Machine Learning Approaches to Better Utilize HM450K Data in Cancer Analysis

Sam Coleman and Jacob Tye

Methylation dysregulation is a hallmark of many cancers, offering critical insights into tumor biology and progression. Leveraging methylation data for tasks such as tumor classification and subtype identification has been an active area of research. In this study, we attempt to address these tasks through machine learning techniques utilizing data from TCGA and cBioPortal, comprising 2,294 samples from four cancer types (BRCA: 769, HNSC: 523, THCA: 499, PRAD: 498) with 238,284 CpG sites. We implemented machine learning pipelines to address two fundamental problems: (1) identifying tumor state and (2) classifying breast cancer subtypes. For tumor state classification, we evaluated two pipelines: one using Locally Linear Embedding with Logistic Regression and a second using Principal Component Analysis with Gradient Boosting model. For breast cancer subtype classification, we implemented a pipeline using K-Means unsupervised clustering and another using Principal Component Analysis into a supervised neural network. Each approach demonstrated varying levels of success, providing insights into the utility of different methods for methylation-based cancer analysis. Our study illustrates how machine-learning pipelines can effectively utilize high-dimensional methylation data such as HM450K to classify tumor state and identify cancer subtypes. By showcasing the strengths and limitations of different approaches, our work provides a roadmap for the future integration of methylation-based biomarkers into clinical practice.

Introduction

DNA methylation is an essential epigenetic mechanism involving adding a methyl group (CH_3) onto the cytosine in a cytosine guanine pair (CpG). This is biologically relevant for several reasons, including improving genomic integrity and stability, aiding gene expression regulation, and imprinting genes[1]. Because of these critical mechanisms for methylation, it is usually highly regulated through proteins such as DMNT1 and DMNT3A/B.

However, a hallmark of cancers is that methylation becomes dysregulated, evidenced by a global loss of methylation[2,3]. Not only is it widespread, but it is one of the earliest detectible aberrations of cancer[4]. Methylation is also an attractive target for research because it is reversible and could be targeted for therapeutic treatments[3]. These facts have led many researchers to harness methylation as a biomarker to understand cancer better.

One common approach for this is to compare primary tumor methylation levels with that of adjacent normal tissues and identify differentially methylated loci (DML)[5–13]. Each of these studies has identified thousands of significant DMLs. Yet a recent systematic review of 20 studies found no overlap in differentially methylated CpGs across studies, highlighting reproducibility challenges in methylation analysis[14].

One reason for this could be due to the high dimensionality of the data. Whole Genome Bisulfite Sequencing (WGBS) data shows about 28 million CpGs[15]. Alternative methods for analyzing CpGs attempt to reduce this by focusing on CpGs thought to be necessary, such as Illumina Human Methylation 450 (HM450K) and Illumina Human Methylation 27 (HM27), which have data for 450,000 and 27,000 CpGs, respectively. This high dimensionality of methylation leads to false positives and multiple testing errors, making studying and utilizing methylation difficult.

Machine learning has been proposed as a potential solution to this problem as it can handle high-dimensional data and utilize feature reduction to identify CpGs of interest better. For this project, we attempted to highlight this by addressing two questions: (1) can machine learning be used to predict cancer state in samples given only methylation data, and (2) can we identify cancer subtypes in cancers using unsupervised learning methods? We used a variety of different machine learning methods and have highlighted the benefits and drawbacks of each.

Background

Using machine learning with methylation data is an active area of research, with new models being developed constantly. One of the latest promising models is EMethylNET, a machine learning model that can predict tumor state for 13 cancers with a 98.4% accuracy using only methylation data[4]. As input, they utilized HM450K methylation data taken from the Cancer Genome Atlas (TCGA) and XGBoost, which fed into a deep neural network. After filtering, they used 276,016 CpGs as input, and the model filtered this further to 3388. The authors utilized SHAP to make their model explainable and identify key CpGs.

Methylation is also being explored for subtype classification of various cancer types. A recent study on breast cancer utilized DNA methylation data to classify distinct subtypes and cellular lineages of the tumor. In this study, Fleischer et al. employed a Hidden Markov Model (ChromHMM) combined with methylation and gene expression data to identify unique clusters of breast cancer patients, each associated with distinct clinical

phenotypes not well-understood before this analysis[16]. This work demonstrates how DNA methylation profiling can provide crucial insights into tumor heterogeneity, contributing to improved characterization and potentially guiding treatment strategies.

Another example of machine learning being applied to methylation data is demonstrated in a study by Singh et al., which utilized unsupervised learning to identify unique molecular subtypes of colorectal cancer. The researchers employed hierarchical clustering methods to analyze methylation and mutation data from colorectal cancer samples. Their analysis revealed three distinct molecular clusters characterized by unique genetic and epigenetic alterations. These findings highlighted the potential of machine learning approaches to stratify cancers based on molecular features, which could enhance personalized medicine efforts and improve prognostic and therapeutic strategies[17]. Building on these prior studies, our work aims to explore the utility of machine learning for methylation-based cancer analysis in two distinct contexts: predicting tumor states and classifying breast cancer subtypes. To further examine the use of methylation data, we implemented and evaluated multiple machine-learning pipelines using large-scale methylation datasets (**Supplemental Table 1**).

Methods

Data Preparation

We obtained publicly available data from cBioportal and the Cancer Genome Atlas (TCGA) to begin our analysis [18–31]. From cBioportal, we used datasets for BRCA, HNSC, THCA, and PRAD from TCGA Pan-Cancer Atlas. This data included clinical data, such as cancer subtype, survival rates, and methylation data. This methylation data, however, was HM450K data that had been filtered down to HM27 CpGs. Because we were interested in utilizing HM450K data, which contains more CpGs and biological information, we used TCGA's API to download the 450K data for all patients from cBioportal. We obtained 189 additional supplemental solid tissue normal samples from TCGA to further augment the dataset. This data was then filtered to only include CpGs that were common between the cBioPortal and supplemental Normal samples. We further filtered the CpGs to only those CpGs that had data in 80% or more of the samples. In total, we obtained 2729 samples split as follows: BRCA: 769, HNSC: 523, THCA: 499, PRAD: 498 (**Figure 1A**).

The first aim of this project was to test and develop possible machine-learning pipelines that can distinguish tumor states given only methylation data. For these aims, we used the entire dataset, which included 2289 tumor samples and 440 normal tissue samples (**Figure 1B**). Following filtering the CpGs for the datasets, we left data for 238,284 CpGs.

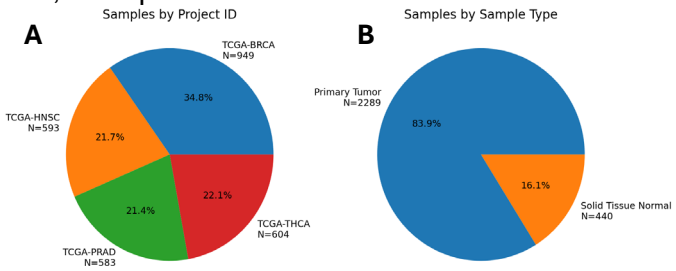


Figure 1 Pie charts showing the split of the datasets used for this project. (A) Dataset split by project id. (B) Dataset split by sample type.

The next aim of this project is to identify cancer subtypes using machine learning, focusing specifically on breast cancer (**Figure 2**). This choice is motivated by multiple well-established subtypes, including Luminal A, Luminal B, HER2-enriched, Triple-Negative Breast Cancer (TNBC, also referred to as Basal-like), and Normal-like. Notably, TNBC is recognized as a heterogeneous group, with evidence suggesting the existence of clinically significant subclassifications[8]. These characteristics make breast cancer an excellent candidate for machine learning classification, both supervised and unsupervised clustering, as established subtypes can serve as benchmarks for validation. At the same time, novel subtypes with potential clinical relevance may also be discovered.

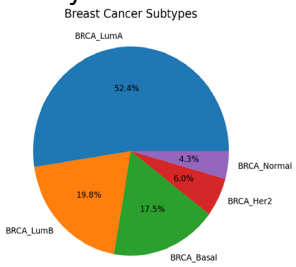


Figure 2 Pie chart depicting the split of the breast cancer subtypes in the filtered dataset: Luminal A, Luminal B, HER2, Normal-like, and Basal-like (TNBC).

Problem 1: Predicting Tumor State

The first problem we addressed was predicting tumor state using only methylome data. To approach this, we developed two distinct machine learning pipelines using Scikit-learn, both of which incorporated dimensionality reduction techniques to address the high dimensionality of the dataset. These reduced-dimensional features were subsequently input into supervised classification models. The data was partitioned into 80% for

training and 20% for testing in each pipeline to ensure robust evaluation.

Solution 1.1: Locally Linear Embedding with Logistical Regression

The first pipeline we developed employed Locally Linear Embedding (LLE) for dimensionality reduction and Logistic Regression for classification. LLE was selected for its ability to preserve local relationships, which aligns well with the tendency of neighboring CpG sites to exhibit high correlation[32,33]. This pipeline took in the raw data of the CpGs and used a simple imputer using the mean method for any missing data. This imputed data was then fed into the LLE layer and reduced into 1500 locally linear embeddings, which were then used by the logistical regressor to predict whether the sample was tumor or healthy tissue (**Figure 3**). For this model, we utilized Optuna with 5-fold cross-validation tune a variety of hyperparameters *(**Supplemental Table 2**)*.

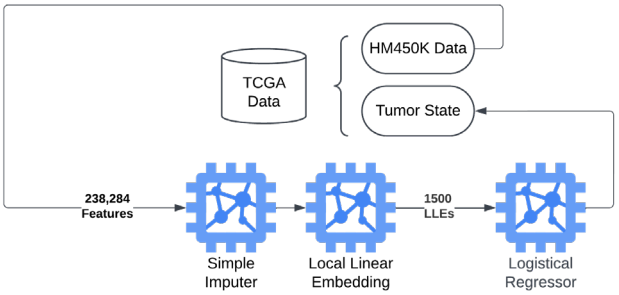


Figure 3 Diagram depicting the first machine learning pipeline developed for identifying tumor vs normal tissue. The model took in the raw data from the 238,284 CpGs, imputed any missing data using the mean method. This data was then further reduced to 1500 locally linear embeddings. This was finally used by the Logistical Regressor to determine tumor state of the sample.

Solution 1.2: Principal Component Analysis with Gradient Boosting

The second pipeline developed for predicting tumor state focused on gradient boosting. First, CpG sites that had missing data in 50% or more samples were removed from the dataset. Any missing data in the remaining features was imputed through the median strategy in Sklearn's SimpleImputer. Next, a principal component analysis (PCA) was applied to the training dataset, and the top 1000 components were retained. The remaining components were then used as feature input into a gradient-boosting classifier. This classifier then predicted whether a sample belonged to the normal or the tumor category (**Figure 4**). Optuna was used in addition to 5-fold cross-validation to conduct a randomized search for hyperparameters (**Supplemental Table 2**). The model was then assessed using multiple statistical descriptors and a SHAP analysis for explainability.

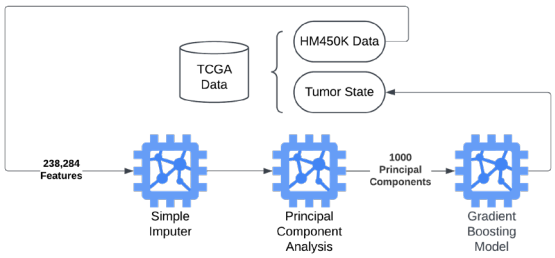


Figure 4 Diagram depicting the machine learning pipeline for **Solution 1.2** developed for identifying tumor vs normal tissue. The model took in the raw data from the 238,284 CpGs, imputed any missing data using the median method. This data was then further reduced to 1000 principal components. This was finally used by the Gradient Booting model to determine tumor state of the sample.

Problem 2: Identification of Breast Cancer Subtypes

The next problem we sought to address was identifying breast cancer subtypes using machine learning. Breast cancer was selected for this analysis due to its well-established subtypes, including Triple-Negative Breast Cancer (TNBC), which is highly heterogeneous—a characteristic with significant clinical implications for prognosis and treatment strategies[8]. To explore this, we applied supervised and unsupervised machine learning techniques to identify meaningful patient clusters based on methylation data. For unsupervised clustering, known breast cancer subtypes provided a benchmark to evaluate the effectiveness of the clustering. At the same time, discovering new clusters—particularly subclusters within TNBC—could offer valuable insights into its heterogeneity and potential clinical applications. Demonstrating that methylation data can effectively stratify patients into existing subtypes or reveal novel groupings would highlight its potential as a powerful tool for advancing precision medicine.

Solution 2.1: K-means Clustering to Identify Subtypes

2.1.1: K-means clustering analysis

The first pipeline we developed for identifying breast cancer subtypes utilized a K-means unsupervised clustering model. For this model, we provided raw methylation data to Sklearn's SimpleImputer using the mean strategy to impute any missing data. This imputed data was then provided to the K-means clustering model, which was told to identify 5 clusters, one for each of the known subtype labels in this dataset. This was done to see if the K-means clustering model would naturally segregate by known subtypes. The normalized mutual info score was calculated to quantify how well the K-means clustering aligned with known subtypes.

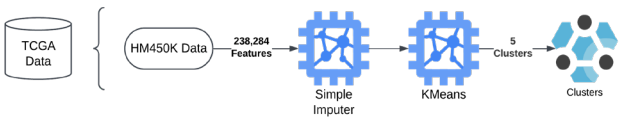


Figure 5 *Diagram depicting the first machine learning pipeline for identifying breast cancer subtypes. Methylome data for 238,284 CpGs was fed into a simple imputer which imputed missing data using a mean strategy. This data was then fed into a K-means clustering model to identify 5 clusters.*

2.1.2: Kernal Principal Component Analysis for Visualization of Clusters

To display the results of this clustering, dimensionality reduction was performed by developing another machine-learning pipeline that imputed the raw methylation data in the same way outlined in **2.1.1**. This data was then passed into a Kernal Principal Component (KPCA) model to identify three principal components that could be used in plotting. We chose to use KPCA because of its ability to find principal components with nonlinear data.

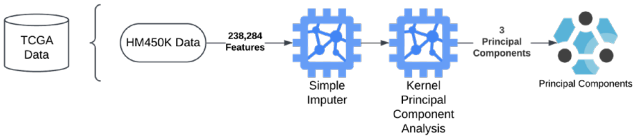


Figure 6 *Diagram depicting the machine learning pipeline utilized for visualization of the k-means clusters. Methylome data was fed into a simple imputer that imputed missing data using a means strategy. This data was then used by a Kernel Principal Component model to identify 3 principal components.*

2.2: Principal Component Analysis with Neural Network Classification

The second pipeline developed for classifying breast cancer subtypes was built around a neural network implemented in PyTorch. This supervised approach was initiated with CpG sites missing in 50% or more samples being removed. Training data was passed through the pipeline, and any missing information was imputed through the median strategy of Sklearn's SimpleImputer. Optuna, combined with 5-fold cross-validation, was utilized to search for effective hyperparameters to modify the downstream pipeline (**Supplemental Table 2**). Following imputation, the pipeline consisted of a PCA dimensionality reduction of input CpG features which was passed into a PyTorch neural network. The resulting model was then assessed for accuracy using multiple statistical descriptors, while SHAP was used for explainability. Finally, the normalized mutual information score was calculated for comparison with other models.

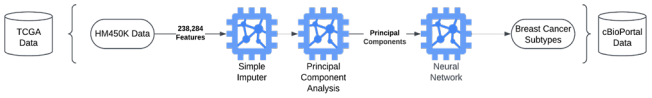


Figure 7 *Diagram depicting the machine learning pipeline utilized for **Solution 2.2**. Methylome data was fed into a simple imputer that imputed missing data using a median strategy. This data was then used in Principal Component analysis to identify principal components. This data was then used by a Neural Network for supervised learning to identify the breast cancer subtypes from the cBioPortal data.*

Results

Problem 1: Tumor State Identification Results

Solution 1.1: LLE with Logistical Regression Results

Our first pipeline, which utilized an LLE to perform dimensionality reduction, was able to predict tumor state with an accuracy of 97%. When we plotted the receiver operating characteristic curve (ROC), with the tumor being defined as positive and healthy being negative, we found the area under the curve (AUC) to be 98.814% (**Figure 8B**). The average precision (AP) obtained from the precision-recall curve was also high at 99.773% (**Figure 8C**).

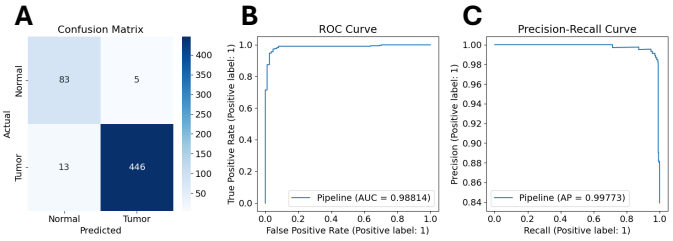


Figure 8 *Performance plots for **Solution 1.1**. (A) Confusion matrix of the model using the testing dataset. (B) ROC curve of pipeline with tumor defined as positive. (C) Precision-Recall Curve of the pipeline with tumor again being defined as positive.*

We next performed SHapley Additive exPlanations (SHAP) value analysis on the pipeline. Due to memory constraints, which we will discuss later in this paper, we were only able to calculate the importance of the locally linear embeddings and not the original CpG features (**Supplemental Figure 1**).

Solution 1.2 Principal Component Analysis with Gradient Boosting

The second pipeline we developed utilized a combination of PCA dimensionality reduction with gradient boosting to classify samples as either tumor or normal. We found this to be another robust classifier, with an overall model accuracy of 97.62%, AUC 99.30%, and AP 99.86%. After calculating the overall performance metrics, a SHAP value analysis was performed on the model. Due to the conversion of features into principal

components in the pipeline, SHAP cannot directly infer the values of individual CpGs. Instead, we utilized SHAP to identify important principal components in decision-making. We then used the loading values for these important principal components to determine what CpG sites were essential in contributing to that component (**Supplemental Figures 2 and 3**).

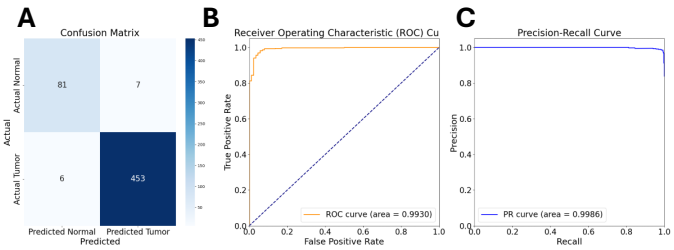


***Figure** 9: The performance plots for **Solution 1.2**. (A) Performance matrix of the model using the test dataset. (B) The ROC curve of the pipeline with the tumor is defined as positive. (C) The precision-recall curve of the pipeline with the tumor is defined as positive.*

Problem 2: Identification of Breast Cancer Subtypes

The identification of breast cancer subtypes through unsupervised clustering presented unique challenges due to the heterogeneity of the disease. We aimed to determine whether methylation data could effectively stratify patients into meaningful clusters that align with established subtypes. This analysis provides insight into the potential of methylation data for subtype identification and explores whether novel clusters may reveal additional clinical or biological significance.

Solution 2.1: K-means Clustering Results

Our first pipeline for identifying breast cancer obtained mixed results. Using K-means clustering, we identified five distinct clusters (**Figure 10A**) However, when we compared these clusters to known breast cancer subtypes, we found no clear overlap (**Figure 10B**, **Supplemental Figures 4** and **5**).

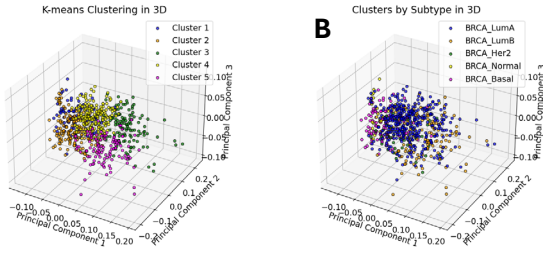


***Figure 10** Scatter plot using KPCA components for the plotting of the results of **Solution 2.1**. (A) Scatter plot of Clusters identified by K-Means Clustering. (B) Scatterplot of known breast cancer subtypes.*

This observation was further confirmed by a normalized mutual information (NMI) score of 0.2902, indicating that while K-means could detect distinct patterns within the methylation data, these clusters did not align with established subtype classifications.

Solution 2.2 Principal Component Analysis with Neural Network Classification

Our second pipeline was centered on using a neural network to classify breast cancer samples into their subtype based on methylation data. While our model obtained an overall accuracy of 81%, the performance by class varied, with the Basal class having the highest F1 score at 0.96 while the Normal class had an F1 score of 0.4. We calculated that the NMI score of this model was 0.58. As with **Solution 1.2**, a SHAP analysis was conducted where the informative principal components were converted into important CpG sites based on the loading values (**Supplemental Figure 6**).

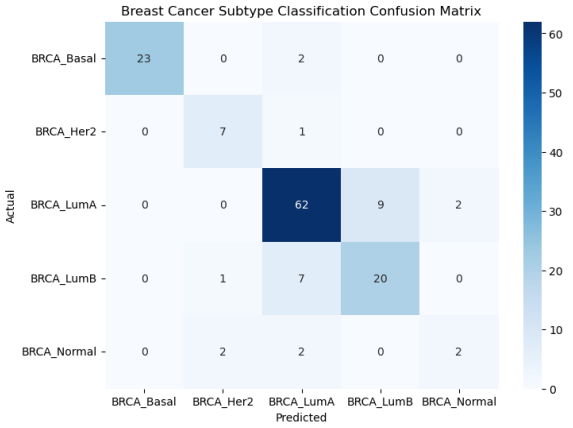


***Figure** 11: A confusion matrix displaying the performance of **Solution 2.2** in classifying breast cancer samples based on subtype.*

Discussion

All pipelines utilizing methylome data to predict tumor state demonstrated exceptionally high accuracy (Accuracy > 0.97). These findings suggest that there is great potential for methylation data as a candidate for early cancer detection, as it represents one of the earliest molecular aberrations in cancer development. Notably, the model's ability to achieve high accuracy across multiple cancer types highlights its potential for broad applicability in early screening. This raises the exciting possibility of employing methylation-based methods, such as blood-derived methylation data, to screen multiple cancers in clinical settings simultaneously.

However, addressing the potential for overfitting is essential before considering clinical applications, despite efforts to mitigate it by using training and test datasets. Additional validation on independent

external datasets will be critical to confirm these findings and ensure generalizability.

The identified clusters from **Solution 2.1** did not align with known breast cancer subtypes. While we anticipated uncovering unique methylation signatures between subtypes, this result opens up opportunities for further investigation. It would be valuable to explore the shared characteristics within the clusters we identified to understand their biological significance. Additionally, an avenue for future work could involve reattempting clustering with the inclusion of normal tissue samples to determine whether unsupervised methods can segregate samples based on tumor state.

Our neural network model also reflected this challenge of classifying breast cancer subtypes based on methylation signatures (**Solution 2.2**). While this supervised method did perform better, with an NMI of 0.58 compared to the NMI of 0.2902 for **Solution 2.1**, overall performance declined compared to the more straightforward task of classifying tumors as normal. This is not unexpected based on the results seen in **Solution 2.1**. As these subtypes are not defined with methylation context in mind, epigenetic heterogeneity within a subtype likely causes more misclassifications.

We ran into memory issues repeatedly when we attempted to perform SHAP value analysis on all of our pipelines. Due to the high dimensionality of our data, SHAP would fail due to needing 83 TB of memory, which is unfeasible for this project. To address this, we modified the pipeline for **Solution 1.1** in the following ways: perform SHAP on only 1 example, remove the LLE layer, perform imputing outside pipeline, and utilize HM27 data rather than HM450K. Both removing the LLE and utilizing HM27 data allowed SHAP analysis to work (**Supplemental Figures 6**, **7**,**8**, **9**, **10**, and **11**). Since the classifier performed very well without the LLE and is more interpretable, dimensionality reduction may not be necessary or desired when utilizing methylation data. While using HM27 data is also a potential solution to making this project more interpretable, it comes at the cost of missing hundreds of thousands of CpGs that could have biological significance.

Conclusions

As our knowledge of the genetics behind cancer continues to grow, so does our understanding of the epigenetics of cancer. In this project, we sought to leverage the methylation data available in cBioPortal and TCGA to address critical areas of clinical interest, such as tumor status and breast cancer subtype classification. We utilized multiple machine learning methods to accomplish these goals. The first problem addressed was categorizing samples as normal or tumor based on methylation data. The two approaches were based on local linear embedding with logistic regression and principal component analysis with gradient boosting. Both models achieved high accuracy (> 97%) in the testing dataset. The next area of interest was the classification of breast cancer samples into subtypes based on methylation data. An unsupervised and supervised approach was taken. The unsupervised approach utilized K-Means clustering and achieved a normalized mutual information score of 0.2902. The supervised approach utilized a neural network following the feature dimensionality approach and achieved a normalized mutual information score of 0.58. Overall, these models demonstrate the applicability of methylation and machine learning for differentiating tumors from normal cells and could inform a roadmap for early detection and diagnosis methods. While the models underperform on breast subtype classification, it could serve as an indicator of the epigenetic heterogeneity that exists within the current categorization system. Thus, there may be an opportunity for an improvement to the current categorization system that considers the epigenetic status of samples.

Data Availability

Data for this project is public and can be obtained from cBioPortal and the Cancer Genome Atlas Research Network [18–31]. The samples used in this study can be found on our GitHub at: https://github.com/jacob-tye/20232110_TCGA_METHYLATION_CLINICAL_ML/blob/master/data/methylation/all_samples_450K.tsv.

Code Availability

All code used for this project is available on our public GitHub at: https://github.com/jacob-tye/20232110_TCGA_METHYLATION_CLINICAL_ML.

1. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).

2. Denis, H., Ndlovu, 'Matladi N & Fuks, F. Regulation of mammalian DNA methyltransferases: a route to new mechanisms. *EMBO Rep.* **12**, 647–656 (2011).

3. Lakshminarasimhan, R. & Liang, G. The Role of DNA Methylation in Cancer. *Adv. Exp. Med. Biol.* **945**, 151–172 (2016).

4. Newsham, I., Sendera, M., Jammula, S. G. & Samarajiwa, S. A. Early detection and diagnosis of cancer with interpretable machine learning to uncover cancer-specific DNA methylation patterns. *Biol. Methods Protoc.* **9**, bpae028 (2024).

5. Kim, Y. *et al.* Hypomethylation of ATP1A1 Is Associated with Poor Prognosis and Cancer Progression in Triple-Negative Breast Cancer. *Cancers* **16**, 1666 (2024).

6. Chen, L. *et al.* Identifying Methylation Pattern and Genes Associated with Breast Cancer Subtypes. *Int. J. Mol. Sci.* **20**, 4269 (2019).

7. Szyf, M., Pakneshan, P. & Rabbani, S. A. DNA methylation and breast cancer. *Biochem. Pharmacol.* **68**, 1187–1197 (2004).

8. Lin, L. H. *et al.* DNA Methylation Identifies Epigenetic Subtypes of Triple-Negative Breast Cancers With Distinct Clinicopathologic and Molecular Features. *Mod. Pathol.* **36**, 100306 (2023).

9. Joo, J. E. *et al.* Heritable DNA methylation marks associated with susceptibility to breast cancer. *Nat. Commun.* **9**, 867 (2018).

10. Szczepanek, J., Skorupa, M., Jarkiewicz-Tretyn, J., Cybulski, C. & Tretyn, A. Harnessing Epigenetics for Breast Cancer Therapy: The Role of DNA Methylation, Histone Modifications, and MicroRNA. *Int. J. Mol. Sci.* **24**, 7235 (2023).

11. Wang, L. *et al.* Identifying subtypes and developing prognostic models based on N6-methyladenosine and immune microenvironment related genes in breast cancer. *Sci. Rep.* **14**, 16586 (2024).

12. Almeida, B. P. de, Apolónio, J. D., Binnie, A. & Castelo-Branco, P. Roadmap of DNA methylation in breast cancer identifies novel prognostic biomarkers. *BMC Cancer* **19**, 219 (2019).

13. Calanca, N. *et al.* Inflammatory breast cancer microenvironment repertoire based on DNA methylation data deconvolution reveals actionable targets to enhance the treatment efficacy. *J. Transl. Med.* **22**, 735 (2024).

14. Ennour-Idrissi, K., Dragic, D., Durocher, F. & Diorio, C. Epigenome-wide DNA methylation and risk of breast cancer: a systematic review. *BMC Cancer* **20**, 1048 (2020).

15. Sun, Z., Cunningham, J., Slager, S. & Kocher, J.-P. Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. *Epigenomics* **7**, 813–828 (2015).

16. Fleischer, T. *et al.* DNA methylation at enhancers identifies distinct breast cancer lineages. *Nat. Commun.* **8**, 1379 (2017).

17. Singh, M. P., Rai, S., Gupta, S. K., Singh, N. K. & Srivastava, S. Unsupervised machine learning-based clustering identifies unique molecular signatures of colorectal cancer with distinct clinical outcomes. *Genes Dis.* **10**, 2270–2273 (2023).

18. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov.* **2**, 401–404 (2012).

19. Bruijn, I. de *et al.* Analysis and Visualization of Longitudinal Genomic and Clinical Data from the AACR Project GENIE Biopharma Collaborative in cBioPortal. *Cancer Res.* **83**, 3861–3867 (2023).

20. Gao, J. *et al.* Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013).

21. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**, 400-416.e11 (2018).

22. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291-304.e6 (2018).

23. Gao, Q. *et al.* Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. *Cell Rep.* **23**, 227-238.e3 (2018).

24. Taylor, A. M. *et al.* Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell* **33**, 676-689.e3 (2018).

25. Bonneville, R. *et al.* Landscape of Microsatellite Instability Across 39 Cancer Types. *JCO Precis. Oncol.* **2017**, 1–15 (2017).

26. Lakbir, S. *et al.* Tumour break load is a biologically relevant feature of genomic instability with prognostic value in colorectal cancer. *Eur. J. Cancer* **177**, 94–102 (2022).

27. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst.* **6**, 271-281.e7 (2018).

28. Ding, L. *et al.* Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell* **173**, 305-320.e10 (2018).

29. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173**, 321-337.e10 (2018).

30. Bhandari, V. *et al.* Molecular landmarks of tumor hypoxia across cancer types. *Nat. Genet.* **51**, 308–318 (2019).

31. Poore, G. D. *et al.* RETRACTED ARTICLE: Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579**, 567–574 (2020).

32. 2.2. Manifold learning — scikit-learn 1.5.2 documentation. https://scikit-learn.org/1.5/modules/manifold.html.

33. Wu, C., Mou, X. & Zhang, H. Gbdmr: identifying differentially methylated CpG regions in the human genome via generalized beta regressions. *BMC Bioinform.* **25**, 97 (2024).

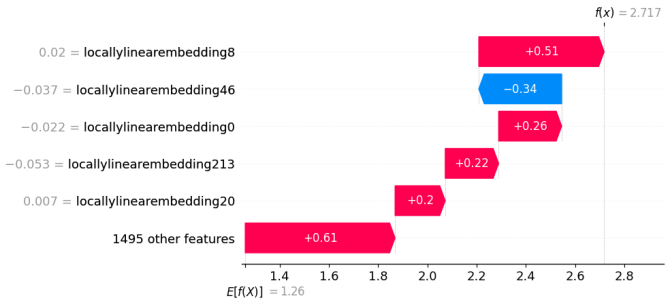
Appendix

Problem	Solution	Unsupervised	Supervised
Cancer State Detection	1.1	LLE	Logistical Regressor
Cancer State Detection	1.2	PCA	Gradient Booster
Breast Cancer Subtype Identification	2.1.1	K-means	
Breast Cancer Subtype Identification	2.1.2	KPCA	
Breast Cancer Subtype Identification	2.2	PCA	Neural Network

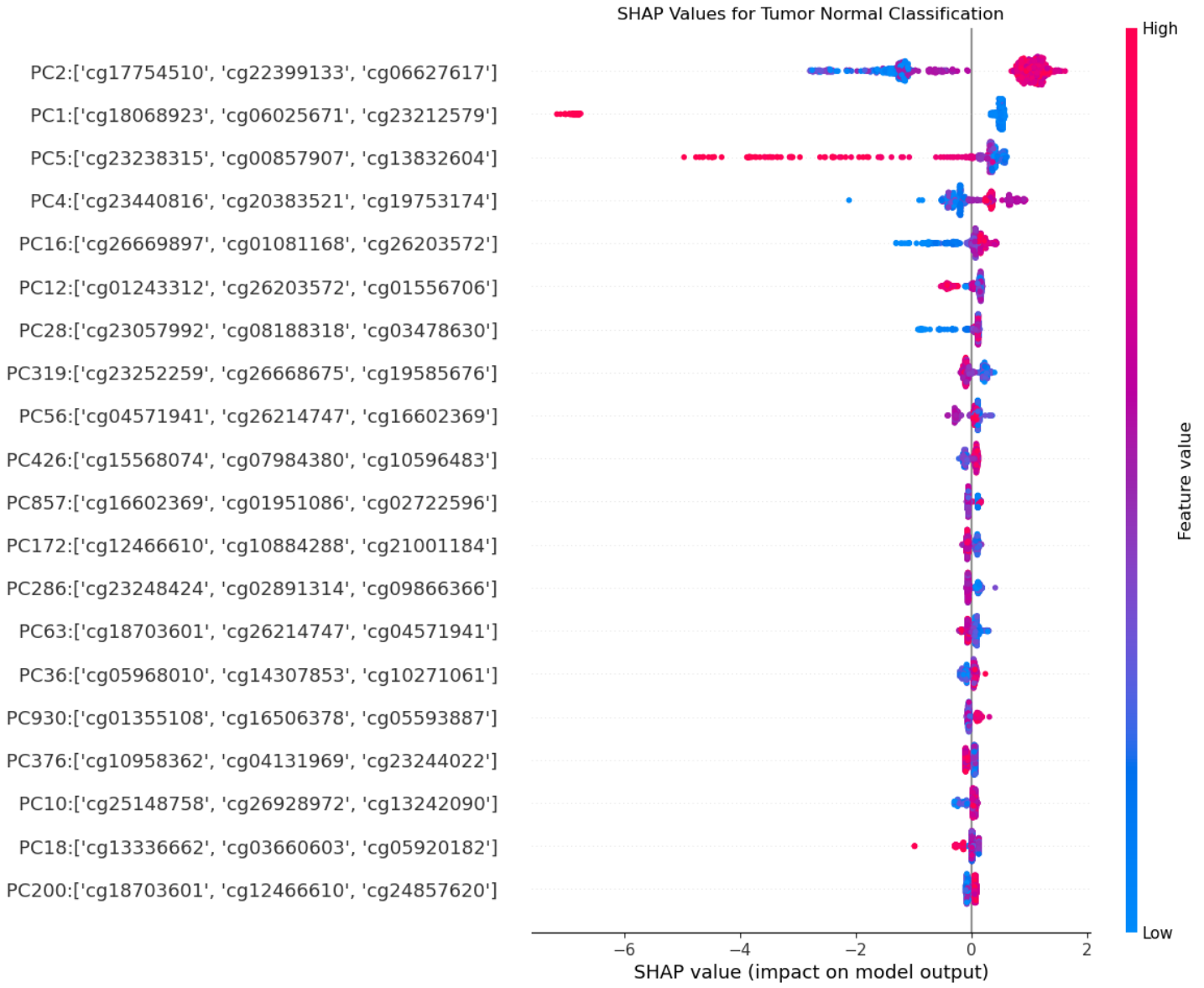
Supplemental Table 1 *Table summarizing all the different types of models used in our attempts*

Solution	Parameter	Component	Range
1.1	tol	Logistical Regressor	0.000001 – 0.001
1.1	l1_ratio	Logistical Regressor	0 - 1
1.1	C	Logistical Regressor	2 – 22
1.1	max_iter	Logistical Regressor	75 – 150
1.1	n_components	Locally Linear Embedding	500 – 2000
1.1	n_neighbors	Locally Linear Embedding	2 – 97
1.1	tol	Locally Linear Embedding	0.000001 – 0.001
1.1	max_iter	Locally Linear Embedding	75 – 150
1.1	hessian_tol	Locally Linear Embedding	0.000001 – 0.001
1.2	n_estimators	Gradient Boosting	50 - 200
1.2	learning_rate	Gradient Boosting	0.01 - 0.3
1.2	max_depth	Gradient Boosting	3 - 10
1.2	min_samples_split	Gradient Boosting	2 - 10
1.2	min_samples_leaf	Gradient Boosting	1 - 20
2.2	pca_n_components	PCA	1 - 300
2.2	num_layers	Neural Network	1 - 3
2.2	hidden_dims	Neural Network	10 - 100
2.2	epochs	Neural Network	10 - 50
2.2	learning_rate	Neural Network	0.0001 - 0.01

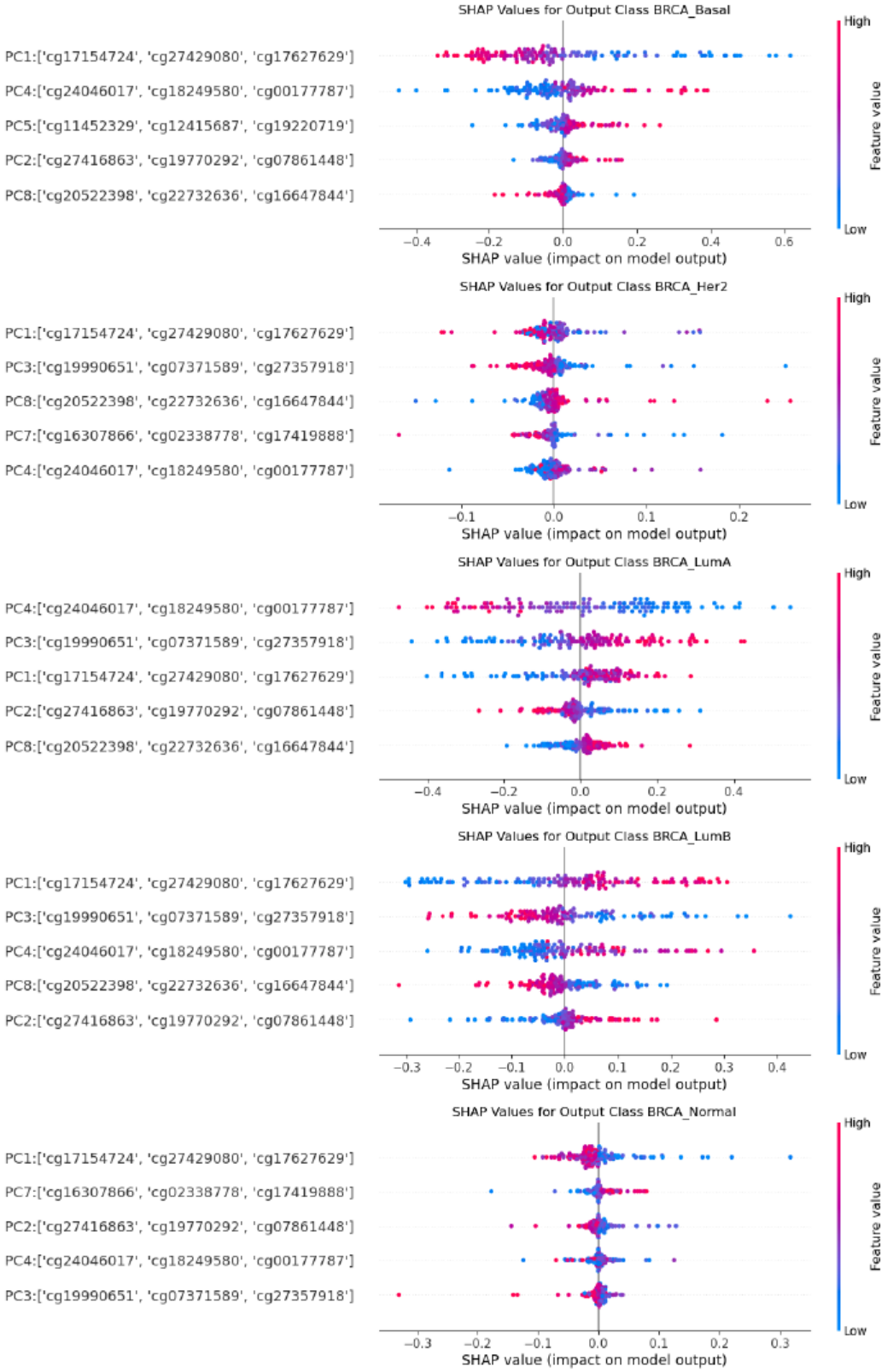
Supplemental Table 2 *The hyperparameter space that was searched by Optuna for* ***Solution 1.1, Solution 1.2*** *and* ***Solution 2.2****.*



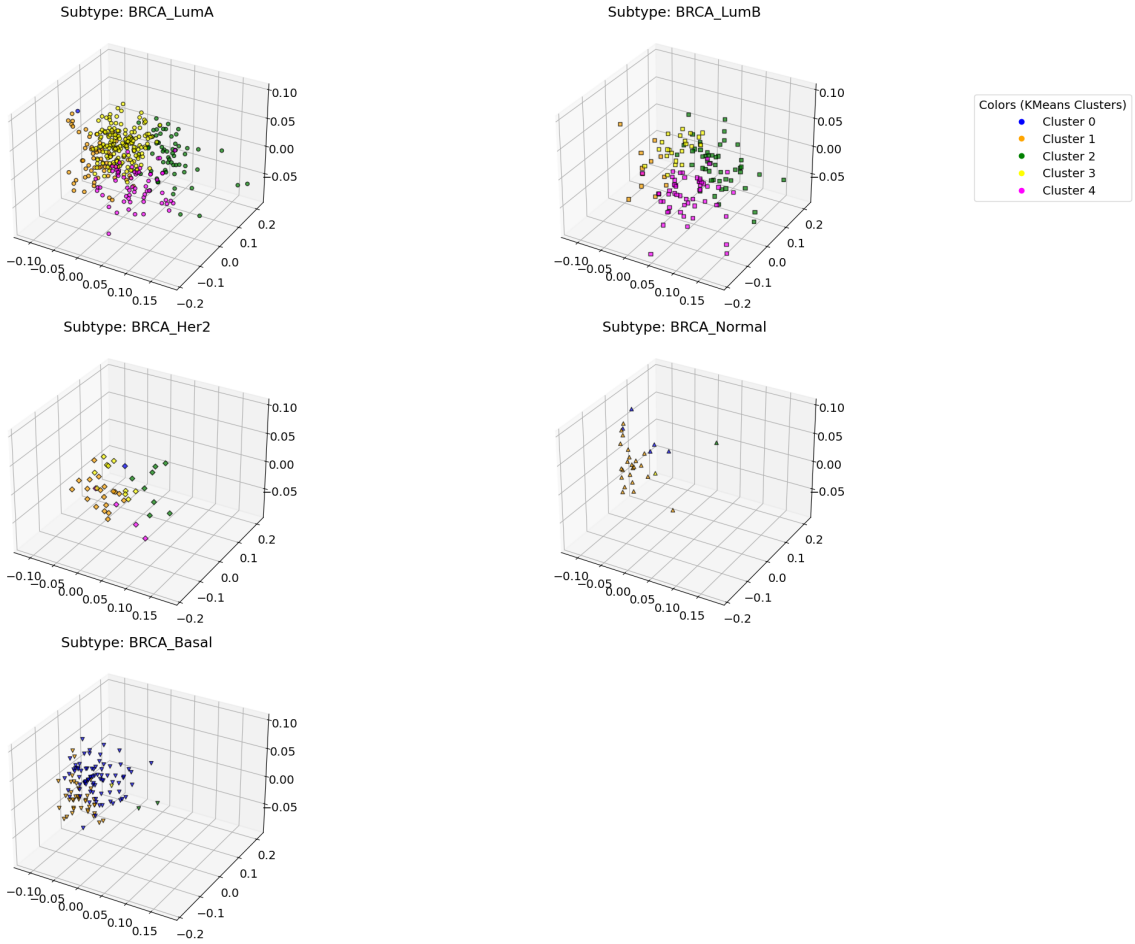
Supplemental Figure 1 *Waterfall plot of Shap values for* ***Solution 1.1*** *for a random example from the testing dataset.*



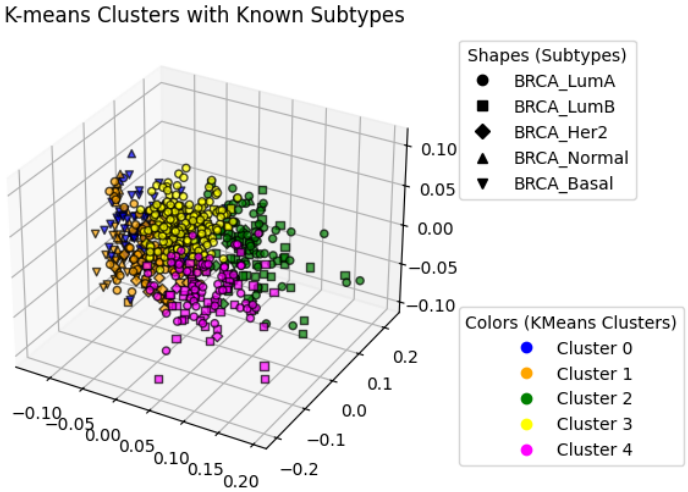
Supplemental Figure 2 *SHAP values for the top principal components that contributed to* ***Solution 1.2****. The three most important CpG sites, as determined by the loading values for the principal component, are shown on the Y-axis.*



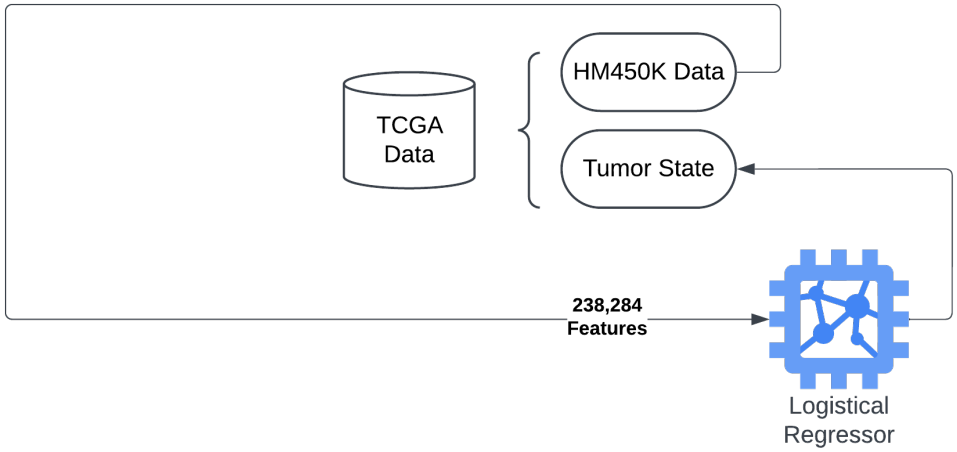
Supplemental Figure 3 *The SHAP summary plots for* ***Solution 2.2*** *show each breast cancer subtype's five most critical principal components. Each principal component is displayed with the three most important CpG sites as determined by the loading values.*



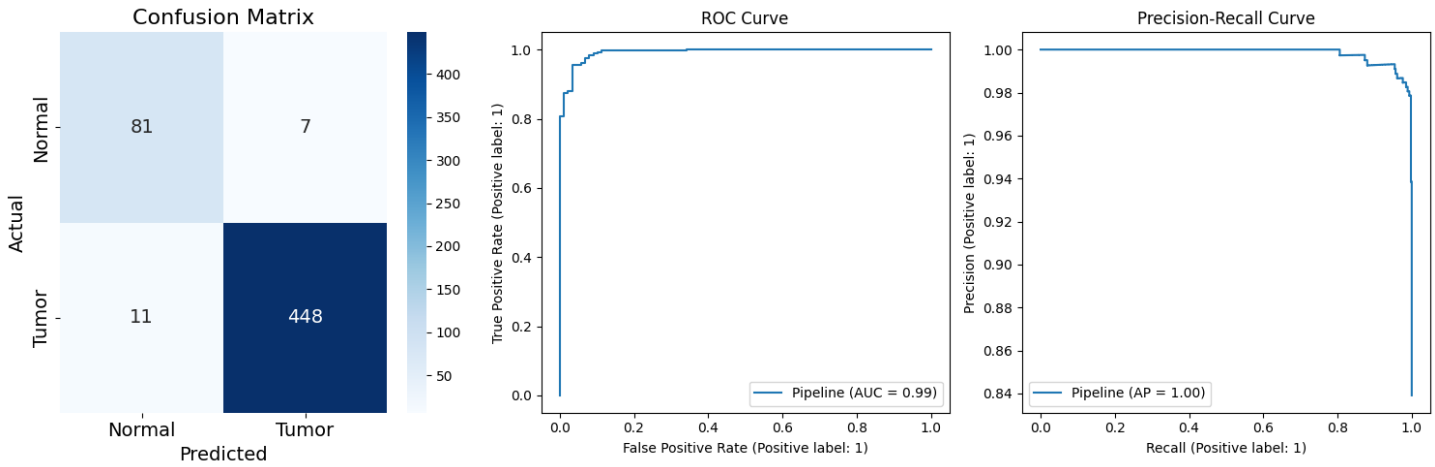
Supplemental Figure 4 *Scatter plots* ***Solution 2.1*** *clusters separated by subtype.*



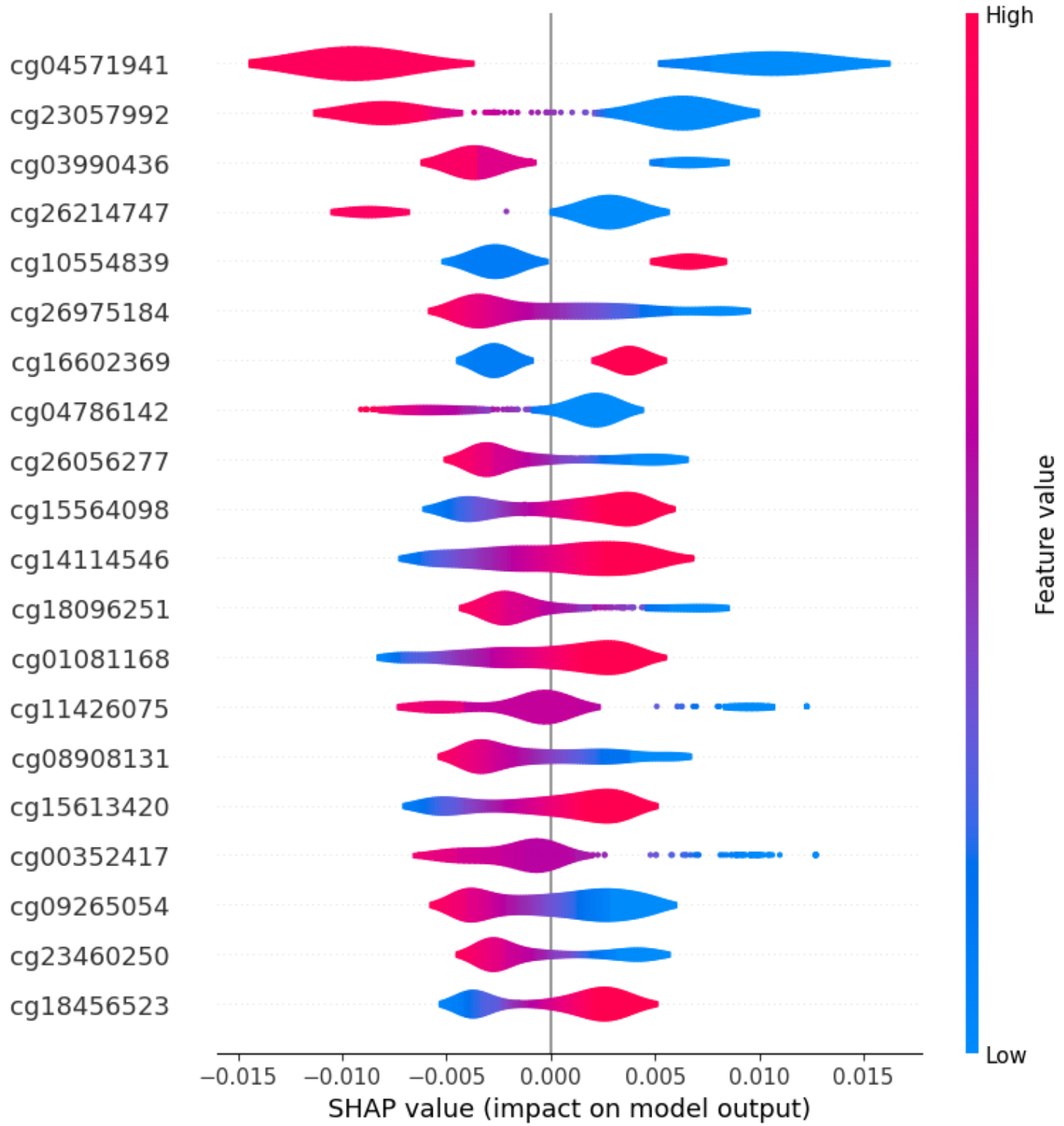
Supplemental Figure 5 *Figure showing the clusters as colors from* ***Solution 2.1*** *overlayed with breast cancer subtypes as the different shapes.*



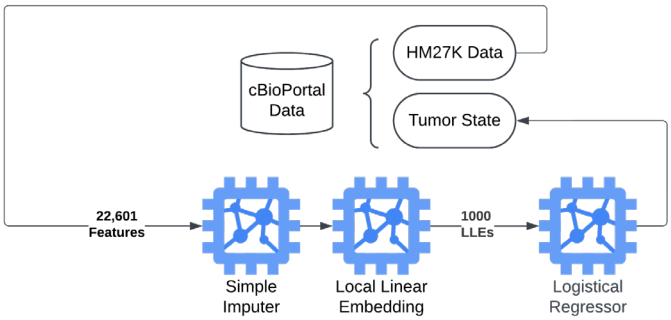
Supplemental Figure 6 *Pipeline from* ***Solution 1.1*** *with imputing outside the pipeline and without dimensionality reduction*



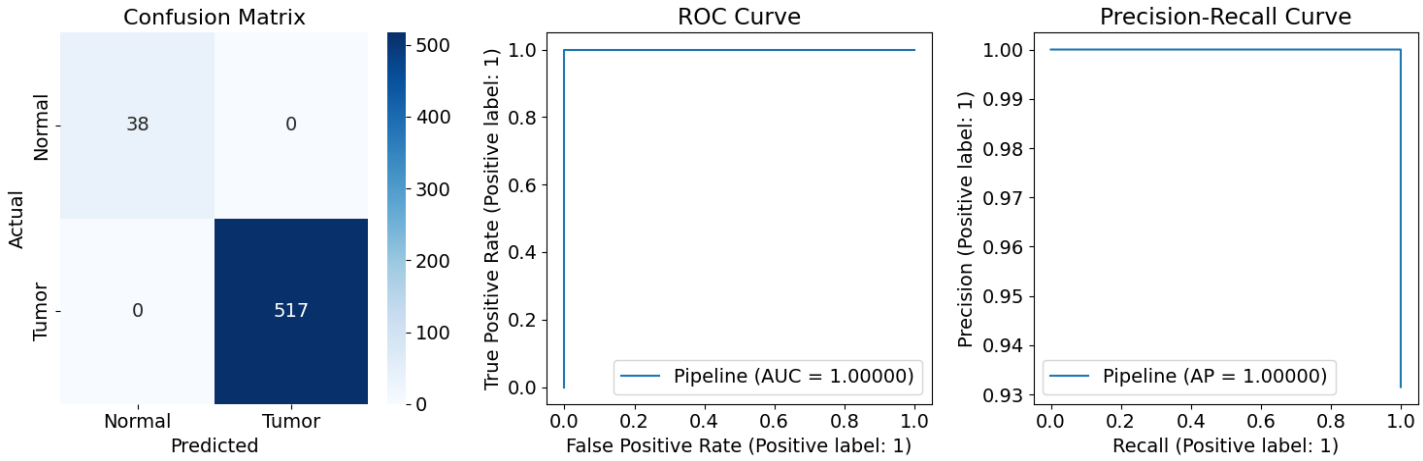
Supplemental Figure 7 *Performance of alternative solution for* ***Solution 1.1*** *without utilizing dimensionality reduction with LLE.*



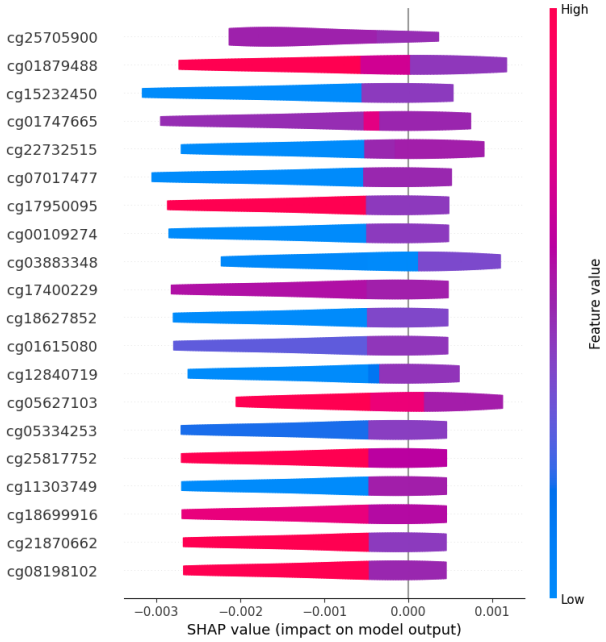
Supplemental Figure 8 *SHAP value performance for alternative implementation of* ***Solution 1.1*** *without using dimensionality reduction with LLE*



Supplemental Figure 9 *Pipeline from* ***Solution 1.1*** *using HM27 data instead if HM450K*



Supplemental Figure 10 *Performance of the pipeline for* ***Solution 1.1*** *using HM27 data rather than HM450K*



Supplemental Figure 11 *Shap values of the top CpGs when using HM27 data with a machine learning pipeline from* ***Solution 1.1***