overlap, important pathways included for 27K and for 450K focused on altered cell signaling and cancer. The overlap between the 27K, 450K, and oncogene/tumor suppressor gene list included seven genes (RTN4IP1, MYO18B, ANP32A, BRF1, SETBP1, NTRK1, IGF2R). This list of seven genes highlights those genes that are known to be important in cancer development and happen to be on both Illumina chips. There is one study by Savci-Heijink et al. [43] observing a relationship between RTN4IP1 gene expression and breast cancer. A study by Koo et al. [44] demonstrated a decrease of BRF2 methylation with exposure to soy isoflavone daidzein in breast cancer cells. A study by Di Emidio et al. [45] linking breast cancer treatment, Cyclophosphamide, to altered methylation of IGF2R in mouse offspring suggests that identification of this gene may be a result of treatment rather than cancer process.

The strengths of this study include a large sample size, established, and standardized pre-processing of the methylation data, using simulation to balance an unbalanced dataset, and imputing missing data. There is great overlap between the 27K and 450K datasets in terms of overlapping site-specific markers (94% of 27K loci appear in the 450K set) and correlation of methylation values ($R^2 = 0.95$) [46]. This means that any overlapping genes appearing in both datasets represent an independent validation of those overlapping markers as samples will only have either 27K or 450K data but not both in TCGA. All these components help to improve the accuracy and reliability of our analysis. Methylation markers have been associated with genes based on sequence relationships, however, there are likely to be methylation marker effects in more distant genes or based on spatial relationships formed during cell cycle phases that have not been captured within these datasets. Therefore, we expect that some of our gene enrichment interpretations may be missing or inaccurate. The samples were not grouped based on disease features and methylation is a dynamic process that may fluctuate over time, which limits our ability to determine which methylation changes are responsible for the development of breast cancer vs changes that are the result of the presence of breast cancer.

5. Conclusions

We proposed a deep learning framework that can capture the most significant biomarkers responsible for breast cancer. Our model is capable of handling a high volume of data with missing values and class imbalance. We observed that reduced features with a balanced class performed better in predicting outcomes than features with an imbalanced dataset. We also performed Gene Set Enrichment Analysis on the sets of genes reduced by our model. To evaluate the efficacy of our model, we compared the reduced sets of genes with several cancer resources. The results seem to support the notion that deep learning methodologies for cancer prediction can be extended for use in the prediction of different types of cancer which will form the basis of our future work. Incorporating methylation data into this story is important from a public health standpoint providing a potential point of prevention and from a treatment standpoint for a potential point to target the P53 or cellular senescence pathways. Deep learning models can be computationally expensive as shown in this research, but to provide accurate results there is a need to handle more diverse datasets as well as take less time to train. Further research will focus on expanding the training dataset to incorporate other clinical variables in the decision-making process. We will also incorporate tumor sub-types and grade information followed by a web-based API to enhance the efficacy of the proposed approach. The feasibility of label-specific weights while training a deep learning model as an alternative to the application of SMOTE will also be explored to verify if that is a more robust technique compared to generating synthetic data for addressing data imbalance. Our focus on methylation markers rather than gene expression has discovered some novelty in breast cancer-specific markers. In addition, these methylation markers have been annotated to be associated with specific genes based on distance, however, these methylation markers may in fact alter or affect other more distant genes in the genome. To supplement this approach, we will explore modifications in the feature selection so that it can accept microarray gene expression data alongside methylation values. This enables the evaluation of specific genes based on their differentially expressed values and their previous association with a cancer type. Identification of these core genes will further reduce methylation markers that are being analyzed by the deep learning model thereby establishing a more robust and targeted approach. By fulfilling these five limitations, we will continue to develop functionality and test its use to enhance the utility of this workflow for the cancer research community.

Author Contributions: Conceptualization, methodology, and supervision, R.G. and R.J.J.; validation, N.P., R.J.J. and R.G. data curation and software, N.P., A.F.H. and N.H.; writing—original draft preparation, N.P., R.G. and R.J.J.; writing—review and editing, R.G. and R.J.J.; visualization, N.H., A.F.H. and N.P.; funding acquisition, R.J.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by NIH NDSU COBRE Center for Diagnostic and Therapeutic Strategies in Pancreatic Cancer (P20GM109024). The computational resources of the study were provided by the NDSU CCAST through the NSF Major Research Instrumentation (MRI) Program (2019077) and the Blugold Center for High-Performance Computing under NSF MRI Program (1920220).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Links to our scripts used for analysis can be found here: https://github.com/rahulgomes19/Deep_Learning_Methylation (accessed on 10 August 2022).

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

Abbreviations

The following abbreviations are used in this manuscript:

GSEA	Gene Set Enrichment Analysis
FDR	False Discovery Rate
SMOTE	Synthetic Minority Over Sampling Technique
ANOVA	Analysis of Variance
KNN	K-Nearest Neighbor
TCGA	The Cancer Genome Atlas

Appendix A

The following section contains the results of the deep learning model that were applied on six different methylation datasets.

Table A1. Results of deep learning model on 27K dataset.

Dataset	Trials	Accuracy	Precision	Recall	F1 Score	Cohen Kappa	RUC
All markers	1	0.02694	0	0	0	0	0.5
	2	0.97306	0.97306	1	0.98635	0	0.5
	3	0.02694	0	0	0	0	0.5
	4	0.97306	0.97306	1	0.98635	0	0.5
	5	0.97306	0.97306	1	0.98635	0	0.5
	Avg.	0.59461	0.58384	0.6	0.59181	0	0.5
	St.dev.	0.51822	0.53297	0.54772	0.54025	0	0
Anova_RF	1	0.50168	1	0.48789	0.65581	0.04882	0.74395
	2	0.92593	0.99628	0.92734	0.96057	0.36216	0.90117
	3	0.52525	1	0.51211	0.67735	0.05352	0.75606
	4	0.49158	1	0.47751	0.64637	0.04692	0.73875
	5	0.63636	1	0.62630	0.77021	0.08281	0.81315
	Avg.	0.61616	0.99926	0.60623	0.74206	0.11885	0.79061
	St.dev.	0.18252	0.00166	0.18904	0.13165	0.13679	0.06854
Smote	1	0.98370	0.96842	1	0.98396	0.96739	0.98370
	2	0.99457	1	0.98913	0.99454	0.98913	0.99457
	3	0.97826	1	0.95652	0.97778	0.95652	0.97826
	4	0.97283	0.97802	0.96739	0.97268	0.94565	0.97283
	5	0.97283	0.98876	0.95652	0.97238	0.94565	0.97283
	Avg.	0.98043	0.98704	0.97391	0.98027	0.96087	0.98043
	St.dev.	0.00909	0.01385	0.01975	0.00926	0.01819	0.00909

Table A2. Results of deep learning model on 450K dataset using the base model.

Dataset	Trials	Accuracy	Precision	Recall	F1 Score	Cohen Kappa	ROC AUC
All markers base model	1	0.04231	0	0	0	0	0.5
	2	0.95769	0.95769	1	0.97839	0	0.5
	3	0.95769	0.95769	1	0.97839	0	0.5
	4	0.04231	0	0	0	0	0.5
	5	0.95769	0.95769	1	0.97839	0	0.5
	Avg.	0.59154	0.57461	0.6	0.58703	0	0.5
	St.dev.	0.50137	0.52455	0.54772	0.53588	0	0
AnovaRF base model	1	0.79831	0.99814	0.79087	0.8825	0.23336	0.87877
	2	0.72355	0.99589	0.71429	0.8319	0.15957	0.82381
	3	0.72496	1	0.71281	0.83233	0.17359	0.85641
	4	0.83216	0.99822	0.82622	0.90411	0.27686	0.89644
	5	0.79408	1	0.78498	0.87954	0.23603	0.89249
	Avg.	0.77461	0.99845	0.76583	0.86608	0.21588	0.86958
	St.dev.	0.04828	0.00169	0.05027	0.03242	0.04845	0.03
AnovaRF- SMOTE base model	1	0.98889	1	0.97778	0.98876	0.97778	0.98889
	2	0.97778	1	0.95556	0.97727	0.95556	0.97778
	3	0.96889	1	0.93778	0.96789	0.93778	0.96889
	4	0.97778	1	0.95556	0.97727	0.95556	0.97778
	5	0.96222	0.99524	0.92889	0.96092	0.92444	0.96222
	Avg.	0.97511	0.99905	0.95111	0.97442	0.95022	0.97511
	St.dev.	0.01011	0.00213	0.01886	0.01057	0.02022	0.01011

Table A3. Results of deep learning model on 450K dataset using the larger model.

Dataset	Trials	Accuracy	Precision	Recall	F1 Score	Cohen Kappa	ROC AUC
All markers large model	1	0.95769	0.95769	1	0.97839	0	0.5
	2	0.04231	0	0	0	0	0.5
	3	0.04231	0	0	0	0	0.5
	4	0.95769	0.95769	1	0.97839	0	0.5
	5	0.04231	0	0	0	0	0.5
	Avg.	0.40846	0.38307	0.4	0.39135	0	0.5
	St.dev.	0.50137	0.52455	0.54772	0.53588	0	0
AnovaRF large model	1	0.93089	0.99685	0.93078	0.96268	0.50331	0.93206
	2	0.89563	0.99672	0.89396	0.94255	0.39113	0.91365
	3	0.86601	1	0.86009	0.92478	0.3422	0.93004
	4	0.95628	1	0.95435	0.97664	0.63885	0.97717
	5	0.91537	0.99839	0.91311	0.95385	0.45727	0.93989
	Avg.	0.91283	0.99839	0.91046	0.9521	0.46656	0.93856
	St.dev.	0.03431	0.00161	0.0359	0.01971	0.11432	0.0236
AnovaRF-SMOTE large model	1	0.98222	0.99543	0.96889	0.98198	0.96444	0.98222
	2	0.99556	1	0.99111	0.99554	0.99111	0.99556
	3	0.98667	1	0.97333	0.98649	0.97333	0.98667
	4	0.98	1	0.96	0.97959	0.96	0.98
	5	0.99333	1	0.98667	0.99329	0.98667	0.99333
	Avg.	0.98756	0.99909	0.976	0.98738	0.97511	0.98756
	St.dev.	0.00678	0.00204	0.0128	0.00693	0.01355	0.00678

Table A4. Confusion matrices of deep learning model on 27K and 450K datasets showing average outcome of five trials. Columns represent prediction and rows represent actual values. It can be observed that without SMOTE application and deep learning, majority of samples have been predicted as cancer.

		Original		AnovaRF		SMOTE	
		Normal	Cancer	Normal	Cancer	Normal	Cancer
27K	Normal	3.2	4.8	7.8	0.2	90.8	1.2
	Cancer	115.6	173.4	113.8	175.2	2.4	89.6
Prediction Sample Size		297		297		184	
		Original		AnovaRF		SMOTE	
		Normal	Cancer	Normal	Cancer	Normal	Cancer
450K base model	Normal	12	18	28.2	1.8	224.8	0.2
	Cancer	271.6	407.4	150	529	11	214
Prediction Sample Size		709		709		450	
		Original		AnovaRF		SMOTE	
		Normal	Cancer	Normal	Cancer	Normal	Cancer
450K larger model	Normal	18	12	28.4	1.6	224.8	0.2
	Cancer	407.4	271.6	86	593	5.4	219.6
Prediction Sample Size		709		709		450	

Appendix B

The following section lists the enrichment analysis outcome from 27K and 450K datasets before and after SMOTE application.

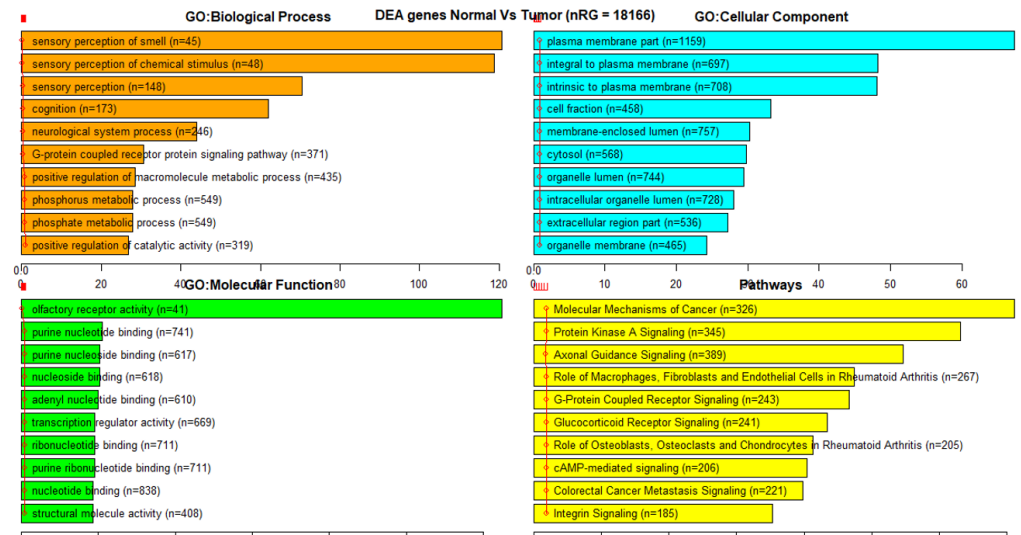


Figure A1. Gene Set Enrichment Analysis on all markers of 27K.

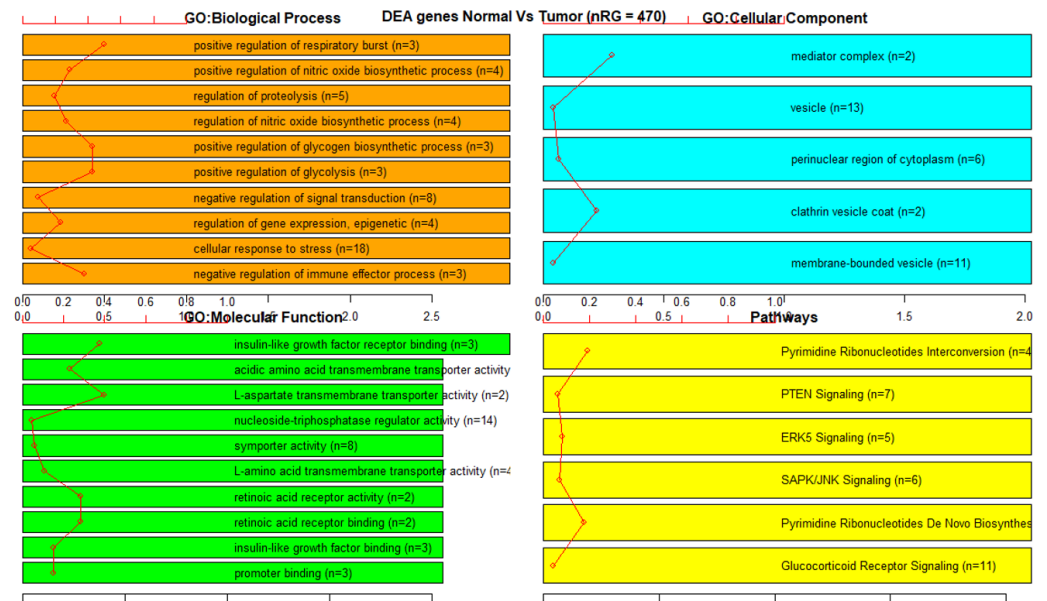


Figure A2. Gene Set Enrichment Analysis on reduced markers of 27K without SMOTE.

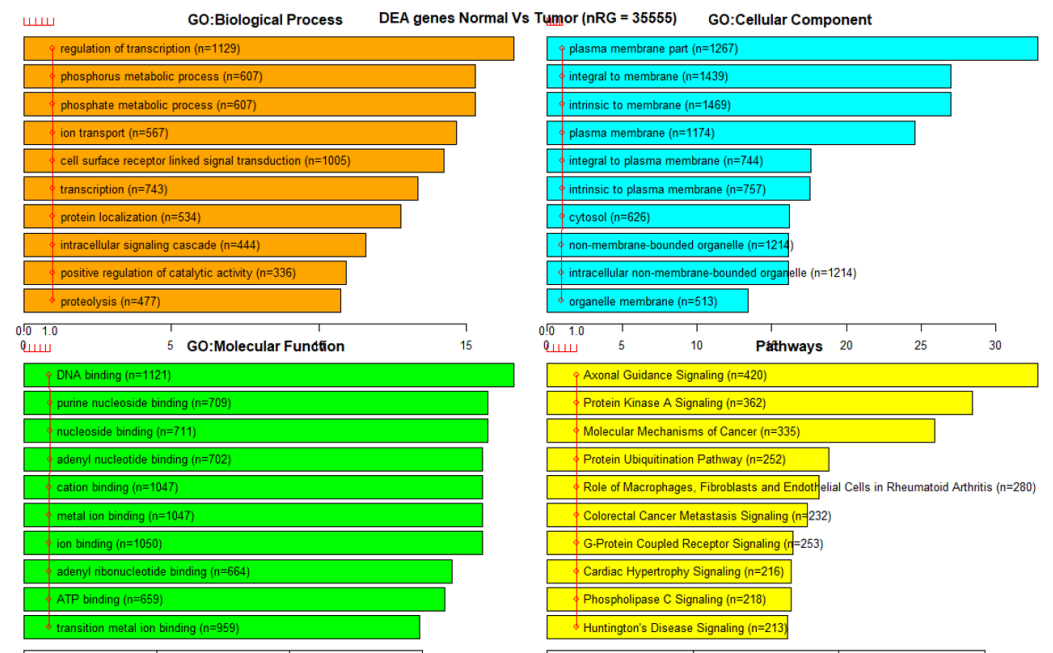


Figure A3. Gene Set Enrichment Analysis on all markers of 450K.

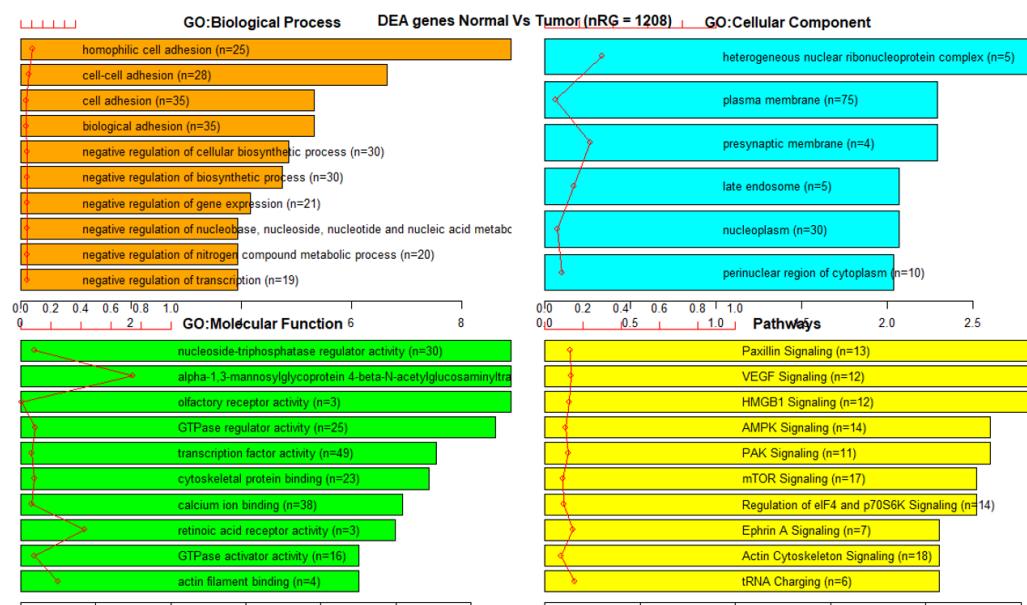


Figure A4. Gene Set Enrichment Analysis on reduced markers of 450K without SMOTE.

References

- Xiao, C.L.; Zhu, S.; He, M.; Chen, D.; Zhang, Q.; Chen, Y.; Yu, G.; Liu, J.; Xie, S.Q.; Luo, F.; et al. N6-methyladenine DNA modification in the human genome. *Mol. Cell* **2018**, *71*, 306–318. [[CrossRef](#)] [[PubMed](#)]
- Gardiner-Garden, M.; Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **1987**, *196*, 261–282. [[CrossRef](#)]
- Levin, J.Z.; Yassour, M.; Adiconis, X.; Nusbaum, C.; Thompson, D.A.; Friedman, N.; Gnirke, A.; Regev, A. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* **2010**, *7*, 709–715. [[CrossRef](#)] [[PubMed](#)]
- IlluminaHumanMethylation450kmanifest: Annotation for Illumina's 450k Methylation Arrays. Available online: <https://bioconductor.org/packages/release/data/annotation/html/IlluminaHumanMethylation450kmanifest.html> (accessed on 10 August 2022).
- O'Shea, K.; Nash, R. An introduction to convolutional neural networks. *arXiv* **2015**, arXiv:1511.08458.
- Zaremba, W.; Sutskever, I.; Vinyals, O. Recurrent neural network regularization. *arXiv* **2014**, arXiv:1409.2329.
- Halevy, A.; Norvig, P.; Pereira, F. The unreasonable effectiveness of data. *IEEE Intell. Syst.* **2009**, *24*, 8–12. [[CrossRef](#)]
- Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*. Available online: <https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html> (accessed on 10 August 2022). [[CrossRef](#)]
- Johnson, R.; Zhang, T. Effective use of word order for text categorization with convolutional neural networks. *arXiv* **2014**, arXiv:1412.1058.
- Verleysen, M.; François, D. The curse of dimensionality in data mining and time series prediction. In *International Work-Conference on Artificial Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 758–770.
- Ahsan, M.; Gomes, R.; Chowdhury, M.; Nygard, K.E. Enhancing Machine Learning Prediction in Cybersecurity Using Dynamic Feature Selector. *J. Cybersecur. Priv.* **2021**, *1*, 199–218. [[CrossRef](#)]
- Longadge, R.; Dongre, S. Class imbalance problem in data mining review. *arXiv* **2013**, arXiv:1305.1707.
- Wang, Y.; Liu, T.; Xu, D.; Shi, H.; Zhang, C.; Mo, Y.Y.; Wang, Z. Predicting DNA methylation state of CpG dinucleotide using genome topological features and deep networks. *Sci. Rep.* **2016**, *6*, 19598. [[CrossRef](#)]
- Angermueller, C.; Lee, H.J.; Reik, W.; Stegle, O. DeepCpG: Accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* **2017**, *18*, 67.
- Hou, Y.; Guo, H.; Cao, C.; Li, X.; Hu, B.; Zhu, P.; Wu, X.; Wen, L.; Tang, F.; Huang, Y.; et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.* **2016**, *26*, 304–319. [[CrossRef](#)]
- Smallwood, S.A.; Lee, H.J.; Angermueller, C.; Krueger, F.; Saadeh, H.; Peat, J.; Andrews, S.R.; Stegle, O.; Reik, W.; Kelsey, G. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **2014**, *11*, 817–820. [[CrossRef](#)]
- Ni, P.; Huang, N.; Zhang, Z.; Wang, D.P.; Liang, F.; Miao, Y.; Xiao, C.L.; Luo, F.; Wang, J. DeepSignal: Detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics* **2019**, *35*, 4586–4595. [[CrossRef](#)]
- Liu, B.; Liu, Y.; Pan, X.; Li, M.; Yang, S.; Li, S.C. DNA methylation markers for pan-cancer prediction by deep learning. *Genes* **2019**, *10*, 778. [[CrossRef](#)]
- Tian, Q.; Zou, J.; Tang, J.; Fang, Y.; Yu, Z.; Fan, S. MRCNN: A deep learning model for regression of genome-wide DNA methylation. *BMC Genom.* **2019**, *20*, 192. [[CrossRef](#)]

20. Heath, A.P.; Ferretti, V.; Agrawal, S.; An, M.; Angelakos, J.C.; Arya, R.; Bajari, R.; Baqar, B.; Barnowski, J.H.; Burt, J.; et al. The NCI genomic data commons. *Nat. Genet.* **2021**, *53*, 257–262. [[CrossRef](#)]
21. Di Lena, P.; Sala, C.; Prodi, A.; Nardini, C. Missing value estimation methods for DNA methylation data. *Bioinformatics* **2019**, *35*, 3786–3793. [[CrossRef](#)]
22. Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R.B. Missing value estimation methods for DNA microarrays. *Bioinformatics* **2001**, *17*, 520–525. [[CrossRef](#)]
23. Tipping, M.E. Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **2001**, *1*, 211–244.
24. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
25. Kim, T.K. Understanding one-way ANOVA using conceptual figures. *Korean J. Anesthesiol.* **2017**, *70*, 22. [[CrossRef](#)]
26. Gomes, R.; Ahsan, M.; Denton, A. Random forest classifier in SDN framework for user-based indoor localization. In Proceedings of the 2018 IEEE International Conference on Electro/Information Technology (EIT), Rochester, MI, USA, 3–5 May 2018; pp. 0537–0542.
27. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
28. Abadi, M.; Barham, P.; Chen, Z.; Chen, J.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. {TensorFlow}: A System for {Large-Scale} Machine Learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
29. Xu, B.; Wang, N.; Chen, T.; Li, M. Empirical evaluation of rectified activations in convolutional network. *arXiv* **2015**, arXiv:1505.00853.
30. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
31. Ge, S.X.; Jung, D.; Yao, R. ShinyGO: A graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **2020**, *36*, 2628–2629. [[CrossRef](#)]
32. Ein-Dor, L.; Kela, I.; Getz, G.; Givol, D.; Domany, E. Outcome signature genes in breast cancer: Is there a unique set? *Bioinformatics* **2005**, *21*, 171–178. [[CrossRef](#)]
33. Colaprico, A.; Silva, T.C.; Olsen, C.; Garofano, L.; Cava, C.; Garolini, D.; Sabedot, T.S.; Malta, T.M.; Pagnotta, S.M.; Castiglioni, I.; et al. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **2016**, *44*, e71. [[CrossRef](#)]
34. Silva, T.C.; Colaprico, A.; Olsen, C.; D’Angelo, F.; Bontempi, G.; Ceccarelli, M.; Noushmehr, H. TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. *F1000Research* **2016**, *5*, 1542. [[CrossRef](#)]
35. Mounir, M.; Lucchetta, M.; Silva, T.C.; Olsen, C.; Bontempi, G.; Chen, X.; Noushmehr, H.; Colaprico, A.; Papaleo, E. New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. *PLoS Comput. Biol.* **2019**, *15*, e1006701. [[CrossRef](#)] [[PubMed](#)]
36. Forbes, S.; Clements, J.; Dawson, E.; Bamford, S.; Webb, T.; Dogan, A.; Flanagan, A.; Teague, J.; Wooster, R.; Futreal, P.; et al. COSMIC 2005. *Br. J. Cancer* **2006**, *94*, 318–322. [[CrossRef](#)] [[PubMed](#)]
37. Zhao, M.; Sun, J.; Zhao, Z. TSGene: A web resource for tumor suppressor genes. *Nucleic Acids Res.* **2013**, *41*, D970–D976. [[CrossRef](#)] [[PubMed](#)]
38. Zhao, M.; Kim, P.; Mitra, R.; Zhao, J.; Zhao, Z. TSGene 2.0: An updated literature-based knowledgebase for tumor suppressor genes. *Nucleic Acids Res.* **2016**, *44*, D1023–D1031. [[CrossRef](#)]
39. Luo, W.; Brouwer, C. Pathview: An R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* **2013**, *29*, 1830–1831. [[CrossRef](#)]
40. Kanehisa, M.; Furumichi, M.; Sato, Y.; Ishiguro-Watanabe, M.; Tanabe, M. KEGG: Integrating viruses and cellular organisms. *Nucleic Acids Res.* **2021**, *49*, D545–D551. [[CrossRef](#)]
41. Liang, Z.Z.; Zhang, Y.X.; Zhu, R.M.; Li, Y.L.; Jiang, H.M.; Li, R.B.; Chen, Q.X.; Wang, Q.; Tang, L.Y.; Ren, Z.F. Identification of epigenetic modifications mediating the antagonistic effect of selenium against cadmium-induced breast carcinogenesis. *Environ. Sci. Pollut. Res.* **2022**, *29*, 22056–22068. [[CrossRef](#)]
42. Kominsky, S.L.; Argani, P.; Korz, D.; Evron, E.; Raman, V.; Garrett, E.; Rein, A.; Sauter, G.; Kallioniemi, O.P.; Sukumar, S. Loss of the tight junction protein claudin-7 correlates with histological grade in both ductal carcinoma in situ and invasive ductal carcinoma of the breast. *Oncogene* **2003**, *22*, 2021–2033. [[CrossRef](#)]
43. Savci-Heijink, C.; Halfwerk, H.; Koster, J.; Horlings, H.; Van De Vijver, M. A specific gene expression signature for visceral organ metastasis in breast cancer. *BMC Cancer* **2019**, *19*, 333. [[CrossRef](#)]
44. Koo, J.; Cabarcas-Petroski, S.; Petrie, J.L.; Diette, N.; White, R.J.; Schramm, L. Induction of proto-oncogene BRF2 in breast cancer cells by the dietary soybean isoflavone daidzein. *BMC Cancer* **2015**, *15*, 905. [[CrossRef](#)]
45. Di Emidio, G.; D’Aurora, M.; Placidi, M.; Franchi, S.; Rossi, G.; Stuppia, L.; Artini, P.G.; Tatone, C.; Gatta, V. Pre-conceptional maternal exposure to cyclophosphamide results in modifications of DNA methylation in F1 and F2 mouse oocytes: Evidence for transgenerational effects. *Epigenetics* **2019**, *14*, 1057–1064. [[CrossRef](#)]
46. Bibikova, M.; Barnes, B.; Tsan, C.; Ho, V.; Klotzle, B.; Le, J.M.; Delano, D.; Zhang, L.; Schroth, G.P.; Gunderson, K.L.; et al. High density DNA methylation array with single CpG site resolution. *Genomics* **2011**, *98*, 288–295. [[CrossRef](#)]