Methylation data analysis using PyMethylProcess

Shubhangi Kaushik

Full list of author information is available at the end of the article

Abstract

The goal of the project: The goal of this project was to compare the results of papers on prostrate tumors and sinonasal tumors, with the results obtained using PyMethylProcess as the data pre-processing step for Methylation data.

Main results of the project: PyMethylProcess was used in the Pre-processing step for the Methylation data. Following this, Machine learning was applied and the result was compared to that of the paper used.

Estimated working hours: 1 week(5 days :- making PyMethylProcess work, 6 hours :- writing code, 1 day :- writing report)

Goal

The goal of the project was to analyze two methylation datasets to highlight the significance of utilizing clinical epigenetics data in precision medicine. The project for the first dataset focuses on prostate cancer, aiming to analyze DNA methylation patterns in blood samples from patients with prostate cancer and healthy controls[1]. On the other hand, Project for dataset 2 focuses on sinonasal tumors and seeks to identify DNA methylation-based molecular subtypes for it[4].

Both projects share a common objective, starting from feature extraction. Relevant features were first extracted using PyMethylProcess and machine learning classifiers were modeled based on these features. Their performance was evaluated using various metrics and compared with the existing study. Additionally, a feature importance analysis was also conducted. By leveraging the epigenetic markers of interest, the projects aim to advance diagnostics and classification methods for prostate cancer and sinonasal tumors, respectively.

Data

The data sets for prostate cancer[1] and sinonasal tumors[4] were used in this project. They were obtained from the NCBI GEO database using accession IDs, GSE112047 and GSE189778.

Prostate Cancer

The data retrieved from GEO accession id ”GSE112047” consisted of 47 prostate tissue samples, where 31 samples are of tumor tissue, and the remaining 16 samples are of normal tissue. Hence, the dataset has a binary classification.

Sinonasal Tumor

The data retrieved from GEO accession id ”GSE189778” was made up of 461 tumor samples with a total of 8 tumor classes. hence the dataset got multi-class classification.

For the Sinonasal Tumor dataset, The research paper identified 18 unique and consistent epigenetic classes of sinonasal tumors by analyzing DNA methylation data. The classes were 5 HPV-related cancer, 5 Non-HPV-related cancer, 2 NEC-like carcinoma, Adenocarcinoma, Neuroendocrine carcinoma (NEC), Intestinal-type adenocarcinoma, Sinonasal undifferentiated carcinoma (SNUC), SMARCB1-deficient carcinoma and Adenoid cystic carcinoma (ACC). Each class was then assigned a unique identifier (ID) based on its DNA methylation profile. When the preprocessed methylation data was obtained after using PyMethylProcess was analyzed, only 8 classes were present. The subsequent analysis focused specifically on those 8 classes.

Data Preprocessing

For the preprocessing analysis. docker was used for loading the data directly from GEO databases into the pymethylprocess. It is important that the dataset chosen to be used contains clinical data. After loading the data, the clinical data is merged with the actual dataset. Following it, the preprocessing step is done on the dataset using minfi and noob, which are the libraries from R used for the differently methylated cg sites analysis of the methylation data. Non-autosomal CpGs and missing values are removed from the dataset. As the next step, the columns are imputed using KNN, and feature selection is conducted using mean absolute deviation. As the last step of this process, the dataset was split into 20 percent test,70 percent train, and 10 percent validation sets.

Methods

For Sinonasal Tumor data and Prostate cancer FFPE samples were restored using the Illumina Infinium HD FFPE DNA Restore Kit, followed by bisulfite conversion using the EpiTect Bisulfite Kit (Qiagen)[4][1]. The converted DNA for sinonasal tumor data was imaged using the Illumina iScan system with default settings. The data for both datasets were normalized using minfi package of R.

Various tools and machine learning processes were used to build the models and evaluate the results.

PyMethylProcess

The analysis started with generating the pre-processed data using PyMethylProcess, a Python package designed specifically for the processing and analysis of DNA methylation data. It provides a set of methods for tasks like data preprocessing, normalization, differential methylation analysis, visualization, etc.

As the first step, the raw file along with all the idat files and clinical data files are downloaded. the data is formatted using the idat files, clinical file, and sample sheet. The preprocessing is conducted just as stated in the data preprocessing section and the chosen number of features are obtained.

Gradient boosting for GSE112047

Gradient Boosting[2] trains weaker algorithms like Decision Trees sequentially, such that every later model corrects the mistakes made by the former model. The optimization in Gradient Boosting is performed in the function space, where the function estimate $\hat{f}_i(x)$ is parameterized in an additive functional form:

$$\hat{f}(x) = \hat{f}_M(x) = \sum_{i=0}^{M} \hat{f}_i(x)$$

where M is the number of iterations, $\hat{f}_i(0)$ is the initial guess and $\{f_i\}_{i=1}^{M}$ are the function increments or "boosts."

Random Forest for GSE189778

Random Forest[5] is a bagging method that uses Decision Trees. Decision Trees are created for the combination of random subsets of observations and random subsets of features. The final prediction in Random Forest is based on the majority vote from each tree.

The choice of predictor variables used to split nodes in decision trees is determined by specific optimization criteria. In classification, one commonly used criterion is entropy(E),

$$E = -\sum_{i=1}^{c} p_i \log(p_i)$$

where c represents the number of unique classes and p_i is the prior probability of each class. The goal is to maximize the entropy value, aiming to gain the most information with each split in the decision tree.

Evaluation

The machine learning model was evaluated using the prediction probabilities calculated using the K-Fold cross-validation technique where the number of folds was set to 10. Based on the prediction probabilities, the class was predicted for both train and test datasets. These predictions were evaluated using the following:

Confusion Matrix

A confusion matrix is a table that summarizes the performance of a classification model by giving the counts of true positive, true negative, false positive, and false negative predictions (Table 2).

Predicted/Actual	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Table 1: Confusion Matrix

Since the Sinonasal Tumor data model was a multi-label model, Macro-average values for the sensitivity and specificity were calculated using the confusion matrix.

ROC Curve

(Receiver Operating Characteristic) Curve[3] for each class was also generated using the probabilities. It plots the true positive rate against the false positive rate at different classification thresholds.

Results and Discussion

Prostate Cancer binary classification

For the prostate tumor dataset,the number of classes was 2, where one was 'Adjacent normal prostate tissue' and the other one was 'Prostate tumor'. Hence, binary classification was conducted on top of this dataset. After performing preprocessing using PyMethylProcess, Random Forest was used to identify the top 5000 features. These selected features were subsequently employed to train the final model using Gradient Boosting. The entire dataset except the labels was the 'X' dataset and the labels were put in dataset 'Y', which was encoded using LabelEncoder. The datasets were later split into test and train sets and the class imbalance was resolved using SMOTE(synthetic minority oversampling technique). Following this hyperparameter tunning was conducted and 500 was chosen as the optimal number of trees for model evaluation.

For the training dataset, the model gave 100 percent accuracy, which could also be inferred from the confusion matrix, as there was no false positive or false negative.

Predicted/Actual	Positive	Negative
Positive	20	0
Negative	0	20

Table 2: Confusion Matrix for training dataset

For the test dataset, the accuracy was around 87.5 percent, and the confusion matrix had no false positives.

Based on these results the ROC curve was built for both training and test dataset. The AUC score was obtained to be 100 percent in the case of the training dataset, whereas was 90 percent in the case of the test data set.

Predicted/Actual	Positive	Negative
Positive	5	0
Negative	2	9

Table 3: Confusion Matrix

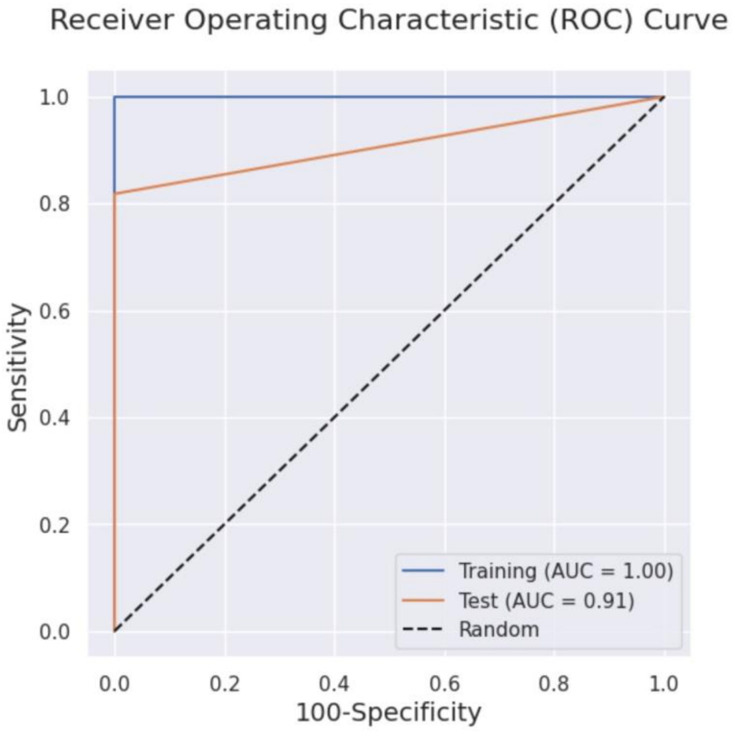


Figure 1: The orange line indicates the ROC curve for the test dataset whereas the blue line indicates the ROC curve for the training set. The performance of the training dataset was ideal, and the test dataset gave decent results. Moreover, the performance of the test could be improved by changing the number of trees in the model.

When compared to the previous study, the AUC score generated for the training dataset was also 100 percent. And for the test dataset, also known as the validation set in the paper yielded a score of 97 percent[1].

After the ROC curve analysis, important features were extracted and the graph was plotted for the top 10 features.

When comparing this project with the previous study, the previous study performed better than this study for the testing dataset, whereas, the performances were the same as for the training dataset. This difference could be due to the fact that the model in this project was trained on the top 5000 features, instead of training it on the entire 30,000 features. To test if the performance could be improved when the entire feature set is used instead of a subset, the entire 30,000 features were planned to be used to train the model. However, before it could be

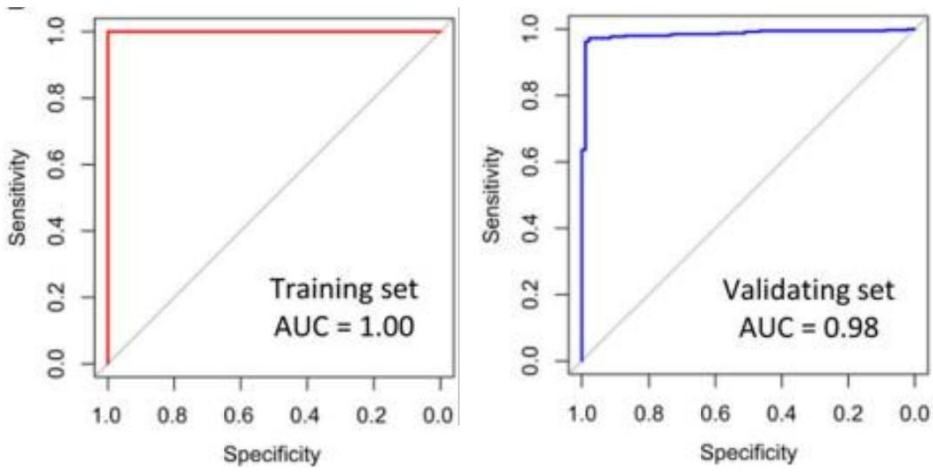


Figure 2: The graph on the left suggests the ROC curve for the training dataset, whereas the right side is for the test dataset. According to the paper, the model performed very well on both the train and test datasets.

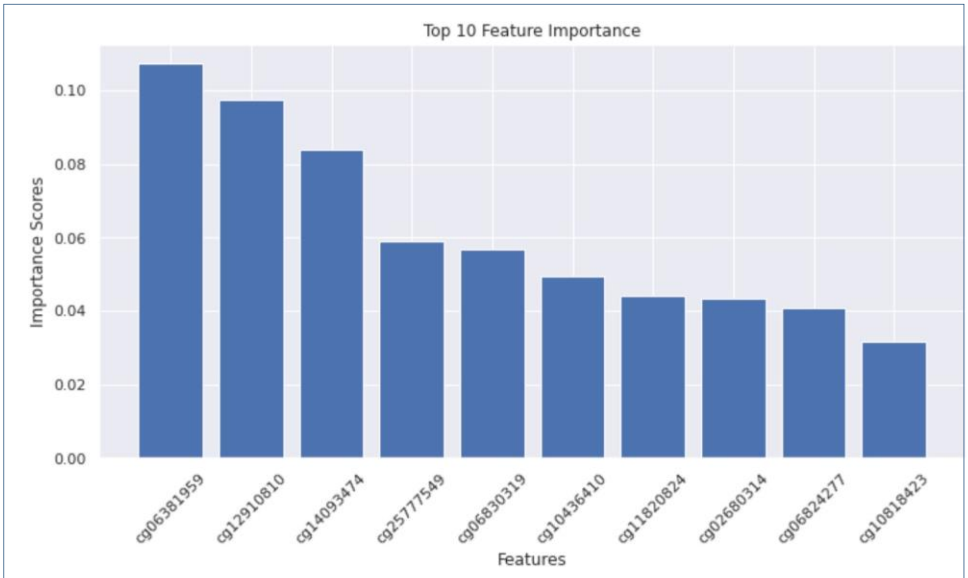


Figure 3: The bar graph is plotted for important features obtained using the Gradient boost model The X axis is made up of the top 10 CPG IDs obtained from the analysis, and the Y axis showcases the importance found for each IDs.

done, the kernel on which the Python code to obtain the probability score for the 30,000 features in pymethylprocess was running, crashed. It was believed that this could have happened due to the update in the Python version while SMOTE was imported into the Python file for fixing class imbalance. Later when trying to reinstall the pymethylprocess with Python 3.6 version, the time taken was long. Hence, the reinstallation idea was given up and the analysis was concluded here.

The previous paper didn't specifically give a list of important features that they obtained hence it was not possible to compare the list of features obtained in this study to that of the previous study.

To obtain the biological context for the CG IDs obtained, I started by searching for the IDs in research papers related to prostate tumors, however, desired results were not found. The next option was going through the GEO database which also didn't yield any good results. At last CHAMP library in R was used to obtain the gene and enzyme names. Based on the CHAMP analysis the genes and enzymes I found to be linked to the CG IDs obtained were A1BG, A2M A2ML1, A4GALT, AAAS.

Gene Names	Biological Context
A1BG	Altered gene expression found in prostate cancer
A2M	It is a protease inhibitor involved in regulating tumor growth and invasion, hence could play a role in prostate cancer
A2ML1	Inhibits proteases and suppresses tumor growth, any alteration could influence tumor progression
A4GALT	It is an enzyme involved in glycosylation processes. alterations in glycosylation patterns influence tumor progression
AAAS	protein involved in nuclear transport and stress response. It has potential involvement in prostate cancer development and progression

Table 4: Biological context for the gene, enzymes, and proteins found based on the CG IDs, using CHAMP library in R for Prostate tumor dataset

Sinonasal Tumor multiclass classification:

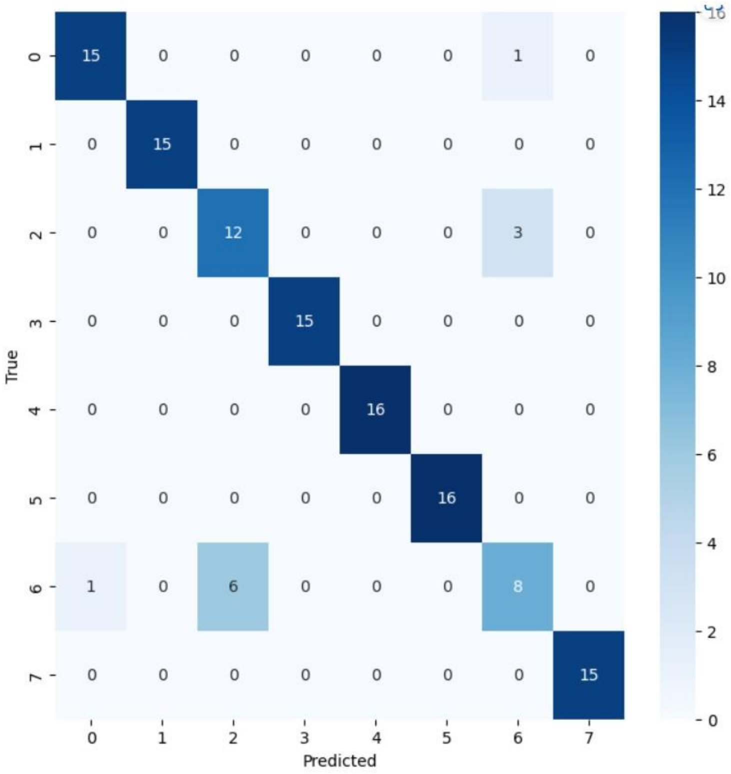
Since the 8 classes in the dataset had a big difference in the number of samples, it created an imbalance. To tackle this problem, ”LabelEncoder” was used. Balancing of the dataset improved the accuracy of the analysis drastically.

The construction of the model involved the utilization of a Random Forest classifier algorithm. The dataset was split into a training set comprising 70% of the data and a test set containing the remaining 30%. Confusion matrix in Figure 4 and scores in Table 5 were calculated for both the train and the test data after building the model on top of the tuned hyperparameters.

Model	Accuracy	Sensitivity	Specificity
Random Forest(Train)	90.20%	90.28%	89.84%
Random Forest(Test)	91.05%	90.88%	90.88%
Random Forest(Paper)	100%	90.4%	98.2%

Table 5: Accuracy, Sensitivity, and Specificity of the Random Forest model for both the training and test sets. Additionally, these performance metrics were compared with the results reported in the original research paper by Jurmeister et al.

The model's performance was assessed by calculating Sensitivity, Specificity, and Accuracy (as shown in Table 5). The results were compared with the results from the original paper(Table 5 and Figure 7 from Supplementary Image). The results



True \ Predicted	0	1	2	3	4	5	6	7
0	15	0	0	0	0	0	1	0
1	0	15	0	0	0	0	0	0
2	0	0	12	0	0	0	3	0
3	0	0	0	15	0	0	0	0
4	0	0	0	0	16	0	0	0
5	0	0	0	0	0	16	0	0
6	1	0	6	0	0	0	8	0
7	0	0	0	0	0	0	0	15

Figure 4: Confusion matrix heatmap for the test data for the 8 classes of sinonasal tumor. While most of the classes showed accurate results with true positive predictions, class 2 and class 6 exhibited slightly deviated outcomes. However, the majority of the classes demonstrated true positive predictions without any discrepancies.

indicated that the model performed exceptionally well when applied to the data processed using PyMethylProcess. However, the model exhibited even better performance for the analysis conducted in the original research paper.

The paper includes a ROC curve illustrating the Area Under the Curve (AUC) for outlier detection, which aimed to differentiate sinonasal tumors from non-sinonasal tumors using SVM analysis. However, this specific analysis was not the focus of our study. Nonetheless, I plotted ROC graphs, Figure 5, to visually represent the performance of our model for the 8 classes in both the test and training set.

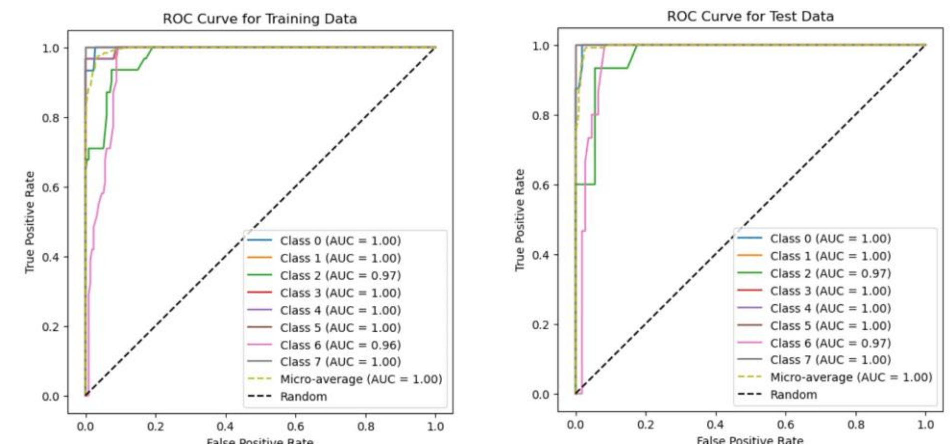


Figure 5: The ROC Curve depicted in the figure illustrates the performance for each class of sinonasal tumor in both the training and test data. Consistent with the findings from the confusion matrix, classes 2 and 6 exhibited lower scores for AUC in the ROC Curve analysis for both datasets.

Although the paper did not provide a comprehensive list of the most significant CpG sites for class separation, there was a mention of it. However, due to the paper's emphasis on additional analysis aspects, a detailed list was not included. In our analysis, I identified the most important CpG islands and created a bar graph to visualize the top 10 sites(Figure 6). The genes associated with the CpG islands were then mapped using Bioconductor Packages(Table 7).

CpG Island	Chromosome	Ranges	Strand	ENTREZID	SYMBOL
cg06317989	chr4	23755041-24472951	-	10891	PPARGC1A
cg14499563	chr13	36998816-37025881	+	11340	EXOSC8
cg01889461	chr7	34346512-34871582	-	404744	NPSR1-AS1
cg06592160	chr13	41132928-41236686	+	101929140	LOC101929140
cg15439294	chr19	56507850-56530221	+	57573	ZNF471
cg00640147	chr17	61942605-62065278	-	9969	MED13

Table 6: Genes that overlap to the genomic coordinates of the top 10 CpG islands. The table provides a concise overview of chromosome-related information, including the genomic positions, gene identifiers or their Entrez ID, and gene symbols for various chromosomes.

A thorough examination is needed to better understand the relationship between the six mapped genes, PPARGC1A, EXOSC8, NPSR1-AS1, LOC101929140, ZNF471, and MED13, and sinonasal carcinoma. The precise function of these genes in sinonasal carcinoma is yet unknown, despite some early research suggesting potential correlations.

Even though the precise role of PPARGC1A, EXOSC8, and MED13 in sinonasal cancer has not been studied so far, there is a potential connection between the genes and sinonasal cancer. This is because they have been linked to a number of other cancers, including head and neck cancer. Similarly ZNF471, a zinc finger protein recognized for its function in transcriptional control, has also been related to cancer, however research on how it specifically relates to sinonasal cancer is also lacking.

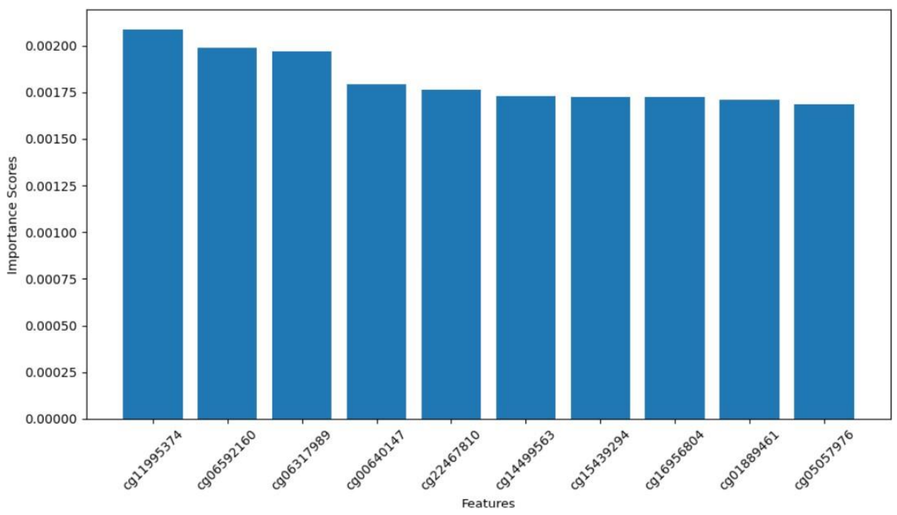


Figure 6: A bar graph was generated using the random forest model, showcasing the top 10 important CpG islands for the classification. The graph illustrates the importance scores assigned to each feature, providing insights into their relative significance in the analysis.

The relationship between NPSR1-AS1 and LOC101929140 and sinonasal carcinoma is not well understood and little is known about it.

In conclusion, both studies overall highlight the possibility to predict patient outcomes and tailor treatment for prostate and sinonasal cancer with more research into the genes and molecular characteristics of the important epigenetic markers. The importance of cutting-edge computational approaches in improving tumor categorization and patient treatment is emphasized in this work.

Appendix

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Blessy Rajan:-Worked on making PymethylProcess run, and ran machine learning analysis on GSE112047 dataset. Worked on report for GSE112047 including all the graphs the abstract,discussion, results, methods images, data, and improving the overall report and report structure.

Shubhangi Kaushik:-Worked on making PymethylProcess run, and ran machine learning analysis on GSE189778 dataset. Worked on report for GSE189778 including all the graphs abstract, goal, discussion, data, results, methods images, and improving the overall report and report structure

...

Acknowledgements

Special thanks to Professor Tim Conrad for helping us in the projects, and Aakash Nepal from the team ’BigOcean’ for helping us install the r dependencies.

Supplementary image

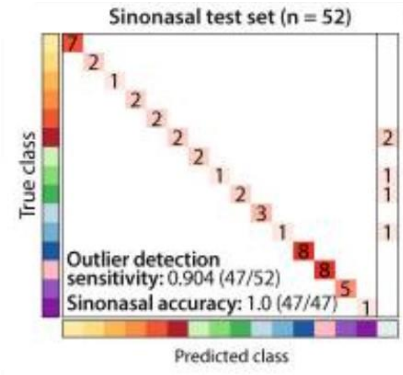


Figure 7: Confusion matrix for the classifier from the paper. The classifier for the sinonasal test set with 52 samples showed an accuracy of 1.0 for classification of sinonasal tumors.

...

Author details

References

1. Peter Ainsworth Hanxin Lin David I Rodenhiser Jean-Claude Cutz Bekim Sadikovic Erfan Aref-Eshghi, Laila C Schenkel. Genomic dna methylation-derived algorithm enables accurate detection of malignant prostate tissues. *NCBI*, 2022.
2. Jerome H Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.
3. Karimollah Hajian-Tilaki. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, 4(2):627, 2013.
4. Ren´ee Roller Maximilian Leitheiser Simone Schmid Liliana H Mochmann Emma Pay´a Capilla Rebecca Fritz Carsten Dittmayer Corinna Friedrich Anne Thieme Philipp Keyl Armin Jarosch Simon Schallenberg Hendrik Bl¨aker Inga Hoffmann Claudia Vollbrecht Annika Lehmann Michael Hummel Daniel Heim Mohamed Haji Patrick Harter Benjamin Englert Stephan Frank Ju¨rgen Hench Werner Paulus Martin Hasselblatt Wolfgang Hartmann Hildegard Dohmen Ursula Keber Paul Jank Carsten Denkert Christine Stadelmann Felix Bremmer Annika Richter Annika Wefers Julika Ribbat-Idel Sven Perner Christian Idel Lorenzo Chiariotti Rosa Della Monica Alfredo Marinelli Ulrich Schu¨ller Michael Bockmayr Jacklyn Liu Valerie J Lund Martin Forster Matt Lechner Sara L Lorenzo-Guerra Mario Hermsen Pascal D Johann Abbas Agaimy Philipp Seegerer Arend Koch Frank Heppner Stefan M Pfister David T W Jones Martin Sill Andreas von Deimling Matija Snuderl Klaus-Robert Mu¨ller Erna Forgo´ Brooke E Howitt Philipp Mertins Frederick Klauschen David Capper Philipp Jurmeister, Stefanie Gl¨oß. Dna methylation-based classification of sinonasal tumors. *NCBI*, 2022.
5. Matthias Schonlau and Rosie Yuyan Zou. The random forest algorithm for statistical learning. *The Stata Journal*, 20(1):3–29, 2020.