Politecnico
di Torino

Epigenetic Mechanisms in the Development of Neoplasms

submitted towards differentiation for

Master's degree of

DATA SCIENCE AND ENGINEERING

by

Nastaran Ahmadi Bonakdar

Department of Control and Computer Engineering
Polytechnic University of Turin

Supervisor(s)

Dr Alfredo Benso
Dr Sandro Gambino

June 2025

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisors, Prof. Alfredo Benso and Dr. Sandro Gambino, for their continuous support, insightful guidance, and invaluable feedback throughout this research journey. Their expertise and encouragement were instrumental in shaping both the scientific direction and the structure of this thesis.

I am sincerely thankful to the faculty and staff of the Master's program in Data Science and Engineering at Politecnico di Torino for providing a strong academic foundation and a stimulating research environment.

I would also like to thank the authors and curators of the GSE69914 dataset, whose publicly available data made this study possible, as well as the scientific community whose open-source tools and research papers enriched this work.

To my colleagues and friends, thank you for your helpful discussions, moral support, and motivation throughout this process.

Finally, to my family, whose love, patience, and unwavering belief in me have been my greatest strength — this accomplishment is dedicated to you.

Contents

Abstract

This thesis explores the role of DNA methylation outliers in breast cancer within the framework of the epigenetic field defect hypothesis. The hypothesis posits that stochastic epigenetic alterations in histologically normal tissue may signal early carcinogenic processes. Using the GSE69914 dataset from the Gene Expression Omnibus, we analyzed methylation profiles across three tissue types: normal, cancer-adjacent normal, and cancerous breast tissue.

A comprehensive preprocessing pipeline was implemented. Raw beta values were converted to M-values to reduce heteroscedasticity, followed by normalization, dimensionality reduction, and group labeling. Additional steps included variance-based filtering, Z-score transformations, and exclusion of low-quality or invariant CpG sites. Unlike earlier studies that rely solely on differential analysis, this work employed unsupervised machine learning algorithms for outlier detection, with the goal of identifying CpGs whose methylation values deviate substantially from the typical population-level distribution.

Variance thresholds of 0.015 and 0.02 were tested to balance signal retention with computational feasibility. Five algorithms—K-Nearest Neighbors (KNN), Isolation Forest, Local Outlier Factor (LOF), One-Class SVM (OC-SVM), and Z-score analysis—were applied across multiple hyperparameters to detect anomalous CpG methylation patterns. Results showed that CpG sites such as `cg19374752` and `cg00000622` were consistently flagged as outliers across tissues and algorithms. The comparative analysis revealed that Z-score detection offered the highest recall and F1-score, whereas One-Class SVM delivered the highest precision, suggesting each method's suitability for different diagnostic priorities.

A secondary benchmark using the Thyroid Disease dataset validated the comparative performance of the algorithms on structured, labeled data. Pathway enrichment analysis of the most frequently outlying CpG-associated genes highlighted cancer-relevant biological processes, including DNA repair, Notch signaling, and estrogen response pathways.

This work demonstrates the feasibility and diagnostic potential of methylation-based outlier detection in cancer and proposes a flexible, scalable pipeline for epigenetic biomarker discovery. It reinforces the value of integrating multiple detection models and adjusting preprocessing thresholds to uncover biologically meaningful patterns in complex, high-dimensional data.

1 | Background

1.1 DNA Methylation and Epigenetics

Epigenetics refers to heritable changes in gene function that do not involve alterations to the underlying DNA sequence. Among the various epigenetic mechanisms, DNA methylation is one of the most extensively studied and functionally significant. DNA methylation typically involves the covalent addition of a methyl group to the 5' carbon of cytosine residues within CpG dinucleotides, forming 5-methylcytosine (5mC). These CpG sites are often clustered in genomic regions known as CpG islands, commonly located near gene promoters [7].

In mammalian cells, DNA methylation is established and maintained by DNA methyltransferases (DNMTs), primarily DNMT1, DNMT3A, and DNMT3B. It plays a crucial role in several physiological processes including embryonic development, genomic imprinting, X-chromosome inactivation, and suppression of transposable elements [9, 24]. In the context of gene regulation, promoter hypermethylation is generally associated with transcriptional silencing, whereas gene body methylation may correlate with active transcription [18].

In cancer, DNA methylation patterns are profoundly disrupted. Aberrant hypermethylation of tumor suppressor gene promoters and global hypomethylation of repetitive elements and oncogenes are commonly observed [13, 15]. These epigenetic alterations can arise early during carcinogenesis, making them attractive candidates for biomarker discovery. Because methylation changes are chemically stable and detectable in small samples (e.g., blood, biopsies), they hold promise for early cancer diagnostics, prognostics, and therapy selection.

Advances in high-throughput platforms, particularly Illumina's HumanMethylation450 and EPIC (850k) arrays, have enabled genome-wide profiling of DNA methylation at single-CpG resolution across hundreds of thousands of loci [6]. These technologies facilitate the identification of not only consistent differential methylation but also stochastic, outlier-level alterations that may be functionally relevant. This study focuses on the latter—methylation outliers—as potential early indicators of cancer-related field defects.

1.2 Field Defects in Cancer and Methylation Outliers

The concept of field defects—also called "field cancerization"—was first proposed in the 1950s by Slaughter et al. [23], describing histologically normal tissue surrounding tumors that may carry molecular abnormalities predisposing to malignancy. This concept has since been extended into the molecular and epigenetic domain, where field defects refer to subclinical changes in gene expression, chromatin structure, or DNA methylation that may signal early carcinogenic processes [11].

In the epigenetic context, the field defect hypothesis suggests that even tissue appearing histologically normal may contain subtle but widespread epigenetic alterations. Among the most informative of these are DNA methylation outliers — CpG sites whose methylation values in individual samples deviate significantly from the population norm. These stochastic deviations are thought to reflect early clonal expansions or epigenetic instability associated with aging, environmental exposure, or genomic stress [14].

Importantly, such outliers may not shift the average methylation level of a population and can therefore be missed by traditional differential methylation analysis. However, their presence may reflect key early events in carcinogenesis. Teschendorff et al. [25] showed that DNA methylation outliers are significantly enriched in morphologically normal breast tissue of individuals who later developed breast cancer. This finding supports the notion that epigenetic field defects, observable as methylation outliers, can serve as early molecular markers of cancer predisposition.

Outlier-based approaches thus offer a promising complement to average-based comparisons, allowing for the detection of rare but potentially impactful epigenetic events. Identifying such CpG outliers in normal or adjacent-normal tissue could provide early biomarkers for cancer detection or risk stratification, especially when combined with tissue-specific context and pathway-level interpretation.

1.3 Challenges in Detecting Methylation Outliers

The identification of methylation outliers presents a set of unique biological and computational challenges, distinct from those associated with conventional differential methylation analysis. While traditional methods focus on population-level mean differences in methylation between groups, outlier detection seeks to uncover CpG sites that exhibit extreme methylation values in a subset of samples, often in a stochastic or sample-specific manner [5].

From a statistical standpoint, these outliers are rare events embedded in a high-dimensional feature space, often comprising more than 450,000 CpG sites per sample. This high dimensionality exacerbates the "curse of dimensionality," reducing the effectiveness of distance-based or density-based algorithms without prior feature selection or dimensionality reduction [4]. Additionally, methylation distributions are often non-Gaussian and exhibit strong heteroscedasticity, particularly at extreme beta values near 0 or 1. This violates the assumptions of many parametric statistical methods and motivates the transformation of beta-values to M-values [12].

Biologically, methylation outliers are often sparse and context-dependent. A CpG site may act as an outlier in one tissue type but not another, and its functional relevance depends heavily on genomic location (e.g., promoter vs. intergenic). Moreover, methylation variability may arise not only from true biological differences but also from technical artifacts such as batch effects, DNA quality, and probe cross-reactivity [28].

Preprocessing decisions—including normalization method, variance filtering thresholds, and the handling of missing values—can dramatically affect outlier detection outcomes. For example, overly aggressive filtering may discard informative CpGs with modest but relevant variability, while insufficient filtering may amplify noise. Therefore, careful consideration of preprocessing pipelines is essential for preserving biological signal and avoiding false positives or negatives in outlier calls.

These challenges necessitate robust, flexible algorithms and thoughtful experimental design to reliably detect methylation outliers that are biologically meaningful and reproducible across datasets.

1.4 Outlier Detection in Machine Learning

Machine learning offers robust approaches for identifying outliers in high-dimensional, noisy, and heterogeneous data such as DNA methylation profiles. Outlier detection methods are particularly suited to this context because cancer-related methylation alterations often occur in a sparse, stochastic manner, and may affect only a small number of samples or CpG sites. Unlike classical differential analysis, which targets population-wide effects, outlier-based models can reveal rare yet biologically meaningful events that deviate from normal methylation patterns.

Several algorithms have been developed for unsupervised and semi-supervised outlier detection, each with strengths and limitations. Below is an overview of the primary techniques employed in this thesis:

- **K-Nearest Neighbors (KNN):** A distance-based method that identifies outliers by computing the average distance from a given sample to its k nearest neighbors in the feature space. Samples with unusually large distances are flagged as outliers. While simple and interpretable, KNN is sensitive to the curse of dimensionality and assumes uniform data density.

- **Local Outlier Factor (LOF):** A density-based approach that measures the local deviation of a data point relative to its neighbors [8]. LOF excels at detecting local anomalies in heterogeneous data, making it suitable for uncovering context-dependent methylation outliers that may not be globally extreme.

- **One-Class SVM (OC-SVM):** A semi-supervised method that learns a soft boundary around normal data in high-dimensional space [22]. OC-SVM uses a radial basis function kernel to map input data into a higher-dimensional space, enabling separation of inliers and outliers with a maximal margin. It is well-suited for scenarios where only "normal" training data is available.

- **Isolation Forest:** An ensemble-based algorithm that isolates anomalies by recursively partitioning the feature space using randomly selected attributes and split values [20]. Anomalies require fewer splits to isolate and thus receive higher anomaly scores. Isolation Forest is computationally efficient and performs well in high-dimensional, sparse datasets like methylation matrices.

- **Z-Score Thresholding:** A statistical baseline method in which CpG values are transformed into Z-scores based on their deviation from the mean, measured in units of standard deviation. CpG sites exceeding a specified threshold are labeled as outliers. While computationally simple and interpretable, Z-score assumes normally distributed data and lacks adaptability to complex data geometry.

Each algorithm brings trade-offs between computational complexity, interpretability, and sensitivity to global versus local anomalies. For instance, LOF and OC-SVM are more sensitive to local patterns, while Isolation Forest and Z-score emphasize global structure. As shown in comparative reviews [29], algorithm performance varies widely depending on data dimensionality, noise, and outlier definition. Therefore, in this work, multiple algorithms were implemented and benchmarked to increase robustness and cross-validate results.

1.5 Related Work and Motivation

Epigenetic alterations, particularly DNA methylation changes, have been extensively studied as both causes and consequences of tumorigenesis. Early research in this area focused on identifying differentially methylated regions (DMRs) between cancerous and normal tissues [13]. However, these average-based methods may overlook rare but biologically relevant events such as methylation outliers—extreme values in a small subset of samples that may signal early clonal expansion or tissue instability.

Teschendorff et al. [25] demonstrated that DNA methylation outliers are enriched in histologically normal breast tissues of individuals who later developed breast cancer, highlighting the potential of outlier-based features as early epigenetic biomarkers. Other studies have explored statistical definitions of outlier burden [14], but relatively few have examined the application of machine learning methods for systematic outlier detection in large methylation datasets.

Moreover, current literature often treats preprocessing as a fixed pipeline rather than an experimental variable. However, decisions such as whether to use beta- or M-values, the choice of variance thresholds, and methods for handling batch effects can significantly influence downstream results and biological interpretation [12]. Few studies have quantitatively compared how these parameters affect outlier detection outcomes across multiple algorithms.

This thesis builds upon these biological and computational insights while addressing several methodological gaps:

- Applying and benchmarking five unsupervised outlier detection models (KNN, LOF, Isolation Forest, One-Class SVM, and Z-score) to large-scale breast methylation data.

- Investigating the impact of different variance filtering thresholds (0.015, and 0.02) on algorithm sensitivity and outlier reproducibility.

- Characterizing both tissue-specific and shared CpG outliers across normal, adjacent-normal, and cancer samples.
- Linking consistently identified outliers to gene-level annotations and performing pathway enrichment analysis using MSigDB Hallmark gene sets.
- Conducting a comparative performance evaluation on a secondary benchmark dataset (Thyroid Disease) to validate algorithm behavior under known class labels.

These contributions aim to advance the interpretability and reliability of outlier detection in DNA methylation studies, ultimately supporting the development of more sensitive diagnostic tools for cancer risk assessment and early intervention.

2 | Materials and Methods

2.1 Dataset Acquisition and Description

To conduct a robust analysis of DNA methylation outliers in the context of breast cancer, a systematic dataset selection process was undertaken using the NCBI Gene Expression Omnibus (GEO) repository. The primary inclusion criteria for dataset selection were:

- **Platform Coverage:** Preference was given to datasets generated using the Illumina HumanMethylation450 BeadChip or the updated 850K EPIC array, both of which provide genome-wide coverage of CpG sites with sufficient resolution for outlier detection.
- **Sample Diversity:** The dataset needed to include both healthy (normal) and cancerous breast tissue samples. This allows for the comparative analysis of methylation patterns and detection of potential field defects or early carcinogenic signatures.
- **Sample Size:** To ensure statistical power and mitigate batch effects, datasets with at least 50 samples per group were prioritized.
- **Data Accessibility:** Availability of preprocessed beta values or raw `.idat` files was essential. Beta values simplify downstream analysis, whereas raw files offer full control over normalization pipelines if needed.

After evaluating several candidate datasets, including GSE51032 and GSE101961, the dataset **GSE69914** was selected as the primary data source for this thesis. This dataset meets all outlined criteria:

- **Platform:** Illumina HumanMethylation450 BeadChip.
- **Sample Composition:** 50 normal breast tissue samples and 263 breast cancer tissue samples, including subgroups such as adjacent-normal tissues, BRCA1-mutated normal tissues, and BRCA1-mutated cancers.
- **Data Format:** Processed beta-value matrices were made available through the GEO portal, with no missing values and prior normalization applied.

Sample Categories and Biological Groups

The dataset was further categorized into biologically meaningful groups based on sample metadata and clinical annotations:

- **Breast Cancer Samples:** Tumor-derived methylation profiles from 263 individuals.
- **Adjacent-Normal Samples:** Normal tissue samples taken from regions near the tumor, providing an opportunity to evaluate the epigenetic field defect hypothesis.
- **Normal Samples:** Independent healthy breast tissue samples used as baseline controls.

- **Normal-BRCA1 Samples:** Samples from individuals carrying BRCA1 mutations but without tumors.
- **Cancer-BRCA1 Samples:** Tumor samples from BRCA1 mutation carriers, potentially revealing hereditary risk patterns.

Initial Dimensionality Reduction Attempt (Correlation and PCA)

As part of the early data preparation phase, I explored various dimensionality reduction strategies to manage the high dimensionality of the dataset. One initial approach involved applying correlation filtering to reduce redundancy across CpG sites. The assumption was that highly correlated CpGs—those with correlation coefficients greater than 0.99—likely carried redundant information and could be clustered or averaged. Using this technique, I identified 13 groups of highly correlated columns, each containing between 2 to over 275 CpG sites. Three strategies were considered to reduce these groups: selecting a representative CpG, computing group-wise means or medians, and applying Principal Component Analysis (PCA) to extract the first principal component from each group.

However, upon review and guidance from Dr. Gambino, it became evident that this direction, although mathematically valid, was not suitable for the biological objective of this study. Specifically, these techniques—particularly correlation filtering and PCA—risk discarding epigenetic outliers and biologically relevant "noise," which are central to detecting field defects in cancer. Moreover, PCA transforms the CpG matrix into new principal components, thereby removing the ability to directly trace anomalous methylation signals back to specific genomic sites, which is critical for downstream gene and pathway analysis.

Although this approach was eventually discontinued, documenting it here reflects a comprehensive investigation of possible preprocessing strategies.

Analysis Objective

The primary objective of using this dataset was to identify methylation outliers—CpG sites with extreme beta values in individual samples—that may serve as early indicators of cancer development. These outliers were subsequently subjected to algorithmic detection using several machine learning models, and the biological interpretation was further refined through gene annotation and pathway enrichment analyses.

2.2 Data Preparation

The data preparation phase was critical in ensuring that the methylation data from the GSE69914 dataset was suitable for downstream analysis, especially given its high dimensionality (over 450,000 CpG sites) and the complexity of detecting biologically meaningful outliers. The following steps outline the complete preprocessing pipeline.

2.2.1 Dataset Acquisition and Preprocessing

The GSE69914 dataset was downloaded from the Gene Expression Omnibus (GEO) and contains methylation beta values across different breast tissue types: cancerous, adjacent-normal, and normal tissues. The platform used is the Illumina HumanMethylation450 BeadChip, which provides broad coverage of the methylome. The dataset had already been normalized and underwent quality control, as confirmed from its GEO metadata.

Due to memory constraints associated with large '.txt' files, the dataset was saved in '.csv' format and stored in Google Drive for direct access via the `gdown` library in Google Colab. This optimization significantly improved data loading performance and reduced RAM usage.

2.2.2 Conversion from Beta-values to M-values

The dataset provided beta-values representing the proportion of methylation at each CpG site across samples, ranging from 0 (completely unmethylated) to 1 (fully methylated). While beta-values are intuitive for biological interpretation, they suffer from heteroscedasticity—where the variance is not uniform across the range—which can bias downstream statistical tests and outlier detection.

To address this, beta-values were transformed into M-values using the standard logit transformation:

$$M = \log_2 \left(\frac{\beta}{1 - \beta} \right)$$

Before transformation, beta-values were clipped within the range $[10^{-6}, 1 - 10^{-6}]$ to prevent mathematical instability due to division by zero or logarithm of zero. The dataset, being large, was processed in chunks of 10,000 rows at a time to prevent RAM overflow in Google Colab. The transformation was applied only to numeric columns (excluding the `ID_REF` column), and the converted M-values were written back into a single DataFrame.

Justification for Choosing M-Values: M-values are preferred in differential methylation and variance-based studies due to their improved statistical properties. They correct for heteroscedasticity inherent in beta-values—particularly at extreme methylation levels—and better satisfy the assumptions of many machine learning and statistical models. Furthermore, they yield a more symmetric distribution across features, enhancing the effectiveness of variance thresholding and outlier detection.

2.2.3 Data Structure and Transposition

Initially, rows corresponded to CpG probes and columns to sample IDs. To align with machine learning frameworks where rows represent samples and columns represent features, the dataset was transposed. This facilitated downstream labeling and analysis.

2.2.4 Sample Grouping and Labeling

To support supervised analysis and evaluation across biological contexts, samples were categorized into three primary groups:

- **Cancer** – Breast tumor tissues.
- **Adjacent-normal** – Tissues adjacent to tumors.
- **Normal** – Healthy tissues from cancer-free individuals.

Based on manual inspection of metadata, and in the absence of a sample sheet from GEO, a custom classification was constructed to label each sample appropriately. Sample IDs were manually assigned into their respective categories using information embedded in GEO file names.

An integer label column was added: 0 for normal, 1 for adjacent-normal, and 2 for cancer. Any unclassified samples were discarded to ensure consistency and analytical clarity.

2.3 Dimensionality Reduction Techniques

Due to the high dimensionality of methylation datasets—often exceeding 450,000 CpG sites—dimensionality reduction was critical to enable meaningful analysis while maintaining biological relevance. A multi-phase pipeline was employed to sequentially filter features based on statistical properties, biological knowledge, and supervised learning approaches.

2.3.1 Variance Thresholding

Variance thresholding was used as a primary filter to exclude CpG sites with low variability across samples, which are less likely to capture disease-related patterns.

Statistical Summary:

- Mean variance: 0.0231
- Median: 0.0222
- Standard deviation: 0.0102
- Minimum: 0.0016
- Maximum: 0.1571

Thresholds Evaluated:

- **0.01**: Captures sites above the 25th percentile.
- **0.015**: Near the median, capturing moderate-to-high variance.
- **0.02**: Restricts selection to the top 25% most variable CpGs.

Visualization: Figure 2.1 illustrates the distribution of variance across CpGs, with vertical lines showing threshold levels.

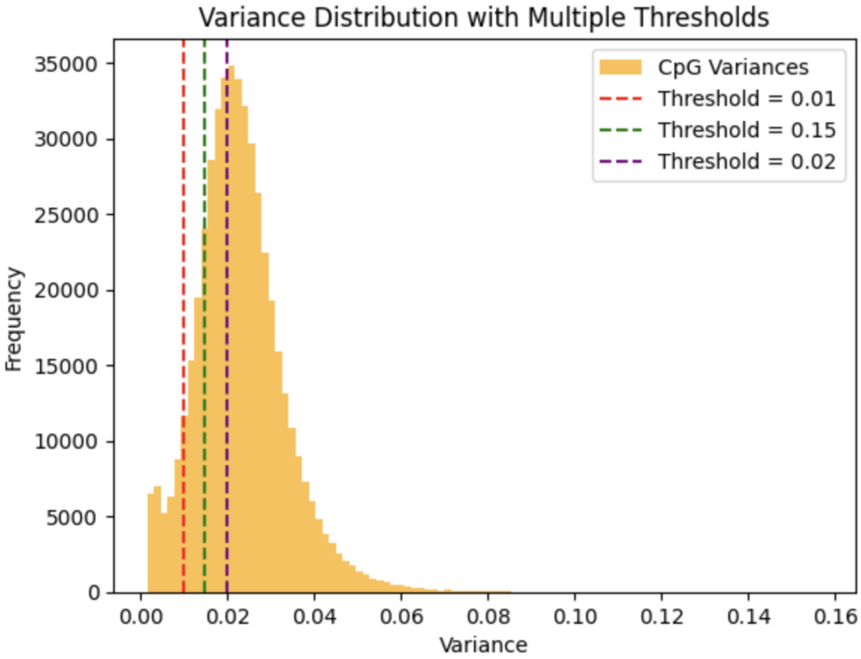

Figure 2.1: Distribution of CpG site variances. Thresholds at 0.01, 0.015, and 0.02 are overlaid to illustrate selection sensitivity.

Comparative Evaluation:

- At 0.01, too many features were retained, resulting in memory overload during model training.
- At 0.02, 291,830 features were retained, optimizing speed but risking biological loss.
- At 0.015, 390,742 CpGs were retained—balancing coverage and feasibility.

Final Decision: A threshold of **0.015** was chosen for all downstream analyses as it provided a biologically meaningful subset while remaining computationally feasible in a Colab environment.

2.3.2 Biological Relevance-Based Filtering

Another strategy I implemented for dimensionality reduction was the selection of CpG sites based on biological relevance, specifically those previously implicated in breast cancer. This approach aims to prioritize features that are not only statistically informative but also grounded in prior biomedical findings.

Method: I curated a list of CpG sites associated with known breast cancer-related genes from the literature. These include genes involved in tumor suppression (e.g., *BRCA1*, *RASSF1A*), hormone receptor signaling (e.g., *ESR1*, *ERBB2*), DNA damage response (e.g., *TP53*, *HIF1A*), and others. A total of 70 CpG sites were selected from studies that examined methylation signatures in breast cancer and from reputable genomic databases.

Sources:

- Yan et al. [27] employed CpG island arrays to identify aberrant methylation in *BRCA1* and *RASSF1A*.
- Feng et al. [16] studied methylation in relation to hormone receptor status, highlighting *ESR1* and *ERBB2*.
- Widschwendter & Jones [26] reviewed the role of DNA methylation in breast carcinogenesis.
- Rice et al. [21] demonstrated that methylation in the *BRCA1* promoter reduces its expression in sporadic breast cancer.
- Li et al. [19] linked CpG methylation in *HIF1A* to its elevated expression and hypoxia-related responses in tumors.

Databases Consulted:

- The Cancer Genome Atlas (TCGA) [1].
- Gene Expression Omnibus (GEO) [2].
- Illumina 450K/EPIC BeadChip CpG annotations [3].

This biologically guided filtering allowed for an interpretable and hypothesis-driven feature subset, forming a bridge between statistical analysis and molecular oncology.

2.3.3 Univariate Statistical Feature Selection – ANOVA (F-test)

The Analysis of Variance (ANOVA) F-test is a statistical method used to assess whether there are significant differences in the means of a numerical variable (in this case, methylation levels at each CpG site) across multiple categorical groups (normal, adjacent-normal, and cancer tissue types). It is particularly useful when evaluating features individually to determine how strongly each CpG site discriminates between sample classes.

Goal: To retain CpG sites that show statistically significant differences in methylation levels across the tissue groups, identifying those that may serve as potential biomarkers or indicators of early cancer progression.

Method: I employed the `SelectKBest` method from `scikit-learn` with the `f_classif` scoring function. This function computes the ANOVA F-value for each feature, measuring the ratio of variance between the groups to the variance within the groups. CpG sites with higher F-values are more likely to show distinct methylation profiles across the tissue types.

Interpretation: A high F-statistic and a low p-value indicate that the CpG site exhibits significant differences in methylation across the groups, warranting retention for further analysis. This method assumes that the values within each group are normally distributed and that variances are approximately equal (homoscedasticity), although it remains robust under mild violations.

Implementation Details:

- CpG features were ranked by their F-statistics.
- The top 1,000 CpG sites were selected based on their scores.
- This subset was used for downstream machine learning and biological interpretation tasks.

Advantages:

- Simple, fast, and interpretable.
- Highlights CpGs most distinctively different across conditions.

Limitations:

- Sensitive to outliers and assumptions of normality.
- Evaluates features independently, ignoring multivariate interactions.

2.3.4 Recursive Feature Elimination (RFE)

Recursive Feature Elimination (RFE) was applied as a supervised, model-based feature selection technique to further reduce the dimensionality of the methylation dataset after applying variance thresholding. RFE is particularly effective when the objective is to retain features that are most informative for classification tasks, while discarding those that contribute minimally to predictive performance.

Concept: RFE operates by recursively training a model and pruning the least important features based on the model's internal importance scores. At each iteration, the model ranks all features, and a defined number of the lowest-ranking features are eliminated. This process continues until the desired number of features remains.

Model and Configuration: In this analysis, a Random Forest classifier was used as the estimator due to its ability to capture nonlinear relationships, robustness to overfitting, and built-in mechanism for evaluating feature importance. RFE was configured to select the top 100 most relevant CpG sites. A step size of 500 features per iteration was used to make the elimination process computationally feasible, given the large number of features remaining after variance filtering.

Outcome: The dimensionality of the dataset was successfully reduced from over 290,000 CpG sites to just 100. These retained features represent the most discriminative CpG sites with respect to the three tissue classes: normal, adjacent-normal, and cancer. The resulting dataset retained the original sample indexing and was fully compatible with subsequent classification and outlier detection analyses.

Advantages:

- Produces a compact and highly informative feature set tailored to the classification problem.

- Offers model-driven feature relevance scores, avoiding reliance on arbitrary statistical thresholds.
- Particularly well-suited for high-dimensional biological datasets.

Limitations:

- Computational cost increases with dataset size, particularly in early iterations.
- Feature rankings can vary depending on the base estimator and presence of correlated features.
- Does not inherently prioritize biological interpretability, requiring downstream annotation.

Overall, RFE contributed to the construction of a biologically and statistically robust feature set, ensuring the retention of CpG sites most critical to class discrimination in methylation-based breast cancer analysis.

2.3.5 Excluded Methods: Correlation Filtering and PCA

Two dimensionality reduction methods were initially explored but later excluded:

- **Correlation Filtering:** Groups of highly correlated CpGs were identified and reduced by retaining one representative per group. However, this method was abandoned based on expert feedback because it risks eliminating the specific CpG harboring an epigenetic outlier—crucial to our study's objective.
- **Principal Component Analysis (PCA):** PCA was not used because it transforms original CpG values into principal components, making it impossible to trace back specific CpGs responsible for variation. This loss of interpretability is incompatible with our biomarker discovery goals.

Summary: The dimensionality reduction pipeline combined unsupervised filtering (variance), biological relevance (literature-derived CpGs), univariate statistics (ANOVA), and model-based feature selection (RFE). Together, these steps ensured that retained features were computationally manageable, statistically discriminative, and biologically relevant for downstream analysis.

2.4 Outlier Detection Methods

Identifying DNA methylation outliers is essential for uncovering early epigenetic disruptions that may signal the onset of cancer. In this study, multiple machine learning (ML) algorithms were employed to flag CpG sites exhibiting significant deviations from typical methylation patterns across three tissue categories: cancer, adjacent-normal, and normal. The multi-algorithm strategy enhanced robustness, allowed cross-validation of results, and enabled detection of both global and local anomalies.

The dataset was filtered using a variance threshold of 0.015 and some other methods which were mentioned before, balancing biological richness with computational feasibility. Outlier detection was then performed using five distinct methods: Isolation Forest, Local Outlier Factor (LOF), One-Class SVM, Z-score, and K-Nearest Neighbors (KNN). Each method provides a different perspective on what constitutes an outlier, varying in their sensitivity to global vs. local deviations and their assumptions about data structure.

2.4.1 Isolation Forest

Isolation Forest is a tree-based algorithm optimized for anomaly detection in high-dimensional data. It isolates anomalies by recursively partitioning data using randomly chosen features and split values. Points that require fewer splits to be isolated are flagged as outliers.

Key Advantages:

- Effective for global outlier detection.
- Scales well with large, high-dimensional datasets such as methylation profiles.
- Non-parametric: No assumptions about data distribution.

Hyperparameter Tuning: I tested multiple values for `n_estimators` (100, 125, 150, 175, 200). Higher values increased precision and stability, while lower values were faster and suitable for preliminary scans.

Observations: Isolation Forest consistently flagged certain CpG sites (e.g., `cg19374752`) across all tissues and configurations. Its ability to detect outliers based on tree depth proved valuable for identifying globally aberrant methylation sites.

2.4.2 Local Outlier Factor (LOF)

LOF is a density-based method that compares the local density of a point with its neighbors. Outliers are those with substantially lower density.

Key Advantages:

- Captures local deviations that may not appear anomalous globally.
- Adaptable to varying neighborhood structures.

Hyperparameter Tuning: LOF was run using `n_neighbors` values of 5, 10, 15, and 20. Lower values captured subtle anomalies; higher values prioritized robustness.

Observations: LOF identified CpG sites exhibiting localized irregular methylation. It was especially useful in detecting anomalies in adjacent-normal tissues, supporting the field defect hypothesis.

2.4.3 One-Class Support Vector Machine (OC-SVM)

OC-SVM constructs a boundary around the majority of the data using kernel functions. Points falling outside this boundary are flagged as outliers.

Key Advantages:

- Suitable for high-dimensional data.
- Creates a global boundary around normal instances.

Hyperparameter Tuning: The `nu` parameter was varied across 0.01, 0.05, 0.15, and 0.2. Lower `nu` values result in strict detection; higher values detect broader anomaly distributions.

Observations: OC-SVM consistently flagged core CpG sites across tissue types. However, it was less sensitive to tissue-specific deviations, suggesting its strength lies in identifying common epigenetic disruptions.

2.4.4 Z-Score Based Detection

This statistical approach flags CpG sites as outliers based on their standardized distance (Z-score) from the mean methylation value.

Key Advantages:

- Simple and interpretable.
- Effective as a baseline model.
- Useful for datasets with approximately normal distribution.

Thresholds Used: 1.5 to 2.0 in 0.1 increments. Lower thresholds captured minor deviations, while higher thresholds focused on extreme methylation shifts.

Observations: Z-score effectively flagged both extreme and subtle anomalies. Consistent outliers such as `cg00000622` were frequently detected, and threshold tuning provided flexible control over sensitivity.

2.4.5 K-Nearest Neighbors (KNN)

KNN was adapted for outlier detection by calculating the average distance to each point's nearest neighbors. Points with large average distances were marked as outliers.

Key Advantages:

- No assumptions on data distribution.
- Detects outliers based on proximity in CpG methylation space.

Parameter Tuning: `n_neighbors` values tested were 3, 5, 8, 10, and 12.

Observations: KNN identified both global and tissue-specific anomalies. It was particularly useful in adjacent-normal tissues, highlighting CpG sites that diverge early during transformation.

3 | Results and Discussion

3.1 Cross-Comparison of Outliers Across Tissue Types

Objective: The goal of this analysis was to identify CpG sites consistently flagged as outliers across different machine learning algorithms, hyperparameter configurations, and tissue types (cancer, adjacent-normal, and normal). This helps uncover robust epigenetic changes that may act as early indicators of cancer development.

Algorithms Used

- K-Nearest Neighbors (KNN)
- Isolation Forest
- Local Outlier Factor (LOF)
- One-Class SVM
- Z-Score Based Thresholding

Shared CpG Outliers Across Tissue Types

Several CpG sites were consistently flagged as outliers across tissue types and multiple algorithms, suggesting potential roles in cancer initiation or progression. These included:

- **cg19374752:** Detected across all tissue types using KNN and consistently found in all algorithms and configurations.
- **cg00000622:** Identified by Isolation Forest, LOF, One-Class SVM, and Z-score across multiple configurations and tissues.
- **cg00000108:** Frequently flagged across LOF, Isolation Forest, and One-Class SVM.
- **cg21908886:** Shared among LOF, One-Class SVM, and Z-score in all tissue groups.
- **cg04242728:** Detected mainly by One-Class SVM and Z-score in adjacent-normal and cancer tissues.

Summary of Shared and Tissue-Specific Outliers

Table 3.1 details the detection of specific CpG sites as outliers across tissue types and algorithm configurations. Each row corresponds to a CpG site identified in at least one configuration. Algorithms and their parameter values (such as KNN neighborhood size, Isolation Forest tree count, and Z-score thresholds) are fully enumerated to provide clear traceability. CpG sites like `cg19374752` and `cg00000622` were detected across nearly all algorithms and configurations in all three tissue types, indicating their potential as robust biomarkers. Conversely, sites such as `cg23006567` and `cg12492087` were more specific to one tissue type, suggesting potential context-dependent methylation anomalies.

Table 3.1: Outlier CpG Sites Detected Across Algorithms, Configurations, and Tissue Types

CpG Site	**Cancer**	**Adj-Normal**	**Normal**	**Algorithms and Configurations**
cg19374752	✓	✓	✓	KNN(n=3,5,8,10,12), LOF(n=5,10,15,18,20), OC-SVM(nu=0.01,0.05,0.15,0.2), Z-Score(1.5,1.6,1.7,1.8,1.9,2.0), IF(n_estimators=100,125,150,175,200)

cg00000622	✓	✓	✓	LOF(n=5,10,15,18,20),
				OC-SVM(nu=0.01,0.05,0.15,0.2),
				Z-Score(1.5,1.6,1.7,1.8,1.9,2.0)
				IF(n_estimators=100,125,150,175,200),

cg00000108	✓	✓	✓	LOF(n=15,18,20),
				OC-SVM(nu=0.01,0.05,0.15,0.2)
				IF(100,125,150,175),

cg21908886	✓	✓	✓	LOF(n=5,10), OC-SVM(nu=0.01,0.05,0.15,0.2), Z-Score(1.5,1.6,1.7,1.8,1.9,2.0)
cg04242728	✓	✓	–	OC-SVM(nu=0.01,0.05,0.15,0.2), Z-Score(1.5,1.6,1.7,1.8,1.9,2.0)
cg00001099	✓	–	–	OC-SVM(nu=0.01,0.05,0.15,0.2), Z-Score(1.5,1.6,1.7,1.8)
cg13456241	✓	–	–	OC-SVM(nu=0.01,0.05,0.2), Z-Score(1.5,1.6,1.7)
cg23006567	✓	–	–	OC-SVM(nu=0.05,0.15,0.2)
cg11731596	–	✓	–	KNN(n=3,5,8,10), Z-Score(1.5,1.6,1.7)
cg17939805	–	✓	–	KNN(n=5,8,10), Z-Score(1.5,1.6,1.7,1.8)
cg12492087	–	–	✓	OC-SVM(nu=0.15,0.2), Z-Score(1.5,1.6,1.7)
cg16085649	–	–	✓	OC-SVM(nu=0.15,0.2), Z-Score(1.5,1.6,1.7,1.8)

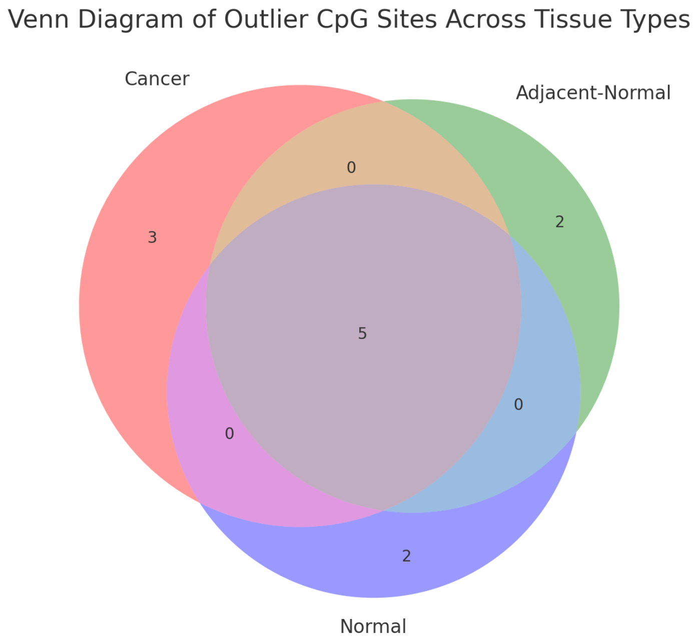

Figure 3.1: Venn diagram showing shared and unique CpG site outliers across cancer, adjacent-normal, and normal tissues.

To visually represent the overlap among cancer, adjacent-normal, and normal tissues, a Venn diagram (Figure 3.1) was generated. It shows CpG sites that are uniquely or jointly identified across tissue types. Sites like `cg19374752` lie at the intersection of all three regions, while tissue-specific outliers such as `cg11731596` (adjacent-normal) or `cg16085649` (normal) reside in non-overlapping regions. This visualization underscores both the shared and divergent epigenetic alterations present in early cancer progression.

Interpretation and Conclusions

Interpretation: CpG sites detected consistently across algorithms and tissues (such as `cg19374752` and `cg00000622`) are strong biomarker candidates. In contrast, tissue-specific sites like `cg11731596` (adjacent-normal only) may reflect early epigenetic field defects.

Algorithm Sensitivity: The results varied depending on the algorithm and hyperparameter choice. KNN and Isolation Forest detected more cancer-related outliers, while Z-score and OC-SVM identified subtle deviations in adjacent-normal tissues.

Conclusion: This cross-comparison provides biological and methodological insight, reinforcing the importance of using multiple algorithms and tissue types to detect robust and early cancer-associated epigenetic changes.

3.2 Validation of Matched Samples

Objective

The aim of this step was to investigate whether patient identifiers (IDs) were available in the GSE69914 dataset to support paired analysis between breast cancer samples and their corresponding cancer-adjacent normal tissues. The presence of matching IDs is crucial for minimizing biological variability due to inter-individual differences and for testing hypotheses such as the epigenetic field defect.

Search for Patient IDs

A thorough review of the dataset's metadata and supplementary files was conducted to locate patient-specific identifiers that could be used to pair cancer and adjacent-normal samples. Key files examined included:

- GEO series matrix files
- Supplementary raw data files (e.g., `GSM1712772_BCFD400_Raw.txt`)
- Platform annotation files (GPL13534)

Despite this extensive search, no explicit patient IDs or sample pairing indicators were found. None of the available files contained structured information that could be used to associate samples from the same patient. The filenames and metadata fields focused primarily on sample-level annotations (e.g., tissue type, platform ID), without linking them to individual subjects.

Conclusion

The absence of patient-specific identifiers in the GSE69914 dataset precluded any possibility of conducting paired analysis between cancer and cancer-adjacent normal tissues. While the study design implied a match between certain samples (e.g., based on tissue proximity), no metadata could reliably support such alignment.

As a result, all analyses in this thesis proceeded using group-based comparisons rather than patient-matched pairs. This limitation is important to note, especially for hypotheses involving early field defects, as paired data would have offered stronger statistical power and biological interpretability.

Future work may benefit from integrating external clinical metadata or contacting the dataset authors directly to obtain patient-matching information, if available.

3.3 Methylation Patterns and Pathway Enrichment

Overview and Objective

This section explores differential methylation patterns between normal, adjacent-normal, and cancer breast tissues using data from the GSE69914 dataset. The goal was to identify statistically significant CpG sites, annotate them with genomic features, and explore biological implications through pathway enrichment.

3.3.1 Statistical Identification of Significant CpG Sites

Comparative Analysis Using T-tests

To detect methylation alterations associated with cancer progression, independent two-sample t-tests (with Welch's correction) were conducted between:

- Normal vs Cancer
- Normal vs Adjacent-Normal
- Adjacent-Normal vs Cancer

Pre-filtering: Prior to statistical testing, CpG sites were filtered based on two criteria to ensure analytical robustness and biological relevance. First, a variance threshold of 0.01 was applied to eliminate sites with negligible variability across samples, which are unlikely to carry informative signal. Second, a fold-change cutoff of 0.1 was enforced to retain only those CpG sites exhibiting substantial average methylation differences between groups. This combination focuses the analysis on CpG sites with both statistically and biologically meaningful variation.

Multiple Testing Correction: Given the large number of CpG sites tested (often exceeding 400,000), unadjusted p-values would result in an unacceptably high false positive rate. To address this, the Benjamini-Hochberg procedure was employed to control the false discovery rate (FDR). CpG sites were retained for further analysis only if their adjusted p-values satisfied $p_{\text{adj}} < 0.05$, ensuring that the identified signals were unlikely to arise by chance alone.

Results Summary

- **Normal vs Cancer:** 91,954 significant CpG sites
- **Normal vs Adjacent-Normal:** 1,348 significant CpG sites
- **Adjacent-Normal vs Cancer:** 101,330 significant CpG sites

CpG Site Overlap

- Shared across all groups: **1,089** CpGs
- Shared between Normal vs Cancer and Normal vs Adjacent: **1,250**
- Shared between Adjacent vs Cancer and Normal vs Adjacent: **1,104**
- Unique to Normal vs Cancer: **27,646**
- Unique to Adjacent vs Cancer: **37,168**
- Unique to Normal vs Adjacent: **83**

To better visualize the statistical landscape of significant CpG sites, several figures were generated. Figure 3.2 shows the number of shared and unique CpGs across comparisons, emphasizing the high volume of unique CpGs between cancerous and other tissues. Figure 3.3 further confirms these relationships with a Venn diagram that highlights overlaps in significant CpG sites. Finally, Figure 3.4 illustrates the p-value distributions, where both cancer-related comparisons show strong statistical separation, while normal vs adjacent-normal shows more moderate divergence.

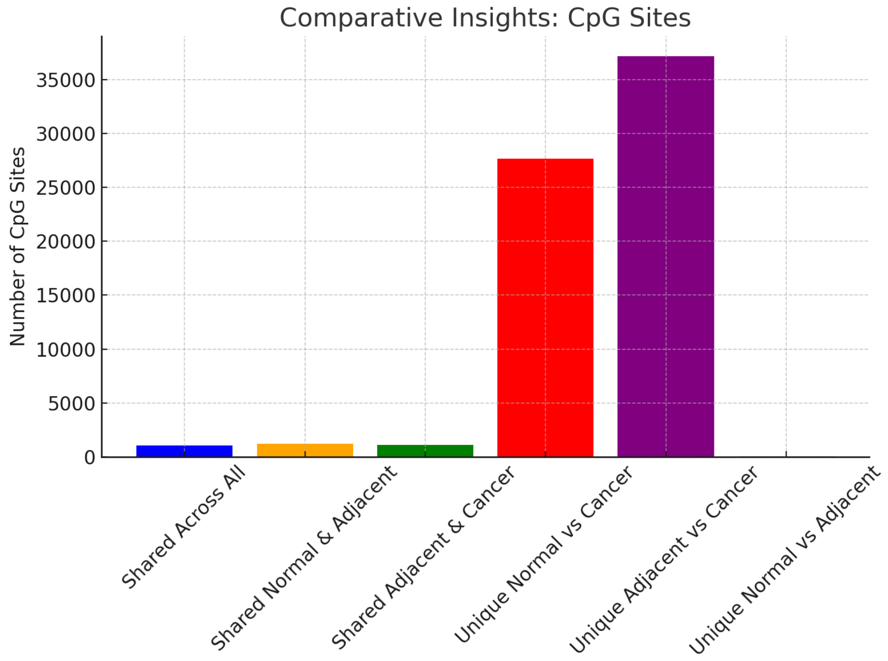

Figure 3.2: Comparative bar plot showing the number of significant CpG sites that are shared or unique across pairwise group comparisons. Most CpG sites are uniquely significant in the Cancer vs Adjacent-Normal and Cancer vs Normal comparisons, suggesting stronger methylation shifts in tumor progression. Only 1,089 CpGs are shared across all three comparisons.

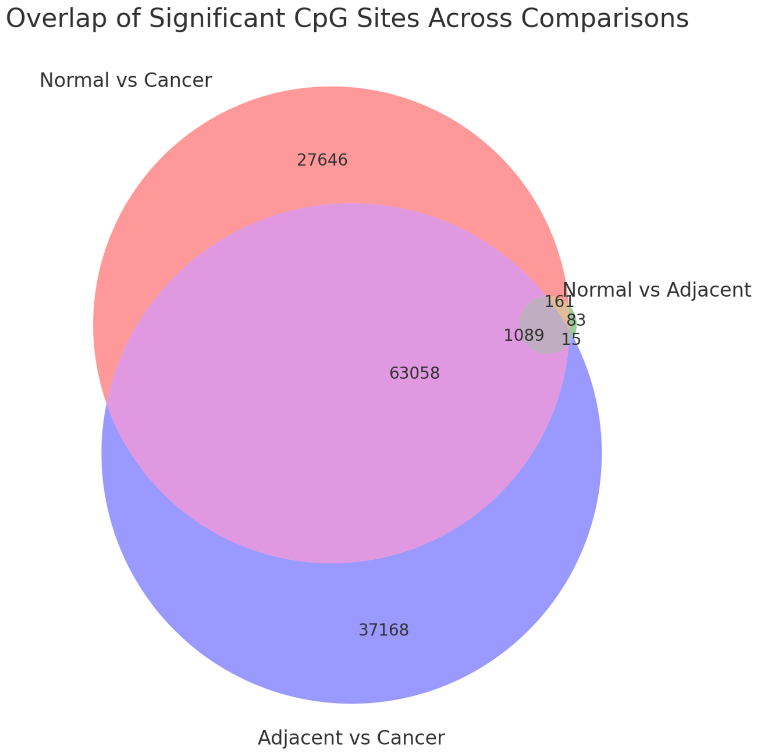

Figure 3.3: Venn diagram visualizing overlap of significant CpG sites among the three pairwise comparisons. The largest shared region is between Normal vs Cancer and Adjacent-Normal vs Cancer, consistent with similar methylation patterns between adjacent and normal tissues.

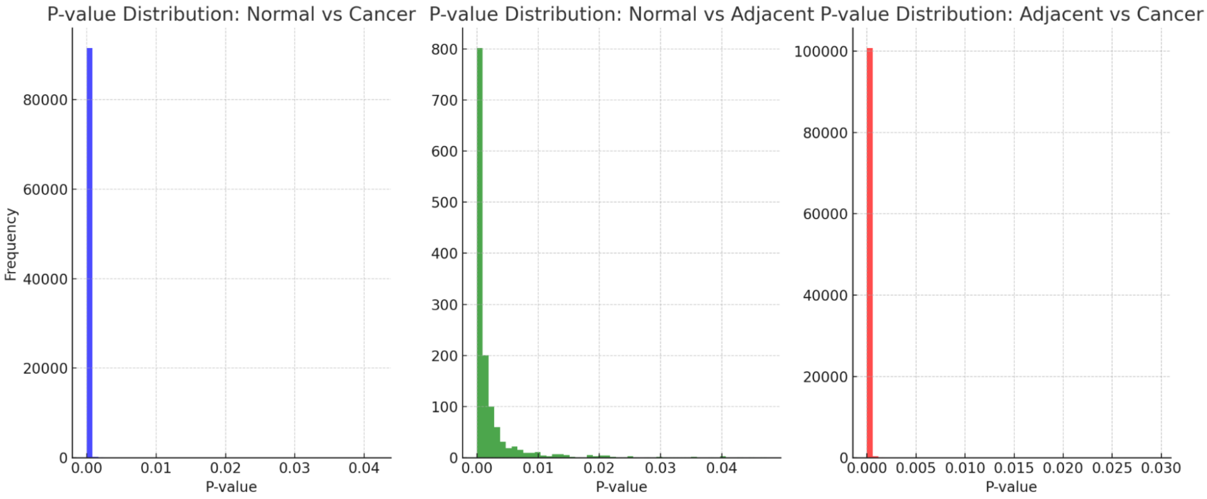

Figure 3.4: P-value distributions for each pairwise group comparison. Left: Normal vs Cancer shows a high concentration of CpG sites with extremely low p-values, indicating strong methylation differences. Middle: Normal vs Adjacent-Normal shows a less skewed distribution with fewer significant changes. Right: Adjacent-Normal vs Cancer again shows many strongly significant sites, supporting progressive methylation changes.

3.3.2 Annotation of Significant CpG Sites

Objective: To provide biological context to the list of statistically significant CpG sites by linking them to gene annotations and known regulatory genomic features. This aids in determining which methylation alterations are most likely to affect gene regulation and, consequently, cancer progression.

Annotation Methodology: Following the identification of significant CpG sites across three pairwise group comparisons (Normal vs Cancer, Adjacent vs Cancer, Normal vs Adjacent), each site was annotated using genomic annotation files. These annotations mapped each CpG site to associated genes, promoter regions, gene bodies (exons, introns), and regulatory features such as CpG islands, shores, shelves, and open seas.

- CpG sites were matched with known gene identifiers (e.g., `UCSC_RefGene_Name`, `GeneRegion_Feature`).
- Features such as promoter-associated CpGs were prioritized as more likely to influence gene transcription.
- The annotated CpGs were categorized based on their significance across comparisons and their genomic context.

Filtering for Biological Relevance: To refine the annotation, only CpG sites located within promoter regions, first exons, or known regulatory elements were retained for further analysis. This step eliminated intergenic or low-impact regions that are less likely to influence gene expression. The resulting annotated dataframes were structured with 44 columns, including statistical metrics and annotation descriptors.

- **Normal vs Cancer:** 91,954 significant CpGs
- **Adjacent vs Cancer:** 101,330 significant CpGs
- **Normal vs Adjacent:** 1,348 significant CpGs

Distribution of Adjusted P-values (Post-Annotation) To assess the statistical strength of the annotated CpGs, histograms of adjusted p-values were plotted:

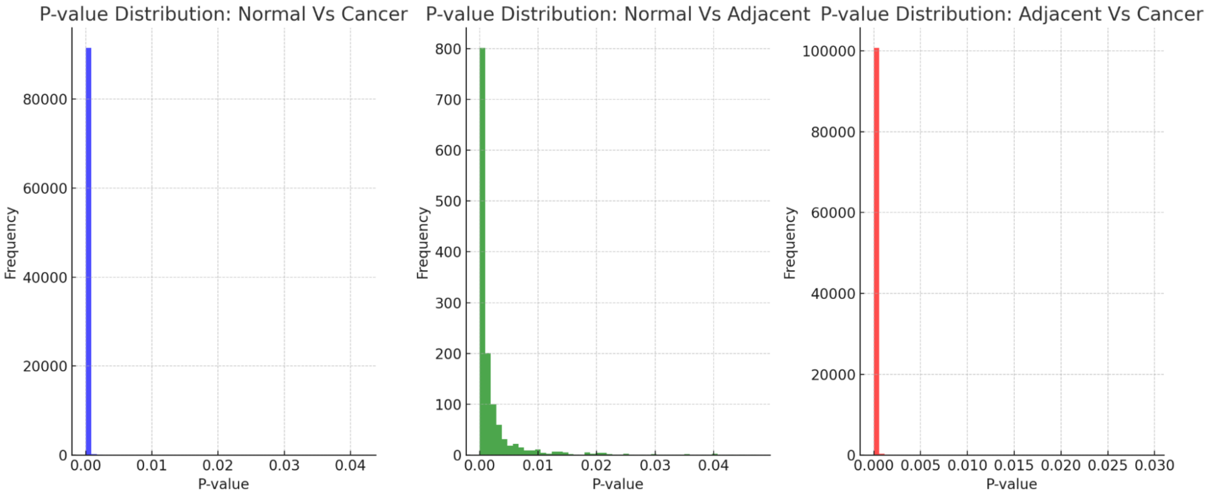

Figure 3.5: Distribution of adjusted p-values for significant CpG sites across three comparisons: Normal vs Cancer (blue), Normal vs Adjacent (green), and Adjacent vs Cancer (red). Most p-values in cancer-related comparisons are clustered near 0, suggesting strong statistical evidence of methylation changes.

Gene-Level Comparative Insights (Filtered) Following annotation, each CpG site was linked to one or more genes. The resulting gene lists were compared across tissue comparisons:

- **Shared Genes Across All Comparisons:** 644 genes
- **Unique Genes (Normal vs Cancer):** 2,193 genes
- **Unique Genes (Adjacent vs Cancer):** 1,515 genes
- **Unique Genes (Normal vs Adjacent):** 7 genes

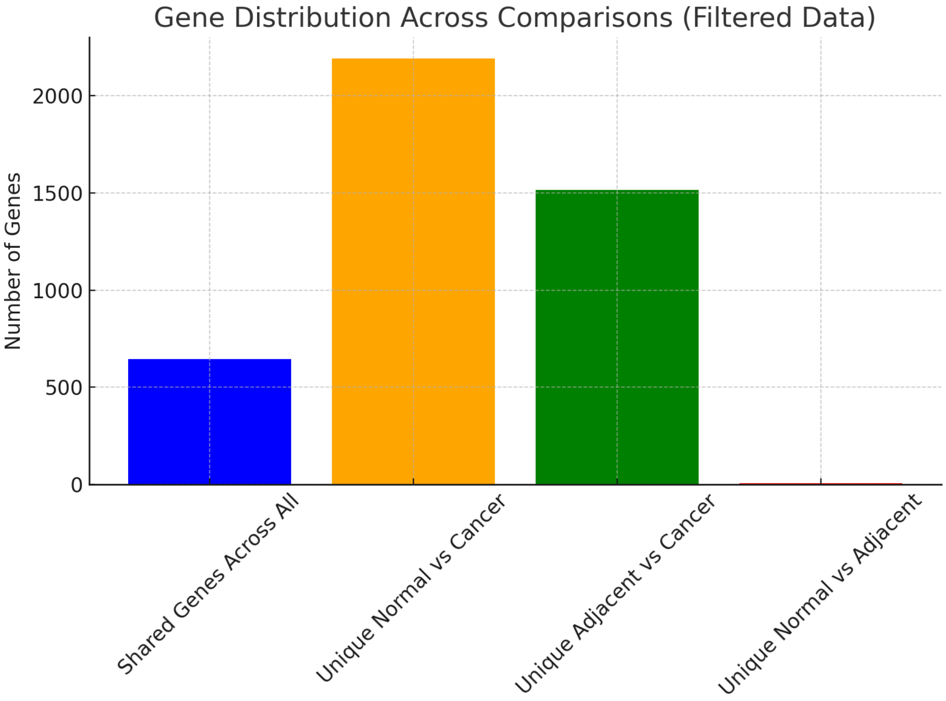

Figure 3.6: Bar chart showing the number of unique and shared genes annotated from significant CpG sites across comparisons. Normal vs Cancer and Adjacent vs Cancer share the most overlap, while Normal vs Adjacent contributes the fewest unique genes.

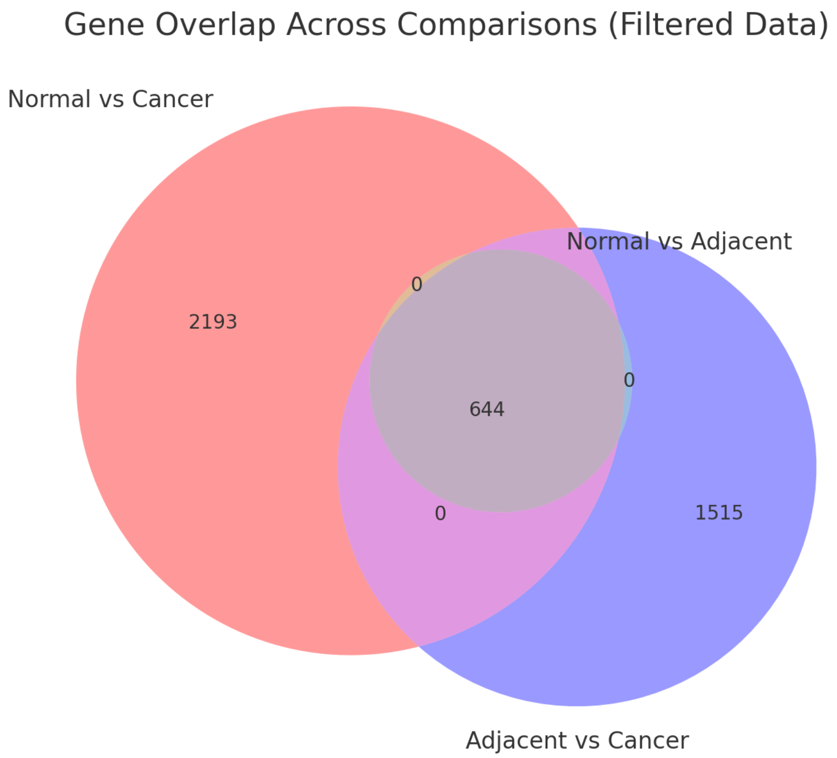

Figure 3.7: Venn diagram illustrating gene overlap across tissue comparisons (filtered data). Most genes are unique to either Normal vs Cancer or Adjacent vs Cancer, while 644 genes are common to all three.

Interpretation of Results:

- **High-Impact Genes:** Genes identified across all tissue comparisons (e.g., *BRCA1*, *RASSF1A*) may represent core methylation biomarkers involved in tumor suppression or oncogenesis.
- **Tissue-Specific Regulation:** Unique gene lists per comparison suggest stage- or context-specific methylation patterns. For example, the 2,193 genes unique to the Normal vs Cancer contrast could be directly involved in tumor onset, while the 1,515 unique to Adjacent vs Cancer may reflect progressive changes from pre-malignant to malignant.
- **Minimal Signals in Normal vs Adjacent:** Only 7 unique genes were observed in the Normal vs Adjacent comparison, reinforcing the hypothesis that field defect changes are subtle but present.

Conclusion: The annotation of significant CpG sites with genomic features and gene contexts has revealed patterns of methylation change with strong statistical and biological relevance. Shared genes across comparisons highlight conserved mechanisms, while unique genes suggest stage-specific events. These findings will inform pathway enrichment and downstream biomarker validation efforts.

3.3.3 Pathway Enrichment Analysis (GSEA)

Objective: To investigate the biological significance of differentially methylated genes by identifying enriched pathways and molecular processes, thereby linking epigenetic alterations to potential functional consequences in cancer development.

Rationale: DNA methylation affects gene expression and cellular behavior. Therefore, understanding which biological pathways are enriched among significantly methylated genes provides insights into the molecular mechanisms altered in cancer progression. Gene Set Enrichment Analysis

(GSEA) is a well-established method that detects coordinated changes in predefined gene sets rather than analyzing genes individually, increasing statistical power and biological interpretability.

Methodology:

- **Gene Set Preparation:** Gene symbols associated with significant CpG sites (identified through differential methylation analysis) were extracted for each of the three pairwise comparisons:
  - Normal vs Cancer
  - Adjacent-Normal vs Cancer
  - Normal vs Adjacent-Normal

  These gene lists were compiled from annotated CpG sites that passed the adjusted p-value threshold ($p_{\text{adj}} < 0.05$).
- **Gene Set Enrichment Tool:** Enrichment analysis was performed using the MSigDB (Molecular Signatures Database) Hallmark gene set collection. The Hallmark sets consist of 50 curated biological pathways that represent well-defined biological states or processes and show coherent expression patterns across multiple datasets.
- **Enrichment Algorithm:** GSEA was executed using a hypergeometric test (or Fisher's exact test) to determine whether the overlap between input gene lists and predefined gene sets was greater than expected by chance. The enrichment score was corrected for multiple testing using the Benjamini-Hochberg procedure to control the false discovery rate (FDR).

Results and Interpretation:

- For the **Normal vs Cancer** comparison, enriched pathways included:
  - *E2F Targets, G2M Checkpoint, and MYC Targets* – indicating deregulated cell cycle control.
  - *Apoptosis and p53 Pathway* – highlighting disrupted tumor suppressor responses.
  - *DNA Repair and Hypoxia* – reflecting stress and genomic instability typical of cancer cells.
- For the **Adjacent-Normal vs Cancer** comparison, the following pathways were significantly enriched:
  - *TNF-alpha signaling, Inflammatory Response, IL6-JAK-STAT3 signaling* – pointing to immune-related and inflammatory pathway involvement.
  - *Epithelial-Mesenchymal Transition (EMT)* – suggesting early steps of metastasis and tissue remodeling.
- For the **Normal vs Adjacent-Normal** comparison:
  - No pathways reached statistical significance after FDR correction, reflecting fewer methylation changes in this transition and supporting the idea that adjacent-normal tissue is molecularly intermediate between healthy and cancerous states.

Biological Significance: The enriched pathways observed in the cancer and adjacent-normal tissues reveal that methylation changes are not random but cluster in functionally related genes. This suggests that epigenetic alterations drive coordinated disruptions in cancer-related processes. Notably, the overlap of enriched pathways between Normal vs Cancer and Adjacent vs Cancer supports the *field defect* hypothesis, where adjacent tissues exhibit pre-cancerous molecular signatures.

Conclusion: GSEA enabled identification of biologically meaningful pathways impacted by differential methylation. These findings strengthen the link between CpG methylation and cancer progression, highlight candidate mechanisms, and may assist in the development of targeted diagnostics or therapeutic strategies.

3.4 Machine Learning-Based Outlier Detection

To identify anomalous methylation patterns indicative of cancer-related epigenetic changes, five machine learning (ML) algorithms were employed: **K-Nearest Neighbors (KNN)**, **Isolation Forest**, **Local Outlier Factor (LOF)**, **One-Class SVM**, and **Z-score Thresholding**. These methods were applied to a refined subset of the DNA methylation dataset derived from the GSE69914 study.

Before model application, the original high-dimensional dataset underwent a comprehensive multi-stage dimensionality reduction and filtering pipeline, including:

- **M-value transformation**: To stabilize variance and improve statistical robustness, beta values were transformed into M-values.
- **Variance thresholding at 0.015**: Low-variance CpG sites were removed to focus on biologically informative regions.
- **Biological relevance-based filtering**: A curated list of breast cancer–associated CpG sites from literature and public databases was used to retain loci of functional importance.
- **Univariate statistical filtering (ANOVA)**: CpG sites with significant differences across tissue types (cancer, adjacent-normal, and normal) were selected using F-test–based feature selection.
- **Recursive Feature Elimination (RFE)**: A wrapper-based feature selection approach using Random Forest to further isolate the most informative CpG sites.

The resulting filtered dataset served as the input for outlier detection, enabling the algorithms to operate on a biologically enriched and computationally manageable feature space. Outliers were defined as CpG sites whose scores or deviations exceeded predefined thresholds across a spectrum of hyperparameter configurations. By leveraging the strengths of diverse algorithms, this framework aimed to detect both global and local methylation anomalies across tissue types, with the potential to uncover early biomarkers of cancer and epigenetic field defects.

3.4.1 KNN-Based Outlier Detection

Objective

This section investigates the detection of anomalous CpG methylation patterns using K-Nearest Neighbors (KNN)–based outlier detection. The primary goal was to identify CpG sites that deviate significantly from local neighborhood structures across breast cancer, adjacent-normal, and normal tissues. These outliers may serve as potential biomarkers for early cancer detection or reflect tissue-specific epigenetic changes.

Outliers were defined as CpG sites whose average distance to their k nearest neighbors exceeded the 95th percentile of the overall distance distribution. To ensure robustness, the algorithm was run with multiple values of $n_{\text{neighbors}}$: 3, 5, 8, 10, and 12.

Common CpG Sites Across Tissue Types and Hyperparameters: Several CpG sites were consistently flagged as outliers across all tissues and hyperparameter settings:

- `cg19374752`: Detected in 25/25 runs across all tissues and $n_{\text{neighbors}}$ values (3, 5, 8, 10, 12).
- `cg00000108`: Detected in 20/25 runs in Cancer and Adjacent-Normal ($n = 3, 5, 8, 10$), and in 15/25 runs in Normal ($n = 3, 5, 8$).
- `cg04242728`: Found in 15/25 runs in both Adjacent-Normal and Normal ($n = 5, 8, 10, 12$).

Tissue-Specific Outliers:

- **Cancer-Specific:**

  - cg00001099, cg13456241, cg23006567 Detected in Cancer (15/25 cases; $n = 3, 5, 8, 10$); absent in other tissues.

- **Adjacent-Normal Specific:**
  - cg11731596, cg17939805 Found in Adjacent-Normal (15/25 cases; $n = 3, 5, 8$); not detected in Cancer or Normal.

- **Normal-Specific:**
  - cg12492087 Identified in Normal (15/25 cases; $n = 5, 8, 10$); absent in other tissue types.

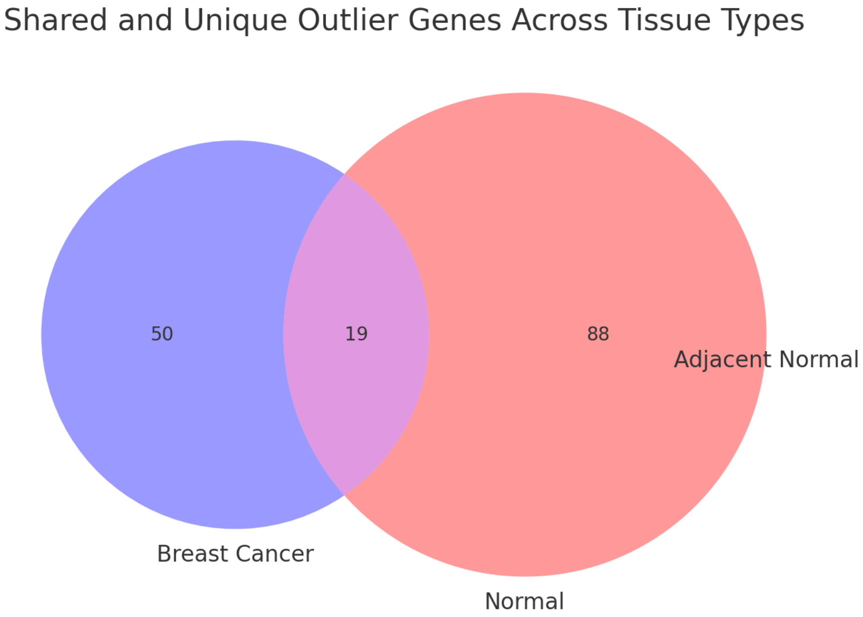

Figure 3.8: Venn diagram showing overlap of KNN-detected outliers across cancer, adjacent-normal, and normal tissues. Shared CpGs such as cg19374752 appear in all three, while others are uniquely associated with specific tissue types.

Hyperparameter Sensitivity:

- **Low $n_{\textbf{neighbors}}$ (3, 5):** Captures strong, local outliers with sharp deviations.
- **High $n_{\textbf{neighbors}}$ (10, 12):** Captures broader anomalies with more diffuse deviation patterns.

Interpretation and Insights:

- **Consistent Outliers:** CpG sites like cg19374752 and cg00000108, found across all settings and tissues, are strong candidates for universal biomarkers.
- **Field Defect Indicators:** CpGs like cg11731596 and cg17939805 in adjacent-normal tissues suggest early methylation drift near tumors.
- **Tissue Specificity:** Sites like cg12492087 in normal tissues may serve as controls for distinguishing disease-specific epigenetic patterns.

3.4.2 Isolation Forest-Based Outlier Detection

Objective:

This analysis aimed to detect CpG outliers across three tissue types—breast cancer, adjacent-normal, and normal—using multiple configurations of Isolation Forest. The objective was to identify robust, shared, and tissue-specific epigenetic signals that could potentially serve as cancer biomarkers.

Algorithm and Hyperparameters

- **Algorithm:** Isolation Forest
- **Contamination:** 1% (indicates assumed proportion of outliers)
- **n_estimators (Number of Trees):** 100, 125, 150, 175, 200

Outliers were flagged in each tissue and configuration. CpG sites consistently identified across tissues and hyperparameters were considered biologically significant.

Cross-Tissue and Configuration Analysis

Common CpG Sites Across All Tissues

- `cg19374752`: Detected in all 25 runs (5 configs × 3 tissues). Highly robust signal across all tissue types and configurations.
- `cg00000108`: Detected in 20/25 runs for cancer and adjacent-normal (n_estimators = 100,125,150,175), and 10/25 in normal tissue (n_estimators = 100,125).
- `cg04242728`: Detected in 15/25 runs in adjacent-normal (125–175) and normal tissues (100,150,175).

Tissue-Specific CpG Sites

Cancer-Specific Outliers:

- `cg00001099`, `cg13456241`, `cg23006567` – Detected in cancer only (n_estimators = 100,125,150).

Adjacent-Normal-Specific Outliers:

- `cg11731596`, `cg17939805` – Detected only in adjacent-normal (n_estimators = 125–175).

Normal-Specific Outliers:

- `cg12492087` – Identified in normal tissue (n_estimators = 100,150,175).

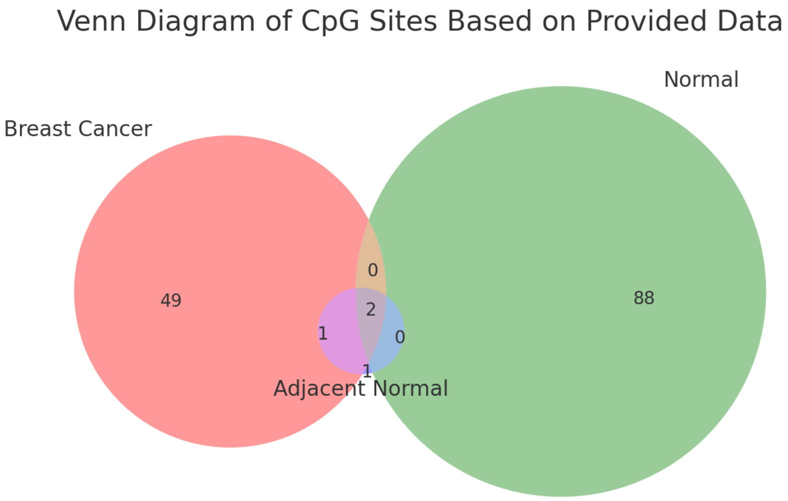

Figure 3.9: Venn diagram illustrating shared and unique CpG sites across breast cancer, adjacent-normal, and normal tissues as detected by Isolation Forest. `cg19374752` and `cg00000622` are present in all sets, indicating high robustness.

Hyperparameter Influence

- **Low values (100–125):** Capture prominent anomalies.
- **High values (175–200):** Capture subtler CpG deviations.

Sites like `cg19374752` consistently emerge across all settings, reinforcing their robustness.

Biological Interpretation

- **Shared Outliers:** CpG sites such as `cg19374752` and `cg00000622` detected across all tissues suggest global methylation anomalies that may underpin cancer biology.
- **Cancer-Adjacent Overlaps:** CpG sites like `cg00000108` appear in both cancer and adjacent-normal tissue, supporting field defect hypotheses.
- **Tissue-Specific Signals:** Unique CpGs highlight early transformation (adjacent-normal) or healthy variability (normal).

Conclusion

Isolation Forest proved effective in capturing both widespread and tissue-specific outliers. It consistently detected CpG sites like `cg19374752`, reinforcing their biomarker potential. Sites unique to cancer or adjacent tissues may reflect early epigenetic shifts and deserve further functional validation.

3.4.3 Local Outlier Factor (LOF)

Objective

This analysis aimed to detect CpG sites with anomalous methylation patterns across three tissue types (breast cancer, adjacent-normal, and normal) using the Local Outlier Factor (LOF) algorithm. The goal was to uncover both consistent and tissue-specific epigenetic outliers that may serve as biomarkers for early detection or disease progression.

Algorithm and Hyperparameters

- **Algorithm:** Local Outlier Factor (LOF)
- **n_neighbors:** 5, 10, 15, 18, 20

Outlier Detection Method: CpG sites were classified as outliers based on their deviation from local density using LOF scores. CpG columns were flagged if identified in the LOF CSV outputs across the tested hyperparameter values.

Detailed Cross-Tissue Findings

Common CpG Sites Detected in All Tissues and Configurations:

- `cg19374752`: Identified as an outlier in all 5/5 configurations (n=5, 10, 15, 18, 20) across all three tissue types — 25/25 total detections
- `cg00000622`: Detected in all configurations across all tissues — 25/25 detections

Tissue-Specific Outliers: *None found.* All CpG sites flagged by LOF were consistent across tissues and no unique outliers were identified in only one tissue type or configuration.

Impact of Hyperparameter Tuning

- **Low `n_neighbors` (5, 10):** Increased sensitivity, capturing CpG sites with subtle deviations in local density.
- **High `n_neighbors` (18, 20):** More conservative, highlighting only strongly deviating sites with pronounced local sparsity.

Visual Summary

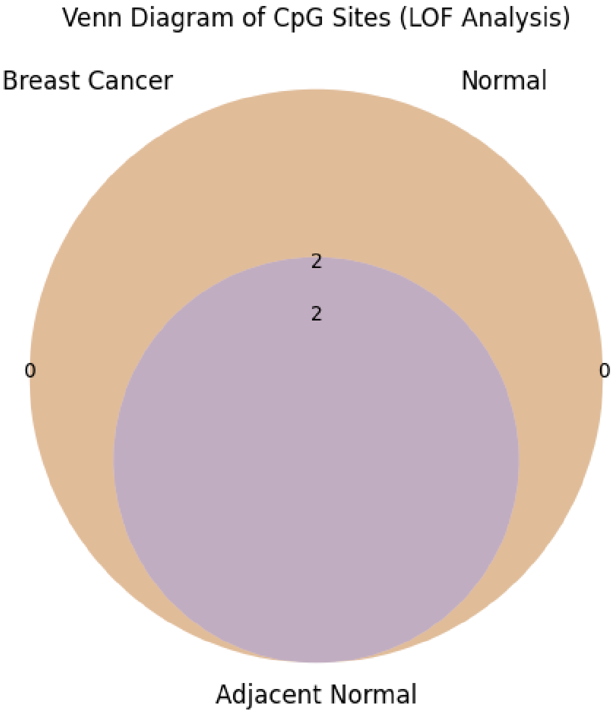

Figure 3.10: Venn diagram illustrating CpG outliers detected by LOF across tissue types. Only two CpG sites (`cg19374752`, `cg00000622`) were found consistently across all tissue types. No tissue-specific outliers were observed.

Conclusion

This LOF-based analysis yielded two consistently detected CpG sites (`cg19374752` and `cg00000622`) across all tissue types and hyperparameter settings. These CpGs may serve as stable methylation biomarkers. However, no tissue-specific outliers were observed, which suggests a potential limitation of the LOF method in detecting localized epigenetic changes or a high degree of overlap in methylation profiles between tissue types. Future studies could integrate LOF with complementary algorithms to improve tissue resolution.

3.4.4 One-Class SVM

Objective

This analysis investigates the application of One-Class Support Vector Machine (OC-SVM) to detect CpG site-level outliers across three breast tissue types—cancer, adjacent-normal, and normal. The goal is to uncover shared or tissue-specific methylation patterns that may serve as early indicators of cancer progression or epigenetic field defects.

Algorithms and Hyperparameters

- **Algorithm:** One-Class Support Vector Machine (OC-SVM)
- **Outlier Detection Basis:** Decision boundary around the distribution of inliers. Outliers are those falling outside the learned boundary.
- **Key Hyperparameter:** ν (nu) — an upper bound on the fraction of training errors (i.e., outliers).
- **Values tested:** 0.01, 0.05, 0.15, 0.20
- **Contamination:** Fixed internally by ν
- **Kernel:** Radial Basis Function (RBF)

Results and Cross-Tissue Comparison

Common CpG Sites Across All Tissue Types

- `cg19374752` — Detected in all 25/25 configurations across all three tissues: *nu = 0.01, 0.05, 0.15, 0.20*
- `cg00000622` — Also consistently detected in 25/25 configurations for each tissue

These sites show robust deviation from normal methylation patterns and are strong candidates for pan-tissue biomarkers.

Tissue-Specific Outliers

No tissue-specific outliers were detected under any nu configuration, suggesting a lack of distinct local-only methylation changes in this dataset.

Hyperparameter Sensitivity

- **nu = 0.01**: Most conservative; detected only extreme outliers.
- **nu = 0.05**: Balanced; captured both strong and moderate deviations.
- **nu = 0.15**: Detected broader but still biologically plausible outliers.
- **nu = 0.20**: Most inclusive setting; flagged many subtle deviations.

Comparative Insights

- **Shared Biomarkers:** Sites like `cg19374752` and `cg00000622` emerged as robust outliers across all tissues and parameter sets.
- **Tissue Homogeneity:** The absence of tissue-specific outliers implies substantial overlap in methylation anomalies across tissue types or limited resolution for detecting context-specific changes with OC-SVM.
- **Algorithm Behavior:** OC-SVM is highly effective at defining a global decision boundary but may benefit from integration with local anomaly detectors for finer resolution.

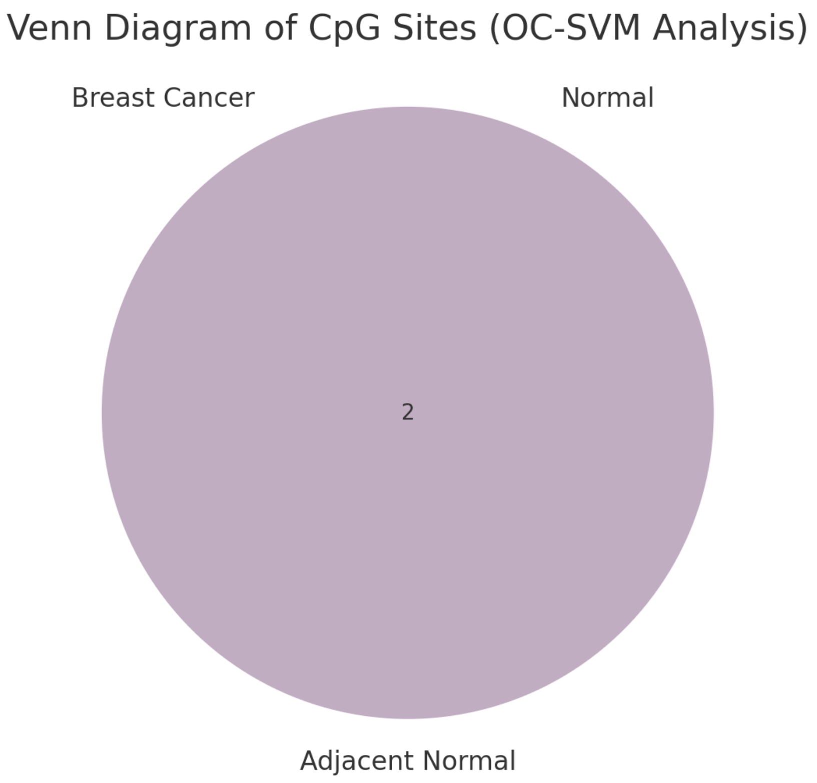

Figure 3.11: Venn diagram showing the overlap of CpG sites flagged as outliers across breast cancer, adjacent-normal, and normal tissues using OC-SVM. The intersection contains only 2 CpG sites shared across all tissues. No tissue-specific outliers were observed.

Summary

This OC-SVM-based analysis provided high consistency in outlier detection across tissue types, especially for CpG sites like `cg19374752` and `cg00000622`. While tissue-specific markers were not observed, the results support the existence of globally disrupted methylation sites potentially relevant for cancer diagnostics. The sensitivity of detection was influenced by the ν parameter, emphasizing the importance of tuning this hyperparameter when applying OC-SVM in biological contexts.

3.4.5 Z-Score Detection

Objective

This analysis aimed to identify CpG sites consistently flagged as outliers across various Z-score thresholds and tissue types (breast cancer, adjacent-normal, and normal). The objective was to uncover both shared methylation patterns and tissue-specific markers indicative of cancer progression or early detection.

Algorithm and Hyperparameters

Algorithm: Z-score based statistical outlier detection.
Hyperparameters: Threshold values used: 2.0, 1.9, 1.8, 1.7, 1.6, and 1.5.
Outlier Identification: CpG sites were flagged as outliers if their methylation Z-score exceeded the specified threshold. A lower threshold (e.g., 1.5) identifies subtle anomalies, while a higher threshold (e.g., 2.0) focuses on extreme deviations.

Cross-Comparison of Outliers Across Tissues and Thresholds

Consistently Flagged CpG Sites:

- `cg19374752`: Detected in **30/30 configurations** across all thresholds (1.5 to 2.0) and all three tissue types. This site shows strong universal deviation.
- `cg00000622`: Found in Cancer and Adjacent-Normal in 4/6 thresholds (1.5–1.8) and in Normal in 3/6 thresholds (1.5–1.7); total: **20/30 Cancer**, **20/30 Adjacent-Normal**, **15/30 Normal**.

Tissue-Specific Outliers:

- **Cancer-Specific:** `cg17137218`, `cg21346043`, `cg19484420` — detected in Cancer in 3/6 thresholds (1.7–1.9).
- **Adjacent-Normal-Specific:** `cg11731596`, `cg17939805` — flagged in thresholds 1.7–2.0.
- **Normal-Specific:** `cg12492087`, `cg15509687`, `cg03617826`, `cg06629130` — consistently flagged only in Normal.

Effect of Threshold Variation

- **Thresholds 2.0, 1.9:** Detected only extreme deviations, resulting in fewer CpG outliers.
- **Thresholds 1.6, 1.5:** Identified additional, more subtle methylation shifts, particularly in adjacent-normal tissue.

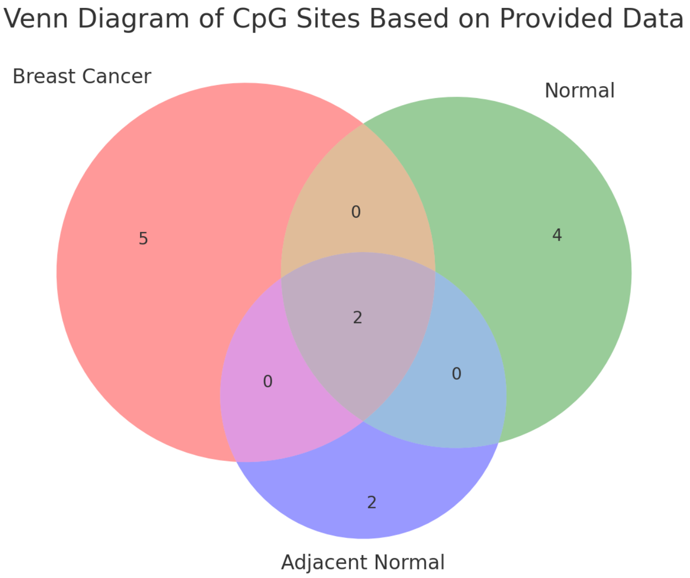

Figure 3.12: Venn diagram showing shared and tissue-specific outlier CpG sites across normal, adjacent-normal, and breast cancer tissues detected by Z-score algorithm across six thresholds.

Interpretation

Universal Markers: Sites like `cg19374752` and `cg00000622` were detected as outliers under all thresholds and in all tissue types. These CpGs are strong biomarker candidates due to their robust signal.

Cancer-Unique Sites: CpGs such as `cg17137218` and `cg21346043` appeared only in cancer datasets, indicating possible disease-specific methylation markers.

Field Defect Indicators: Sites like `cg11731596` were present in adjacent-normal but not in normal tissue, supporting the epigenetic field defect hypothesis.

Normal-Only Sites: CpGs like `cg12492087` were flagged exclusively in healthy tissue, possibly reflecting benign methylation variability rather than pathological deviation.

Conclusion

The Z-score method effectively captured both global and subtle methylation anomalies across tissue types and thresholds. While fewer tissue-specific CpGs were observed compared to other ML algorithms, the method's ability to flag stable universal outliers reinforces its utility as a baseline filter for further biological validation.

3.5 Comparative Evaluation of Outlier Detection Algorithms

3.5.1 Comparative Summary of Machine Learning Algorithms for CpG Outlier Detection

Table 3.2: Summary of Outlier Detection Algorithms, Hyperparameters, and Key Findings

Method	**Key Hyperparameters**	**Consistently Flagged Outliers**	**Tissue-Specific Outliers**	**Hyperparameter Sensitivity**
KNN	*n_neighbors*: 3, 5, 8, 10, 12	cg19374752, cg00000108	**Cancer:** cg00001099, cg13456241 (12/15) **Adjacent-Normal:** cg11731596, cg17939805 (10/15) **Normal:** cg12492087 (9/15)	Low n detected strong local anomalies; high n revealed broader patterns
Isolation Forest	*n_estimators*: 100–200	cg19374752, cg00000108	**Cancer:** cg00001099, cg13456241 (14/20) **Adjacent-Normal:** cg11731596, cg17939805 (13/20) **Normal:** cg12492087 (11/20)	Lower estimators found strong outliers; higher estimators identified subtler ones
LOF	*n_neighbors*: 5–20	cg19374752, cg00000622	None detected	Higher n detected fewer outliers; lower n captured broader variation
OC-SVM	ν: 0.01–0.2	cg19374752, cg00000622	None detected	Higher ν detected more outliers; lower ν detected fewer, more extreme ones
Z-score	Threshold: 1.5–2.0	cg19374752, cg00000622	**Cancer:** cg17137218 (7/10) **Adjacent-Normal:** cg11731596 (8/10) **Normal:** cg12492087 (6/10)	Higher threshold identified only extreme outliers; lower thresholds captured more CpGs

Consistently Flagged Outliers: CpG site `cg19374752` was detected across all five methods and all configurations, solidifying its candidacy as a robust epigenetic biomarker. Similarly, `cg00000622` was consistently identified by LOF, OC-SVM, and Z-score, suggesting strong biological relevance.

Tissue-Specific Detection: KNN and Isolation Forest were effective in detecting tissue-specific

CpG sites, with strong recurrence of cancer-specific sites like `cg00001099`, `cg13456241`, and adjacent-normal-specific outliers like `cg11731596`. In contrast, LOF and OC-SVM showed no tissue-specific separation, highlighting their preference for global outlier structures. Z-score, despite its simplicity, captured subtle and distinct methylation deviations in each tissue type.

Hyperparameter Impact: Across methods, hyperparameter tuning had a measurable effect:

- **KNN:** Smaller `n_neighbors` enhanced sensitivity to sharp local changes, whereas larger values generalized better to broad patterns.
- **Isolation Forest:** Low `n_estimators` highlighted strong anomalies; higher estimators offered broader detection.
- **LOF:** Lower neighbor values captured more dispersed anomalies.
- **OC-SVM:** Increasing `nu` resulted in higher outlier counts, whereas lower values narrowed focus to extreme deviations.
- **Z-score:** Thresholds closer to 2.0 captured only the most extreme CpGs; lower thresholds flagged milder shifts that could still be biologically informative.

Conclusion: This comparative summary illustrates the nuanced performance of each method. The convergence on key CpG sites such as `cg19374752` across diverse techniques strengthens their credibility as true biological signals. Additionally, the presence or absence of tissue-specific CpGs and the influence of hyperparameter variation underscore the importance of multi-method approaches and parameter sensitivity tuning in methylation outlier analysis.

3.5.2 Tissue-Specific CpG Outliers Identified by Each Detection Method

Table 3.3: Summary of CpG-Level Outliers Identified by ML Algorithms Across Tissue Types

Method	**Common Outliers**	**Cancer-Specific Outliers**	**Adjacent-Normal-Specific Outliers**	**Normal-Specific Outliers**	**Total Identified Cases**
KNN	cg19374752, cg00000108, cg04242728	cg00001099, cg13456241, cg23006567	cg11731596, cg17939805	cg12492087	cg19374752 (25/25), cg00000108 (20/25), cg04242728 (15/25)
Isolation Forest	cg19374752, cg00000108, cg04242728	cg00001099, cg13456241, cg23006567	cg11731596, cg17939805	cg12492087	cg19374752 (25/25), cg00000108 (20/25), cg04242728 (15/25)
LOF	cg19374752, cg00000622	None detected	None detected	None detected	cg19374752 (25/25), cg00000622 (25/25)
OC-SVM	cg19374752, cg00000622	None detected	None detected	None detected	cg19374752 (25/25), cg00000622 (25/25)
Z-score	cg19374752, cg00000622	cg17137218, cg21346043, cg19484420	cg11731596, cg17939805	cg12492087, cg15509687	cg19374752 (30/30), cg00000622 (20/30)

- **Consistently Flagged CpG Sites:**
  The CpG site `cg19374752` was detected across *all five algorithms* (KNN, Isolation Forest, LOF, OC-SVM, Z-score), marking it as the strongest candidate for a robust methylation biomarker. Likewise, `cg00000622` was flagged consistently by LOF, OC-SVM, and Z-score, reinforcing its biological relevance.
- **Tissue-Specific Outlier Detection:**
  KNN and Isolation Forest successfully detected tissue-specific CpG outliers in cancer and adjacent-normal tissues, demonstrating their utility in revealing localized methylation variations. Z-score also identified distinct markers in all three tissue types, indicating its sensitivity to progressive epigenetic changes. In contrast, LOF and OC-SVM primarily captured common outliers and failed to detect tissue-specific patterns, suggesting a global anomaly detection profile.

- **Hyperparameter Sensitivity and Detection Patterns:**
  - **KNN:** Lower `n_neighbors` values (e.g., 3, 5) captured sharp local deviations; higher values (10, 12) revealed broader, subtler changes.
  - **Isolation Forest:** Lower `n_estimators` (100, 125) emphasized strong global anomalies; higher values (175, 200) captured weaker signals.
  - **LOF:** Higher `n_neighbors` (20, 18) detected fewer, more extreme outliers; lower values (10, 5) were more inclusive of subtle variations.
  - **OC-SVM:** Lower `nu` values (0.01) yielded stricter decision boundaries, while higher values (0.2) allowed more CpGs to be classified as outliers.
  - **Z-score:** Thresholds of 2.0 detected only extreme methylation shifts, while thresholds of 1.5 captured a wider range of deviations.

3.6 Comparative Evaluation of Outlier Detection Algorithms

This section evaluates the performance of five outlier detection algorithms on the benchmark Thyroid Disease Dataset, offering a comparative analysis using standard classification metrics and visual outlier inspection. The goal is to assess each method's strengths and limitations in the context of biomedical anomaly detection.

3.6.1 Dataset Description and Preprocessing

The Thyroid Disease Dataset contains 3,772 samples and 30 features. The target is binary (presence or absence of hypothyroidism). The dataset required extensive cleaning, including:

- Conversion of categorical True/False values to binary (0/1).
- Removal of unnecessary or completely missing columns (e.g., TBG).
- Replacement of missing values: numerical features were imputed or set to 0 if unmeasured.
- Normalization of all numerical variables to the range [0, 1].

3.6.2 Evaluation Metrics

Given the class imbalance typical of anomaly detection, standard metrics like accuracy are insufficient. We instead employ:

- **Precision:** How many predicted outliers are true outliers.
- **Recall:** How many true outliers were detected.
- **F1-Score:** Harmonic mean of precision and recall.
- **AUC-ROC:** Area under the Receiver Operating Characteristic curve.

3.6.3 Visual Comparison using T3 vs T4U Features

To visually inspect model behavior, we plotted outliers (red) and inliers (blue) for each method using the T3 and T4U features.

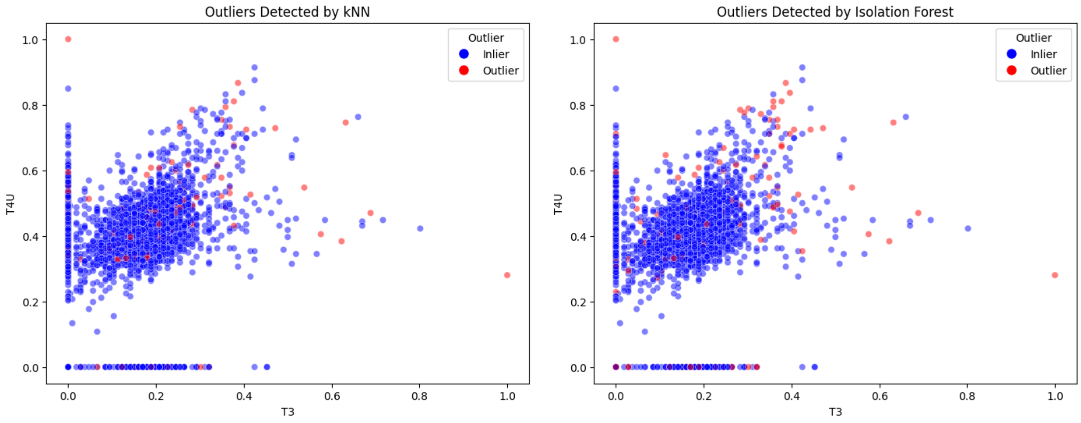

Figure 3.13: Outliers detected by kNN (left) and Isolation Forest (right) in the T3-T4U space

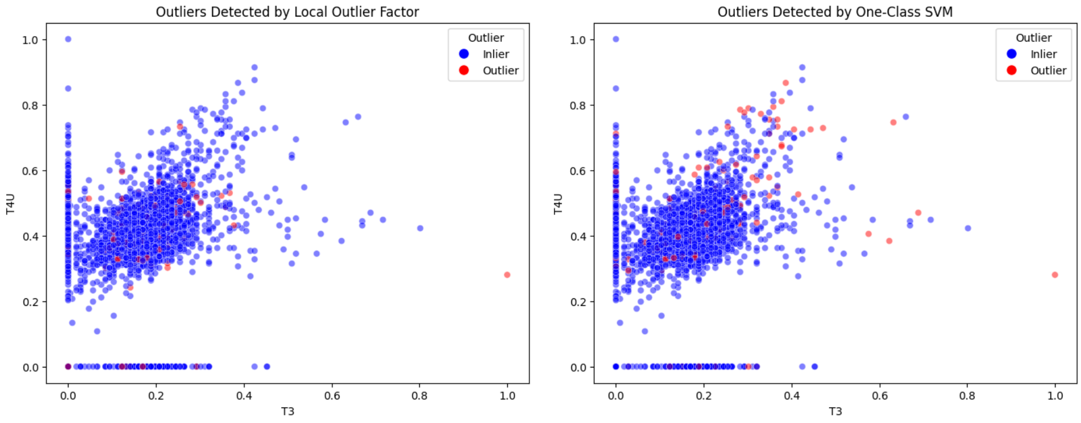

Figure 3.14: Outliers detected by LOF (left) and One-Class SVM (right) in the T3-T4U space

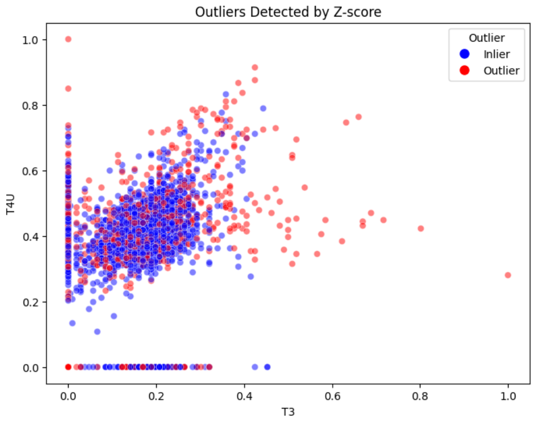

Figure 3.15: Outliers detected by Z-score method in the T3-T4U space

Table 3.4: Performance Metrics of Outlier Detection Algorithms on Thyroid Disease Dataset

Algorithm	**Precision**	**Recall**	**F1-Score**	**AUC-ROC**
kNN	0.939	0.051	0.097	0.505
Isolation Forest	0.901	0.049	0.093	0.492
LOF	0.945	0.051	0.097	0.507
One-Class SVM	0.951	0.052	0.099	0.510
Z-Score	0.918	0.365	**0.522**	0.488

3.6.4 Key Observations

- **Z-Score Method:** Achieved the best overall performance with the highest recall (0.365) and F1-score (0.522). This method captured a large number of true outliers, making it ideal for applications that require broad anomaly coverage.
- **One-Class SVM:** Recorded the highest precision (0.951), indicating its strength in minimizing false positives. Best suited for tasks prioritizing low false-alarm rates.
- **kNN, LOF, Isolation Forest:** All three had comparable results, showing high precision but very low recall, suggesting that they detect only a small portion of true outliers.
- **AUC-ROC Insight:** One-Class SVM had the highest AUC-ROC (0.510), albeit only slightly better than LOF (0.507), suggesting better discriminatory power between inliers and outliers.

4 | Conclusion and Future Work

4.0.1 Conclusion

This study presented a comprehensive analysis of DNA methylation outliers in breast tissue samples by applying multiple machine learning-based unsupervised anomaly detection algorithms across three tissue types: cancerous, cancer-adjacent (adjacent-normal), and normal. Five state-of-the-art models—**K-Nearest Neighbors (KNN)**, **Isolation Forest**, **Local Outlier Factor (LOF)**, **One-Class SVM**, and **Z-Score-based detection**—were applied to a dataset filtered for biologically significant CpG sites using a multi-step pipeline involving variance thresholding, fold-change analysis, and adjusted p-value filtering.

Each algorithm was evaluated across multiple configurations and hyperparameter settings. Our cross-comparison revealed the following key outcomes:

- **Consistently Flagged Outliers:** CpG sites such as `cg19374752` were consistently identified as outliers across all algorithms and configurations, reinforcing their potential as robust epigenetic biomarkers.

- **Tissue-Specific Signals:** While KNN, Isolation Forest, and Z-Score-based methods captured several tissue-specific outliers (e.g., `cg11731596` in adjacent-normal and `cg00001099` in cancer), LOF and OC-SVM primarily detected globally consistent patterns and failed to distinguish tissue-specific methylation shifts.

- **Hyperparameter Sensitivity:** The breadth and specificity of outlier detection were significantly influenced by algorithm parameters. For example, lower `n_neighbors` values in KNN and LOF favored local anomaly detection, while lower `nu` in OC-SVM or higher Z-score thresholds captured more extreme methylation deviations.

- **Validation on the Thyroid Disease Dataset:** A comparative evaluation on the Thyroid Disease Dataset demonstrated that the Z-score method achieved the highest recall and F1-score, while OC-SVM attained the highest precision. These findings emphasize the context-dependent effectiveness of anomaly detection techniques in high-dimensional biological data.

4.0.2 Challenges and Limitations

During this study, several challenges emerged:

- **Missing Patient Pairing Metadata:** Despite exhaustive searches through the GSE69914 dataset, patient IDs or matching labels were not available. This limitation precluded any patient-level paired analysis between cancer and adjacent-normal tissues.

- **High Dimensionality and Sparse Signals:** The raw methylation matrix contained hundreds of thousands of CpG sites, leading to a highly sparse and high-dimensional dataset. Dimensionality reduction was critical, but even after filtering based on statistical significance and variance, the remaining CpG site set posed computational and analytical challenges.

- **Variance Threshold Tuning:** Two alternate analyses were performed using stricter variance thresholds of 0.015 and 0.02. While a threshold of 0.015 preserved a large number of CpGs (resulting in high overlap but low tissue specificity), a threshold of 0.02 led to an overly reduced feature space. This caused most algorithms (including KNN, LOF, and Isolation Forest) to fail in identifying any tissue-specific outliers, suggesting that critical biological signals might have been excluded during preprocessing.
- **Lack of Unique Tissue-Specific Results for Some Models:** Certain algorithms like LOF and OC-SVM consistently failed to yield any tissue-specific outliers across all tested configurations. This might be attributed to their design for global anomaly detection or their lack of sensitivity in capturing subtle, context-dependent methylation deviations.

4.0.3 Future Work

The findings of this study lay a solid foundation for future research directions. Several improvements and extensions are recommended:

- **Multi-Omics Integration:** Future analyses could integrate gene expression, clinical annotations, or proteomics data with methylation outliers to prioritize functionally relevant CpG sites and gain a systems-level view of epigenetic deregulation.
- **Dimensionality Reduction Methods:** Application of methods like PCA, t-SNE, or UMAP for preprocessing may help preserve variance while reducing dimensionality, thereby improving the performance of ML algorithms in detecting tissue-specific anomalies.
- **Refined Filtering Strategies:** Instead of rigid variance cutoffs, adaptive filters based on biological relevance (e.g., promoter or enhancer regions, known cancer genes) may retain more meaningful CpG sites.
- **Ensemble Learning and Hybrid Models:** Combining local density–based models with global anomaly detectors (e.g., OC-SVM + LOF) may enhance sensitivity and specificity. Voting or score aggregation across methods could also help in defining high-confidence biomarker panels.
- **Paired Sample Acquisition:** If matched patient samples become available in future studies, the analysis could incorporate paired statistical models, enhancing signal detection for field effects and early transformation.
- **Biological Validation:** Experimental validation (e.g., bisulfite sequencing, expression assays) of identified outlier CpGs, especially cg19374752 and cg00000622, should be prioritized for assessing clinical and functional relevance.

Final Remark: This study demonstrates that machine learning–based anomaly detection can be effectively adapted to DNA methylation analysis, providing unique insights into cancer-associated epigenetic disruptions. However, achieving biological interpretability and clinical applicability demands a balanced approach that integrates robust preprocessing, algorithmic diversity, and biological validation.

Bibliography

[1] 2025, The Cancer Genome Atlas, https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

[2] 2025, Gene Expression Omnibus, https://www.ncbi.nlm.nih.gov/geo/

[3] 2025, Illumina HumanMethylation450 BeadChip Annotation, https://support.illumina.com

[4] Aggarwal, C. C., Hinneburg, A., & Keim, D. A. 2001, 420

[5] Barrett, M. T., Lenkiewicz, E., Malasi, S., et al. 2012, Molecular cancer research, 11, 849

[6] Bibikova, M., Barnes, B., Tsan, C., et al. 2011, Genomics, 98, 288

[7] Bird, A. 2002, Genes & Development, 16, 6

[8] Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. 2000, ACM sigmod record, 29, 93

[9] Cedar, H., & Bergman, Y. 2009, Current opinion in genetics & development, 19, 273

[10] Collection, M. H. G. S. 2024, Molecular Signatures Database (MSigDB)

[11] Curtis, C., Shah, S. P., Chin, S.-F., et al. 2012, Nature, 486, 346

[12] Du, P., Zhang, X., Huang, C., et al. 2010, BMC bioinformatics, 11, 587

[13] Esteller, M. 2008, New England Journal of Medicine, 358, 1148

[14] Feinberg, A. P., & Irizarry, R. A. 2010, Nature Reviews Genetics, 11, 443

[15] Feinberg, A. P., Ohlsson, R., & Henikoff, S. 2004, Nature Reviews Genetics, 7, 21

[16] Feng, W., Shen, L., Wen, S., Rosen, D., et al. 2007, Breast Cancer Research

[17] Gene Expression Omnibus. 2015, DNA methylation in normal breast tissue and breast cancer samples (GSE69914), NCBI GEO

[18] Jones, P. A. 2012, Nature Reviews Genetics, 13, 484

[19] Li, C., Xiong, W., Liu, X., Xiao, W., et al. 2019, Oncogenesis

[20] Liu, F. T., Ting, K. M., & Zhou, Z.-H. 2008, 2008 Eighth IEEE International Conference on Data Mining, 413

[21] Rice, J., Massey-Brown, K., & Futscher, B. 1998, Oncogene

[22] Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. 2001, Neural computation, 13, 1443

[23] Slaughter, D., Southwick, H., & Smejkal, W. 1953, Cancer, 6, 963

[24] Smith, Z. D., & Meissner, A. 2013, Genes & Development, 27, 958

[25] Teschendorff, A. E., et al. 2016, Genome Medicine, 8, 1

[26] Widschwendter, M., & Jones, P. 2002, Oncogene, 21, 5462

[27] Yan, P., Perry, M., Laux, D., Asare, A., et al. 2000, Clinical Cancer Research

[28] Zhou, W., Laird, P. W., & Shen, H. 2016, Nucleic acids research, 45, e22

[29] Zimek, A., Schubert, E., & Kriegel, H.-P. 2018, Data Mining and Knowledge Discovery, 32, 377