Cancer Classification Using DNA Methylation Profiles

Ayaan Shaikh, Emerging Diagnostic and Investigative Technologies, Department of Pathology
Dartmouth Hitchcock Medical Center - Levy Lab

Abstract

- **DNA methylation influences cancer by reducing methylation in oncogenes and increasing it in tumor suppressor genes. This study used machine learning to classify cancer types based on DNA methylation profiles. Data from eight cancers in the Cancer Genome Atlas was used to train random forest and SVM models.**
- **The Stacked Ensemble model performed best, with 99.49% accuracy. Our results indicate that Stacked Ensemble model composed of logistic regression, SVM, and Random Forest model, can effectively classify cancer types from DNA methylation profiles, though further work is needed to reduce overfitting.**

Introduction

Role of DNA Methylation in Cancer:

- DNA methylation influences cancer by decreasing methylation in oncogenes and increasing it in tumor suppressor genes.
- This process is an epigenetic modification where methyl groups are added to DNA, altering gene activity without changing the DNA sequence.

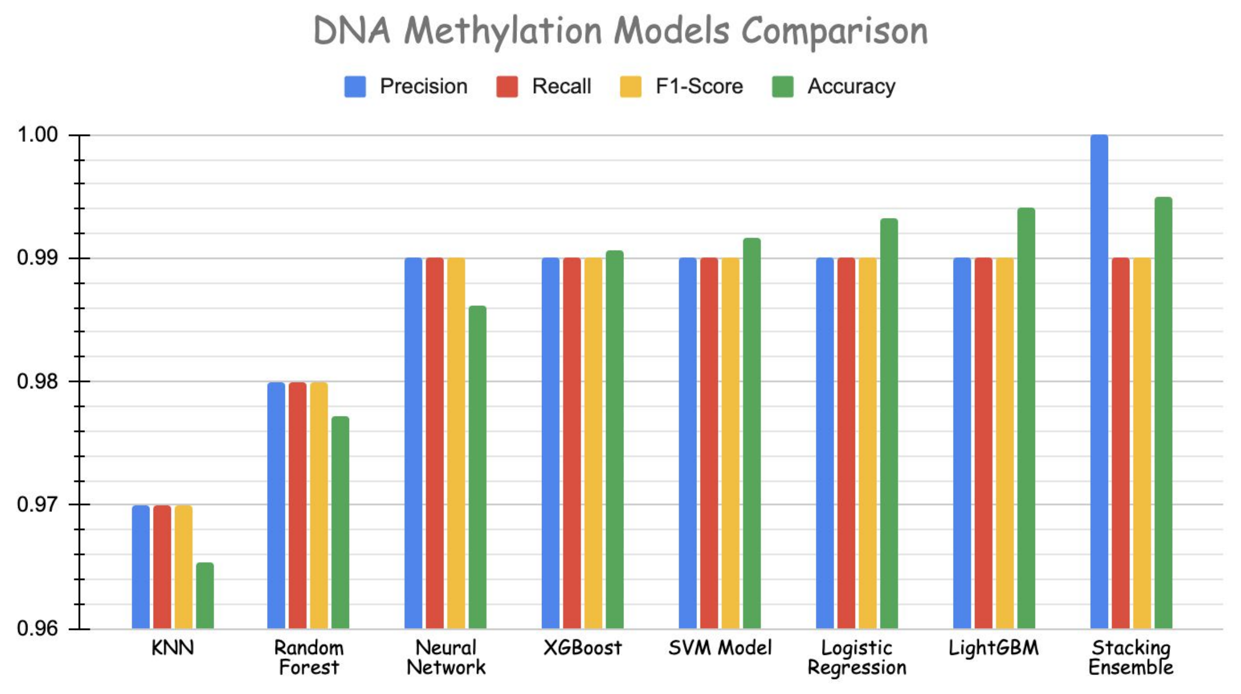
Diagnostic Potential of DNA Methylation:

- Methylation can serve as a biomarker for detecting various cancers, including colon, breast, ovarian, and cervical cancer.
- Identifying cancer through DNA methylation profiles is minimally invasive and can aid in diagnosis.

Enhancing Cancer Treatment with Machine Learning:

- Applying machine learning to classify cancers based on DNA methylation can enhance treatment accuracy and reduce costs.
- DNA methylation plays a crucial role in normal growth and disease development, often silencing genes by blocking their activity.

RESULTS



*KNN = K-Nearest Neighbors

*Stacking Ensemble = (SVM + Random Forest + Logistic Regression)

Figure 1: The bar chart compares various DNA methylation models, showing that the Stacking Ensemble outperforms other models with the highest precision, recall, F1-score, and accuracy, while KNN lags behind in accuracy.

Data & Methods

Data Collection and Sources

- **Data Origin**: The data for this project was sourced from the National Cancer Institute's Cancer Genome Atlas (TCGA), which includes 20,000 primary cancer and matched normal samples covering 33 different cancer types.

Cancer Types and Methylation Profiles

- **DNA methylation profile data was analyzed for eight specific types of cancer:**
  - [BRCA]: Breast Invasive Carcinoma
  - [PAAD]: Pancreatic Adenocarcinoma
  - [BLCA]: Bladder Urothelial Carcinoma
  - [HNSC]: Head and Neck Squamous Cell Carcinoma
  - [KICH]: Kidney Chromophobe
  - [SKCM]: Skin Cutaneous Melanoma
  - [LIHC]: Liver Hepatocellular Carcinoma
  - [LUAD]: Lung Adenocarcinoma

Methylation Data Analysis

- **The DNA methylation sites were evaluated using beta values, which range from 0 to 1.**
  - Beta values near 0 indicate low levels of DNA methylation.
  - Beta values near 1 indicate high levels of DNA methylation.

Machine Learning Models

- All models utilized a linear kernel with default parameters for classification tasks.
- Eight different models were employed in the study.
  - Support Vector Machine (SVM)
  - Random Forest Model
  - Logistic Regression
  - Stacked Ensemble
    - Logistic Regression
    - SVM
    - Random Forest
  - Neural Network
  - K-Nearest Neighbors (KNN)
  - XGBoost
  - LightGBM

RESULTS

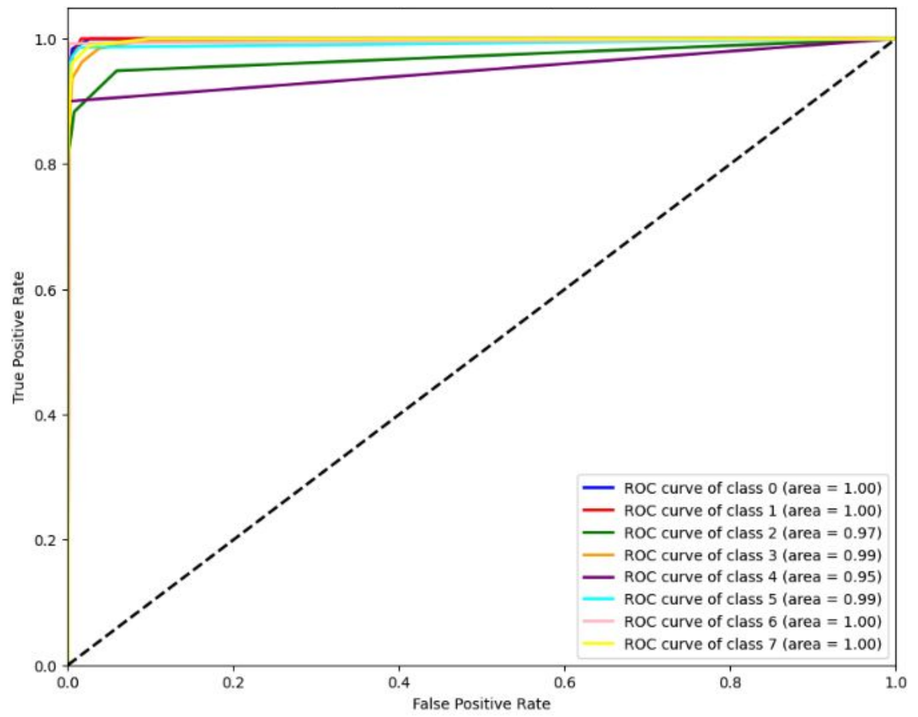
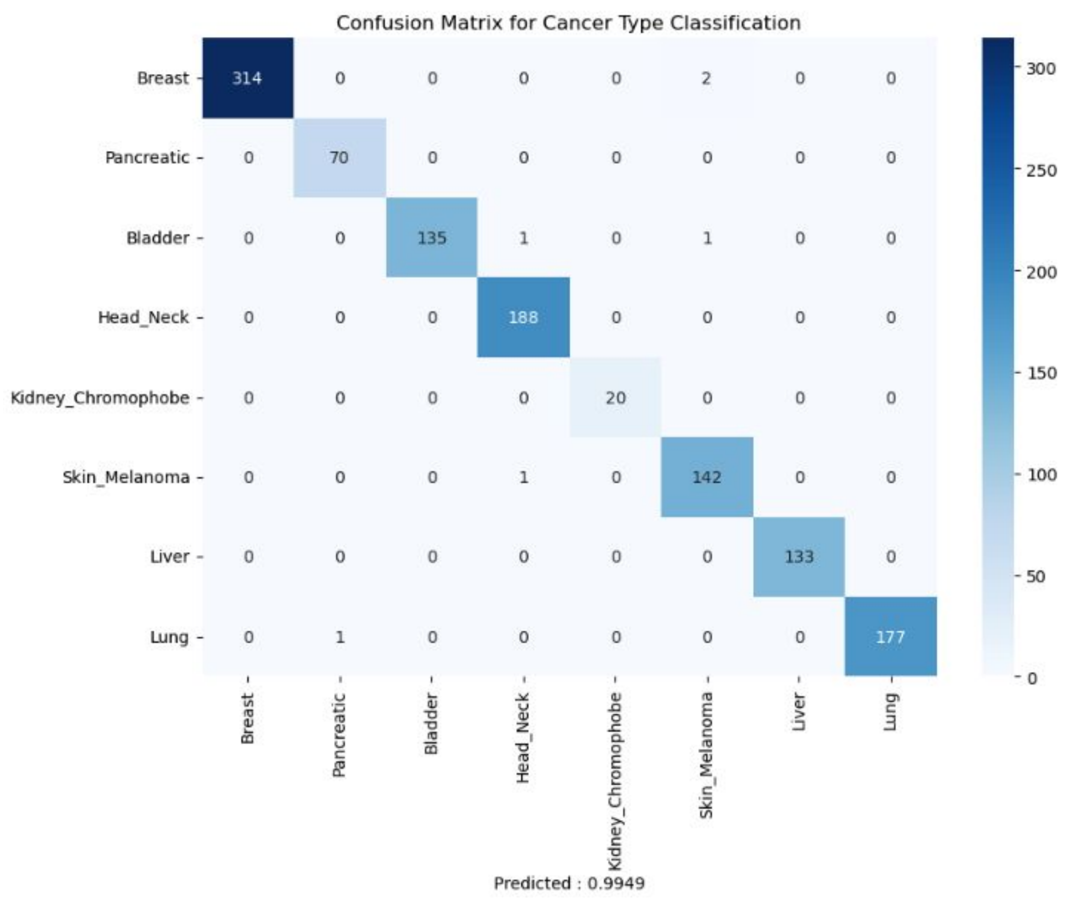


Figure 2: The ROC curve figure demonstrates that the KNN model achieves near-perfect classification for most classes, with AUC values close to or at 1.00, indicating strong overall performance.

RESULTS



	Breast	Pancreatic	Bladder	Head_Neck	Kidney_Chromophobe	Skin_Melanoma	Liver	Lung
Breast	314	0	0	0	0	2	0	0
Pancreatic	0	70	0	0	0	0	0	0
Bladder	0	0	135	1	0	1	0	0
Head_Neck	0	0	0	188	0	0	0	0
Kidney_Chromophobe	0	0	0	0	20	0	0	0
Skin_Melanoma	0	0	0	1	0	142	0	0
Liver	0	0	0	0	0	0	133	0
Lung	0	1	0	0	0	0	0	177

Figure 3: The confusion matrix for the Stacking Ensemble model demonstrates its strong performance in cancer type classification, with the majority of samples being correctly classified, as shown by the high counts along the diagonal. Only a few misclassifications are present, such as one pancreatic cancer sample being misclassified as lung cancer and a couple of other minor errors, reflecting the model's overall accuracy and reliability in distinguishing between different cancer types.

Discussion & Conclusion

Model Performance

- The Stacked Ensemble (Logistic Regression, SVM Model, Random Forest) model outperformed the other machine learning models in terms of precision, recall, accuracy, and F1 scores.
- All models showed strong performance metrics, though there is a possibility of overfitting.

Overfitting Concerns

- Potential overfitting was noted, particularly due to the uneven distribution of DNA methylation sites among cancer types (e.g., fewer sites in KICH and PAAD compared to BRCA).
- The high scores across models might be influenced by this imbalance, suggesting the need for caution in interpreting results.

Future Directions

- Future studies should aim to reduce overfitting by using fewer methylation sites or employing oversampling and undersampling techniques.
- Exploring additional machine learning models and analyzing a wider variety of cancer types could lead to further improvements in classification accuracy.

Acknowledgements

I would like to thank Dr. Joshua Levy, Dr. Khang Le, Onyi Owo, Akshat Alok, & ZoëFaith Caraballo-Bobea!