Politecnico di Torino

Master's Degree in Mathematical Engineering

Material for Thesis

a– Dataset Construction, Storage and Metadata acquisition

Elisabetta Roviera s328422

Contents

Abstract

This document provides a complete description of the dataset construction and metadata acquisition pipeline developed for the thesis project *"Advanced Study of Epigenetic Mechanisms in Neoplasms"*. It details the retrieval, ingestion, transposition, labeling, and storage of three large GEO methylation datasets — **GSE69914**, **GSE225845**, and **GSE287331** — all focused on breast tissue methylation across normal, adjacent-normal, and tumor states. The resulting harmonized Parquet files and phenotype tables constitute the reproducible foundation for all subsequent preprocessing, statistical, and machine-learning analyses.

All notebooks and resources are publicly available in the `thesis repository`:

- `00-dataset-preparation-GSE69914.ipynb` — GSE69914 β-value matrix ingestion and labeling.
- `00-dataset-metadata-GSE69914.ipynb` — Metadata extraction for GSE69914.
- `00-dataset-preparation-GSE225845.ipynb` — GSE225845 β-value streaming ingestion and merging.
- `00-dataset-metadata-GSE225845.ipynb` — Metadata extraction for GSE225845.
- `00-dataset-preparation-GSE287331.ipynb` — GSE287331 β-value matrix ingestion.
- `00-dataset-metadata-GSE287331.ipynb` — Metadata extraction for GSE287331.

All raw and processed methylation data are available from the NCBI Gene Expression Omnibus (GEO):

- GSE69914 — Illumina 450K, normal vs adjacent vs tumor.
- GSE225845 — EPIC array, breast tissue methylation (normal, adjacent, tumor).
- GSE287331 — EPIC array, tumor-proximity axis (HDB–CUB–OQ–AN–TU).

1 Data Source and Rationale

Gene Expression Omnibus (GEO). I reviewed several public portals for DNA methylation arrays (e.g., GEO, ArrayExpress, GDC/TCGA). To reduce heterogeneity in formats and metadata, I decided to use a single source: the **NCBI Gene Expression Omnibus (GEO)**. GEO is both a website and a curated repository: it assigns stable accessions (GSE/GSM/GPL), preserves sample-level metadata, and hosts links to processed β-value matrices and raw IDAT files. Using one source simplifies downstream harmonization.

Continuation of prior thesis work. This study builds upon the master's thesis by Nastaran Ahmadi Bonakdar [1], which focused on the GEO series **GSE69914** [2]. Building on that foundation, I continue using the Gene Expression Omnibus (GEO) as the unified repository for DNA methylation data and extend the analysis to two additional series: **GSE225845** [3] and **GSE287331** [4]. All datasets were selected following consistent criteria to ensure comparability across platforms and tissue types.

1.1 Inclusion criteria

The datasets were screened according to the following requirements, ensuring technical compatibility and adequate cohort representation:

- **Platform:** Illumina HumanMethylation450 (450K) or MethylationEPIC (850K), providing broad CpG coverage and cross-study comparability.
- **Tissue groups:** availability of *normal/healthy*, *adjacent-normal*, and *tumor* samples to enable comparative analyses.
- **Sample size:** preference for at least ∼50 samples per group to improve statistical power and reduce sampling noise.
- **Data availability:** presence of *processed* methylation matrices with β-values (Series Matrix or supplementary files). Raw `.idat` files are optional and not required for the present objectives.

1.2 Processed β-values versus raw IDAT files

Raw `.idat` files contain the fluorescence intensities measured for each CpG probe in the methylated ($y^{(M)}$) and unmethylated ($y^{(U)}$) channels. In contrast, β-values represent a normalized proportion of methylation, defined as [5]:

$$\beta := \frac{\max(y^{(M)}, 0)}{\max(y^{(M)}, 0) + \max(y^{(U)}, 0) + \alpha}, \qquad \beta \in [0, 1],$$

where α is a small offset (typically 100) added to stabilize the ratio and prevent division by very small denominators.

β-values provide an intuitive and interpretable measure of methylation: values close to 0 indicate unmethylated CpG sites, whereas values close to 1 correspond to highly methylated loci. Because the goal of this thesis is to identify possible alterations in DNA methylation in normal tissues and investigate whether they may contribute to tumor transformation, working with already normalized β-values is fully appropriate. Reprocessing raw IDAT files would not add relevant information for this objective, as it involves low-level signal calibration that has already been performed by the original authors. Using the processed matrices allows me to focus on the analytical and comparative aspects of the study — applying statistical and machine learning methods to detect biologically meaningful methylation differences.

2 Candidate GEO Series

Three GEO datasets were selected according to the inclusion criteria described above. All of them focus on breast tissue methylation and collectively cover normal, adjacent-normal, and tumor samples across the two Illumina array generations (450K and EPIC). This combination provides both historical continuity and higher-resolution coverage for comparative analyses.

GSE69914 – Genome-wide DNA methylation profiling of normal breast, adjacent-normal, and breast cancer tissue.

- **Public release:** Jun 18, 2015.
- **Platform:** Illumina HumanMethylation450 BeadChip (GPL16304).
- **Sample composition:** 407 samples in total — 50 normal, 42 matched adjacent-normal/tumor pairs (84 samples), 263 tumors, plus 7 normal and 3 tumor samples from BRCA1 carriers.

- **Data availability:** I downloaded the processed β-values directly from the `Series Matrix file(s) (.txt)` provided on GEO. These files contain already normalized methylation levels for all CpG sites across the available samples.

- **Notes:** This dataset forms the cornerstone of the previous thesis by Nastaran Ahmadi Bonakdar [1] and remains a key reference in field-defect and outlier methylation studies. Despite some incomplete metadata, it provides an ideal baseline for comparison across tissue states.

- **GEO link:** GSE69914.

GSE225845 – High neighborhood deprivation impacts DNA methylation and gene expression in cancer-related genes.

- **Public release:** Oct 15, 2023.

- **Platform:** Illumina Infinium MethylationEPIC (GPL21145).

- **Sample composition:** 402 breast tissue samples from 289 women, including 185 tumors, 113 paired adjacent-normal samples, and 104 normal tissues obtained from reduction mammoplasty.

- **Data availability:** I downloaded the processed β-value matrices available on GEO, one file for each tissue group (normal, adjacent-normal, and tumor). These tables provide normalized methylation levels across all CpG sites. `GSE225845_normal_normalized_betas.txt.gz`, `GSE225845_adjnorm_normalized_betas.txt.gz`, `GSE225845_tumors` `normalized_betas.txt.gz`.

- **Notes:** The dataset explores the effect of neighborhood deprivation on epigenetic regulation and includes clearly defined tumor, adjacent-normal, and normal groups, making it suitable for tissue-type comparisons in this study.

- **GEO link:** GSE225845.

GSE287331 – DNA methylation patterns in breast cancer, paired benign tissue, and healthy controls.

- **Public release:** Jun 18, 2025.

- **Platform:** Illumina Infinium MethylationEPIC v1.0 (GPL21145).

- **Sample composition:** A "tumor-proximity axis" design including 69 tumor (TU), 60 adjacent-normal (AN), 67 ipsilateral opposite quadrant (OQ), 68 contralateral unaffected breast (CUB), and 182 healthy-donated breast (HDB) samples.

- **Data availability:** I downloaded the processed β-values provided as compressed CSV files on GEO `GSE287331_betas_processed.csv.gz`.

- **Notes:** This recent large-scale EPIC study explicitly models the transition from healthy to tumor tissue along the tumor-proximity axis. Its structure is highly relevant for assessing progressive methylation alterations in the early stages of tumorigenesis.

- **GEO link:** GSE287331.

3 Dataset preparation

Before applying any preprocessing or modelling steps, I first constructed standardized working methylation matrices and their corresponding phenotype tables for each dataset (GSE69914, GSE225845, GSE287331). All matrices were converted to a common `sample` × `CpG` layout and stored as compressed Parquet files, with harmonized label encodings to enable downstream comparative analyses.

3.1 GSE69914 — Dataset preparation

Beta-value matrix construction. The working methylation matrix for GSE69914 (`GSE69914.parquet`) was constructed in three main stages, prioritizing speed, low memory usage, and reproducible I/O.

1) **Ingestion and transposition.** The GEO Series Matrix file was parsed directly, skipping the 73-line metadata header. The source table is organized as `CpG × sample`, with `ID_REF` entries representing CpG probe identifiers. All sample columns were coerced to numeric values (invalid entries set to `NaN`) and explicitly cast to `float32` to minimize memory usage while preserving numerical precision. The matrix was then transposed to the analysis layout (`sample × CpG`), promoting the GSM accession identifiers to a dedicated column named `id_tissue`.

2) **Label derivation and attachment.** Class labels were derived from the GSM-level metadata by parsing the field `status(0=normal, 1=normal-adjacent, 2=breast cancer, 3=normal-BRCA1, 4=cancer-BRCA1)`, resulting in integer codes within {0,1,2,3,4}. To avoid performing a costly in-memory join on a very wide table (∼485k CpGs), labels were streamed and appended row-wise, preserving sample order and ensuring constant memory usage.

3) **Columnar storage and typed schema.** For long-term access and efficient downstream analysis, the labeled table was written to Parquet format using a fixed schema: `id_tissue` (string), `label` (`Int8`), and all CpG probe intensities as `Float32`. The dataset was compressed using lossless LZ4 and dictionary encoding, providing fast full-table reads and efficient column projection in both Polars and pandas. This design yields a compact, typed matrix suitable for high-throughput preprocessing (technical filtering, normalization) and machine-learning–based modeling.

Precision and data representation. Because β-values are strictly bounded within $[0, 1]$ [6] and biologically meaningful methylation differences typically occur at magnitudes of 10^{-2}–10^{-3}, single-precision floating point (`float32`, machine $\epsilon \approx 10^{-7}$) offers more than adequate numerical accuracy while reducing memory consumption and I/O time. This representation aligns with recent large-scale genomics frameworks that process molecular features, including DNA methylation data, entirely in `float32` precision [7].

Phenotypic metadata extraction. A dedicated phenotype table (`pheno_GSE69914_lz4.parquet`) was constructed from the GEO metadata using the `GEOparse` API. For each sample (GSM), I extracted the following attributes:

- `sample_id` — GEO accession identifier;
- `label` — numeric class code (0–4) consistent with the β-matrix;
- `group` — categorical tissue class (Normal, Adjacent, Tumor, Normal_BRCA1, Tumor_BRCA1);
- `sentrix_id`, `slide_id` — Illumina BeadChip and batch identifiers;
- `er`, `pr`, `her2`, `ki67` — binary immunohistochemical markers (0=negative, 1=positive);
- `batch` — internal batch field replicated from `slide_id`.

All variables were stored in a compact Parquet table with `Int8` encoding for categorical and binary fields and LZ4 compression for efficient downstream joins and batch-effect correction steps.

3.2 GSE225845 — Dataset preparation

Beta-value matrix construction. The working methylation matrix for GSE225845 (`GSE225845.parquet`) was built from the three tissue-specific normalized β-value tables provided on GEO (normal, adjacent-normal, and tumors), prioritizing streaming I/O, low memory usage, and a consistent typed schema across tissue groups.

1) **Streaming ingestion of normal and adjacent-normal matrices.** The normalized β-value tables for normal and adjacent-normal breast tissue were ingested directly from the original tab-delimited TXT files using a lazy Polars `scan_csv` pipeline. Only the metadata columns `accession_num` and `basenames` were preserved as strings, while all remaining columns (CpG probes) were cast to `Float32`. The resulting tables were written in a fully streaming fashion to Parquet with LZ4 compression, without ever loading the full matrix into memory.

2) **Chunked conversion of the tumor matrix.** Owing to the larger size of the tumor file (EPIC array, wide CpG dimension), the same conversion strategy could not be applied directly without exceeding the available RAM. Instead, the tumor TXT file was processed in row-wise chunks using `pandas.read_csv` with an explicit `dtype` map: `accession_num` and `basenames` as strings, and all CpG probes as `float32`. Each chunk was converted to an Arrow table with a pre-defined schema and streamed to a Parquet file via a `pyarrow.ParquetWriter`, preserving the original `sample` × `CpG` layout and avoiding any intermediate transposition or dense in-memory representation.

3) **Vertical concatenation and label attachment.** The three Parquet files (normal, adjacent-normal, tumor) were then loaded lazily using `pl.scan_parquet`. For each table, a numeric `label` column was added to encode the tissue class (`0 = normal`, `1 = normal-adjacent`, `2 = tumor`), and the columns were reordered to enforce a uniform schema of the form `[accession_num, basenames, label, CpG_1, ..., CpG_p]`. The three datasets were finally concatenated vertically and written to a single Parquet file with LZ4 compression using `sink_parquet`, yielding a unified, labeled methylation matrix covering all 595 samples.

Phenotypic metadata extraction. A dedicated phenotype table (`pheno_GSE225845.parquet`) was constructed from the GSE225845 SOFT record using the `GEOparse` API. For each GSM entry, the following attributes were extracted or derived:

- `geo_accession` — GEO sample accession (GSM identifier);
- `sample_name` — human-readable title of the sample;
- `source_name` — free-text description of the tissue source (e.g., frozen normal breast tissue);
- `sample_type_raw` — raw sample type string from `characteristics_ch1` (e.g., `normal`, `adj_norm`, `tumor`);
- `tissue_type_raw` — additional tissue descriptor when available (e.g., explicit "tissue:" fields);
- `age_at_surgery` — age at surgery, parsed from text and cast to `Float32`;
- `race` — reported race/ethnicity (e.g., European American, African American);
- `sex` — sex, normalized to uppercase codes (F/M);
- `idat_basename` — Illumina IDAT basename (`methylation id (basenames)`), linking each sample to the corresponding raw array files;
- `label_3class` — compact 3-class code derived from `sample_type_raw`, with `0 = normal`, `1 = normal-adjacent (adj_norm)`, `2 = breast cancer / tumor`.

All categorical labels were cast to `Int8`, and the resulting phenotype table was stored as an LZ4-compressed Parquet file to support efficient joins with the methylation matrix and downstream stratified analyses.

Biological replicates and sample-count discrepancy. The unified Parquet matrix contains 595 arrays (231 normal, 140 adjacent-normal, 224 tumors), exceeding the 402 samples reported in the original study description (185 tumors, 113 adjacent, 104 normal). To investigate this discrepancy, I first searched for potential *technical* duplicates by checking for repeated `idat_basename` values, replicated `sample_name` entries, and duplicate (`sample_type_raw`, `idat_basename`) combinations; no such duplicates were detected. Subsequently, I inspected the sample titles and extracted embedded numeric case identifiers using a regular expression, yielding a field `case_id_guess`. Many of these inferred identifiers appeared exactly twice, typically associated with different tissue states (e.g., normal vs. tumor or normal vs. adjacent-normal), indicating the presence of *biological* replicates (paired specimens from the same individual) rather than technical replication.

All arrays were retained in the working dataset to preserve the full normal–adjacent–tumor structure of the cohort. However, these pairings must be handled with particular care in downstream machine-learning experiments: training, validation, and test splits must be defined at the patient level (e.g., grouping by `case_id_guess`) to avoid data leakage between splits and overly optimistic performance estimates.

3.3 GSE287331 — Dataset preparation

Beta-value matrix construction. The working methylation matrix for GSE287331 (`GSE287331_lz4.parquet`) was constructed from the processed β-value CSV provided by the authors, using a memmap-based pipeline to handle the very high CpG dimensionality while keeping memory usage under control.

1) **Schema discovery and layout definition.** The processed matrix was supplied as a wide, tabular file with CpG probes as rows and samples as columns. The separator was inferred automatically, and the header row was read once with `pandas.read_csv` to identify the probe column (CpG identifiers) and the full set of sample columns. The total number of samples and CpGs was computed by inspecting the header and counting the remaining data rows, respectively.

2) **Memmap-based transposition in blocks.** To avoid materializing the full CpG × sample matrix in RAM, a `numpy.memmap` array was allocated on disk with shape `(samples, CpGs)` in Fortran (`column-major`) order and `Float32` dtype. The input file was then streamed in row blocks (50 000 CpGs at a time) using `pandas.read_csv` with the probe column kept as string and all sample columns parsed directly as `float32`. For each chunk, CpG identifiers were collected, the numeric block was transposed to `samples × block`, and written into the appropriate slice of the memmap. This strategy implements an out-of-core transposition, ensuring that at no point the full matrix needs to fit in memory.

3) **Arrow table construction and Parquet export.** After the memmap had been fully populated, an Arrow table was built column-by-column in a zero-copy fashion. The first column, `id_sample`, stores the sample identifiers derived from the original column names of the processed CSV; all remaining columns correspond to CpG probes, each backed directly by a contiguous `float32` slice from the memmap. The resulting table was written to a single Parquet file with LZ4 compression, using a fixed schema (`id_sample` as string, all CpG intensities as `Float32`) and disabling dictionary encoding for the floating-point columns. The temporary memmap file was removed at the end of the process, yielding a compact, columnar representation in the standard `sample` × `CpG` layout used throughout the thesis.

Phenotypic metadata extraction. A dedicated phenotype table (`pheno_GSE287331.parquet`) was constructed from the GSE287331 SOFT record using the `GEOparse` API. For each GSM sample, the following attributes were extracted:

- `geo_accession` — GEO sample accession (GSM identifier);
- `sample_name` — title of the sample, as provided in the GEO record;
- `tissue_type_raw` — raw tissue code parsed from `characteristics_ch1` entries of the form tissue: `HDB/CUB/OQ/AN/TU`;
- `idat_basename` — Illumina IDAT basename extracted from the description field (e.g., `202234810048_R01C01`), enabling linkage to the raw EPIC arrays;
- `label_3class` — numeric three-class label derived from `tissue_type_raw` via a deterministic mapping:
  - `0 = HDB` (healthy donated breast, normal controls),
  - `1 = CUB/OQ/AN` (contralateral unaffected breast, ipsilateral opposite quadrant, adjacent-normal; tumor-proximity axis),
  - `2 = TU/TUMOR` (breast cancer tissue).

The `label_3class` field was cast to `Int8` to minimize storage, and the full phenotype table was stored as an LZ4-compressed Parquet file. This compact representation provides a consistent three-level encoding of the tumor-proximity axis (normal, benign/adjacent, tumor), harmonized with the labeling schemes adopted for GSE69914 and GSE225845 and ready for downstream comparative analyses.

3.4 Harmonization and unified schema

All working methylation matrices were normalized to the same backbone: `id_tissue` (sample identifier), `label` (0=normal, 1=adjacent, 2=tumor), and genome-wide CpG columns stored as `Float32` in LZ4-compressed Parquet (layout: sample × CpG). Each dataset also has a phenotype table aligned one-to-one with the matrix, minimally containing `id_tissue` and `label`, plus dataset-specific attributes (e.g., `idat_basename`, clinical and technical fields). This unified schema guarantees clean joins, consistent preprocessing, and seamless cross-cohort comparisons across 450K and EPIC platforms.

3.5 Outlook and modelling strategy

The unified methylation matrices constructed in this stage provide the foundation for all subsequent preprocessing and modelling tasks. In the next phase, feature filtering, normalization, and bias correction will be applied consistently across datasets to ensure comparability. Machine-learning models will then be trained primarily on the largest available datasets (GSE225845 and GSE287331), while GSE69914 will serve as an independent baseline for cross-dataset validation.

To ensure unbiased generalization, training, validation, and test splits will be carefully defined at the *patient level*—particularly for datasets containing paired normal/adjacent/tumor samples—to prevent information leakage. Cross-validation strategies will be used within the training sets, and both within-dataset and across-dataset evaluations will be performed to study the reproducibility and transferability of methylation signatures between cohorts.

A multi-dataset analytical design will be adopted, inspired by recent large-scale methylation studies that evaluate both within-dataset performance and cross-cohort generalizability [8]. Following that approach, each dataset will first be analyzed individually—optimizing preprocessing and feature selection pipelines per cohort—before performing a combined meta-analysis to identify reproducible epigenetic alterations across studies.

References

[1] N. A. Bonakdar, "Epigenetic mechanisms in the development of neoplasms," Master's thesis, Data Science and Engineering. Supervisors: Alfredo Benso, Sandro Gambino, M.S. thesis, Politecnico di Torino, Department of Control and Computer Engineering, Turin, Italy, Jun. 2025. [Online]. Available: https://webthesis.biblio.polito.it/36339/1/tesi.pdf

[2] National Center for Biotechnology Information (NCBI), *Gse69914 – dna methylation profiles in breast tissue samples*, Processed using minfi v1.8.9 and BMIQ v1.4 as reported in the GEO metadata, Gene Expression Omnibus (GEO), 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE69914

[3] National Center for Biotechnology Information (NCBI), *Gse225845 – high neighborhood deprivation impacts dna methylation and gene expression in cancer-related genes*, Methylation profiles measured using the Illumina Infinium MethylationEPIC 850K BeadChip; processed β-values available for normal, adjacent-normal, and tumor samples., Gene Expression Omnibus (GEO), 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE225845

[4] National Center for Biotechnology Information (NCBI), *Gse287331 – dna methylation patterns in breast cancer, paired benign tissue, and healthy controls*, Profiles obtained using the Illumina Infinium MethylationEPIC v1.0 BeadChip; processed β-values and sample metadata available., Gene Expression Omnibus (GEO), 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE287331

[5] P. Du, X. Zhang, C. Huang, N. Jafari, W. A. Kibbe, L. Hou, and S. M. Lin, "Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis," *BMC Bioinformatics*, vol. 11, no. 1, p. 587, 2010. DOI: 10.1186/1471-2105-11-587 [Online]. Available: https://doi.org/10.1186/1471-2105-11-587

[6] A. E. Teschendorff, F. Marabita, M. Lechner, T. Bartlett, D. Tegner, J. Gomez-Cabrero, and S. Beck, "A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450k dna methylation data," *Bioinformatics*, vol. 29, no. 2, pp. 189–196, 2013. DOI: 10.1093/bioinformatics/bts680 [Online]. Available: https://academic.oup.com/bioinformatics/article/29/2/189/199637

[7] B. P. de Almeida, G. Richard, H. Dalla-Torre, C. Blum, L. Hexemer, P. Pandey, S. Laurent, M. Lopez, A. Laterre, M. Lang, and et al., "A multimodal conversational agent for dna, rna and protein tasks," *Nature Machine Intelligence*, 2025. DOI: 10.1038/s42256-025-01047-1 [Online]. Available: https://www.nature.com/articles/s42256-025-01047-1

[8] A. V. Sokolov and H. B. Schiöth, "Decoding depression: A comprehensive multi-cohort exploration of blood dna methylation using machine learning and deep learning approaches," *Translational Psychiatry*, vol. 14, no. 1, p. 326, 2024. DOI: 10.1038/s41398-024-02992-y [Online]. Available: https://www.nature.com/articles/s41398-024-02992-y