MACHINE LEARNING TECHNIQUES AND THEIR ANALYSIS USING HYBRID DNA DIFFERENTIAL METHYLATION ANALYSIS FOR PREDICTING BREAST CANCER

[1]*Remyamol K M, [2] Philip Samuel
[1]Department of Information Technology, School of Engineering,CUSAT,Kerala, India
E-mail:*rems84@gmail.com
[2]Department of Computer Science,CUSAT,Kerala, India. philipcusat@gmail.com

Abstract
Breast cancer is a leading cause of death for women in many parts of the world. The disease is often misdiagnosed until it has progressed beyond the point of effective treatment. Therefore, early detection of the disease would aid in reducing mortality and other associated risks. Microarray gene expression data is difficult to identify and interpret, making it challenging to evaluate and choose the most relevant set of genes for use as breast cancer markers. Our research utilized Matlab 2018a and the Breast Cancer Wisconsin Diagnostic dataset to develop a mixed machine-learning model for fast breast cancer prediction. Here, we apply various machine learning techniques, including random forest classification, logistic regression, Support Vector Machine, and Naive Bayes, as well as to a dataset to make predictions about their development and eventual size. The success percentage for the eXtreme Gradient Boosting classifier comparing to various machine learning approaches is 99.78%. This new approach to prediction has the potential to revolutionize the detection, analysis, and prognosis of breast cancer.

Keywords: Logistic regression, Breast cancer, Naïve Bayes, random forest, Support Vector Machine, XGBoost.

Introduction

Breast cancer (BC) has become the most common malignancy in females worldwide and the second leading cause of death in both developing and industrialized nations.Patients with cancer have a lower risk of dying from the disease if they receive an early diagnosis, which may be achieved through the use of comprehensive screening programs [1]. Also, the efficiency of cancer diagnosis and treatment could be greatly enhanced by a deeper comprehension of the disease's pathophysiology and underlying mechanism. Most cases of breast cancer are diagnosed as ductal carcinoma in situ (DCIS). Mutations in the DNA or RNA of a cell can cause it to transform from normal to cancerous. Free radicals, DNA/RNA aging, entropy, nuclear radiation, fungus spores, parasites, and increased atmospheric chemical levels can all cause mutations [2]. Normal cells can mutate into cancerous ones. As soon as feasible, and as efficiently as possible, tumors need to be treated. An inaccurate diagnosis of a malignant tumor in its early stages might have devastating consequences for the patient's health.

Most individuals have tumors that aren't particularly large, but they'll still seem strange on a

mammogram [3]. The rapid development of a breast lump, when none existed before, is the most common sign of cancer. However, it is not correct to say that every new lump found has the potential to develop into cancer. Multiple subtypes of BC have emerged, each characterized by a unique set of signs and symptoms. Patients who experience breast pain or a lump may have a benign cyst. However, testing and research can be used to discount BC as a possibility [4].

There are two broad types of BC, and the phrases "invasive" and "non-invasive" refer to them, respectively [5]. Breast cancer that has not spread beyond the original tissue is called non-invasive breast cancer or BC in situ. However, invasive cancer begins in the ducts or glands of the breast and then spreads throughout the body. The aforementioned groups include the following BC subtypes: DCIS: In the so-called DCIS scenario, there is no invasion. DCIS occurs when breast cancer cells are confined to the milk ducts and have not spread to the surrounding tissue.

(ii) Lobular carcinoma in situ (LCIS) is a subtype of DCIS that develops in the milk-producing ducts of the breast. The cancer cells have not spread into neighboring tissue like they do in DCIS. The most common kind of breast cancer is invasive ductal carcinoma or IDC. Breast cancer develops in the milk ducts, moves to nearby tissues, and eventually metastasizes to other organs. Breast cancer that begins in the breast's lobules and then metastasizes is known as invasive lobular carcinoma (ILC).

If the patient's cancer stages can be determined in advance, treatment can be optimized [4]. Methods grounded on machine learning (ML) have been used extensively to discover how changes in genes and regulatory areas translate to observable traits [5]. Changes in personality, health, and well-being are only a few examples. Enhancers, promoters, and gene sequence levels are just a few examples of genomic components where DL-based approaches are being used to make predictions about their structure and function [6, 7]. Conventional machine learning methods, such as Decision Trees and Support Vector Machines, were often utilized in early computational approaches to studying gene expression. Feature engineering is a critical part of computational approaches to studying gene expression. Challenges like the high dimensionality of gene expression data and a relatively limited number of samples are overcome by this technique. To determine the extent to which a given property of the data is correlated with the target anticipated variable, filtering methods rely on quantitative analysis. Wrapper methods employ a classification algorithm to evaluate the significance of the data attributes in question. Following this sorting, a search algorithm is used to zero in on the most relevant characteristics of the assessed data. Feature engineering is incorporated into the classifier's learning process via embedded methods [7]. This allows us to pinpoint the specific features that boost the efficiency of a classification system. In most cases, the processing time and computational complexity of filter algorithms are very low. However, the efficiency of the applicable classification algorithm is typically improved through the use of wrapper and embedding strategies [8]. However, more development is needed to make the techniques generalizable and long-lasting, as the success of the current approaches depends on a variety of factors. Another drawback of current methods is that they are complicated to comprehend and can only be partially integrated with a wide variety of other data types and processes. Although doctors have shown a preference for more traditional approaches, it may be to their benefit to perform variable identification using ML algorithms.The purpose of this study is to

use ML models to foresee the critical analytic aspects that impact BC patients' survival rates.

Proposed Methodology

Breast cancer is an epidemic that has a global reach.It starts when the growth of breast cells is no longer regulated. Cancers often originate in these cells. Finding and identifying the tumor requires either an X-ray or a physical examination of the breast for abnormalities. Differentiating between benign and cancerous tumors, often known as malignant and benign, can be challenging. The data is collected from Breast Cancer Wisconsin Diagnostic (WDBC) dataset. It is difficult to utilize the data in the dataset as input since it is in a format that the computer cannot read directly.

The input is obtained in through a median filter using local window size $w*w$. This is described as (1).

$$U\ (i, j) = H\ [x, y]\ *V1\ w*w\ [x, y] \qquad (1)$$

where$U(i, j)$ is the convolved dataset, $V1\ [x, y]$ is the result obtained from input section and the convolution mask $H[x, y]$ is a local median filter. The difference data(x, y) is then computed. This is described as (2).

$$D\ (x, y) = U\ (i, j) - V1\ [x, y] \qquad (2)$$

The segmented data$S$data is obtained as (3).

$$S\text{data}\ (x, y) = \{0, \text{if}\ D\ (x, y) \leq T\ (x, y)\ ,$$

1, otherwise (3)

where$T(x, y)$ = K

A new dataset is generated by preliminary data cleaning and examination following an analysis of the existing datasets. After the data have been cleaned and organized, they are imported, and a procedure is used to identify any missing values that pertain to a specified feature. The algorithms have never been able to discover a reasonably near match, even when a value is absent. It is possible to construct a biased machine learning model if missing values are not handled correctly, which might lead to inaccurate findings. After that, segmentation methods based on an adjustable threshold are implemented. Careful outlier detection and management is an essential part of any data processing procedure.

Algorithm: Adaptive Threshold

Step 1: Input gene dataset
Step 2: Compute neighborhood size
Step 3: Compute the threshold using local mean intensity around the neighborhood of the pixel
Step 4: If the pixel value is below the threshold set the pixel to background else foreground
Step 5: Clear the border
Step 6: Remove small objects of fewer than 100 pixels
Step 7: Fill holes

The distance from the information has been used to set a beginning threshold value for outlier detection. An outlier is a data point overall a hybrid algorithmic learning framework that is significantly different in magnitude from the surrounding data points.Then, for all points in the experimental facts, we measured how far they were from the average of every cluster while

contrasting those numbers to the overall average. If the amount exceeds the predetermined limit, it is considered an anomaly. Figure 1 depicts the recommended methodological approach.

Figure 1: Schematic representation of the proposed methodology

Grayscale gene datasets are used as input for adaptive thresholding, with the resulting binary picture representing the categorization. The threshold for each gene dataset pixel is determined using a locally adaptable method. Each pixel's threshold is calculated using the local mean intensity of its neighboring pixels and a sensitivity factor that is defined by sensitivity. Scalar value (between 0 and 1) that characterizes how sensitive the thresholding method is to freshly identified foreground pixels. If the pixel score was lower than the threshold, it was assumed that the foreground was more important; nonetheless, the background value was still calculated. In the steps that followed, the dataset was put through a variety of different classification techniques, such as logistic regression (LR), random forest (RF), naive Bayes (NB), support vector machine (SVM) and extreme gradient boosting (XGBoost). Group learning using decision trees is the basis of the XGBoost algorithm. Prediction issues with unstructured data (pictures, text, etc.) are tackled using a gradient-boosting architecture. Artificial neural networks produce far more effective results than conventional machine learning methods. However, if the data is tabular or similarly organized, decision tree-based algorithms will perform better than XGBoost. XGBoost and gradient boosting machines (GBMs) are examples of ensemble learning systems that use the gradient descent technique and hence will apply the rule of boosting weak learners, such as CARTs. However, XGBoost enhances the functionality of the GBM framework by optimizing the underlying systems and introducing new algorithmic improvements. XGBoost excels at problems that need supervised learning, such as predicting the value of a target variable from a given dataset.

Results and Discussion

The characteristics acquired with the adaptive threshold method are trained and evaluated using five different supervised classification approaches: NB, LR, SVM, RF, and XGBoost. Additionally, classification models are trained and evaluated using both the training dataset and the test dataset using a technique known as stratified five-fold cross-validation. By evaluating the classification model that best distinguishes the target variable labels using the subset of gene biomarkers that is most beneficial overall, this work aims to aid in the early diagnosis of BC patients. Researchers will be able to find patients much more rapidly if they do this.

Table1: Performance evaluation values

Sl. no	Performance analysis	NB	SVM	LR	RF	XGBoost
1	Accuracy	86.3184	89.5436	92.6586	97.3711	99.7797

2	Error	5.6816	4.9959	4.5734	4.0217	3.7453
3	Sensitivity	99.9999	99.8731	99.4643	99.2754	98.3871
4	Specificity	34.2105	40.4853	46.2795	57.4211	68.4211
5	Precision	95.2015	95.8749	96.5876	97.2781	99.6751
6	False positive rate	65.7895	62.5446	57.8759	54.7568	31.5789
7	F1_score	95.5418	96.6213	97.6587	97.922	99.699
8	Mathews Correlation Coefficient	57.0692	60.0765	63.0749	67.4562	70.3457
9	Kappa	49.1351	54.2531	61.7586	69.2435	72.5688

There are a few different ways to describe how effectively we can predict outcomes: sensitivity (SE), accuracy (AC), and specificity (SP). The ability of a diagnostic test to positively identify the presence of a disease is measured by its sensitivity. A test's specificity is measured by how effectively it rules out a certain condition. It is possible to evaluate the accuracy of the categorization by counting the number of correctly labeled data. Nine performance metrics are used to systematically evaluate the classifying efficacy of the predicted method values of adaptive threshold segmentation using different classifiers, as shown in Figure 2. We calculate the F1 Score, the Matthews Correlation Coefficient, the error rate, the specificity rate, the sensitivity rate, the false positive rate, and the kappa coefficient. The goal of these evaluations is to find the classifiers that produce the most reliable results.

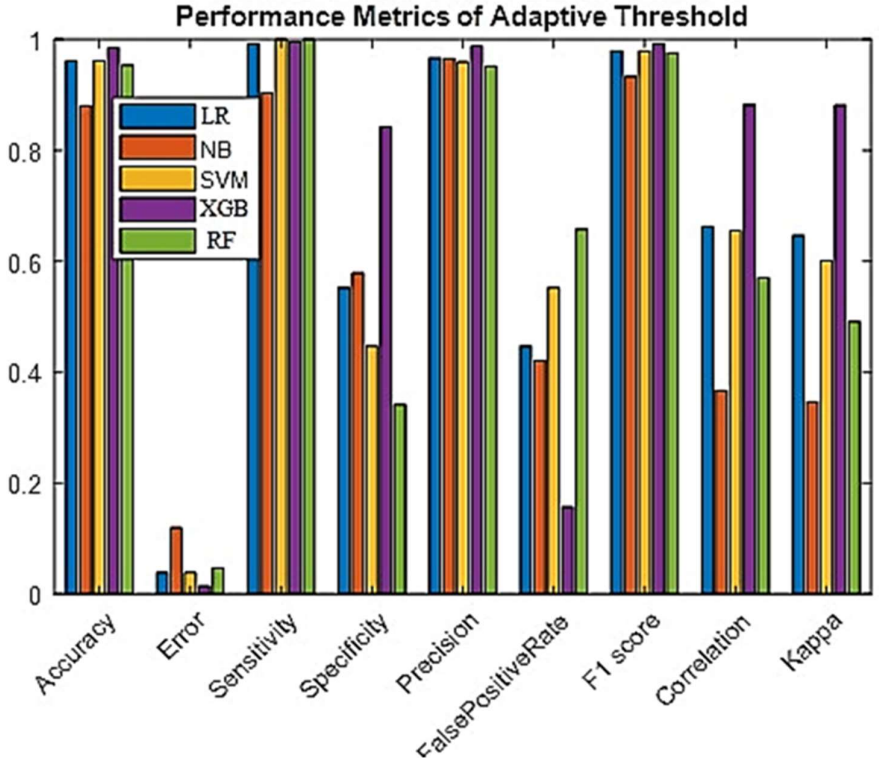

Figure 2: Performance analysis of the proposed methodology

The histogram in Figure 3 compares two sets of data (normal breast and primary breast cancer samples) to highlight the distribution of the most significant gene biomarkers found using the adaptive threshold approach. Both healthy and breasts with primary breast cancer were used for the samples. No breast tissue samples were identified as having originated from women who had been diagnosed with breast cancer. All of the women in the samples were in good health. Newly found gene biomarkers such as NM_152426, NM_138957, and NM_001008493 outperform previously used features in properly differentiating primary breast cancer samples from normal breast samples. This is because a new set of gene biomarkers, including NM_152426, NM_138957, and NM_001008493, has been discovered. Other genes in this group include NM_001008493. In addition, the XGBoost-based model that was constructed with the optimal feature subset obtained by the adaptive threshold technique had the best performance. An earlier diagnosis of breast cancer (BC) is possible if the primary breast cancer can be distinguished from normal breast samples. The primary breast cancer detector uses the levels of expression of APOBEC3B, MAPK-1, and ENAH actin regulator (ENAH) to distinguish between breast samples that include primary breast cancers and breast samples that contain normal breast tissue. This is done by comparing the levels of expression of ENAH. Because of this, our breast tumor predictor might potentially be utilized by medical professionals to diagnose the disease at an earlier stage.

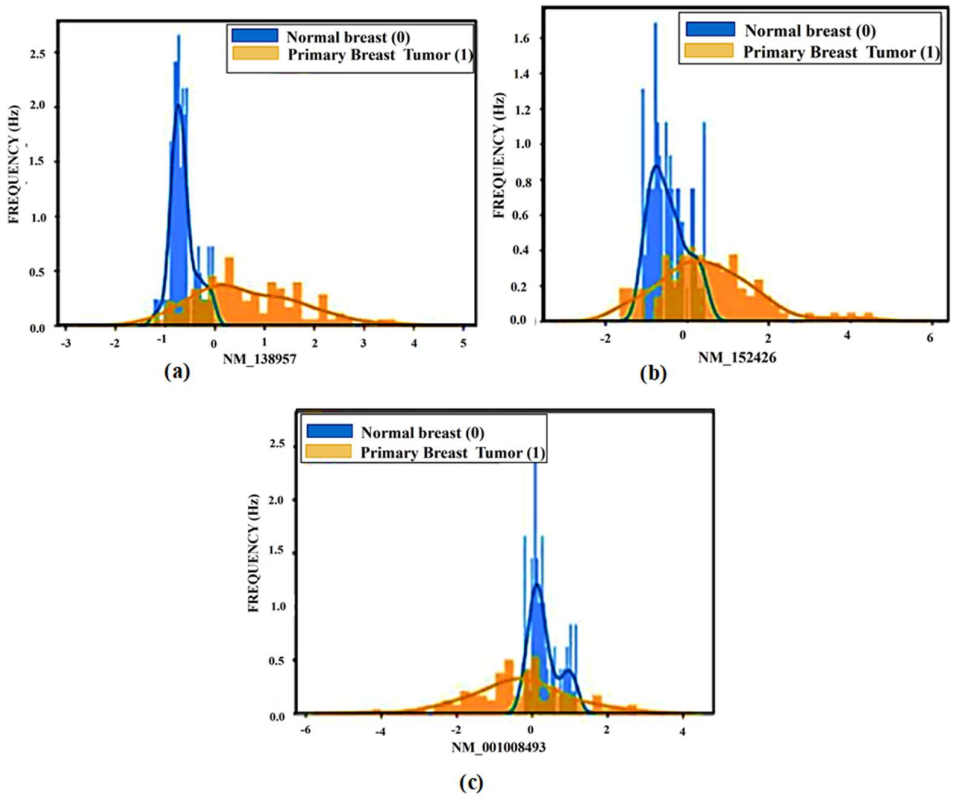

Figure 3: The histogram-based frequency distribution of the three best gene biomarkers is depicted visually

It has been difficult and time consuming to find reliable gene biomarkers for the early detection of breast cancer due to the small sample size and multi-dimensionality qualities of the data from gene microarrays. In this research, a hybrid learning in feature selection framework was used to look at possible gene biomarkers. In this study, the ability of gene biomarkers to consistently diagnose early-stage breast cancer utilizing the hybrid-based FS pipeline was evaluated. The results of the analysis are shown in the table below. Previous research has linked these gene biomarkers to breast cancer development. The results of the XGBoost-based classification model applied to independent test data are displayed in a confusion matrix (Figure 4). Primary breast cancer represents the "positive" class in the matrix, whereas a sample of normal breast tissue represents the "negative" class. Using this matrix has additional benefits.

The accuracy is the proportions of successfully predicted pixels [5]. It is expressed in the equation below.

$$Acc = (TP + TN)/ (TP + FP + TN + FN)$$

Sensitivity is the percentage of nodule variables that are clearly predicted, and precision is the percentage of input images that are clearly predicted that are measured below.

$$Se = TP/(TP + FN)$$

$$Sp = TN/ (TN + FP)$$

FPR is the proportion of falsely described as nodule pixels as well as the fraction of wrongly described as pixel values that are described below seems to be the false negative ratio (FNR) [2].

$$FPR = FP/(TP + TN)$$

$$FNR = FN /(TP + TN)$$

Overlapping score is a similarity measurements, which reflects how the subdivision outcome of the principles matches the ground truth.

$$Overlap = TP/(TP + FP + FN)$$

where true positive = exactly found number as nodule pixels. False positive = incorrect found number as nodule pixels. True positive negative = number of exact identification as background pixels. False Negative = number of incorrect identification as background . The values of 5 calculation measures ranging from 0 to 1. The lesser FPR and FNR the good is segmentation performance.

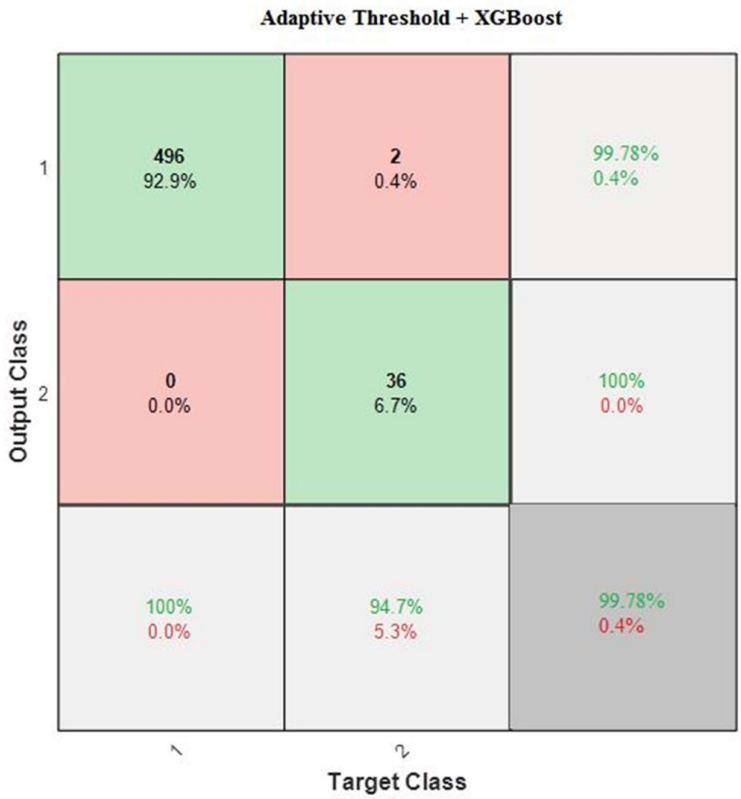

Figure 4: Confusion matrix

Ontology-based research has uncovered MAPK1's involvement in the mitogen-activated protein kinase (MAPK) intracellular signaling pathway. More so than other intracellular signaling systems, the MAPK pathway promotes cell proliferation, differentiation & survival, angiogenesis, as well as tumor spread in breast cancer. This is due to the intracellular location of the MAPK pathway. To determine the most effective feature collection of predictors such as LR, NB, XGBoost, SVM, and RF that differentiate primary breast tumors from healthy breast expressed genes microarray statistics, several supervised approaches to classification are compared and contrasted.These formulas are employed in the feature subset screening process. This method is an exhaustive study of a hybrid approach to the early diagnosis of primary breast cancers. This strategy employs a filter-based mechanism for successively selecting gene features. When compared to other hybrid feature selection models, our proposed XGBoost classification model achieved the highest total accuracy (99.78%).

Conclusion

Modern feature selection algorithms might be used to categorize primary breast tumors, which would aid in the search for possible gene biomarkers. Breast cancer survival rates can only be improved by the adoption of more effective treatment methods if the disease is detected at an early stage while it is still curable. Combining the APOBEC3B, MAPK1, and ENAH gene biomarkers is a reliable and sufficient method for detecting early breast cancer in patients. These considerations all contribute to the current limitations of the study. Due to the binary nature of the dataset utilized in this study, the outcomes might differ if the same methods were applied to a dataset with more than two classifications. Furthermore, the proposed framework only makes limited use of the entire training set. Because of this, we'll need to test the

robustness of the novel hybrid feature selection method on a bigger training sample size in the future. We need a bigger training sample size to do this. This methodology employs metaheuristics, statistical methods, and a filter-based strategy to identify actionable genetic traits for early breast cancer detection.

References

1. Barret, J. E., Herzog, C., Jones, A., Leavy, O. C., Evans, I., Knapp, S., ...&Widschwendter, M. (2022). The WID-BC-index identifies women with primary poor prognostic breast cancer based on DNA methylation in cervical samples. Nature Communications, 13(1), 449.

2. Hsu, P. C., Kadlubar, S. A., Siegel, E. R., Rogers, L. J., Todorova, V. K., Su, L. J., &Makhoul, I. (2020). Genome-wide DNA methylation signatures to predict pathologic complete response from combined neoadjuvant chemotherapy with bevacizumab in breast cancer. PLoS One, 15(4), e0230248.

3. de Almeida, B. P., Apolónio, J. D., Binnie, A., &Castelo-Branco, P. (2019). Roadmap of DNA methylation in breast cancer identifies novel prognostic biomarkers. BMC cancer, 19(1), 1-12.

4. Orozco, J. I., Le, J., Ensenyat-Mendez, M., Baker, J. L., Weidhaas, J., Klomhaus, A., ...&DiNome, M. L. (2022). Machine learning-based epigenetic classifiers for axillary staging of patients with ER-positive early-stage breast cancer. Annals of Surgical Oncology, 29(10), 6407-6414.

5. Ensenyat-Mendez, M., Rünger, D., Orozco, J. I., Le, J., Baker, J. L., Weidhaas, J., ...&DiNome, M. L. (2022). Epigenetic signatures predict pathologic nodal stage in breast cancer patients with estrogen receptor-positive, clinically node-positive disease. Annals of surgical oncology, 29(8), 4716-4724.

6. Panagopoulou, M., Karaglani, M., Manolopoulos, V. G., Iliopoulos, I., Tsamardinos, I., &Chatzaki, E. (2021). Deciphering the methylation landscape in breast cancer: diagnostic and prognostic biosignatures through automated machine learning. Cancers, 13(7), 1677.

7. Gupta, S. (2022, January). Receptor Status Prediction in Breast Cancer Patients Using Machine Learning Pipeline on DNA Methylation Data. In 2022 12th International Conference on Bioscience, Biochemistry and Bioinformatics (pp. 38-43).

8. Karaglani, M., Panagopoulou, M., Baltsavia, I., Apalaki, P., Theodosiou, T., Iliopoulos, I., ...&Chatzaki, E. (2022). Tissue-Specific Methylation Biosignatures for Monitoring Diseases: An In Silico Approach. International Journal of Molecular Sciences, 23(6), 2959.