Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Department of Mathematics and Physics

Master of Data Science

A Bioinformatics and Deep Learning Approach Using Biological Databases to Evaluate Cell Physiological Function-Associated Genes in Medulloblastoma Mortality

---

THESIS to obtain the **DEGREE** of
MASTER OF DATA SCIENCE

A thesis presented by:
Andrés Rivera Chávez

Thesis Advisors:
Dr. Alba Adriana Vallejo Cardona
Dr. Rocío Carrasco Navarro

Tlaquepaque, Jalisco, August, 2026

Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Department of Mathematics and Physics

Master of Data Science Approval Form

Thesis Title: **A Bioinformatics and Deep Learning Approach Using Biological Databases to Evaluate Cell Physiological Function-Associated Genes in Medulloblastoma Mortality**

Author: **Andrés Rivera Chávez**

Thesis Approved to complete all degree requirements for the Master of Science Degree in Data Science.

Thesis Advisor, **Dr. Alba Adriana Vallejo Cardona**

Thesis Co-Advisor, **Dr. Rocío Carrasco Navarro**

Thesis Reader, **Dr. Jorge Isaac Chairez Oria**

Thesis Reader, **Dr. David José Mendoza Aguayo**

Academic Advisor, **Dr. Rocío Carrasco Navarro**

Tlaquepaque, Jalisco, August, 2026

A Bioinformatics and Deep Learning Approach Using Biological Databases to Evaluate Cell Physiological Function-Associated Genes in Medulloblastoma Mortality

Andrés Rivera Chávez

Abstract

Medulloblastoma is the most common malignant pediatric brain tumor, currently stratified into four molecular subgroups (WNT, SHH, Group 3, and Group 4) that differ substantially in clinical behavior, therapeutic response, and prognosis. Accurate transcriptomic classification of these subgroups remains methodologically demanding, requiring the integration of high-dimensional gene-expression data with classification architectures capable of exploiting subgroup-specific molecular signatures while controlling for the curse of dimensionality inherent to microarray feature spaces. This thesis presents a reproducible computational pipeline for the molecular subgroup classification of medulloblastoma from microarray transcriptomic data, extending the methodological framework of Reveles et al. (2025) to a substantially larger and independently constructed dataset. A total of 794 samples assembled from the GSE85217 and GSE124814 cohorts were harmonized across the GPL22286 and GPL16977 annotation platforms through a shared-genome intersection retaining approximately 89% of gene identifiers, then processed through stringent differential expression analysis (FDR < 0.005, $|\log_2 \mathrm{FC}| > 2.32$) yielding 1,139 unique discriminative genes as input features. Five feedforward artificial neural networks (one five-class classifier and four binary one-vs-rest classifiers) were implemented in MATLAB with three hidden layers, batch normalization, dropout, L_2 weight decay, and the Adam optimizer, and validated under stratified ten-fold cross-validation. The configuration achieved cross-validated accuracy at or above 97.99% across all five networks, with mean AUC values between 0.988 and 1.000, meeting or exceeding the current state of the art for transcriptomic medulloblastoma classification. Residual errors concentrated systematically along the Group 3 / Group 4 boundary, in agreement with the documented transcriptomic overlap between these subgroups and with the structure observed in unsupervised UMAP projections of the feature space, providing cross-method evidence that the residual error is biologically interpretable rather than methodological. Implemented under a deliberate commitment to reproducibility through open-source bioinformatics tooling, the pipeline is positioned as a methodological template for accessible computational oncology research, particularly in resource-limited research environments such as the Latin American context from which it emerges.

Keywords: medulloblastoma; molecular subgroup classification; transcriptomic data; differential expression analysis; artificial neural networks; deep learning; cross-platform microarray harmonization; UMAP; reproducible bioinformatics; computational oncology.

A Bioinformatics and Deep Learning Approach Using Biological Databases to Evaluate Cell Physiological Function-Associated Genes in Medulloblastoma Mortality

Andrés Rivera Chávez

Resumen

El meduloblastoma es el tumor cerebral pediátrico maligno más frecuente, y actualmente se estratifica en cuatro subgrupos moleculares (WNT, SHH, Grupo 3 y Grupo 4) que difieren sustancialmente en comportamiento clínico, respuesta terapéutica y pronóstico. La clasificación transcriptómica precisa de estos subgrupos sigue siendo metodológicamente exigente, ya que requiere la integración de datos de expresión génica de alta dimensionalidad con arquitecturas de clasificación capaces de aprovechar las firmas moleculares específicas de cada subgrupo, controlando al mismo tiempo la maldición de la dimensionalidad inherente a los espacios de características derivados de microarreglos. La presente tesis propone una metodología computacional reproducible para la clasificación por subgrupo molecular del meduloblastoma a partir de datos transcriptómicos de microarreglos, extendiendo el marco metodológico de Reveles et al. (2025) a un conjunto de datos sustancialmente mayor y construido de manera independiente. Un total de 794 muestras, ensambladas a partir de los cohortes GSE85217 y GSE124814, fueron armonizadas entre las plataformas de anotación GPL22286 y GPL16977 mediante una intersección de genes compartidos que retuvo aproximadamente el 89% de los identificadores génicos, y posteriormente procesadas mediante un análisis de expresión diferencial estricto (FDR < 0.005, $|\log_2 \mathrm{FC}| > 2.32$), del cual se obtuvieron 1,139 genes discriminativos únicos como características de entrada. Cinco redes neuronales artificiales de tipo *feedforward* (una clasificadora de cinco clases y cuatro clasificadoras binarias *uno-contra-resto*) fueron implementadas en MATLAB con tres capas ocultas, normalización por lotes, regularización por *dropout*, decaimiento de pesos L_2 y el optimizador Adam, y validadas bajo validación cruzada estratificada de diez pliegues. La configuración alcanzó una exactitud validada cruzadamente igual o superior al 97.99% en las cinco redes, con valores medios de AUC entre 0.988 y 1.000, igualando o superando el estado del arte actual en clasificación transcriptómica de meduloblastoma. Los errores residuales se concentraron sistemáticamente a lo largo de la frontera Grupo 3 / Grupo 4, en concordancia con el solapamiento transcriptómico documentado entre estos subgrupos y con la estructura observada en las proyecciones no supervisadas de UMAP del espacio de características, lo que aporta evidencia *cross-method* de que el error residual es biológicamente interpretable y no de naturaleza metodológica. Implementada bajo un compromiso deliberado con la reproducibilidad mediante herramientas bioinformáticas de código abierto, la metodología se plantea como una plantilla metodológica para la investigación accesible en oncología computacional, particularmente en entornos de investigación con recursos limitados, como el contexto latinoamericano del que emerge.

Palabras clave: meduloblastoma; clasificación por subgrupo molecular; datos transcriptómicos; análisis de expresión diferencial; redes neuronales artificiales; aprendizaje profundo; armonización entre plataformas de microarreglos; UMAP; bioinformática reproducible; oncología computacional.

Contents

Page

1 Introduction 25
- 1.1 Justification 29
- 1.2 Problem Statement 32
- 1.3 Research Question 33
- 1.4 Hypothesis 34
- 1.5 Objectives 34
  - 1.5.1 Specific Objectives 35
- 1.6 Thesis Proposal 35
- 1.7 Novel Contributions 36

2 Theoretical and conceptual framework 39
- 2.1 Medulloblastoma 39
- 2.2 Bioinformatics and Computational Biology 41
  - 2.2.1 Multiomics Data Integration 42
- 2.3 Microarray Technology and Gene Expression Analysis 44
  - 2.3.1 Public Biological Databases and Data Repositories 48
  - 2.3.2 Gene Expression Omnibus (GEO) 48
- 2.4 Microarray Data Analysis 51
- 2.5 Microarray Data Preprocessing 53
- 2.6 Fundamentals of Machine Learning and Deep Learning 62
  - 2.6.1 Artificial Neural Networks 63
  - 2.6.2 Activation Functions 66
  - 2.6.3 Feedforward Neural Network Topology . . . 67
  - 2.6.4 Loss Function 68

3 State of the Art 71
- 3.1 Deep Learning Methods and Applications on Molecular Data 71
- 3.2 Deep Learning Techniques for Survival Prediction in Pancreatic Cancer 73
  - 3.2.1 Deep Learning Architecture: LSTM Networks 75
- 3.3 ANN-Based Molecular Classification of Medulloblastoma Subgroups 78
  - 3.3.1 Artificial Neural Network Architecture 81

4 Methodology . . . 85
4.1 Experimental Design and Data Collection . . . 85
4.1.1 Experimental Design Considerations . . . 87
4.1.2 Data Mining and Repository Search . . . 89
4.2 Metadata Preprocessing and Sample Identification . . . 90
4.3 Gene Expression Data Processing and Differential Expression Analysis . . . 95
4.3.1 Microarray Data Acquisition and Preprocessing 97
4.3.2 Quality Control and Normalization . . . 98
4.3.3 Differential Expression Analysis . . . 101
4.4 Deep Learning Classification . . . 109
4.4.1 Network Architectures . . . 110
4.4.2 Layer Construction and Regularization . . . 110
4.4.3 Training Configuration . . . 111
4.4.4 Cross-Validation Protocol . . . 111
4.4.5 Evaluation Metrics and Outputs . . . 112
5 Results . . . 113
5.1 Overview of Classification Performance . . . 113
5.2 Five-class Classifier: CT vs. WNT vs. SHH vs. G3 vs. G4 114
5.3 Binary Classifiers . . . 116
5.4 Computational Cost Analysis . . . 121
5.5 Comparison with the State of the Art . . . 124
6 Discussions and Conclusions . . . 127
6.1 Summary of Findings . . . 129
6.2 Limitations . . . 132
6.2.1 Subtyping of Medulloblastoma . . . 132
6.2.2 Normalization Strategies . . . 134
6.2.3 Single Platform and Dataset Validation . . . 135
6.2.4 Mono-omic Experimental Design . . . 136
6.2.5 Wet Laboratory Confirmation . . . 137
6.3 Future Work . . . 137
6.3.1 Molecular Subtype of Medulloblastoma´s Subgroups . . . 138
6.3.2 Normalization Strategies Benchmark . . . 139
6.3.3 Molecular Linear Separability Hypothesis . . . 140
6.3.4 Gene Ontology and Gene Set Enrichment Analysis . . . 140
6.3.5 Migration to Open-Source Pipelines . . . 141
6.3.6 Multiomics Data Integration . . . 142
6.4 Final Considerations . . . 142
Bibliography . . . 145
Index . . . 157

List of Figures

Page

1.1 Venn diagram illustrating the interdisciplinary relationships among bioinformatics, information science, artificial intelligence, machine learning, and deep learning. The central intersection represents the convergence and application of AI, ML, and DL across multiple domains, including bioinformatics. 28

1.2 Molecular subgroups of medulloblastoma according to WHO 2016 classification. The four subgroups are defined by transcriptomic profiles: WNT group (10% of cases, best prognosis, characterized by WNT signaling pathway activation), SHH group (30% of cases, associated with age <4 years and >16 years, somatic mutations in SHH pathway), Group 3 and Group 4 (most common subtypes with worst outcomes due to high metastatic capacity and heterogeneity). 31

2.1 Schematic representation of multi-omics data flow across the central dogma of molecular biology, the current scope focus on nucleic acid-based methods (genomics and transcriptomics via microarray analysis). Adapted from Thermo Fisher Scientific [Thermo Fisher Scientific, n.d.]. 43

2.2 Microarray experimental representation of two different cellular lines that derive in tumors with different prognosis. Adapted from reference image on [Herráez, 2012]. 46

2.3 Microarray chips compared to a match for reference. . . 47

2.4 Schema representing relation among the three main subcategories of data in GEO. There is One to many between Platform and Samples whereas a Many to Many relation between Samples and Series. 49

2.5 General overview of microarray processing from experimental design to functional enrichment analysis based on [Federico et al., 2021] & [Dhar and Kallapura, 2022]. 51
2.6 MA plot representation of gene expression distribution between two different experimental conditions using microarray data from GEO Dataset ID: GSE34248[Federico et al., 2021]. 55
2.7 Histogram showing the distribution of gene expression estimates in a microarray experiment [Federico et al., 2021]. 55
2.8 Density plot of gene expression between lesional and non lesional skin samples. The largely overlapping curves reflect the proportion of probes exhibiting a given intensity, with most probes showing low intensity in this experiment[Federico et al., 2021]. 56
2.9 Scatter plot comparing gene expression estimates between two samples from dataset GSE34248 [Federico et al., 2021]. 56
2.10 UMAP plot showing the projection of samples into a low-dimensional space. The circled sample, given its isolated position, represents a candidate outlier [Federico et al., 2021]. 57
2.11 Boxplots showing gene expression distribution of lesional and non lesional skin of atopic dermatitis patients before (A) and after (B) between array normalization [Federico et al., 2021]. 59
2.12 Volcano plot that shows the magnitude of deregulation in terms of log fold changes and the p-values of the genes after differential analysis, the blue dots represent differential expressed genes [Federico et al., 2021]. 61
2.13 General model of an Artificial Neural Network (ANNs) [Montesinos López et al., 2022]. 64
2.14 Feedforward ANN architecture with eight input variables $(x_1, ..., x_8)$, four output variables (y_1, y_2, y_3, y_4), and two hidden layers with three neurons each [Montesinos López et al., 2022]. 65
2.15 Representation of ReLU activation function [Montesinos López et al., 2022]. 67
2.16 Representation of tanh activation function [Montesinos López et al., 2022]. 67
2.17 Simple two-layer feedforward artificial neural network [Montesinos López et al., 2022]. 68

3.1 Procedure following a general methodology towards identifying the survival status in patients with pancreatic cancer based on DL focused on detecting possible markers or therapeutic targets [Salgado et al., 2024]. . . 75
3.2 LSTM general topology used in the design of classification status across a multiomics approach [Salgado et al., 2024]. 76
3.3 Diagram of the general methodology to identify key genes in MB subgroup classification [Reveles-Espinoza et al., 2025]. 79

4.1 Overview of the integrated methodological pipeline. The four sequential stages — data acquisition and experimental design, metadata harmonization, bioinformatics preprocessing and differential expression analysis, and deep learning classification — are connected by well-defined intermediate outputs. Each stage is developed in full in the corresponding section of this chapter. 87
4.2 Distribution of medulloblastoma molecular subgroups in the GSE85217 dataset. All molecular subgroups exceed the minimum calculated sample size ($n = 63$) required for adequate statistical power (shown as dashed line). . 90
4.3 Proportional distribution of samples across annotation versions. GPL22286 (ENSG v19.0.0) accounts for the medulloblastoma samples ($n = 763$) and GPL16977 (ENSG v14.1.0) for the cerebellar controls ($n = 31$), both derived from the same Affymetrix Human Gene 1.1 ST Array chip. 91
4.4 Venn diagram showing the overlap of gene identifiers between GPL22286 (ENSG v19.0.0, used for medulloblastoma samples) and GPL16977 (ENSG v14.1.0, used for cerebellar controls). Both annotation versions are derived from the same physical Affymetrix Human Gene 1.1 ST Array chip, with 20,372 genes (>89%) shared between platforms. 91
4.5 Sample distribution across molecular subgroups and cerebellar controls in the integrated cohort ($n = 794$). Tumor subgroups (WNT, SHH, G3, G4) are derived from GSE85217 and controls (CT) from GSE45878. 93
4.6 Age distribution across the integrated cohort. Median age of 8.0 years is consistent with the known pediatric predominance of medulloblastoma [National Cancer Institute (NCI), 2025]. 94

4.7 Age distribution stratified by condition (tumor vs. normal). The separation between pediatric cancer samples and adult cerebellar controls reflects the distinct cohort origins of GSE85217 and GSE45878. 94

4.8 Gender distribution across molecular subgroups and cerebellar controls. Male predominance of 50-70% per subgroup is consistent with established medulloblastoma epidemiology. 95

4.9 Proportional gender heatmap across all conditions. Values represent the relative frequency of male and female samples within each molecular subgroup and control group. 95

4.10 Metadata preprocessing pipeline for the integrated medulloblastoma cohort ($n = 794$). Cancer samples (GSE85217) and cerebellar controls (GSE45878) are processed through independent tracks before harmonization into a unified dataset for downstream gene expression analysis. 96

4.11 Bioinformatics pipeline for gene expression data processing and differential expression analysis. The pipeline comprises 16 sequential steps organized across six phases: data validation, loading and preprocessing, quality control, LIMMA-based statistical modeling, results extraction and annotation, and output generation. Two hyperparameter configurations were applied at the differential expression stage, corresponding to distinct statistical thresholds established in the literature, yet the main data downstream analysis considered for the present work follows the hyperparameters reported by [Castillo-Rodríguez et al., 2018, Reveles-Espinoza et al., 2025]. 97

4.12 Box-and-whisker plot of mean log2 expression per molecular subgroup and cerebellar control (CT) following RMA normalization. Each box summarizes the distribution of group-averaged expression values across 33,297 gene-level features. Comparable median levels and interquartile ranges across all five groups confirm successful quantile normalization. 99

4.13 Expression value density curves per sample following RMA normalization. Each line represents one of the 794 samples, colored by molecular subgroup. Overlapping unimodal distributions confirm technical consistency across samples. Cerebellar control samples exhibit a sharper density peak, reflecting greater transcriptional homogeneity relative to tumor subgroups. 100
4.14 UMAP projection of 794 samples following RMA normalization (`n_neighbors = 15`, `random_state = 123`). Each point represents one sample, colored by molecular subgroup. WNT and SHH subgroups from tight well-separated clusters, cerebellar controls occupy a distinct isolated region, and Group 3 and Group 4 exhibit partial overlap consistent with their molecular similarity. . . . 100
4.15 Mean-variance trend plot following empirical Bayes moderation. The relatively flat blue trend line cross the expression range confirms that residual variance is not dependent on expression level, validating the standard `limma` modeling approach for this microarray dataset. . 103
4.16 Adjusted p-value distribution across all 17,032 genes following empirical Bayes moderation. Concentration of values below 0.05 with negligible counts across the remainder of the distribution indicates a pervasive differential expression signal across all four subgroup versus cerebellar control contrasts. 103
4.17 Q-Q plot of moderated t-statistics against the theoretical t-distribution. Near-perfect alignment across the central quantile range with pronounced tail divergence — sample quantiles reaching approximately ± 45 against theoretical limits of ± 4 — confirms appropriate model calibration and reflects the presence of strongly differentially expressed genes. 104
4.18 Volcano plot for G3 vs. CT contrast (FDR < 0.005, $|\log_2 \mathrm{FC}| > 2.32$). A total of 534 DEGs were identified (260 upregulated, 274 downregulated). Red points: significantly upregulated genes; blue points: significantly downregulated genes; gray points: non-significant genes. Dashed lines indicate the applied significance thresholds. 105
4.19 Volcano plot for G4 vs. CT contrast (FDR < 0.005, $|\log_2 \mathrm{FC}| > 2.32$). A total of 467 DEGs were identified (251 upregulated, 216 downregulated). Red points: significantly upregulated genes; blue points: significantly downregulated genes; gray points: non-significant genes. Dashed lines indicate the applied significance thresholds. 106

4.20 Volcano plot for SHH vs. CT contrast (FDR < 0.005, $|\log_2 \mathrm{FC}| > 2.32$). A total of 477 DEGs were identified (268 upregulated, 209 downregulated). Red points: significantly upregulated genes; blue points: significantly downregulated genes; gray points: non-significant genes. Dashed lines indicate the applied significance thresholds. 106

4.21 Volcano plot for WNT vs. CT contrast (FDR < 0.005, $|\log_2 \mathrm{FC}| > 2.32$). A total of 708 DEGs were identified (404 upregulated, 304 downregulated), the highest among all four contrasts. Red points: significantly upregulated genes; blue points: significantly downregulated genes; gray points: non-significant genes. Dashed lines indicate the applied significance thresholds. 107

4.22 Venn diagram showing DEG overlap across the four subgroup versus cerebellar control contrasts (FDR < 0.005, $|\log_2 \mathrm{FC}| > 2.32$). Numbers indicate genes meeting both significance criteria in each region. The central intersection of 174 genes represents a core transcriptional signature common to all four molecular subgroups. Subgroup-exclusive DEGs range from 76 (G4–CT) to 271 (WNT–CT). The value 15,893 represents non-significant features outside all four contrasts. 107

5.1 Aggregate confusion matrix for the five-class classifier across all ten cross-validation folds. Row and column summaries report class-wise recall and precision, respectively. 114

5.2 Per-fold classification accuracy for the five-class network. The dashed horizontal line indicates the mean across folds (97.99%). 115

5.3 One-vs-rest ROC curves for the five-class classifier. Per-class AUC values are reported in the legend; the mean AUC across classes is 0.999. 115

5.4 Confusion matrix for the WNT-vs-rest binary classifier aggregated across all ten cross-validation folds. 116

5.5 Per-fold accuracy for the WNT-vs-rest classifier. Eight of ten folds achieved 100% accuracy. 117

5.6 ROC curves for the WNT-vs-rest classifier. Both classes achieved AUC = 1.000, indicating perfect probabilistic separability across the cross-validation procedure. . . . 117

5.7 Confusion matrix for the SHH-vs-rest classifier. A single misclassification occurred across all 794 samples evaluated during cross-validation. 118

5.8 Per-fold accuracy for the SHH-vs-rest classifier. Nine of ten folds achieved perfect classification. 118
5.9 ROC curves for the SHH-vs-rest classifier, with AUC = 0.996 for both the positive and rest classes. 119
5.10 Confusion matrix for the G3-vs-rest classifier. Eleven of 144 G3 samples were misclassified as non-G3, and 5 non-G3 samples were predicted as G3. 120
5.11 Per-fold accuracy for the G3-vs-rest classifier. The highest fold-to-fold variability among all five networks (range 93.75%–100%) reflects the transcriptomic overlap between Group 3 and Group 4. 120
5.12 ROC curves for the G3-vs-rest classifier, with AUC = 0.988 for both classes. 121
5.13 Confusion matrix for the G4-vs-rest classifier. Three G4 samples and four non-G4 samples were misclassified across the ten folds. 122
5.14 Per-fold accuracy for the G4-vs-rest classifier, ranging from 97.47% to 100% across folds. 122
5.15 ROC curves for the G4-vs-rest classifier, with AUC = 0.999 for both classes. 123
5.16 Total training time (red bars, left axis) and mean cross-validated accuracy (blue line, right axis) for each of the five networks. 123

6.1 Sample distribution across the twelve molecular subtypes reported in the GSE85217 metadata, following the schema of Cavalli et al. [2017]. Total $n = 763$. Several subtypes fall below the 63-sample minimum threshold derived from the power analysis, which motivated the restriction of the present work to the four canonical subgroups. 134

List of Tables

Page

3.1 Data types and features retrieved from the GDC repository [Salgado et al., 2024]. 74
3.2 LSTM-based classifier architecture employed by Salgado et al. [Salgado et al., 2024]. 76
3.3 Performance metrics used to evaluate the LSTM-based classifier [Salgado et al., 2024]. TP: true positives; TN: true negatives; FP: false positives; FN: false negatives. . 77
3.4 Classification accuracy and processing time per omics subset for the proposed LSTM classifier [Salgado et al., 2024]. 78
3.5 Confusion matrix for the all-subsets LSTM classification configuration [Salgado et al., 2024]. 78
3.6 Sample distribution across MB subgroups and control . 80
3.7 ANN configuration for binary classifiers [Reveles-Espinoza et al., 2025]. 81
3.8 Binary classifier validation accuracy 82

4.1 Power analysis parameters used for minimum sample size estimation. 88
4.2 GPL platform annotation comparison 90
4.3 Contrast matrix specifying the four subgroup versus cerebellar control comparisons defined for differential expression analysis. Values of $+1$ and -1 indicate the subgroup and reference group respectively; 0 indicates groups not involved in the comparison. 102
4.4 Summary of differentially expressed genes per molecular subgroup versus cerebellar control contrast under the primary configuration (FDR < 0.005, $|\log_2 \mathrm{FC}| > 2.32$). Up: upregulated genes; Down: downregulated genes. . 104
4.5 Feedforward network topologies adapted from Reveles-Espinoza et al. [2025]. Input dimensionality corresponds to the 1,139 DEGs identified in the present pipeline. Hidden-layer widths are given as neurons per layer. . . 110

4.6 Training hyperparameters applied uniformly across all five networks. Validation data correspond to the held-out fold within each cross-validation iteration. 111

5.1 Summary of 10-fold cross-validated performance across all five ANN architectures. Accuracy is reported as mean ± standard deviation across folds; mean AUC corresponds to the average one-vs-rest area under the ROC curve across classes within each network. Total time refers to cumulative wall-clock training across all ten folds on an Intel Core i9-14900HX workstation with 32 GB RAM. 113

5.2 Comparison of cross-validated accuracy obtained in the present work against the reference values reported by Reveles-Espinoza et al. [2025]. Both studies employ the same hierarchical ANN framework with one multi-class network and four one-vs-rest binary networks. . . 124

5.3 Comparison of accuracy, processing time, and computational cost (FLOPs) between the multi-omic ANN classifiers reported by Salgado et al. [2024] for pancreatic cancer and the present medulloblastoma classification pipeline. 125

Dedicated to:

I wish to express my eternal gratitude to my family for always having been by my side when I most needed it. To my mothers: to Aracelia Rivera Chávez, for her infinite patience and absolute love; to Doctor Elba Rivera Chávez, for her guidance and her light throughout the years as my great mentor and role model; and to Professor Estela Guadalupe Rivera Chávez, who, with infinite patience and dedication, encouraged and supported each and every one of my life projects. I deeply appreciate their efforts and their unwavering commitment over the years, and I recognize that, despite the turbulences, they have always been and will always be there for me. Without them I would not be the man I am today, and it is for this reason that I dedicate my personal and professional life to the values they instilled in me from an early age.

To Doctor Alba Adriana Vallejo Cardona, who, in her admirable career, was able to see the light in me and guide me towards a professional purpose devoted to the development of applied science in biomedicine, for the sake of a better future for Mexico. I deeply thank her for her infinite patience and for the trust she placed in me, manifested in every word of encouragement and support she offered me during the most difficult moments of my personal and family life, as well as in the time and space she generously granted me to carry forward the projects and work that sustain my scientific vocation. Thank you for believing in me when I needed it most.

To Doctor Rocío Carrasco Navarro, who accompanied me from the very first steps of my graduate training and never strayed from my path, for the countless hours of guidance, patience, and dedication that she devoted to the development of professional work worthy of a Jesuit university. Her steadfast commitment to my human and professional development, as well as her words of encouragement regarding the matters that concern the historical juncture of Artificial Intelligence and its application in favour of human development, have been a fundamental pillar in my journey. I deeply thank her for having believed in me from the beginning and for having accompanied me with the same firmness through to the end.

To Doctor Jorge Isaac Chairez Oria, who always knew how to guide me and ground me in the rigor of doing science, invariably opening the doors of his institution and his laboratory to orient and motivate me in projects in favor of Mexico's scientific development. I thank him for having supported and encouraged me during the difficult

moments of my life without ever pressuring me, offering me the space and understanding I needed to keep moving forward. His trust and his accompaniment will always remain a reference in my professional path.

To ITESO, whose Jesuit and humanistic formation guided me from a very young age under the spirit of seeking not to be the best in the world, but the best for the world.

To CIATEJ and SECIHTI, who, by placing their trust in me and in the hundreds of young people who seek a better future for Mexico, drive the development of science and technology in favor of the Mexican people.

Finally, to my friends who became brothers, who, without knowing it, were a silent and indispensable companionship during this stage of my life. Their presence on good days and on bad ones — amid laughter, tears, and the simple everyday closeness — made the darker days a little brighter and the heavier ones feel lighter. Thank you for your unconditional support, offered without my ever having to ask for it, and for being there even when you yourselves were not entirely aware of how much you were sustaining me.

From the bottom of my heart, thank you all.

Dedicado a:

Quiero expresar mi eterna gratitud a mi familia por haber estado siempre conmigo cuando más lo necesité. A mis madres: a Aracelia Rivera Chávez, por su infinita paciencia y amor absoluto; a la Doctora Elba Rivera Chávez, por su guía y su luz a lo largo de los años como mi gran mentora y ejemplo a seguir; y a la Maestra Estela Guadalupe Rivera Chávez, quien, con infinita paciencia y dedicación, alentó y apoyó todos y cada uno de mis proyectos de vida. Agradezco profundamente sus esfuerzos y su entrega a lo largo de los años, y reconozco que, a pesar de los tumultos, siempre estuvieron y estarán para mí. Sin ellas no sería el hombre que soy hoy, y es por ello que dedico mi vida personal y profesional a los valores que me inculcaron desde pequeño.

A la Doctora Alba Adriana Vallejo Cardona, quien, en su admirable trayectoria, supo ver la luz en mí y guiarme hacia un propósito profesional dedicado al desarrollo de la ciencia aplicada en biomedicina, en pro de un mejor futuro para México. Le agradezco profundamente su infinita paciencia y la confianza que depositó en mí, manifestadas en cada palabra de aliento y de apoyo que me ofreció durante los momentos más difíciles de mi vida personal y familiar, así como en el tiempo y el espacio que generosamente me brindó para llevar adelante los proyectos y trabajos que sostienen mi vocación científica. Gracias por creer en mí cuando más lo necesité.

A la Doctora Rocío Carrasco Navarro, quien me acompañó desde los primeros pasos de mi formación de posgrado y nunca se apartó de mi camino, por las incontables horas de guía, paciencia y dedicación que entregó al desarrollo de un trabajo profesional digno de una universidad jesuita. Su compromiso constante con mi desarrollo humano y profesional, así como sus palabras de aliento en torno a los temas que conciernen a la coyuntura histórica de la Inteligencia Artificial y su aplicación en favor del desarrollo humano, han sido un pilar fundamental en mi recorrido. Le agradezco profundamente haber creído en mí desde el principio y haberme acompañado con la misma firmeza hasta el final.

Al Doctor Jorge Isaac Chairez Oria, quien siempre supo guiarme y aterrizarme en el rigor de hacer ciencia, abriéndome invariablemente las puertas de su institución y de su laboratorio para orientarme y motivarme en proyectos a favor del desarrollo científico de México. Le agradezco haberme apoyado y alentado en los momentos difíciles de mi vida sin presionarme jamás, ofreciéndome el espacio y la comprensión

que necesité para seguir adelante. Su confianza y su acompañamiento serán siempre una referencia en mi camino profesional.

Al ITESO, cuya formación jesuita y humana me guío desde muy joven bajo el espíritu de no buscar ser los mejores del mundo, sino los mejores para el mundo. Al CIATEJ y la SECIHTI, que, al confiar en mí y en cientos de jóvenes que buscamos un mejor futuro para México, impulsa el desarrollo de la ciencia y tecnología en favor del pueblo de México.

Por último, a mis amigos que se convirtieron en hermanos, quienes, sin saberlo, fueron una compañía silenciosa e indispensable durante esta etapa de mi vida. Su presencia en los días buenos y en los malos, entre risas, lágrimas y la simple compañía cotidiana, hizo que los días oscuros fueran un poco más claros y que las jornadas más pesadas se sintieran más ligeras. Gracias por su apoyo incondicional, ofrecido sin que yo tuviera que pedirlo, y por haber estado ahí incluso cuando ustedes mismos no eran del todo conscientes de lo mucho que me sostenían.

De todo corazón a todos ustedes gracias.

1 Introduction

Contents

1.1 Justification 29
1.2 Problem Statement 32
1.3 Research Question 33
1.4 Hypothesis 34
1.5 Objectives 34
1.5.1 Specific Objectives 35
1.6 Thesis Proposal 35
1.7 Novel Contributions 36

Cancer research presents unique bioinformatics challenges that extend beyond simply managing large databases or applying deep learning algorithms to obtain predictive outcomes. The comprehensive characterization of complex diseases requires a combination of classical statistical methods and advanced computational approaches, particularly when dealing with non-parametric nature of genomic data. In typical cancer genomic studies, researchers face the challenge of analyzing thousands of genes from small patient cohorts, creating a high-dimensional problem where the number of features vastly exceeds the number of samples. This scenario demands not only sophisticated computational resources and biological information systems, but also careful statistical consideration of the data preparation and analysis pipeline[1].

[1] Giuseppe Agapito. Preface. In Giuseppe Agapito, editor, *Microarray Data Analysis*. Springer, 2022

Over the last decades, continuous advancements in omics experimental assays, including microarray technologies which is the chosen tool for the present work, have contributed to producing enormous amounts of data. A standard microarray experiment comprises few samples, often less than 100, but the number of investigated features in a single experiment ranges from 60,000 up to 400,000 DNA spots. Each spot holds a short synthetic, single-stranded DNA sequence from a different human gene, making it possible to carry out a massive number of genetic test on a sample at one time. Microarray data are obtained through case-control studies by collecting information from tissue and cell samples regarding the expression level

of many genes that could be valuable for diagnosing diseases of for distinguishing specific mutation types. This high dimensionality is particularly challenging due to technical and biological noise, making microarray data analysis very appealing for machine learning and statistical researchers [Agapito, 2022].

As these technologies have evolved, data analysis methods have improved as well, especially in bioinformatics to support the storage, management, and analysis of large amounts of data about different aspects of the omics world. With the vast amounts of data now available in public and private repositories, there is an increasing need for the development of fast and accurate analysis methods and algorithms. The availability to improve complex disease understanding is no longer limited by the production of data or its accessibility, but rather by the proper ability to analyze it. This is why microarray data analysis represents an essential step for developing accurate diagnostic and prognostic tools for complex diseases [Agapito, 2022].

The traditional bioinformatics analysis workflow has typically involved downloading public datasets from repositories such as NCBI and ENSEMBL, installing software locally and analyzing data in-house. However modern requirements demand a higher level of integration that enhances both the storage and analysis of bioinformatics big data[2]. Thousands of interesting gene expression datasets are available on the Gene Expression Omnibus (GEO), a public database curated by the US National Institute of Health (NIH). To effectively access and read these datasets, Gentleman and colleagues started the development in 2002 of an open-source R software package platform called Bioconductor[3], which has become popular among bioinformaticians worldwide[4].

To analyze microarray data effectively, data mining techniques borrowed from the statistical and computer sciences worlds represent a practical toolbox with strong theoretical foundations. These techniques, although stemming from different disciplines, result in a complete new set of possibilities that can be adopted in microarray analysis in particular, and bioinformatics in general. The fast growing field of data mining, also known as knowledge discovery from data, can be defined as the results of a natural evolution in information technology. It evolved from the process of data collection and database creation, which have roots mostly in file processing, to the process of database system management. From these foundations, advanced database systems and techniques have arisen, such as web-based databases, data warehouses, and online analytical processing (OLAP). Several of these techniques, as well as the general process of knowledge discovery, have been introduced in the bioinformatics context along with pure computer science techniques such as sequence comparison and alignment. In addition, all these techniques are used in conjunction with machine

[2] Barbara Calabrese. Web and cloud computing to analyze microarray data. In Giuseppe Agapito, editor, *Microarray Data Analysis*, chapter 3. Springer, 2022

[3] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004

[4] Davide Chicco. geneexpressionfromgeo: An r package to facilitate data reading from gene expression omnibus (geo). In Giuseppe Agapito, editor, *Microarray Data Analysis*, chapter 12. Springer, 2022

learning, artificial intelligence, and parallel computing methods, thus expanding their applicability[5].

Machine Learning (ML) is a subdivision of Artificial Intelligence (AI), it enables systems to learn independently from historical data. These learning systems facilitate the development of algorithms and statistical models that allow computers to extract patterns from data and make predictions or decisions based on new inputs. Such techniques applied in molecular biology are able to investigate in depth the hidden mechanisms of gene expression in human genome[6]

In the field of ML, there are two main subcategories. The first is supervised learning, which relies on labeled datasets to train algorithms to classify data and make predictions based on known patterns. This approach is used to develop decision trees, Support Vector Machines (SVMs), and Artificial Neural Networks (ANNs), this last one will be extensively utilized in the methodology presented in this research. The second type is unsupervised learning, which operates without labeled data. Instead, these algorithms autonomously discover underlying patterns and structures within datasets, analogous to certain learning processes in the human brain. Common unsupervised learning methods include clustering and dimensionality reduction techniques.

Deep Learning (DL), a subfield of ML, is based on Artificial Neural Networks (ANNs) that are inspired by the biological structure of neurons in the human brain. DL algorithms consist of multiple layers of interconnected nodes, where each layer performs specific operations on the data it receives, enabling hierarchical feature extraction. Within the field of bioinformatics, ML has been successfully applied to protein structure predictions, biomolecular interaction analysis, gene expression profiling, and disease diagnosis. With the current exponential growth of biomedical data, there is an increasing need to apply ML, DL, and Data Science techniques to address emerging challenges in biology, biotechnology, bioinformatics and clinical research, the relationship between these areas can be seen in Figure 1.1[7]. These techniques enable critical tasks such as feature extraction, feature selection, and the development of computational predictive models capable of analyzing complex biological systems using bioinformatic databases and repositories[8].

ML and DL techniques are frequently integrated with bioinformatics methods, curated databases, and biological networks to enhance model training and validation, identify interpretable features, and ultimately enable comprehensive biological characterization and predictive modeling[9]. One of the most common applications of ML in bioinformatics is the identification of disease-associated genes. This research focuses on identifying genes potentially linked to specific diseases by analyzing gene expression through microarrays. Notably,

[5] Francesco Cauteruccio. Alignment of microarray data. In Giuseppe Agapito, editor, *Microarray Data Analysis*, chapter 14. Springer, 2022

[6] Anna Bernasconi and Silvia Cascianelli. Scenarios for the integration of microarray gene expression profiles in covid-19–related studies. In Giuseppe Agapito, editor, *Microarray Data Analysis*, chapter 19. Springer, 2022

[7] Abhinav Dey, Anshu Malhotra, and Neha Garg. Medullo-blastoma

[8] M. Yousef and J. Allmer. Deep learning in bioinformatics. 47(06)

[9] Noam Auslander, Ayal B. Gussow, and Eugene V. Koonin. Incorporating machine learning into established bioinformatics frameworks. *International Journal of Molecular Sciences*, 22(6):2903, 2021. DOI: 10.3390/ijms22062903

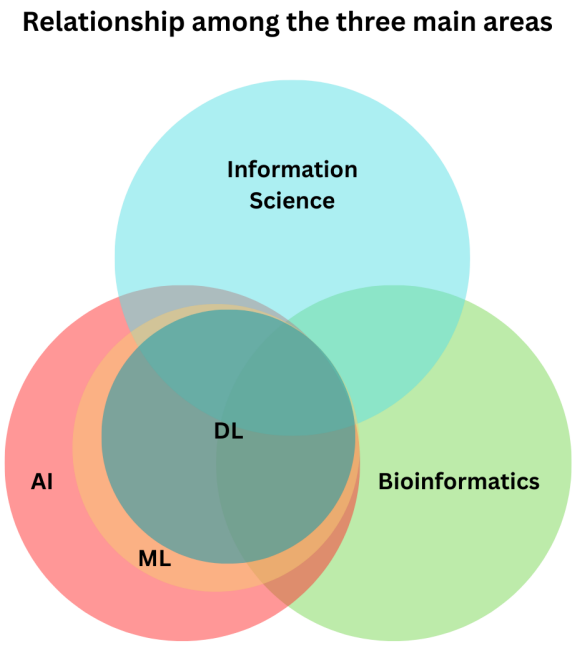

Figure 1.1: Venn diagram illustrating the interdisciplinary relationships among bioinformatics, information science, artificial intelligence, machine learning, and deep learning. The central intersection represents the convergence and application of AI, ML, and DL across multiple domains, including bioinformatics.

gene identification has gained considerable importance in studies aimed at detecting genes likely to contribute to cancer progression and tumor classification at the molecular level.

As microarray data by itself might not be enough to properly analyze and understand the mechanisms or predicting a disease state, a frequent goal in microarray analyses is to understand the molecular mechanism beneath a disease, drug response, and other physiological states. This is where differential expression analysis is able to identify genes involved in the mechanism but does not reveal how these genes work together to reach the physiological state. Then a subsequent stage is required for microarray analysis which is the prediction of current or future physiological state of a sample from gene expression data. An example of this would be predicting whether a patient has cancer, or how they will respond to a drug. For such task, statistical and AI-based machine learning or deep learning methods are used to identify prognostic and predictive gene signatures that are sets of genes whose expression can distinguish physiological states[10].

While bioinformatics has traditionally been considered an independent subfield within biology, it is becoming increasingly evident that future generations of biologist, biotechnologist, and related professionals will view it as an intrinsic discipline, alongside data science, essential to their professional training. Consequently, core areas of data science, including mathematics, statistics, ML, data processing, and visualization, should become fundamental components of curricula

[10] Max Kotlyar, W. H. Wong, Chiara Serene, Pastrello, and Igor Jurisica. Improving analysis and annotation of microarray data with protein interactions. In Giuseppe Agapito, editor, *Microarray Data Analysis*, chapter 5. Springer, 2022

for future professionals in biological sciences[11].

The present work aims to develop a previously established methodology based on DL fundamentals to identify potential biomarkers and therapeutic targets at the genetic level for pancreatic cancer patients. This approach could aid in patient stratification and improve survival outcomes in case of premature death due to pancreatic neoplasms. The methodology involved developing a DL algorithm to analyze gene copy number, gene expression, and protein expression in surviving versus deceased patients using databases such as the Genomic Data Commons (GDC) cohort[12].

Building upon previous experience from an established research group for pancreatic cancer, the current work aims to adapt and develop this approach for Medulloblastoma (MB), a Central Nervous System (CNS) cancer, using different publicly available datasets. By applying this DL algorithms with a methodology based on ANN to a different neoplasm, this research seeks to evaluate whether the approach can achieve comparable precision and accuracy in identifying the most relevant genes for subsequent biomarker studies across different cancer types.

[11] Andreas D. Baxevanis, Gary D. Bader, and David S. Wishart. *Bioinformatics*. Wiley, 2020

[12] I. Salgado, E. Prado Montes de Oca, I. Chairez, L. Figueroa-Yáñez, A. Pereira-Santana, A. Rivera Chávez, et al. Deep learning techniques to characterize the RPS28P7 pseudogene and the Metazoa-SRP gene as drug potential targets in pancreatic cancer patients. 12(02)

1.1 *Justification*

Medulloblastoma (MB) is a primary Central Nervous System (CNS) tumor that originates in the brain or spinal cord, it is the most common malignant pediatric brain tumor, representing 60% of childhood intercranial embryonal tumors[13], and is among the most severe of the CNS malignancies[14]. These tumors are graded based on histological analyses, with al medulloblastomas classified as grade IV, indicating that they are fast growing. Staging procedures are required to determine whether the tumor has spread to other areas of the CNS or beyond. Based on epidemiological data 70% of MB cases occur in children, while 30% occur in adults, typically between 20 and 40 years of age. MB is more prevalent in males than females and with respect to ethnicity it is more common among non-Hispanic whites and Hispanic populations. An estimated 6,070 people are currently living with this tumor in the United States[15]. In Mexico, CNS tumors represent the third leading cause of malignant neoplasm, with MB accounting for 20% of primary CNS tumors[16]. According to the 52th week of 2024 data from the Weekly Epidemiological Bulletin of the Secretary of Health, there were 438 male cases and 448 female cases of malignant brain tumors and other malignancies, which included preliminary information and probable cases[17].

Unlike developed nations with extensive cancer registries, Latin American countries face significant challenges implementing

[13] Abhinav Dey, Anshu Malhotra, and Neha Garg. Medulloblastoma methods and protocols. Springer Nature, 2022. ISBN 978-1-0716-1952-0

[14] Michael Iv, Michelle Zhou, Katie Shpanskaya, Sebastian Perreault, Zhewei Wang, Evan Tranvinh, Brent Lanzman, Sridhar Vajapeyam, Nicholas A. Vitanza, Paul G. Fisher, et al. MR imaging–based radiomic signatures of distinct molecular subgroups of medulloblastoma. *American Journal of Neuroradiology*, 40:154–161, 2019. DOI: 10.3174/ajnr.A5899

[15] National Cancer Institute. Medulloblastoma, 2025. URL https://www.cancer.gov/rare-brain-spine-tumor/tumors/medulloblastoma

[16] Rocío Elizabeth García-Dávila, Sergio Díaz-Bello, Raúl Villanueva-Rodríguez, León René López, Jorge Luis Valencia-Vázquez, and Belén Rivera-Bravo. Utilidad de la tomografía por emisión de positrones/ tomografía computada (pet/ct) en pacientes con diagnóstico de meduloblastoma. 63(1). DOI: 10.22201/fm.24484865e.2020.63.1.06. URL https://doi.org/10.22201/fm.24484865e.2020.63.1.06

[17] Boletín epidemiológico semanal. URL https://www.gob.mx/salud/acciones-y-programas/historico-boletin-epidemiologico. Semana Epidemiológica [52]

comprehensive cancer data collections systems. Consequently, publicly available datasets from the United States that include Latino or Hispanic patient populations provide crucial resources for cancer research in developing nations. Given that MB accounts for a substantial proportion of primary CNS tumors and that its molecular subtypes significantly influence treatment outcomes, it is essential to integrate bioinformatics and deep learning approaches to address these challenges. The present work aims to leverage established technologies and methodologies to analyze multiomics data, improve molecular subtype classification and contribute to more effective and personalized therapeutic strategies. By addressing existing gaps in data collection, integration and precision medicine implementation. The research seeks to enhance patient outcomes and advance more accessible, evidence-based personalized treatments within the Mexican healthcare context.

There are multiple factors such as cancer stage, patient age, and residual tumor presence after surgery that are used to estimate recurrence risk through a process called risk stratification, it guides clinicians in determining optimal treatment approaches. This can be seen as a threshold in which patients can be classified into two risk categories, standard risk or average risk, and high risk. Complementarily, medulloblastoma can also be classified according to its molecular and genetic features. There are at least four distinct molecular subgroups identified in both pediatric and adult patients, they can be identified based on their epidemiological frequency and genetic characteristics which varies by age. The groups are WNT activated, SHH activated (Sonic Hedgehog), Group 3, and Group 4, as observable in Figure 1.2[18]. Notably, Group 3 and 4 often exhibit overlapping characteristics due to high molecular homogeneity between them, this specific characteristic makes their distinction particularly challenging as these two groups encompass most of the diseases mortality cases. Therefore performing a proper molecular distinction a paramount task as each group exhibits different molecular profiles, clinical behaviors and prognoses.

[18] Abhinav Dey, Anshu Malhotra, and Neha Garg. Medulloblastoma methods and protocols. Springer Nature, 2022. ISBN 978-1-0716-1952-0

Although advances in conventional treatments have significantly improved overall survival rates, current therapies still present major limitations in combating the disease. These include severe adverse effects and wide variability in patient response. Additionally, the high complexity and heterogeneity of Medulloblastoma´s molecular biology pose significant challenges for treatment personalization and the identification of effective therapeutic biomarkers.

It is within this context that the integration of a proper data treatment based on bioinformatic methods and Deep Learning (DL) classification algorithms offer an innovative approach to address these challenges by enabling the centralization, standardization and homogenization of

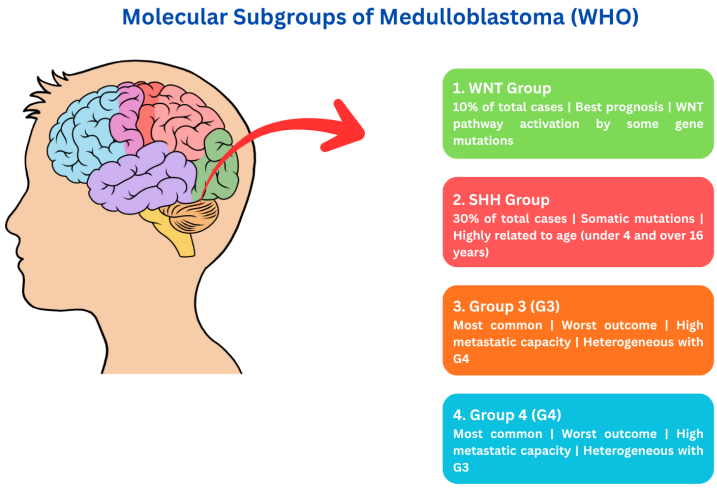

Figure 1.2: Molecular subgroups of medulloblastoma according to WHO 2016 classification. The four subgroups are defined by transcriptomic profiles: WNT group (10% of cases, best prognosis, characterized by WNT signaling pathway activation), SHH group (30% of cases, associated with age <4 years and >16 years, somatic mutations in SHH pathway), Group 3 and Group 4 (most common subtypes with worst outcomes due to high metastatic capacity and heterogeneity).

large volumes of clinical, genomical and transcriptomical data. This integration enables to identify patterns, relation, and features that would be difficult or impossible to discern using traditional analytical methods. Particularly the integration of multiomics data (genomics and transcriptomics) through DL could potentially provide deeper insights into molecular pathways involved in medulloblastoma which can facilitate the identification of key genes associated with disease progression, tumor aggressiveness, and treatment response.

Microarrays as laboratory tools help researcher to detect the expression of thousands of genes at the same time with DNA microscopic slides where in each spot is an specific sequence or gene, these are also known as gene or DNA chips. The molecules printed on each slide serves as probes to detect gene expression, also known as transcriptome or the set of messenger RNA (mRNA)[19]. These are essential characteristics to identify key molecular biomarkers, classifying molecular subtypes, and elucidating the genetic networks involved in tumor progression and aggressiveness.

Gene expression microarrays are one of the most used high-throughput technologies in molecular biology, they have provided some of the most comprehensive and easily usable molecular data. There are hundreds of thousands of profiled samples available in public databases such as Gene Expression Omnibus (GEO) and ArrayExpress. As of 2021, gene expression microarrays were the most frequent data type in GEO, comprising over 1.6 million samples[20]. Within these databases there are comprehensive molecular profiles which come from clinical samples enabling researchers to examine expression patterns across thousands of genes from unified patient cohorts, therefore addressing

[19] Debojyoti Dhar and Gopala Kallapura. Analysis of microarray data from medulloblastoma tissue samples. In *Medulloblastoma: Methods and Protocols*, pages 59–64. Springer, 2022

[20] Max Kotlyar, Serene WH Wong, Chiara Pastrello, and Igor Jurisica. Improving analysis and annotation of microarray data with protein interactions. *Microarray Data Analysis*, pages 51–68, 2021

critical needs in cancer research like extracting maximum biological information from limited tissue samples.

The extensive availability of curated microarray datasets in public repositories enables validation and cross-study analysis, these strategies combined with proper data preprocessing and DL algorithms strengthen the reproducibility and generalization of novel findings. While gene expression microarrays capture individual gene activity levels, their integration with pathway analysis allows research to move beyond isolated gene markers towards understanding complex regulatory cascades and molecular integrations underlying medulloblastoma´s biology.

1.2 *Problem Statement*

Despite the are advances in molecular diagnostics for medulloblastoma, it remains a considerable challenge in low and middle income countries, such as Mexico. Techniques like DNA or gene expression profiling are often limited in terms of logistics and high operational costs. This challenges is particularly acute for Groups 3 and 4 of the molecular classification of MB, which together account for over 60% associated with the poorest prognostics. There is also the overlapping of molecular profiles complicating their discrimination with conventional immunohistochemical markers. To assess these limitations DL approaches have been increasingly applied to categorize and diagnose human diseases, including cancer. However the four molecular MB subgroups have not yet been explored extensively with omic data. With these limitations comes a major obstacle to computational approaches to classify subgroups relying on limited availability of comprehensive datasets encompassing the four subgroups[21].

The identification, integration, and implementation of publicly available datasets into DL models is of critical importance given the increasing availability of omics data from research experiments and clinical trials. When properly processed and analyzed these data can significantly improve both the architecture and performance of predictive algorithms. This study aims to leverage deep learning tools to develop predictive models that enhance molecular subtype classification of medulloblastoma and identify key biomarker genes. Additionally, this work seeks to validate established methodologies, such as those reported by Reveles and colleagues [Reveles-Espinoza et al., 2025], using independent datasets with appropriate preprocessing procedures. Aiming to a more robust molecular classification based on DL algorithms that could become essential for personalized treatment strategies, particularly in low income nations where advance molecular diagnostics and early diagnosis remains limited.

[21] Alicia Reveles-Espinoza, Ulises Villela, Edgar Hernandez-Martinez, Isaac Chairez, Sergio Juárez-Méndez, J Casanova-Moreno, Ma del Pilar Eguía-Aguilar, Luis Figueroa-Yáñez, Adriana Vallejo-Cardona, and Iván Salgado. Machine learning identification of molecular targets for medulloblastoma subgroups using microarray gene fingerprint analysis. *Computational and Structural Biotechnology Journal*, 2025

Unlike developed nations with extensive cancer registries and robust data infrastructure, Latin American countries face significant challenges in implementing comprehensive cancer data collection systems. These limitations create substantial gaps in local datasets available for computational research. Consequently, publicly available datasets from the United States that include Latino or Hispanic patient populations provide crucial resources for cancer research in developing nations, offering valuable molecular information that can inform clinical decision-making in similar demographic contexts.

By addressing existing gaps in data collection, integration, and precision medicine implementation, this work ultimately aims to contribute to more effective prediction of molecular subgroups in medulloblastoma patients from low- and middle-income countries. Through the development of robust deep learning models and validation of computational methodologies, these efforts seek to advance personalized therapeutic strategies that can be adapted to resource-limited settings.

While the identification of differentially expressed genes provides initial insights into molecular alterations, functional enrichment analysis offers a more comprehensive approach towards understanding biological mechanisms by examining collective expression patterns of biological processes rather than individual genes[22]. Gene Ontology (GO), a well-established and curated annotation system, assigns genes to standardized functional classifications across three aspects: Cellular Component (CC), which describes cellular or extracellular localization; Molecular Function (MF), which defines the primary molecular-level activity: and Biological Process (BP), which encompasses the collective processes in which gene products participate. These GO terms, along with pathway database annotations, are associated with gene identifiers in standardized databases[23], enabling systematic exploration of altered biological mechanisms between experimental conditions. Although functional enrichment analysis represents a valuable complementary approach for interpreting gene expression data, the present study focuses primarily on the evaluation metrics and performance of deep learning models for molecular subgroup classification of medulloblastoma.

[22] Marieke L. Kuijjer, Joseph N. Paulson, and John Quackenbush. Expression analysis. In Andreas D. Baxevanis, Gary D. Bader, and David S. Wishart, editors, *Bioinformatics*, chapter 10. John Wiley & Sons, 4th edition, 2020

[23] Andreas D. Baxevanis. Biological sequence databases. In Andreas D. Baxevanis, Gary D. Bader, and David S. Wishart, editors, *Bioinformatics*, chapter 1. John Wiley & Sons, 4th edition, 2020

1.3 *Research Question*

Can a multidataset integrated pipeline combining cross-platform microarray annotation versions with posterior harmonization, differential gene expression analysis, and feed forward artificial neural network classification outperform previously reported state of the art classification benchmarks for Medulloblastoma molecular

subgrouping (WNT, SHH, Group 3, and Group 4) by leveraging biologically validated differentially expressed gene signature derived from a larger harmonized cohort reported of 794 individuals, while simultaneously identifying key genes associated with physiological alterations distinguishing molecular subgroups from normal cerebellar tissue, thereby contributing to improved prognostic stratification between surviving and deceased patients and supporting personalized treatment strategies?

1.4 *Hypothesis*

The harmonization of multiple datasets of microarray data across annotation versions, combined with differential gene expression analysis and feedforward artificial neural network classification, enables the identification of a biologically discriminative gene signature that outperforms previously reported classification metrics in Medulloblastoma molecular subgroup classification, while simultaneously capturing expression patterns associated with physiological alterations distinguishing tumor subgroups from normal cerebellar tissue contributing to improved prognostic stratification of patient survival outcomes based on epidemiological data of survival across subgroups.

1.5 *Objectives*

The general objective of the present work is to develop an integrated bioinformatics and deep learning pipeline capable of assessing differential gene expression patterns associated with physiological alterations in Medulloblastoma, achieving a molecular subgroup classification accuracy that surpasses the previously reported in the state of the art classification metrics, where the minimum reported accuracy was 90% for Group 3 and 97.2% for the multiclass configuration, by delivering classification accuracies ranging from 97.99% for a 5-class classifier to 99.87% for the SHH binary classifier, with AUC values between 0.988 and 1.000 across all subgroups, while simultaneously identifying a biologically validated differentially expressed gene signature derived from a harmonized cohort of 794 individuals distinguishing WNT, SHH, Group 3, and Group 4 molecular subgroups from normal cerebellar tissue, thereby contributing to improved prognostic stratification and personalized treatments supported by survival epidemiological data across molecular subgroups.

1.5.1 Specific Objectives

- To implement preprocessing and normalization techniques for gene expression microarray data and clinical metadata obtained from public databases such as the Gene Expression Omnibus (GEO) of the National Center for Biotechnology Information (NCBI).
- To identify differentially expressed genes associated with Medulloblastoma molecular subgroups and patient survival outcomes.
- To integrate a genomic multi dataset to train and validate a deep learning algorithm for molecular subgroup classification and survival prediction.
- To develop and train a deep learning algorithm for multiclass classification of Medulloblastoma molecular subgroups based on gene expression patterns.
- To characterize molecular expression changed patterns of cell physiological functions related with Medulloblastoma subgroups and evaluate the model´s performance in distinguishing between surviving and deceased patients.
- To evaluate the pipeline´s performance across multiple independent datasets and compare classification accuracy, sensitivity, and specificity metrics.

1.6 Thesis Proposal

The present work proposes the development and validation of an integrated bioinformatics and deep learning pipeline for molecular subgroup classification of Medulloblastoma using gene expression microarray data from public repositories. The approach introduces a cross-platform annotation harmonization strategy between GPL22286 and GPL16977 array versions, enabling the unified RMA normalization of a larger harmonized cohort reported in this context (n=794, comprising 31 CT, 70 WNT, 223 SHH, 144 Group 3, and 326 Group 4 samples), with statistical power analysis performed per molecular subgroup to justify the sample size contribution over previously reported datasets.

The pipeline implements a structured bioinformatics workflow spanning RMA background correction and quantile normalization, distribution based quality controls including sample expression boxplots, density plots and UMAP dimensionality reduction, followed by exploratory data analysis of clinical metadata prior to any molecular intervention. Differential gene expression analysis is

performed under two configurations, a stringent threshold consisted with state of the art reporting FDR < 0.005, |log2FC| > 2.32 and a less stringent bibliographically supported alternative, producing biologically validated subgroup specific gene signatures visualized through MA plots, volcano plots with directional coloring, and Venn diagrams of cross contrasts DEG overlap.

The resulting discriminative gene signature was set to be 1,139 DEGs serving as input to feedforward artificial neural network architectures whose topology is directly adapted from state of the art configurations, implementing both a 5-class multiclass classifier and four binary subgroup versus the rest classifiers with 10-fold stratified cross-validation, dropout regularization, and batch normalization. The pipeline delivers classification accuracies ranging from 97.99% for the multiclass configuration to 99.87% for the SHH binary classifier, with AUC values between 0.988 and 1.000, surpassing the minimum state of the art performance metrics of 90% for Group 3 and 97.2% for the multiclass configuration, with particular advancement in the historically challenging group 3 and Group 4 discrimination.

By leveraging publicly available data under a fully reproducible computational workflow, this work contributes accessible and robust molecular classification tools that can inform personalized treatment strategies in resource-limited settings, while identifying key biomarker genes associated with patient survival outcomes across molecular subgroups that may guide future therapeutic target discovery and clinical decision making in pediatric neuro oncology.

1.7 *Novel Contributions*

The present work introduces a systematic reconciliation of two annotation versions (GPL22286 and GPL16977) sharing the same physical HuGene 1.1 ST array, enabling unified RMA normalization across a harmonized cohort of 794 individuals and tackling down one of the main problems coming from working with microarray data which is to harmonize annotations across different arrays prior to any bioinformatic preprocessing, a methodological contribution not reported in previous Medulloblastoma microarray classification studies taking control samples into consideration that typically operate on single-platform datasets. Building on this harmonization, a formal statistical power analysis is performed per molecular subgroup to explicitly justify the representativeness of each group within the cohort, providing a reproducible framework for future dataset expansion that goes beyond the simple sample distribution reporting characteristic of comparable works in the literature.

Prior to any molecular intervention, a structured epidemiological

characterization of clinical metadata is conducted to establish group-level distributional assumptions and identify potential confounders independently of the expression data, a step systematically absent from comparable computational pipelines. This separation between clinical and molecular analysis layers ensures that the subsequent bioinformatics processing is both statistically grounded and biologically interpretable. Finally, the direct connection established between a stringently filtered differential gene expression signature (FDR < 0.005, |log2FC| > 2.32, n = 1,139 genes) and a feedforward artificial neural network with a non complex architecture represents a biologically grounded feature selection strategy that replaces whole genome input with a molecularly meaningful fingerprint, reducing dimensionality while preserving maximum discriminative power and surpassing state of the art classification benchmarks across all five classification configurations evaluated.

2 Theoretical and conceptual framework

Contents

2.1 Medulloblastoma **39**
2.2 Bioinformatics and Computational Biology . . **41**
2.2.1 Multiomics Data Integration 42
2.3 Microarray Technology and Gene Expression Analysis **44**
2.3.1 Public Biological Databases and Data Repositories 48
2.3.2 Gene Expression Omnibus (GEO) . . 48
2.4 Microarray Data Analysis **51**
2.5 Microarray Data Preprocessing **53**
2.6 Fundamentals of Machine Learning and Deep Learning **62**
2.6.1 Artificial Neural Networks 63
2.6.2 Activation Functions 66
2.6.3 Feedforward Neural Network Topology 67
2.6.4 Loss Function 68

2.1 Medulloblastoma

Medulloblastoma is the most common malignant pediatric brain tumor, despite multimodal advances in therapies over the last decades and particularly in high risk patients mostly pediatrics, the overall 5 years survival rate is less than 70%. First described in 1925, the initial discrimination of medulloblastoma was based on histopathological examination. Until 2007 it was described into three histopathological categories: classic MB, nodular or desmoplastic MB, and large cell or anaplastic MB that carries variable clinical risk[1], histological subtyping provided the means to segregate MB into subgroups, it considered histophenotypic differences between tumors and was not applicable to capture the nuance of intertumoral molecular heterogeneity of the disease. However, most recent advancements are made in global gene expression profiling methods that have allowed scientist to a

[1] Ruth-Mary DeSouza, Benjamin RT Jones, Stephen P Lowis, and Kathreena M Kurian. Pediatric medulloblastoma–update on molecular classification driving targeted therapies. *Frontiers in oncology*, 4:176, 2014

more accurate subclassification and prognosticate the disease based on molecular characteristics. Identification of subtype specific molecular drivers and pathways presents novel therapeutic targets that could lead to a subtype specific treatment[2].

By 2016, the World Health Organization defined two classification systems for the disease; the first histologically defined via histopathological subtypes, and the second which is genetically defined, that as previously mentioned includes four molecular subtypes WNT, SHH, Group 3 and Group 4 that were initially defined by transcriptomic profiling using expression arrays. WNT represents 10% of all cases and has the most favorable prognosis[3]. SHH subgroup comprises 30% of tumors and commonly displays somatic mutations in genes involved in SHH pathway, for this specific case age is a prognostic factor with peak incidence in less that 4 years of age and more than 16 years which represents a bimodal distribution. Group 3 and 4 are the most common subtypes, and at the same time exhibit the worst patient outcomes as these tumors present with metastases at diagnosis[4].

Molecular profiling has allowed a deeper understanding of intertumoral heterogeneity across medulloblastoma tumors and has provided scientist and clinicians the ability to better risk stratify and prognosticate medulloblastoma in different populations. Advances in omic techniques and analysis algorithms in the last decade had allowed scientist to expand the initial four subgroups subdividing them into 7 or 12 subtypes. Particularly focused on a better characterization of Group 3 and 4 by using different machine learning approaches such as tSNE and dbScan to probe methylation patterns reporting eight subclasses[5], other studies have reported up to 3 subgroups; α, β, and γ for the previously defined groups 3 and 4 on gene expression obtained from undefined RNA microarrays and methylation data[6]

A consensus study published in 2019 integrated tumor profiling data from multiple independent cohorts comprising 1,501 medulloblastoma tumors, employing t-SNE and k-means clustering to evaluate concordance across previously proposed molecular classification models and analytical methods[7]. The analysis identified eight molecular subtypes as the most concordant partition of the data, yielding high class definition confidence and reproducibility across cohorts, demonstrating that ML and DL techniques applied to medulloblastoma are of great value for tumor classification and stratification. These subtypes exhibited distinct demographic profiles, molecular features, and survival outcomes, characteristics that collectively advance medulloblastoma research through the integration of multiple bioinformatics approaches. One of the pillars of the present work is the processing of differentially expressed genes derived from gene expression quantification data stored in `.CEL` files, which are

[2] Yujin Suk, William D Gwynne, Ian Burns, Chitra Venugopal, and Sheila K Singh. Childhood medulloblastoma: an overview. *Medulloblastoma: Methods and Protocols*, pages 1–12, 2022

[3] James T Rutka and Harold J Hoffman. Medulloblastoma: a historical perspective and overview. *Journal of neuro-oncology*, 29(1):1–7, 1996

[4] Paul A Northcott, Adrian M Dubuc, Stefan Pfister, and Michael D Taylor. Molecular subgroups of medulloblastoma. *Expert review of neurotherapeutics*, 12(7):871–884, 2012

[5] Paul A Northcott, Ivo Buchhalter, A Sorana Morrissy, Volker Hovestadt, Joachim Weischenfeldt, Tobias Ehrenberger, Susanne Gröbner, Maia Segura-Wang, Thomas Zichner, Vasilisa A Rudneva, et al. The whole-genome landscape of medulloblastoma subtypes. *Nature*, 547 (7663):311–317, 2017

[6] Florence MG Cavalli, Marc Remke, Ladislav Rampasek, John Peacock, David JH Shih, Betty Luu, Livia Garzia, Jonathon Torchia, Carolina Nor, A Sorana Morrissy, et al. Intertumoral heterogeneity within medulloblastoma subgroups. *Cancer cell*, 31(6):737–754, 2017

[7] Tanvi Sharma, Edward C Schwalbe, Daniel Williamson, Martin Sill, Volker Hovestadt, Martin Mynarek, Stefan Rutkowski, Giles W Robinson, Amar Gajjar, Florence Cavalli, et al. Second-generation molecular subgrouping of medulloblastoma: an international meta-analysis of group 3 and group 4 subtypes. *Acta neuropathologica*, 138(2):309–326, 2019

subsequently processed through LIMMA (Linear Models for Microarray Analysis) in R to obtain a molecular fingerprint that aids in the classification of the disease, as further explored in the following sections[8].

[8] Debojyoti Dhar and Gopala Kallapura. Analysis of microarray data from medulloblastoma tissue samples. In *Medulloblastoma: Methods and Protocols*, pages 59–64. Springer, 2022

2.2 *Bioinformatics and Computational Biology*

Bioinformatics has emerged as an interdisciplinary field that integrates computational sciences, mathematics, and information technologies to analyze biological data. It seeks to be the liaison between computational and biological sciences, in its early years it was defined as the subject of genetic data collection, analysis and dissemination. Nowadays it has become a discipline focused on complex mathematical modeling ans simulation[9]. It spans across subfields like proteomics, metabolomics, machine learning and image processing. Some of its applications start from the fact than they handle sequencing data which is abundant, currently, no other omics data is as abundant as high throughput sequencing data, such as transcriptome sequencing for gene expression studies. Such applications can be used to answer biological questions that will continue to increase[10]. From a perspective of information technology, bioinformatics can be seen as a scientific discipline that encompass acquisition, storage, processing, analysis, interpretation and visualization of biological information. While it can be considered an independent subfield of biology, it is likely that next generations of biologist will not consider it as a separated discipline and will consider gaining bioinformatics and data science skills[11].

[9] Gautam B Singh. Fundamentals of bioinformatics and computational biology. *Cham: Springer International Publishing*, pages 159–170, 2015a

[10] Vince Buffalo. *Bioinformatics data skills: Reproducible and robust research with open source tools*. " O'Reilly Media, Inc.", 2015

[11] Andreas D. Baxevanis. Biological sequence databases. In Andreas D. Baxevanis, Gary D. Bader, and David S. Wishart, editors, *Bioinformatics*, chapter 1. John Wiley & Sons, 4th edition, 2020

At is core there are several applications to solve problems related to biological systems which can be solved computationally or algorithmically. One of them is Information retrieval and data mining from biological databases as there has been an explosive growth of sequence biological information which has created challenges for data representation, access, and analysis. Bioinformatics tools are required to process the overwhelming amount of data, such as the Human Genome Project[12]. In the present work microarrays and differential gene expression analysis are the core. These techniques analyze thousands of genes and their expressed products, these expressions lead the way towards detecting genes based on their expression levels, as gene expression is correlated with the tissue type such as comparing tumor samples with healthy samples or control. It can be particularly useful in cancer studies where mutations can amplify or turn off gene expression.

[12] National Human Genome Research Institute. The human genome project. https://www.genome.gov/human-genome-project, 2024

There are several limitations as well, an example of those are data integration from multiple modalities, as research in molecular biology and medicine which rely on progress in data management

an quality. Further advancements in areas like drug development or personalized medicine depend heavily on integrating data from longitudinal studies and levels of detail. An example of these challenges can be the integration of data from proteomics mass spectroscopy and gene sequencing, it requires fulfilling existing gaps between techniques. The new challenges are precisely integration across genomics and proteomics and individual medical data routinely collected in clinics and is the prelude to the so called multiomics.

2.2.1 *Multiomics Data Integration*

Multiomics data are usually experimental measures related to the same set of samples on which multiple molecular experimental results have been performed such as; gene expression, methylation or copy number variation. These different types of data are particularly useful, since they can show complementary aspects related to the same biological process. They can be used to gain better understanding of the overall phenotype(s) under study. Specifically with microarrays there are multiple types of data that can be produced such as the previously mentioned[13]. Multiomics datasets aim to address some of the limitations coming from marker genes and metagenomics analysis, for example measurements of transcripts and protein expression. Examples of addressing some limitations are reflected by similar metagenomic profiles that may conceal significant transcriptional variation in response to medication, temperature shifts, day and night patterns an other environmental parameters[14].

Multiomics approaches are made of different disciplines such as; genomics which analyzes genes and genetic variations through DNA sequencing, epigenomics that studies reversible DNA modifications, transcriptomics examines gene expression through RNA analysis (or the so called functional genomics), proteomics which quantify protein abundance and modification, and metabolomics that measures small molecule metabolites[15]. Several technological advancements have become available across different types of omic data, among them are gene expression, micro-RNA expression, copy number variation, methylation and single nucleotide polymorphism (SNP). This allows for measurement of multiple omics data layers for the same set of samples. These multiomics experimental data are often not highly correlated between each other, thus they provide potentially complementary information and assess different parts of the same complex biological process[16]. This type of approaches have gained an exponential growth in recent years driven by increased accessibility of molecular biology techniques and next generation sequencing technologies, as well as improvements in computational tools for analyzing complex relational

[13] Alisa Pavel, Angela Serra, Luca Cattelani, Antonio Federico, and Dario Greco. Network analysis of microarray data. In *Microarray Data Analysis*, pages 161–186. Springer, 2021

[14] Robert G. Beiko. Metagenomics and microbial community analysis. In Andreas D. Baxevanis, Gary D. Bader, and David S. Wishart, editors, *Bioinformatics*, chapter 16. John Wiley & Sons, 4th edition, 2020

[15] Thermo Fisher Scientific. Supporting multi-omics approaches, n.d. URL https://www.thermofisher.com/mx/es/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/spotlight-articles/supporting-multi-omics-approaches.html

[16] Angela Serra, Luca Cattelani, Michele Fratello, Vittorio Fortino, Pia Anneli Sofia Kinaret, and Dario Greco. Supervised methods for biomarker detection from microarray experiments. In *Microarray Data Analysis*, pages 101–120. Springer, 2021

datasets. Fields as diverse as oncology, neurology, infectious disease research and discovery have taken advantage of multiomic strategies. These integrated strategies probe through the layers of the central dogma of molecular biology, from DNA to RNA to protein, with the addition of metabolites and epigenetic modifications, thereby providing a comprehensive picture of biological processes occurring within cells and tissues, as observable in (Figure 2.1).

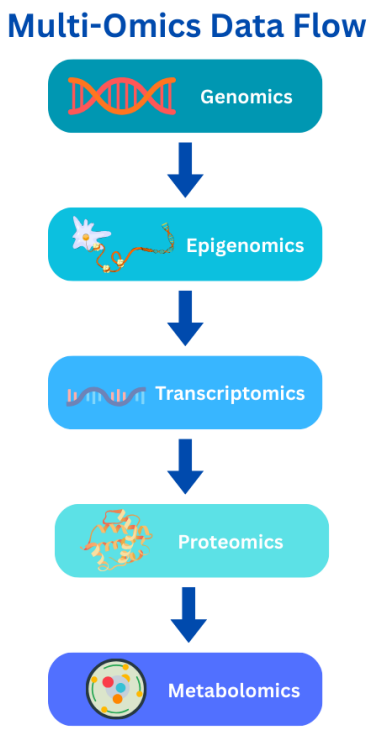

Figure 2.1: Schematic representation of multi-omics data flow across the central dogma of molecular biology, the current scope focus on nucleic acid-based methods (genomics and transcriptomics via microarray analysis). Adapted from Thermo Fisher Scientific [Thermo Fisher Scientific, n.d.].

Therefore integration of omics data is of greater value, especially in complex diseases such as cancer. Many public databases such as The Cancer Genome Atlas (TCGA)[17], collects molecular and clinical data from different types of cancer and thus representing an important source for multiomics integration analysis. However, technical and biological challenges need no be considered. On major limitation is the lack of complete molecular data for all the patients included in a dataset. For instance, in the TCGA the acute myeloid leukemia dataset included information for 197 patients. Among these, 173 have been profiled for mRNA, 194 for methylation, and 188 for miRNA expression. Therefore, when seen as a data frame that collects all data and metadata we could see multiple data entries as null values. Tumor heterogeneity is another relevant challenge, that concerns us for the scope of the present work, within integrative multiomics analysis. It is the result of cancer genomic instability characterized by the acquisition of new mutational

[17] National Cancer Institute, Center for Cancer Genomics. The cancer genome atlas program. https://www.cancer.gov/ccg/research/genome-sequencing/tcga, 2024

events during uncontrolled proliferation of cancer cells. Heterogeneity is crucial for cancer patients with similar molecular profiles, such as medulloblastoma. Another challenge is the choice of normal control to compare cancer alterations at omics levels due to interindividual variability[18].

In the context of medulloblastoma research, genomic and transcriptomic analyses have become essential for molecular subgroup classification, particularly transcriptomic analyses have opened novel avenues for improving its diagnostics as they are simpler, faster and often cost effective and with them comes a benefit as recent studies employing gene expression profiling have yielded novel insights with its prognosis, since genes which are involved in cell cycle regulation, multi-drug resistance, ribosome biogenesis, cerebellar differentiation or encoding extracellular matrix proteins have been associated with survival status[19]. The ultimate goal of large scale transcriptome analyses, such as DNA microarrays, is towards the characterization of molecular alterations underlying certain biological conditions. Transcriptomics analysis allows the identification of a compendium up to hundreds of genes which are deregulated under certain conditions[Pavel et al., 2021]. High throughput microarrays and RNA sequencing have facilitated the identification of genetic fingerprints, which are sets of genes delineating high-risk populations or metastasis-associated prognostic signatures across diverse cancer types. Multiple investigations have indicated that a minimal gene fingerprint may serve as robust biomarkers towards predicting patient outcomes in terms of survival status. Thereby, identifying such biomarker signatures that correspond to a minimal genetic fingerprint could aid in prognostic and prediction in medulloblastoma´s management, eventually leading the way for developing precision medicine. This potential detection is in its early stages for potential detection and characterization of fingerprints in each of the medulloblastoma´s subgroups via transcriptomic data processed by deep learning[20].

[18] Francesca Scionti, Mariamena Arbitrio, Daniele Caracciolo, Licia Pensabene, Pierfrancesco Tassone, Pierosandro Tagliaferri, and Maria Teresa Di Martino. Integration of dna microarray with clinical and genomic data. In *Microarray Data Analysis*, pages 239–248. Springer, 2021

[19] Scott L Pomeroy, Pablo Tamayo, Michelle Gaasenbeek, Lisa M Sturla, Michael Angelo, Margaret E McLaughlin, John YH Kim, Liliana C Goumnerova, Peter M Black, Ching Lau, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–442, 2002

[20] Alicia Reveles-Espinoza, Ulises Villela, Edgar Hernandez-Martinez, Isaac Chairez, Sergio Juárez-Méndez, J Casanova-Moreno, Ma del Pilar Eguía-Aguilar, Luis Figueroa-Yáñez, Adriana Vallejo-Cardona, and Iván Salgado. Machine learning identification of molecular targets for medulloblastoma subgroups using microarray gene fingerprint analysis. *Computational and Structural Biotechnology Journal*, 2025

2.3 *Microarray Technology and Gene Expression Analysis*

Microarrays also known as biochips or DNA chips are the result of the conjunction of molecular biology, information technologies and automatization, since the 1990s they have been revolutionizing molecular diagnostics by easily differentiating the genetic component of multiple diseases[21]. They are composed by a sequence of microscopic DNA spots replaced on a solid surface, it has been exploit by researchers as a platform to analyze thousands of genes simultaneously. As a high throughput technology it produces huge amounts of features (genes) in a very small size of the DNA chip[22].

[21] Ángel Herráez. Hibridación de ácidos nucleicos. In *Biología molecular e ingeniería genética*, chapter 12, pages 163–190. Elsevier Health Sciences, 2012

[22] Agostino Forestiero, Giuseppe Papuzzo, Rosaria De Simone, and Rosa Varchera. A microarray analysis technique using a self-organizing multiagent approach. In *Microarray Data Analysis*, pages 39–50. Springer, 2021

After genome sequencing, DNA microarrays have become the most widely used tool for genome investigation. In the past two decades, oncology research represented one of the major areas of microarray application as they lead the dissection of cancer in subtypes with different molecular characteristics. One of the earliest applications was to compare expression levels of a large number of genes between two biological states, tumor versus normal, to identify key genes whose differential expression patterns are crucial for cancer development and progression[23]. Typically a study that uses microarray technology is able to measure expression levels of genes present in biological samples which result in a matrix of real numbers where the value *(i, j)* represents the expression of the gene *i* on the sample *j*[24].

DNA microarray technology is based on hybridization of nucleic acid molecules with complementary sequences called probes which are fixed on a solid surface, often a glass slide or silicon biochip. Based on probe synthesis method, they can be divided in spotted or cDNA microarrays and oligonucleotide microarrays. In cDNA microarrays the probes are generally amplified by PCR products, while oligonucleotide ranges from 60 to 70 nucleotides in size, or cloned DNA fragments are synthesized prior to deposition on the array and then spotted on the array surface. Before interacting with the actual device, the two different samples have to undergone a preprocessing step on which each specimen is marked with a different fluorophore. Once the microarray probes interact with the samples a light signal can be collected by using a dedicated scanner. Finally, the emitted radiation can be used to generate an image of dark and light spots representing the raw source of data[25].

The proper division of the samples to be analyzed would be cDNA samples as test, and control samples or reference are labeled with different fluorescent dyes, the previously mentioned fluorophores, where Cy3 is a green fluorochrome and Cy5 is a red one, mixed together and hybridized against probe spots on the same array. Then for each sample the fluorescent intensity is calculated as a ratio or log2 ratio between test and reference. In the second type of microarray, oligonucleotide probes which are around 25 nucleotides in size, are synthesized directly *in situ* using photolithography techniques. A visual representation of a microarray can be seen in Figure 2.2.

The Affymetrix GeneChip platform, now Thermo Fisher Scientific (Thermo Fisher Scientific, Inc.), is often the primary example of oligonucleotide microarrays, an example of how they look can be seen in Figure 2.3. For this kind of chips each gene are designed 15 to 20 different 25-mer, or 25 synthetic nucleotide long chain, perfect match (PM) oligonucleotides that are perfectly complementary to the target sequence and additional probes are a mismatch (MM) control

[23] Francesca Scionti, Mariamena Arbitrio, Daniele Caracciolo, Licia Pensabene, Pierfrancesco Tassone, Pierosandro Tagliaferri, and Maria Teresa Di Martino. Integration of dna microarray with clinical and genomic data. In *Microarray Data Analysis*, pages 239–248. Springer, 2021

[24] Fabrizio Marozzo and Loris Belcastro. High-performance framework to analyze microarray data. In *Microarray Data Analysis*, pages 13–27. Springer, 2021

[25] Paolo Zaffino and Maria Francesca Spadea. Algorithms to preprocess microarray image data. In *Microarray Data Analysis*, pages 69–78. Springer, 2021

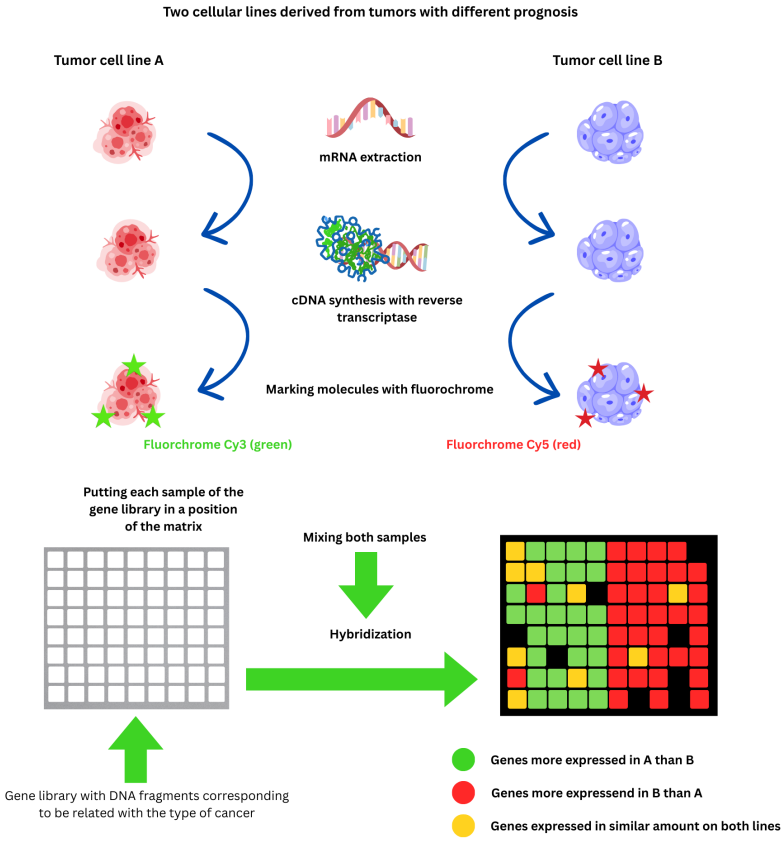

Figure 2.2: Microarray experimental representation of two different cellular lines that derive in tumors with different prognosis. Adapted from reference image on [Herráez, 2012].

oligonucleotides that are identical to PM probes except for a single base near the middle of the probe.

Figure 2.3: Microarray chips compared to a match for reference.

Mismatch oligonucleotide allows to differentiate specific from nonspecific hybridization and to remove background from the PM probe signal. In Gene Chip expression assays the samples both test and reference are labeled with biotin and hybridized to a separate array. Finally the differences between samples are determined comparing the difference values (PM-MM) of each probe pair in the test array to its matching probe pair on the reference array[26].

Microarrays as a high throughput technology that analyzes thousands of gene expressions generate vast amounts of omics data being constantly produced, therefore databases are needed for storage and analysis of these data, necessary for biologist to interpret massive amounts of data from experiments. This poses challenges relative to ample data storage and computing power[27]. That is why analyzing gene expression data, often available in raw forms, implies a huge amount of analytical and computational power, therefore innovative and intelligent mechanisms have to be designed to obtain useful information from this precious data[28].

[26] Francesca Scionti, Mariamena Arbitrio, Daniele Caracciolo, Licia Pensabene, Pierfrancesco Tassone, Pierosandro Tagliaferri, and Maria Teresa Di Martino. Integration of dna microarray with clinical and genomic data. In *Microarray Data Analysis*, pages 239–248. Springer, 2021

[27] Barbara Calabrese. Web and cloud computing to analyze microarray data. In *Microarray Data Analysis*, pages 29–38. Springer, 2021

[28] Agostino Forestiero, Giuseppe Papuzzo, Rosaria De Simone, and Rosa Varchera. A microarray analysis technique using a self-organizing multiagent approach. In *Microarray Data Analysis*, pages 39–50. Springer, 2021

2.3.1 Public Biological Databases and Data Repositories

Public biological databases have become indispensable infrastructure for modern biomedical research, serving not only as repositories for data storage and sharing, but also as essential starting points for scientific discovery. These databases enable researchers to access diverse biological information, including nucleotide sequences, gene expression data, protein sequences, functional annotations, and associated metadata describing sample origin and experimental conditions.

The foundation of public accessible sequence data is maintained through the International Nucleotide Sequence Database Collaboration (INSDC), established by three major entities: the National Center for Biotechnology Information (NCBI) in the United States, the European Molecular Biology Laboratory (EMBL) in Europe, and the DNA Data Bank of Japan (DDBJ). These databases synchronize daily to ensure that nucleotide sequence data generated worldwide remains freely and publicly available to the scientific community. Each member provides specialized tools for data submission, access, and analysis, with NCBI serving as the primary resource for researchers in the Americas[29].

In the case of gene expression, microarrays provide some of the most comprehensive and easily usable high throughput molecular data. Hundreds of thousands of profiled samples are available in public databases such as Gene Expression Omnibus (GEO)[30] and ArrayExpress[31]. By 2021, gene expression microarrays were the most frequent data type in GEO, over 1.6 million samples, covering 30,000 species and thousands of disease states and other physiological conditions, while the second most frequent data type was gene expression profiling by high throughput sequencing with over 1.5 million samples. Even though high throughput sequencing methods have several advantages over typical microarrays, such as detecting novel and low abundance transcripts or higher specificity and sensitivity, microarrays have data easier to analyze, also there is still more data available in that format compare to next generation sequencing data[32].

Since they have lead the way from studies of individual genes or at beast pathways toward a global investigation of cellular activity. The vast amount of data available represents a valuable source of information for solving several biological questions and hypotheses[33].

2.3.2 Gene Expression Omnibus (GEO)

The Gene Expression Omnibus (GEO) is an international public repository with thousands of gene expression datasets, the database is curated by the U.S. National Institutes of Health (NIH). It encompasses millions of samples of microarrays, next generation sequencing and

[29] Mikołaj Danielewski, Marlena Szalata, Jan Krzysztof Nowak, Jarosław Walkowiak, Ryszard Słomski, and Karolina Wielgus. History of biological databases, their importance, and existence in modern scientific and policy context. *Genes*, 16(1):100, 2025

[30] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research*, 41 (D1):D991–D995, 2012

[31] Awais Athar, Anja Füllgrabe, Nancy George, Haider Iqbal, Laura Huerta, Ahmed Ali, Catherine Snow, Nuno A Fonseca, Robert Petryszak, Irene Papatheodorou, et al. Arrayexpress update–from bulk to single-cell expression data. *Nucleic acids research*, 47(D1):D711–D715, 2019

[32] Max Kotlyar, Serene WH Wong, Chiara Pastrello, and Igor Jurisica. Improving analysis and annotation of microarray data with protein interactions. *Microarray Data Analysis*, pages 51–68, 2021

[33] Antonio Federico, Laura Aliisa Saarimäki, Angela Serra, Giusy Del Giudice, Pia Anneli Sofia Kinaret, Giovanni Scala, and Dario Greco. Microarray data preprocessing: from experimental design to differential analysis. In *Microarray Data Analysis*, pages 79–100. Springer, 2021

other forms of high throughput multiomics data, one of its main goals is to provide user friendly mechanisms to allow researchers perform queries, locate, review and download studies and gene expression profiles of interest[34].

GEO records are hierarchically organized into three primary entities; Platform records (GPL), contains information about the microarray chip describing the array or sequencer technology used often being Affymetrix chips. Sample records (GSM) are the fundamental data point for experiments, is where individual hybridization values are recorded describing the conditions under which individual samples were handled and provide abundance measurements for each element. Series records (GSE) are the collection of all related samples collected for a functional genomic study and provide descriptions of the complete study[35]. Within GEO each record is assigned a unique and stable accession number; GPLxxx for platforms, GSMxxx for samples, and GSExxx for series. The relationship among the three entities is illustrated on the schema of Figure 2.4

[34] Gautam B Singh. Microarrays. In *Fundamentals of Bioinformatics and Computational Biology*, chapter 17, pages 159–170. Springer International Publishing, Cham, 2015b

[35] NCBI. Geo overview, 2025. URL https://www.ncbi.nlm.nih.gov/geo/info/overview.html

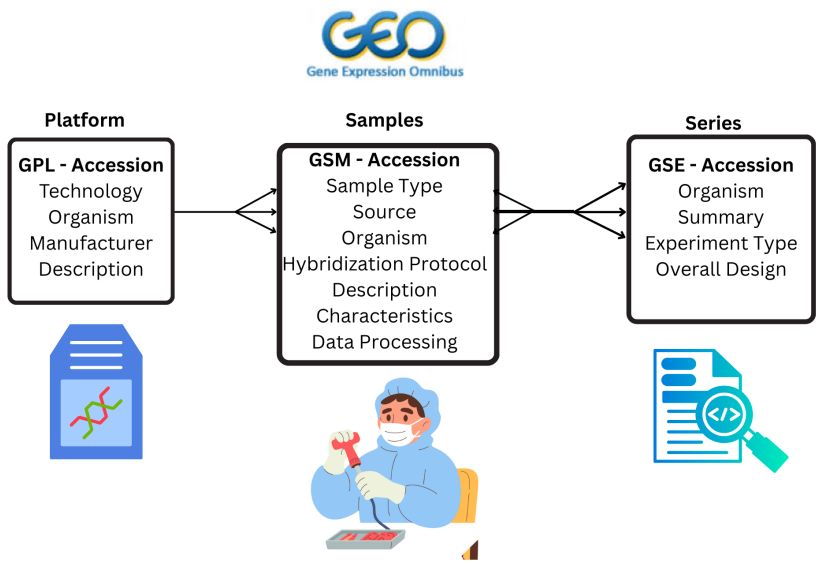

Figure 2.4: Schema representing relation among the three main subcategories of data in GEO. There is One to many between Platform and Samples whereas a Many to Many relation between Samples and Series.

GEO data is available to be downloaded in multiple formats to accommodate different analytical workflows. There are 3 different formats in which GEO data can be downloaded online, such as the Simple Omnibus Format in Text (SOFT) which is a line based and plain text format that organizes data as concatenated records containing both data tables and descriptive information for the three categories which are easy to analyze with programmatic algorithms[36].

Particularly, microarray gene expression data has a layer of

[36] Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F Kim, Alexandra Soboleva, Maxim Tomashevsky, and Ron Edgar. Ncbi geo: mining tens of millions of expression profiles—database and tools update. *Nucleic acids research*, 35(suppl_1): D760–D765, 2007

complexity and its interpretation is strictly related to the biological system under study, experimental conditions and data processing, such limitations were overcome in 2001 when guidelines of standardization were developed, known as MIAME (Minimum Information About Microarray Experiment), they were established for interpretation of microarray results and experimental reproducibility[37].

The guidelines focused on six critical elements enlisted as follows.

1. Raw data for each hybridization.
2. Normalized data.
3. Sample annotation including experimental factors and their values (e.g., phenotype in a dose response experiment).
4. Experimental design including sample data relationships. A raw data file which needs to be linked to the corresponding sample and being classified as technical or biological replicate.
5. Array annotation (e.g., gene identifiers, genomic coordinates, probe oligonucleotide sequence or reference commercial array catalog number).
6. Data processing protocols, like the normalization method.

These are the guidelines that outline the minimum information that should be included when describing a microarray or sequencing study and many journals and founding agencies require microarray data to comply with MIAME standards which also facilitates the further downstream data analysis[38].

The MIAME guidelines prelude is necessary to understand another used data format from GEO which is MIAME Notation in Markup Language (MINiML). It represents an XML rendering of SOFT format, optimized for programmatic access which provides better document structure through XML schema definitions[39].

Finally, series matrix files are compact and tab-delimited value matrix format with one gene per row and one sample per column, particularly these formats are the best one to train deep learning algorithms, these are convenient for direct import into data analysis programs such as Microsoft Excel, R or Python. Additionally, most series records include links to supplementary files containing raw data which can be downloaded individually or as compressed archives[40].

While GEO serves as the primary and only repository for the present work, there are several complementary databases that provide additional resources for cancer genomics and multiomics research. As previously mention ArrayExpress (AE), which is the European counterpart to GEO and is being maintained by the European Bioinformatics Institute (EMBL-EBI). It archives functional genomics data with similar structure and accessibility[41].

There is as well the Genomic Data Common (GDC)[42], a database that

[37] Alvis Brazma, Pascal Hingamp, John Quackenbush, Gavin Sherlock, Paul Spellman, Chris Stoeckert, John Aach, Wilhelm Ansorge, Catherine A Ball, Helen C Causton, et al. Minimum information about a microarray experiment (miame)—toward standards for microarray data. *Nature genetics*, 29(4):365–371, 2001

[38] NCBI. Miame and minseqe guidelines, 2024b. URL https://www.ncbi.nlm.nih.gov/geo/info/MIAME.html

[39] NCBI. Miniml (miame notation in markup language), 2024a. URL https://www.ncbi.nlm.nih.gov/geo/info/MINiML.html

[40] Emily Clough and Tanya Barrett. The gene expression omnibus database. In *Statistical Genomics: Methods and Protocols*, pages 93–110. Springer, 2016

[41] EMBL-EBI. Biostudies arrayexpress, 2025. URL https://www.ebi.ac.uk/biostudies/arrayexpress

[42] GDC Data Portal. Genomic data commons data portal, 2025. URL https://portal.gdc.cancer.gov/

provides a unified repository for large scaled cancer genomics programs including The Cancer Genome Atlas (TCGA)[43], which particularly contains comprehensive multiomic data from over 20, 000 primary cancer samples across 33 cancer types. As of early 2026, TCGA is just one of 91 projects encompassed by GDC, it stretches over 61 primary sites, more than 50,000 cases, 1.2 million files, 22,638 genes and 3.3 million mutations, which currently is one of the largest genomic data repositories in the western hemisphere. Another useful data repository is cBioPortal for Cancer Genomics[44], which similarly to GDC offers interactive exploration and visualization tools for large scale cancer genomic datasets.

[43] NHGRI. The cancer genome atlas program, 2025. URL https://www.genome.gov/Funded-Programs-Projects/Cancer-Genome-Atlas

[44] Memorial Sloan Kettering Cancer Center. cbioportal for cancer genomics. https://about.cbioportal.org/, 2025

2.4 *Microarray Data Analysis*

After the DNA probes of a microarray are passed through a special scanner to measure the expression of each gene printed in the slide, there are several computational steps to be performed after raw data is generated coming from the GeneChips. Raw data comes from a ".CEL" file which contains the gene expression quantified by fluorescence intensities of each probe, a CEL file is generated for each sample under investigation and then the analysis proceeds with bioinformatic processes such a the Linear Models for Microarray package (LIMMA)[45]. This package widely used via R is extensively used for analyzing gene expression data to asses differential expression across samples and screening for Differentially Expressed Genes (DEGs).

Largely the process of analyzing microarray data encompasses four major steps; Raw Data Quality Control, Preprocessing, Estimation of Differentially Expressed Genes (DEGs), and Functional Enrichment Analyses. A further and complementary pipeline overview can be seen in Figure 2.5.

[45] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015

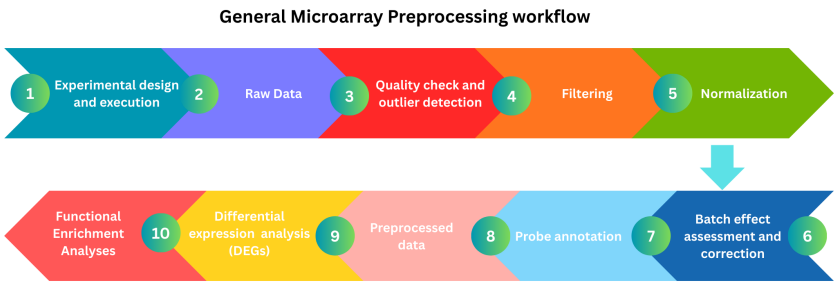

Figure 2.5: General overview of microarray processing from experimental design to functional enrichment analysis based on [Federico et al., 2021] & [Dhar and Kallapura, 2022].

Raw Data and quality controls are formed by multiple plots

representing; sample quality, hybridization quality, signal comparability, biases, and array correlation to mention some. A summary Quality Control table is also generated to support the plots, once the raw data is denoted as good quality, it is then taken further for preprocessing [Ritchie et al., 2015].

Preprocessing takes several steps in the workflow, it eliminates background noise by eliminating genes with little variation and underlines the ones that have significant variations across conditions, particularly, background correction detects expression intensity spots (foreground signal) from the ones surrounding (background signal) and corrects for cross-hybridization and optical detection of the scanner[46]. Then a normalization process is performed in which probe average or spot averaging is taken into consideration to later perform gene annotation. The default normalization strategy for most Affymetrix arrays is the Robust Multiarray Average (RMA), is a method that corrects for systematic bias to make arrays comparable among them[47]

Probe averaging stage continuous with the collapse of all signals generated and recorded from replicated spots within the array aiding only with estimates per gene expression. Once these steps are completed, probes are annotated with unique official gene symbols using the manufacturer annotations or public databases. Finally when these steps are performed with quality checks, now quality data is taken further for estimation of Differentially Expressed Genes (DEGs)[48].

Background corrected, and normalized data across samples are compared to identify the DEGs that are differentially regulated across combinations. LIMMA creates a design matrix, built on a linear modeling function called lmFit, then the intensity values are applied to lmFit and a contrast matrix is generated, representing comparison across all the groups[49]. Finally, the contrast matrix generated is applied to the modeled data to further compute eBayes statistics, and the outputs involving a set of DEGs with several statistics, including Log2Fold Change, P-value, Log odds, Standard Deviation among others, then the output would represent the list of the DEGs and the statistical processes will be further explain.

To identify significant DEGs, LIMMA uses moderated t-test criteria of statistical significance, where rather than estimating within group variability for each gene, it pools the global information from all other genes across replicates of samples under comparison. This bypasses the need for multiple replicates in order to estimate gene specific variance, which otherwise would yield small P-values by chance. Particularly with LIMMA moderated t-test, we could accurately estimate several significantly DEGs across comparing combinations, therefore a gene is considered significantly differentially expressed when its P-value is ≤ 0.01. Then once identified set of DEGs they are further taken for

[46] Matthew E Ritchie, Jeremy Silver, Alicia Oshlack, Melissa Holmes, Dileepa Diyagama, Andrew Holloway, and Gordon K Smyth. A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, 23(20): 2700–2707, 2007

[47] Gordon K Smyth, Joëlle Michaud, and Hamish S Scott. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, 21(9):2067–2075, 2005

[48] Debojyoti Dhar and Gopala Kallapura. Analysis of microarray data from medulloblastoma tissue samples. In *Medulloblastoma: Methods and Protocols*, pages 59–64. Springer, 2022

[49] Belinda Phipson, Stanley Lee, Ian J Majewski, Warren S Alexander, and Gordon K Smyth. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *The annals of applied statistics*, 10(2):946, 2016

various functional enrichment analyses[50].

Functional Enrichment Analyses are performed to gain further insights into biological functions from the previously defined DEGs, most of these analyses are performed using a tool called STRING (Search Tool for Recurring Instances of Neighboring Genes). It retrieves and display the genes from a predefined query that appear within clusters of the genome[51]. The tool performs iterative searches and visualizations of the genomic context coming from associated genes of the genome and the specified query delineates a set of potential functionally associated genes. STRING employs a biological database that quantitatively integrates both direct (physical) and indirect (functional) associations, then it allows to selectively enrich three ontological signatures; Biological Processes (BP), Molecular Functions (MF), and Cellular Components (CC)[52].

STRING processes the set of significantly DEGs from each comparing combinations by enriching each category of the gene ontology and sets a critical cutoff filter of False Discovery Rate (FDR) < 0.05, to ensure only the significant genes of BP, MF, CC are indeed captured. The outputs are recorded as a list of pathways, processes, and functions across several graphs. As a complement of the analyses from STRING, the complete list of genes are further enriched for their involvement in various biological pathways employing KEGG Pathway Enrichment. All the pathways with an FDR < 0.05 are capture through enrichment with relevant pathways cell cycle such as signaling pathways of cancer, hedgehog pathways, and other signaling pathways, which are relevant for the purpose of the present study. The identified pathways and biological processes are further mapped on physical KEGG maps, by highlighting the significantly DEGs on the specified dataset[53].

[50] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015

[51] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, et al. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, 47(D1):D607–D613, 2019

[52] Christian von Mering, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. String: a database of predicted functional associations between proteins. *Nucleic acids research*, 31(1):258–261, 2003

[53] Ganiraju Manyam, Aybike Birerdinc, and Ancha Baranova. Kpp: Kegg pathway painter. *BMC Systems Biology*, 9(Suppl 2):S3, 2015

2.5 *Microarray Data Preprocessing*

Microarrays represent a milestone in biomedical sciences, as this technology opened the doors to omic sciences and revolutionized the way molecular compartments of a cell, such as the transcriptomics, are profiled. However, the nature of data generated from DA microarray experiments introduces several analytical challenges, particularly regarding quality assessment and data transformation.. Countless algorithms and methods have been developed to address these challenges, with preprocessing procedures specifically designed to mitigate biases that may arise across different stages of the experimental workflow .

Preprocessing techniques constitute paramount steps in any microarray data analysis pipeline. According to [Federico et al., 2021], data preprocessing is of critical importance for developing robust

and reproducible analytical pathways, spanning from experimental design to reliable biological interpretation. These steps are organized sequentially, encompassing quality checks, batch effect removal, visualization methods, and data representation, and are sufficiently robust to be applied across different gene expression microarray platforms.

As illustrated in Figure 2.5, the preprocessing of gene expression microarray data follows a well-established pipeline widely recognized as the gold standard in the field. This pipeline can be broadly organized into eight sequential stages: experimental design, quality control, filtering, imputation, normalization, batch effect estimation and correction, probe annotation, and data representation[54]. Each of these stages encompasses multiple methodological approaches and analytical tools, most of which are addressed throughout the development of this work, and are described as follows:

[54] Veer Singh Marwah, Giovanni Scala, Pia Anneli Sofia Kinaret, Angela Serra, Harri Alenius, Vittorio Fortino, and Dario Greco. eutopia: solution for omics data preprocessing and analysis. *Source code for biology and medicine*, 14(1):1, 2019

1. **Experimental Design:** The pipeline begins with proper sample selection to achieve statistical significance for meaningful downstream analysis, including appropriate determination of the number of replicates and control samples. Since microarrays are relatively expensive assays, there is typically a trade-off between the number of sample groups and the number of replicates. Adequate randomization is essential, as it can correct for most technical biases. Accordingly, thorough recording of each sample's features is required to identify additional challenges introduced by uncontrolled technical factors, such as sample processing dates, variability in sample processing or collection procedures, dye usage, and slide or chip area effects.

2. **Quality Control:** The assessment of microarray data quality lacks a universal consensus, as these experiments are highly dependent on sample quality. The use of pure and intact samples is crucial for obtaining reliable biological information. Chip image analysis can additionally address quality irregularities at the individual sample level. Several quality control (QC) approaches have been proposed, many of which rely on the visual inspection of data distribution[55].

   Among the most commonly used visualization tools are MA plots, an example can be seen in Figure 2.6, which represent probe intensity distributions between two samples or groups of samples. The x-axis displays the average probe intensities of the compared samples (A), while the y-axis represents the absolute difference in $\log_2$ scale of the probe intensities (M).

   Heatmaps of mean absolute difference are also widely used for quality assessment. This distance-based method enables comparison across multiple arrays by computing pairwise distances among

[55] Lars MT Eijssen, Magali Jaillard, Michiel E Adriaens, Stan Gaj, Philip J de Groot, Michael Müller, and Chris T Evelo. User-friendly solutions for microarray quality control and pre-processing on arrayanalysis. org. *Nucleic acids research*, 41(W1):W71–W76, 2013

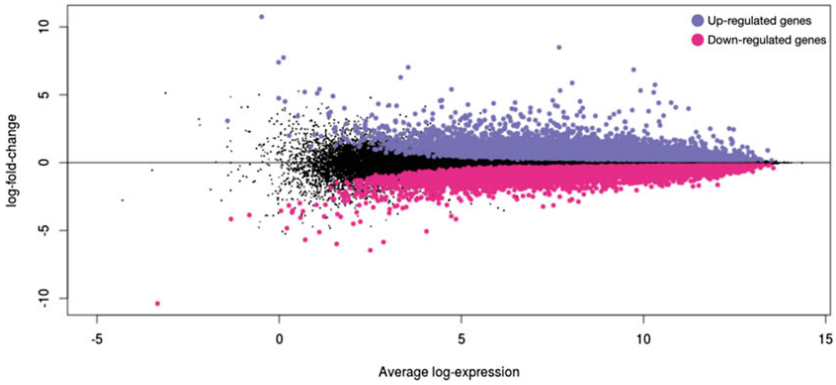

Figure 2.6: MA plot representation of gene expression distribution between two different experimental conditions using microarray data from GEO Dataset ID: GSE34248[Federico et al., 2021].

them. These distances are expressed through a color scale and an accompanying dendrogram that illustrates sample clustering patterns. The method is particularly useful for detecting outliers, as biological replicates tend to cluster together, and deviations from expected groupings may indicate the presence of batch effects[56].

[56] Audrey Kauffmann, Robert Gentleman, and Wolfgang Huber. arrayqualitymetrics—a bioconductor package for quality assessment of microarray data. *Bioinformatics*, 25(3):415–416, 2009

Histograms provide a common representation of probe intensity distributions in a microarray experiment. The underlying assumption is that measured probe intensities are directly proportional to gene expression levels. To better represent this distribution and approximate a normal distribution, expression values are transformed to a logarithmic scale. The x-axis reports the $\log_2$-transformed gene expression estimates, while the y-axis reports the corresponding frequencies, as it can be seen in Figure 2.7.

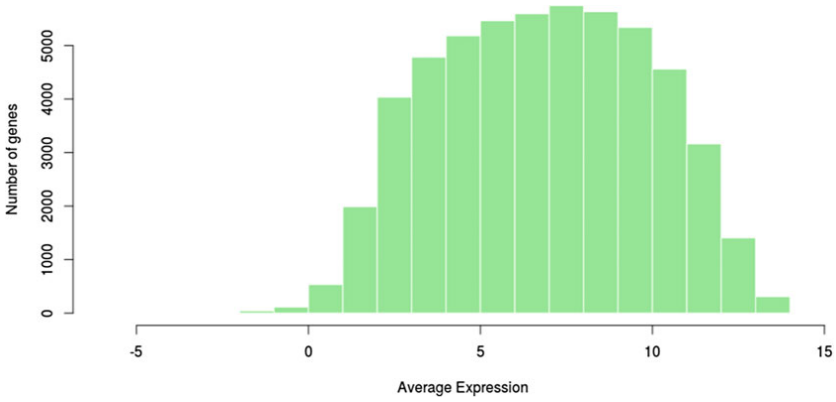

Figure 2.7: Histogram showing the distribution of gene expression estimates in a microarray experiment [Federico et al., 2021].

Density plots serve as a smooth alternative to histograms for visualizing gene expression estimate distributions. Comparing density curves across samples grouped by technical and biological features provides an additional means of assessing sample quality, as shown in Figure 2.8.

Scatterplots highlight variation between arrays by pointing the

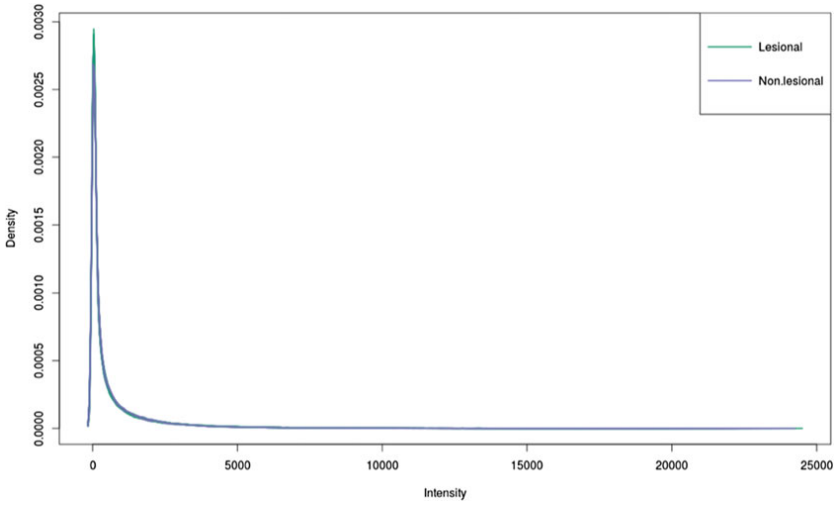

Figure 2.8: Density plot of gene expression between lesional and non lesional skin samples. The largely overlapping curves reflect the proportion of probes exhibiting a given intensity, with most probes showing low intensity in this experiment[Federico et al., 2021].

log-transformed intensity values of two compared samples on the x-and y-axes, respectively. They are effective for evaluating the consistency of technical or biological replicates and identifying potentially problematic arrays, as illustrated in Figure 2.9.

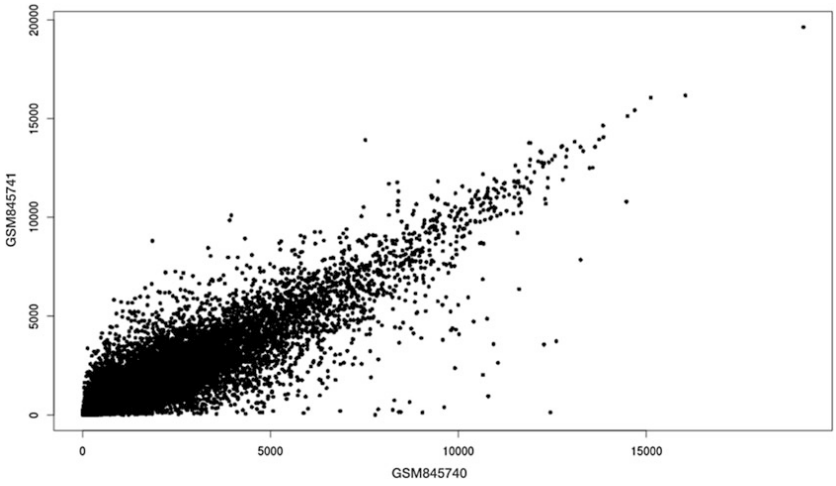

Figure 2.9: Scatter plot comparing gene expression estimates between two samples from dataset GSE34248 [Federico et al., 2021].

Among the most informative quality control approaches are dimensionality reduction techniques, which facilitate the identification of problematic samples or features by projecting high-dimensional data into low-dimensional space and annotating them using known biological or technical groupings. The most commonly employed methods in microarray and broader bioinformatic analysis include Uniform Manifold Approximation and Projection (UMAP), as it can be seen in Figure 2.10 and Principal Component Analysis

(PCA).

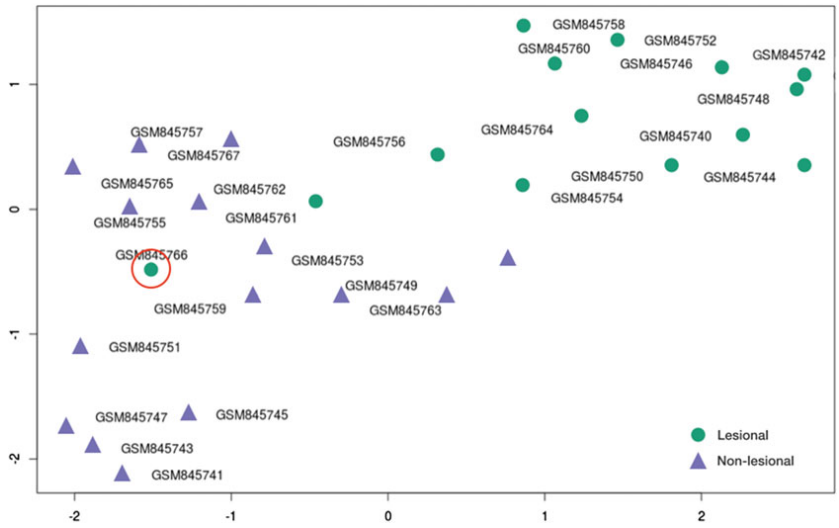

Figure 2.10: UMAP plot showing the projection of samples into a low-dimensional space. The circled sample, given its isolated position, represents a candidate outlier [Federico et al., 2021].

3. **Filtering:** A number of transformations must be applied during microarray analysis to eliminate questionable or low-quality measurements, adjust measured intensities to facilitate comparisons, and select genes that are significantly differentially expressed between sample classes[57]. Probe filtering, in particular, aims to identify and remove probes with spurious values that do not accurately represent the underlying expression state and could lead to erroneous results. This can be addressed by comparing signal intensities against platform´s negative control probes and by examining signal variance across samples.

4. **Imputation:** Missing measurements, arising from absent signals or applied filters, can compromise downstream analyses that require complete input datasets. Depending on the nature of subsequent analyses, a complementary version of the dataset with substituted values values may be produced. Methods reported in the literature for microarray data imputation include K-nearest neighbors (KNN), regression models, and central tendency based approaches[58]

5. **Normalization:** Normalization is a critical step that enables robust and meaningful comparisons across molecular targets, samples, and experiments. Its primary goal is to adjust signal intensities and remove sources of systematic variation that may otherwise confound results. Various normalization strategies exist for microarray data, and the preferred method is often platform-dependent, though several non-dependent platform approaches are also available. Normalization methods can be classified as within-

[57] John Quackenbush. Microarray data normalization and transformation. *Nature genetics*, 32(4):496–501, 2002

[58] Shahla Faisal and Gerhard Tutz. Missing value imputation for gene expression data by tailored nearest neighbors. *Statistical applications in genetics and molecular biology*, 16(2):95–106, 2017; and Pietro Di Lena, Claudia Sala, Andrea Prodi, and Christine Nardini. Methylation data imputation performances under different representations and missingness patterns. *BMC bioinformatics*, 21(1):268, 2020

array or between-array techniques, depending on the data scope considered for transformation, as exemplified in Figure 2.11. Within-array techniques address differences linked to differential chemistries within a single chip, while between-array techniques reduce technical noise arising from the distribution of samples across different chips.

Data standardization is another widely used approach, involving the transformation of expression estimates into Z-scores, whereby values of individual genes are expressed in units of standard deviation (SD) from the normalized mean of zero. This approach renders data from arrays representing distinct biological conditions comparable for differential analysis, with detected differences expressed as Z-score ratios (Z-ratios)[59].

Quantile normalization, the most widespread approach in microarray analysis and also broadly employed in high-throughput sequencing technologies, aims to render probe intensity distributions identical across all arrays in a dataset. This method originates from the quantile-quantile (QQ) plot representation, in which two data vectors sharing the same distribution from a straight diagonal line[60]. By assigning the mean quantile value to each corresponding data point across all arrays, this normalization ensures a uniform distribution while preserving the relative ordering of expression values.

6. **Batch Effect Estimation and Correction:** Systematic variation observed between groups of samples processed under different technical conditions is commonly referred to as a batch effect. If left undressed, batch effects can lead to misleading conclusions, as technical biases may be strong enough to mask the true underlying biological signal[61]. Potentially confounding variables, such as processing date, sample handler, and separate reagent batches, should be carefully documented in the metadata alongside biological variables and exposure details, enabling the identification of non-biological patterns in the data.

   Unknown sources of variation may also arise from hidden technical factors. These can be estimated using Surrogate Variable Analysis (SVA), implemented in the R package *sva*, which identifies latent variables that serve as surrogates for the technical variation observed in the data[62]. Batch effects can be visualized by clustering samples using PCA, which decomposes the data into principal components that capture the dominant sources of variance, thereby allowing quantification of technical variable contributions. Approaches for correcting batch effects include linear models and empirical Bayesian methods[63]. In particular, the empirical Bayes method ComBat, part of SVA framework, has been shown to outperform most alternative batch effect correction methods.

[59] Chris Cheadle, Marquis P Vawter, William J Freed, and Kevin G Becker. Analysis of microarray data using z score transformation. *The Journal of molecular diagnostics*, 5(2):73–81, 2003

[60] Xing Qiu, Hulin Wu, and Rui Hu. The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. *BMC bioinformatics*, 14(1):124, 2013

[61] Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010

[62] Jeffrey T Leek, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, 2012

[63] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8 (1):118–127, 2007

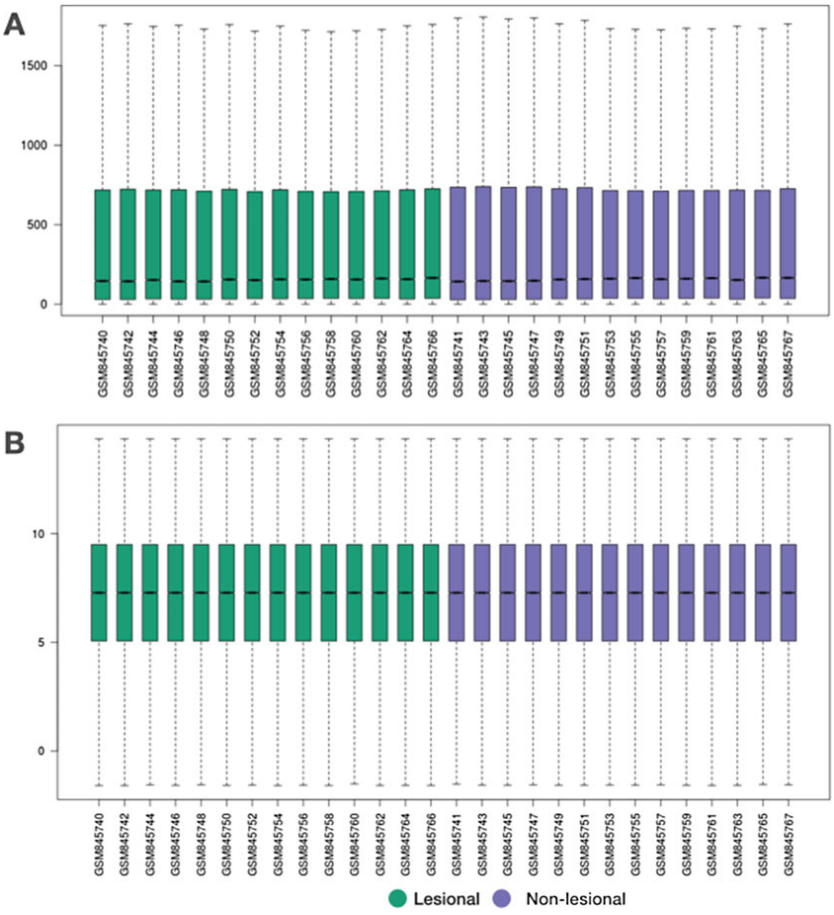

Figure 2.11: Boxplots showing gene expression distribution of lesional and non lesional skin of atopic dermatitis patients before (A) and after (B) between array normalization [Federico et al., 2021].

7. **Probe Annotation:** The probes on a microarray chip represent known biological entities, such as genes or CpG sites, and must be mapped to standardized annotation databases to enable meaningful biological interpretation. Although oligonucleotide probe sequences rarely change, genome assemblies are frequently revised, which can occasionally introduce discrepancies between microarray platforms. Platform-specific annotation packages are available for R such as Bioconductor´s *AnnotationDbi*[64], which maps probes to commonly used molecular identifiers including Ensembl IDs, Entrez Gene IDs, gene symbols.

[64] M Carlson, S Falcon, N Li, et al. Annotationdbi: Manipulation of sqlite-based annotations in bioconductor. *R Package Version 1.54. 1*, 2021

8. **Data Representation:** An important consideration in evaluating gene expression estimates is the manner in which differences across samples are represented. The logarithmic transformation of the expression ratio ($\log_2$(ER)) is almost universally adopted as the standard representation in microarray data analysis. Expression ratios in this scale are intuitively interpretable, as they are symmetrically distributed around zero, upregulated genes yield positive values while downregulated genes yield negative values. Furthermore, this transformation projects the data onto a normal distribution, making it suitable for differential expression analyses based on linear regression models.

Differential testing can be considered the 9th step in data preprocessing, but should be considered another main step as it helps to identify relevant genes across conditions. Once the data has been preprocessed through the preceding stages, it is ready for differential expression analysis, which aims to identify genes whose expression levels show statistically significant differences between two or more experimental conditions. Classical approaches employ linear models to estimate the variability between experimental conditions while accounting for covariate dependencies such as slide, array, processing date or biological variables like tissue type or cell lineage.

A central metric in differential expression analysis is the log fold change (logFC), which provides a quantitative estimation of expression differences between two conditions. It is computed as the $\log_2$ ratio between the mean expression values of samples belonging to the first condition and that of the second condition. A positive logFC indicates increased expression in the first condition relative to the second, while a negative logFC indicates a decrease. The logFC is typically reported alongside a p-value that estimates its statistical significance, and conventional thresholds for identifying differentially expressed genes are set at an absolute logFC > 0.58 and a p-value < 0.05.

Given the high dimensionality of microarray data, where thousands

of genes are tested simultaneously, p-values must be corrected to control the accumulation of false positives. The most stringent correction is the Bonferroni method, which divides the significance threshold by the total number of test performed. A less conservative and more widely adopted alternative is the False Discovery Rate (FDR), which controls the expected proportion of false positives among all significant results, making it better suited for high-dimensional genomic data. Differential expression analysis in microarray studies is commonly conducted using the LIMMA package from the Bioconductor repository[65], which implements linear models and empirical Bayesian moderation of variance estimates to improve statistical power in small sample settings.

A standard graphical representation of differential expression results is the volcano plot, which can be seen in Figure 2.12, which simultaneously displays the magnitude of expression change on the x-axis and the statistical significance on the y-axis as -$\log_{10}$(p-value), allowing for rapid visual identification of biologically and statistically relevant genes.

[65] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015

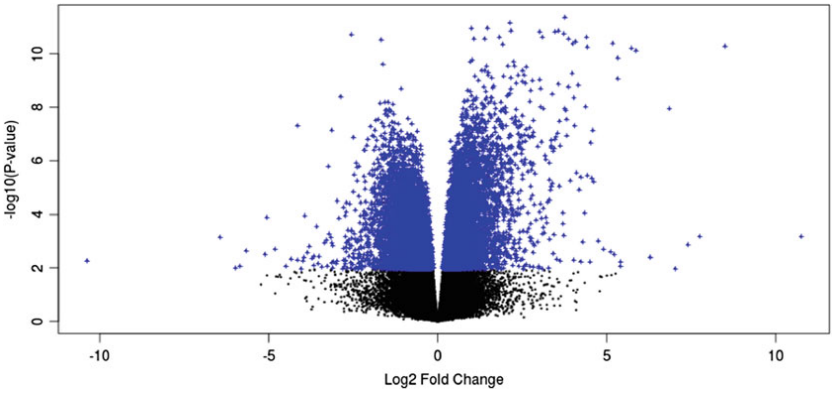

Figure 2.12: Volcano plot that shows the magnitude of deregulation in terms of log fold changes and the p-values of the genes after differential analysis, the blue dots represent differential expressed genes [Federico et al., 2021].

The sequential execution of the preprocessing pipeline described encompassing quality control, filtering, normalization, probe annotation, and differential expression analysis, culminates in a structured, high-confidence gene expression matrix that constitutes the molecular fingerprint of the analyzed samples. In the context of the present work, this pipelines aims to be applied to microarray data processed through `oligo` package in R with RMA normalization, yielding a curated set of DEGs identified through LIMMA across the four molecular subgroups of medulloblastoma relative to cerebellar control tissue. The resulting DEG matrix, exported as a `.csv` file containing $\log_2$ fold change values and FDR-corrected significance estimates, serves as the primary input for the Artificial Neural Network (ANN) classification model developed before by [Reveles-Espinoza et al., 2025] which would be adapted, applied and validated on

new cohorts and datasets with a more robust methodological design. The dimensionality reduction from the full transcriptomic space to a biologically meaningful minimal gene signature provides the feature set upon which the ANN learns to discriminate between molecular subgroups, linking the bioinformatics preprocessing workflow directly to the supervised classification framework described in the following chapters.

2.6 *Fundamentals of Machine Learning and Deep Learning*

Machine learning (ML) and molecular quantification together provide a reproducible computational framework for disease classification based on molecular fingerprints, supported by both statistical robustness and experimental consistency. In recent years, ML and Deep Learning (DL) approaches have been increasingly applied to the categorization and diagnosis of human diseases, including cancer, where high-dimensional molecular data demands methods capable of capturing complex, non-linear patterns beyond the reach of classical statistical techniques. In the context of medulloblastoma, transcriptomic data processed through machine learning pipelines has demonstrated potential for detecting characteristic molecular fingerprints associated with distinct tumor subgroups, underscoring the value of these approaches for improving diagnostic precision in settings where conventional molecular profiling methods may be limited or unavailable[66].

Transcriptomic data are particularly relevant for addressing a number of analytical challenges inherent to molecular profiling. The large volume of molecular data produced by omics-based technologies has enabled the development and application of artificial intelligence (AI) methods at an unprecedented scale, and the continuous growth of publicly available omics datasets is accompanied by an expanding repertoire of computational methods designed to facilitate their analysis, interpretation, and the generation of accurate and stable predictive models. Within this landscape, ML encompasses both unsupervised and supervised learning paradigms. Unsupervised methods, such as clustering, do not require prior classification of samples and instead group them based on similarities across selected features, making them particularly useful for exploratory analyses and the discovery of novel molecular subgroups. Supervised methods, by contrast, require discrete or continuous endpoints and are frequently combined with feature selection strategies to identify an optimal subset of variables capable of discriminating between endpoint values. This reduced feature set can then be leveraged for the prediction of class membership or the estimated effect of a new, unseen sample[67].

[66] Alicia Reveles-Espinoza, Ulises Villela, Edgar Hernandez-Martinez, Isaac Chairez, Sergio Juárez-Méndez, J Casanova-Moreno, Ma del Pilar Eguía-Aguilar, Luis Figueroa-Yáñez, Adriana Vallejo-Cardona, and Iván Salgado. Machine learning identification of molecular targets for medulloblastoma subgroups using microarray gene fingerprint analysis. *Computational and Structural Biotechnology Journal*, 2025

[67] Angela Serra, Michele Fratello, Luca Cattelani, Irene Liampa, Georgia Melagraki, Pekka Kohonen, Penny Nymark, Antonio Federico, Pia Anneli Sofia Kinaret, Karolina Jagiello, et al. Transcriptomics in toxicogenomics, part iii: data modelling for risk assessment. *Nanomaterials*, 10(4):708, 2020

2.6.1 *Artificial Neural Networks*

Artificial Neural Networks (ANNs) emerged from the study of how the human brain performs complex computational processes. The biological brain is capable of highly complex, nonlinear, and parallel processing by organizing its structural constituents, neurons, into interconnected networks that enable task such as accurate prediction and pattern recognition. A defining characteristic of biological neurons is their high degree of interconnectivity, as each neuron receives stimuli from numerous neighboring neurons in the form of electrical signals, integrating this information to produce a response that propagates through the network. ANNs are computational systems designed to replicate this architecture by constructing networks composed of hundreds or thousands of artificial processing units that imitate the behavior of biological neurons. A fundamental property of ANNs is their capacity for experience driven learning, rather than requiring explicit programming of decision rules, an ANN develops its own behavioral rules by exposure to data through a computational learning algorithm. This is achieved through a massively parallel computing structure in which artificial neurons are interconnected and communicate through synaptic weights, which capture and store the strength of connections between units and are progressively adjusted during training to minimize prediction error[68].

[68] Osval Antonio Montesinos López, Abelardo Montesinos López, and Jose Crossa. Fundamentals of artificial neural networks and deep learning. In *Multivariate statistical machine learning methods for genomic prediction*, pages 379–425. Springer, 2022

The fundamental components of an ANN model are illustrated in Figure 2.13. The network receives external information through a set of inputs $x_1, \ldots, x_p$, which represent signals originating either from the external sensory system or from other neurons with which a given unit shares a connection. Each input is modulated by a corresponding synaptic weight, collectively represented by the vector $W = (w_1, \ldots, w_p)$, which modifies the received information in a manner analogous to the biological synapse. These weights function as adaptive gains that can either attenuate or amplify the values propagated towards the neuron. Additionally, each neuron incorporates a bias parameter b_j, also referred to as the intercept or threshold, which shifts the activation independently of the input. In the context of ANNs, learning refers precisely to the iterative modification of these synaptic weights and biases based on the network´s exposure to training data.

The weighted inputs are summated to produce what is known as the net input, expressed mathematically as:

$$v_j = \sum_{j=1}^{p} \omega_{ij} x_j \tag{2.1}$$

This net input v_j determines whether the neuron becomes activated, a decision governed by the activation function g. The output of the

neuron is therefore expressed as $y_i = g(v_j)$, where the choice of activation function defines the nature of the neuron´s response. For instance, a unit step function, also referred to as a threshold function, produces a binary output, yielding $y_j = 1$ if $v_j > 0$ and $y_j = 0$ otherwise [Montesinos López et al., 2022].

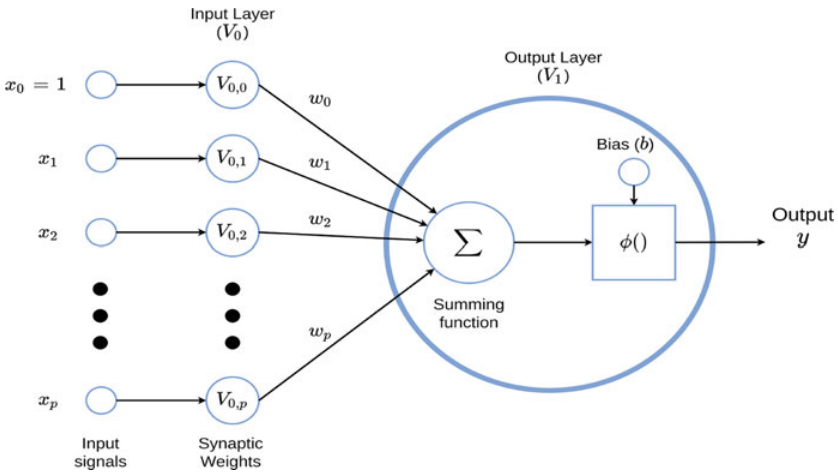

Figure 2.13: General model of an Artificial Neural Network (ANNs) [Montesinos López et al., 2022].

A single layer ANN, such as the one depicted in Figure 2.13, possess limited processing capacity on its own, and its applicability to complex real world problems is consequently restricted. Its true computational power emerges from the interconnection of many of such units into structured networks. An ANN can thus be understood as a system composed of numerous simple processing elements operating in parallel, whose collective function is determined by the network topology and the weights of the connections between neurons[69].

Deep Learning represents an extension of this principle, constituting a subset of ML in which layers are connected in more complex configurations that those found in conventional shallow ANNs. DL models employ larger numbers of neurons and multiple successive layers to capture nonlinear aspects of complex data more effectively, although there is a cost of great computational demands. As illustrated in Figure 2.14, a DL model can be formally represented as a directed graph whose nodes correspond to neurons and whose edges correspond to weighted connections between them, with each neuron receiving as input a weighted sum of the outputs of all neurons connected to its incoming edges[70].

The network in figure 2.14 comprises four layers, V_0, V_1, V_2, V_3, where V_0 denotes the input layer carrying the raw feature information, V_1 and V_2 are hidden layers responsible for successive levels of feature abstraction, and V_3 represents the output layer that produces the final prediction. This hierarchical organization of three trainable layers,

[69] Eduardo Francisco Caicedo Bravo. *Una aproximación práctica a las redes neuronales artificiales*. Universidad del Valle, 2009

[70] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014

excluding the input, is what qualifies the architecture as a deep network, and it is precisely this depth that enables the autonomous extraction of increasingly abstract representation from high dimensional input data [Montesinos López et al., 2022].

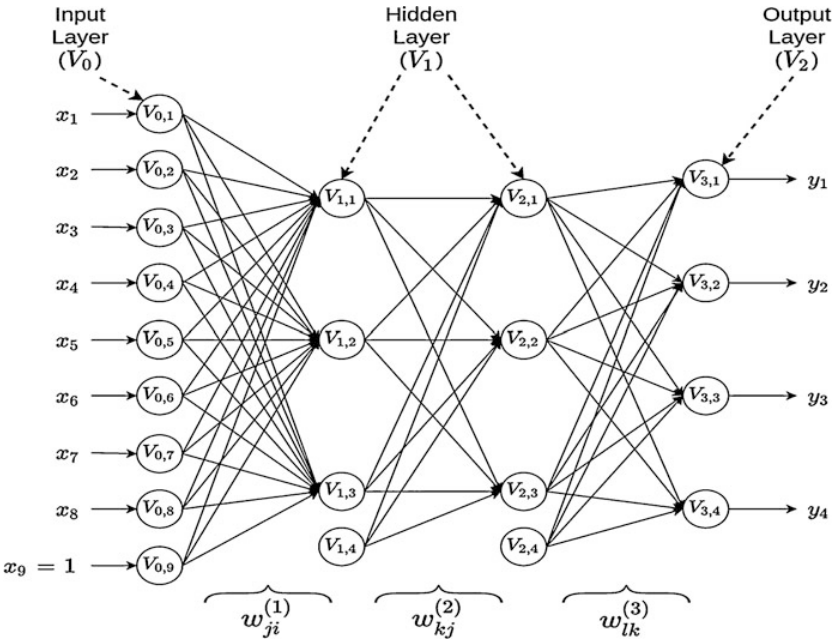

Figure 2.14: Feedforward ANN architecture with eight input variables $(x_1, ..., x_8)$, four output variables (y_1, y_2, y_3, y_4), and two hidden layers with three neurons each [Montesinos López et al., 2022].

The architecture of an ANN is organized into three distinct layer types, each serving a specific functional role in the information processing pipeline. The *Input Layer* constitutes the set of neurons that directly receives information from external sources, and consequently the number of neurons it contains typically corresponds to the number of explanatory variables provided to the network. In feedforward neural networks specifically, the input layer is fully connected to the subsequent hidden layer. establishing the initial flow of information through the network[71]. *Hidden layers* are composed of internal neurons that have no direct contact with the outside environment. A network may contain zero, one, or multiple hidden layers, and it is the presence of multiple hidden layers that defines a deep network architecture. The neurons within each hidden layer collectively process and transform the information received from the preceding layer, and their interconnection patterns, together with their number, determines the overall topology of the ANN. Critically, the knowledge extracted from training data is encoded in the synaptic weight values of the connections between layers, and it is precisely the hidden layers that enable the network to capture complex nonlinear behaviors in the data. The *Output Layer* transfers the processed information to the outside, delivering the final

[71] Josh Patterson and Adam Gibson. *Deep learning: A practitioner's approach*. " O'Reilly Media, Inc.", 2017

answer or prediction of the model. The nature of this output, whether continuous, binary, ordinal, or count, is determined by the activation function specified for the output layer neurons, making the choice of activation function a critical architectural decision that must be aligned with the nature of the prediction task [Bravo, 2009, Patterson and Gibson, 2017].

2.6.2 *Activation Functions*

Activation functions propagate the output of one layer´s neurons forward to the next, up to and including the output layer. They are scalar-to-scalar functions whose primary role is to introduce nonlinearities into the network´s modeling capabilities, enabling it to approximate complex, non-linear relationships that a pure linear model could not capture[72]. Formally, the activation function $g(\cdot)$ defines how a neuron becomes activated given its net input, producing the neuron´s output as $y_j = g(v_j)$. To illustrate this, consider a linear activation function defined as $g(v_j) = v_j$, in which case the neuron´s output reduces to a linear combination of its weighted inputs, equivalent to a standard linear regression model, which severely limits the network´s representational capacity.

[72] Joshua Fredrick Wiley. *R deep learning essentials*. Packt, 2016

For the purpose of this work, two activation functions were employed based on the specific requirements of each stage of the classification architecture.

The **Rectified Linear Unit (ReLU)** is one of the most widely used activation functions in deep learning. It is flat below a defined threshold, typically zero, and linear above it, activating a neuron only when the input exceeds that threshold. Formally, it is defined as $g(v_j) = \max(0, v_j)$. When the input is below zero the output is zero, and when the input rises above the threshold it maintains a linear relationship with the dependent variable, as illustrated in Figure 2.15. Despite its simplicity, ReLU provides nonlinear transformation capabilities, and a sufficient number of linear rectifiers can be used to approximate arbitrary nonlinear functions[Patterson and Gibson, 2017].

ReLU is considered state of the art, having demonstrated reliable performance across a wide range of problems and architectures. However, because of its gradient is either zero or a constant, it does not entirely resolve the vanishing or exploding gradient issue, a limitation commonly referred to as the *dying ReLU* problem. This activation function is most widely employed in hidden layers and in output layers where the response variable is continuous and greater than zero

The **hyperbolic tangent (Tanh)** activation function is defined as:
$g(v_j) = \tanh(v_j) = \frac{\sinh(v_j)}{cosh(v_j)} = \frac{exp(v_j) - exp(-v_j)}{exp(v_j) + exp(-v_j)}$
Tanh produces an S-shaped output bounded in the range (-1, 1), as

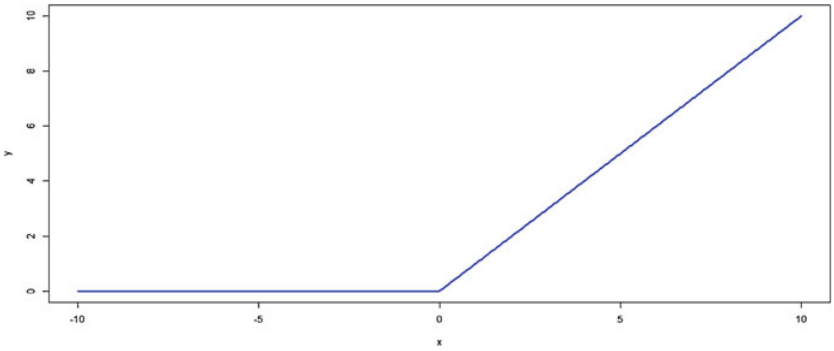

Figure 2.15: Representation of ReLU activation function [Montesinos López et al., 2022].

illustrated in Figure 2.16. A key advantage of this function over the sigmoid activation is that it is less prone to getting stuck during training, owing to its zero-centered output range. Large negative inputs yield large negative outputs, while large positive inputs yield large positive outputs, and its symmetric nature around zero means it can handle negative values more effectively than alternative activation functions.

For these reasons, Tanh is generally preferred over sigmoid for hidden layer activations. In the context of the present work, this property was particularly relevant for the discrimination of medulloblastoma Group 3 (G3) from Group 4 (G4), whose transcriptomic profiles exhibit high molecular homogeneity, as the zero-centered and symmetric output of Tanh provided a more sensitive threshold of separation based on the subtle molecular fingerprint differences between these two subgroups [Reveles-Espinoza et al., 2025].

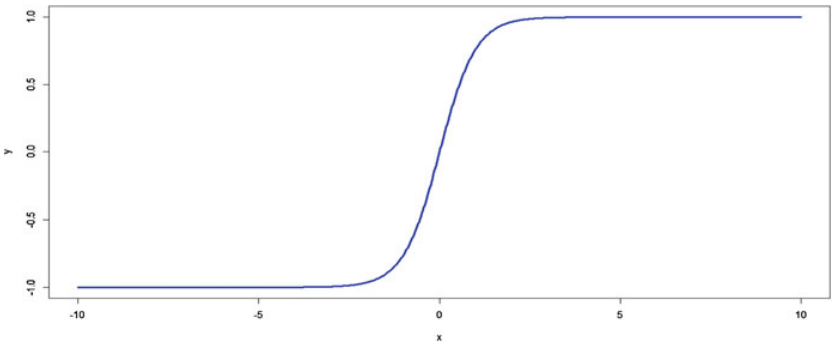

Figure 2.16: Representation of tanh activation function [Montesinos López et al., 2022].

2.6.3 *Feedforward Neural Network Topology*

Neural network topology refers to the structural relationships between neurons established through their connections, and plays a fundamental role in determining the functionality and overall performance of the

network. The terms structure and architecture are used interchangeably as synonyms for the network topology[73]. For the development of the present work, feedforward networks constitute the chosen architectural methodology.

In feedforward networks, information flows strictly in a single direction, from the input layer through one or more processing layers toward the output layer. This unidirectional flow implies the existence of only interlayer connections, meaning that the neurons within the same layer share no connections between them, and no connection transmits information from a higher layer back to a lower one. The absence of intralayer and supralayer connections defines the feedforward topology and distinguishes it from recurrent architectures[74]. Despite this structural simplicity, feedforward networks are not restricted to a single hidden layer, and the inclusion of multiple hidden layers, as illustrated in Figure 2.17, constitutes the basis for deep feedforward architectures.

[73] Osval Antonio Montesinos López, Abelardo Montesinos López, and Jose Crossa. Fundamentals of artificial neural networks and deep learning. In *Multivariate statistical machine learning methods for genomic prediction*, pages 379–425. Springer, 2022

[74] Eduardo Francisco Caicedo Bravo. *Una aproximación práctica a las redes neuronales artificiales*. Universidad del Valle, 2009

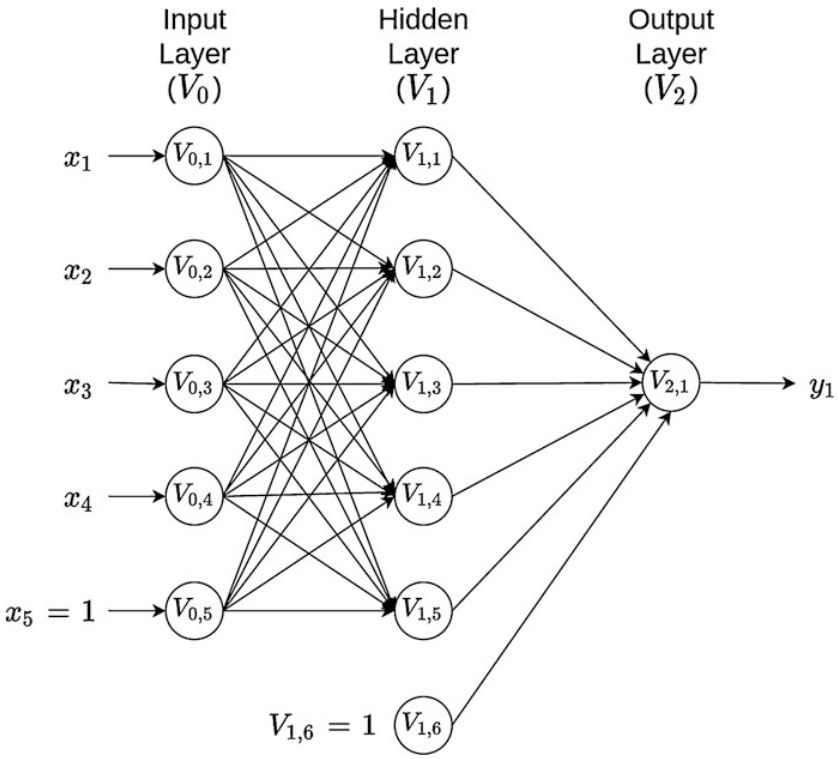

Figure 2.17: Simple two-layer feedforward artificial neural network [Montesinos López et al., 2022].

2.6.4 *Loss Function*

A loss function, also referred to as an objective function, is a function that maps the values of one or more variables onto a real number

representing the cost associated with a given prediction. In the context of statistical machine learning, a loss function quantifies how closely the predicted values produced by an ANN or DL model approximate the true values, measuring the quality of the network´s output by computing a distance score between observed and predicted values[75].

The fundamental principle consist of calculating a metric based on the observed error between true and predicted values across the entire dataset, averaged into a single scalar value that represents the overall performance of the network relative to its ideal. By minimizing this scalar, it becomes possible to identify the parameters (weights and biases) that produce the best possible predictions. Training ANN models with loss functions therefore enables the use of optimization methods to estimate the required network parameters[76].

Although an analytical solution for parameter estimation is generally not attainable, good approximations can be obtained through iterative optimization algorithms. The most fundamental of these is gradient descent, which provides the basis for the parameter update procedures employed during ANN training.

The most fundamental iterative optimization algorithm used to minimize the loss function in ANN training is gradient descent. In the context of feedforward networks, backpropagation is employed to compute the derivatives of the performance function with respect to the weight and bias variables, and each parameter is subsequently adjusted according to the gradient descent update rule. A widely used variant is gradient descent with momentum (`traingdm` in MATLAB´s Deep Learning Toolbox), which extends standard gradient descent by allowing the network to respond not only to the local gradient but also to recent trends in the error surface. Acting as a low-pass filter, momentum enables the network to bypass shallow local minima that would otherwise trap a standard gradient descent algorithm[77].

The parameter update rule for gradient descent with momentum is defined as:

$$\Delta X = mc \cdot \Delta X_{prev} + lr \cdot (1 - mc) \cdot \frac{\partial \text{perf}}{\partial X} \tag{2.2}$$

where ΔX_{prev} is the previous weight or bias adjustment, lr is the learning rate, and mc is the momentum constant, which is set between 0 (no momentum) and values approaching 1 (maximum momentum). A momentum constant of 1 renders the network completely intensive to the local gradient and therefore unable to learn effectively. Training is halted when one of the following conditions is met: the maximum number of epochs is reached, the maximum training time is exceeded, the performance goal is achieved, the performance gradient falls bellow a minimum threshold, or the validation error increases beyond a defined

[75] F Chollet and JJ Allaire. Deep learning with r. manning publications, manning early access program (mea), 2017

[76] Josh Patterson and Adam Gibson. *Deep learning: A practitioner's approach*. " O'Reilly Media, Inc.", 2017

[77] MathWorks. traingdm - gradient descent with momentum backpropagation. https://la.mathworks.com/help/deeplearning/ref/traingdm.html, 2024a. MATLAB Deep Learning Toolbox Documentation

maximum number of consecutive failures.

The theoretical and conceptual framework presented throughout this chapter establishes the foundational knowledge required to understand the computational and biological underpinnings of the present work. Beginning with the molecular basis of transcriptomic profiling and the preprocessing pipeline essential for generating reliable gene expression data, the framework progressed through the statistical methods employed for differential expression analysis, and into the machine learning and deep learning principles that govern the classification approaches developed in this thesis. The characterization of artificial neural network architectures, activation functions, feedforward topologies, and optimization strategies provides the conceptual scaffolding upon which the methodology described in the following chapter is constructed. Together, these elements constitute an integrated computational framework in which molecular data, statistical rigor, and deep learning converge to address the challenge of cancer subgroup classification, with particular emphasis on medulloblastoma, ultimately supporting the development of robust, reproducible, and biologically interpretable predictive models.

3 State of the Art

Contents

3.1 Deep Learning Methods and Applications on Molecular Data **71**
3.2 Deep Learning Techniques for Survival Prediction in Pancreatic Cancer **73**
3.2.1 Deep Learning Architecture: LSTM Networks 75
3.3 ANN-Based Molecular Classification of Medulloblastoma Subgroups **78**
3.3.1 Artificial Neural Network Architecture 81

3.1 Deep Learning Methods and Applications on Molecular Data

Within the ML landscape, Deep Learning (DL) methodology has attracted particular attention as a high-performing alternative for big data analytics. Its rapid advancement has been driven by the development of more powerful GPU hardware, automatic differentiation software, and novel architecture based on the ReLU activation function, which mitigated the vanishing gradient problem first identified in the context of image classification with deep convolutional neural networks[1]. The capacity of DL toe extract abstract representations directly from high-dimensional data has enabled successful applications in transcriptomics, including the prediction of pharmacological properties of drugs and drug repurposing[2]. Unlike classical shallow learning approaches, DL methods are composed of multiple processing layers capable of handling high levels of abstraction and do not necessarily require a prior feature extraction step; instead, their multilayer structure enables the automatic derivation of sophisticated representations directly from raw input data during training[3].

Among the DL architectures that have demonstrated success in transcriptomic applications, feedforward neural networks (FNN)have been applied to tumor type classification using gene expression data[4]

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012; and Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015

[2] Alexander Aliper, Sergey Plis, Artem Artemov, Alvaro Ulloa, Polina Mamoshina, and Alex Zhavoronkov. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Molecular pharmaceutics*, 13(7): 2524–2530, 2016

[3] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015

[4] Boyu Lyu and Anamul Haque. Deep learning based tumor type classification using gene expression data. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 89–96, 2018

and to the analysis of RNA-Seq count data, which shares fundamental structural properties with microarray expression matrices from a bioinformatics processing perspective[5]. Convolutional neural networks (CNN) have been employed through structured transformations of omic data[6], while graph convolutional networks (GCN) have been used to infer cell-to-cell interactions from expression profiles[7]. A representative comparative example is the work of Wang and colleagues[8], who evaluated deep neural networks against Random Forest (RF) and Support Vector Machines (SVM) for predicting chemically induced liver injuries from whole-genome DNA microarray data, demonstrating superior generalization capability of DNN over classical methods across multiple evaluation metrics.

The application of Machine Learning (ML) and Deep Learning (DL) techniques in molecular classification of neoplasms has experienced exponential growth over the past two decades, driven by the increasing availability of high-dimensional multiomics data and the development of progressively more sophisticated computational algorithms. In the context of precision medicine, the identification of minimal gene signatures that enable molecular classification and clinical outcome prediction represents a methodological challenge of paramount importance, particularly in diseases characterized by high molecular heterogeneity.Within the context of the present work there are two recent studies in which most of the methodological framework was based that addresses these challenges through distinct yet complementary computational approaches: the first published by Salgado and colleagues in 2024[9], developing a DL algorithm based on Long Short-Term Memory (LSTM) networks to predict survival in pancreatic cancer through multiomics data integration; the second, published by Reveles-Espinoza and colleagues in 2025[10], implemented feedforward Artificial Neural Network (ANNs) to classify four molecular subgroups of medulloblastoma using gene expression microarray data. Both studies share the fundamental objective of identifying molecular biomarkers with diagnostic and prognostic potential through dimensionality reduction of massive gene sets to biologically meaningful minimal signatures, achieving accuracies exceeding 96% in their respective domains.

A common application of gene expression profiling in clinical and translational contexts is the development of classification models capable of assigning new samples to predefined phenotypic or genotypic groups. The construction of such classifiers requires two fundamental steps: the selection of a feature set, a subset of genes whose expression patterns distinguish between biological classes, and the fitting of model parameters through a training process that enables accurate sample classification. The selection of informative features

[5] Daniel Urda, Julio Montes-Torres, Fernando Moreno, Leonardo Franco, and José M Jerez. Deep learning to analyze rna-seq gene expression data. In *International work-conference on artificial neural networks*, pages 50–59. Springer, 2017

[6] Shiyong Ma and Zhen Zhang. Omicsmapnet: Transforming omics data to take advantage of deep convolutional neural network for discovery. *arXiv preprint arXiv:1804.05283*, 2018

[7] Ye Yuan and Ziv Bar-Joseph. Gcng: Graph convolutional networks for inferring cell-cell interactions. *bioRxiv*, pages 2019–12, 2019

[8] Hao Wang, Ruifeng Liu, Patric Schyman, and Anders Wallqvist. Deep neural network models for predicting chemically induced liver toxicity endpoints from transcriptomic responses. *Frontiers in pharmacology*, 10:42, 2019

[9] Iván Salgado, Ernesto Prado Montes de Oca, Isaac Chairez, Luis Figueroa-Yáñez, Alejandro Pereira-Santana, Andrés Rivera Chávez, Jesús Bernardino Velázquez-Fernandez, Teresa Alvarado Parra, and Adriana Vallejo. Deep learning techniques to characterize the rps28p7 pseudogene and the metazoa-srp gene as drug potential targets in pancreatic cancer patients. *Biomedicines*, 12(02), 2024

[10] Alicia Reveles-Espinoza, Ulises Villela, Edgar Hernandez-Martinez, Isaac Chairez, Sergio Juárez-Méndez, J Casanova-Moreno, Ma del Pilar Eguía-Aguilar, Luis Figueroa-Yáñez, Adriana Vallejo-Cardona, and Iván Salgado. Machine learning identification of molecular targets for medulloblastoma subgroups using microarray gene fingerprint analysis. *Computational and Structural Biotechnology Journal*, 2025

and the subsequent training, validation, and deployment of classifiers necessarily involves a combination of statistical and ML methods, several of which were examined in the previous chapter[11]. The transition from conventional statistical approaches to supervised ML methods in cancer research has been gradual yet transformative. Early attempts at molecular classification relied primarily on hierarchical clustering and principal component analysis (PCA)[12], methods that, while useful for exploratory analysis, lacked the predictive power necessary for clinical decision support.

The application of DL to cancer genomics has since enabled breakthrough studies in tumor classification, biomarker discovery, and survival prediction, while the tension between predictive performance and biological interpretability remains a central consideration in precision medicine pipelines, challenges directly addressed by the studies that forme the methodological foundation of the present work.

[11] Marieke L. Kuijjer, Joseph N. Paulson, and John Quackenbush. Expression analysis. In Andreas D. Baxevanis, Gary D. Bader, and David S. Wishart, editors, *Bioinformatics*, chapter 10. John Wiley & Sons, 4th edition, 2020

[12] Rosa Angélica Castillo-Rodríguez, Víctor Manuel Dávila-Borja, and Sergio Juárez-Méndez. Data mining of pediatric medulloblastoma microarray expression reveals a novel potential subdivision of the group 4 molecular subgroup. *Oncology letters*, 15(5):6241–6250, 2018

3.2 *Deep Learning Techniques for Survival Prediction in Pancreatic Cancer*

Pancreatic cancer (PaCa) represents one of the most lethal malignancies, with a 5 year survival rate below 9% that has remained largely unchanged over the past two decades even in developed nations. The marked heterogeneity in survival times, ranging from 3-9 months for metastatic disease to 20-24 months for resectable tumors, suggests underlying molecular differences that could be leveraged for patient stratification and personalized treatment approaches. The study by Salgado and colleagues[13] addressed a fundamental clinical question: what molecular features distinguish pancreatic cancer patients who die early from those with prolonged survival. The primary objective was to develop a DL algorithm capable of estimating functional relationships between multiomics data (gene copy number, gene expression, and proteome profiles) and vital status (deceased versus alive) in pancreatic cancer patient cohorts from the Genomic Data Commons (GDC) portal[14]. Unlike previous single-omic approaches, this study explicitly aimed to leverage information provided by DNA-level, RNA-level, and protein-level measurements to improve predictive accuracy and identify novel therapeutic targets. The data used for this study was obtained mainly from The Cancer Genome Atlas (TCGA) program[15], specifically from three projects: TCGA-PAAD, CPTAC-3, and HCC, with 1,666 files encompassing 360 samples with comprehensive multi-omics profiling, running on R to retrieve samples corresponding to pancreatic cancer cases with five data modalities: DNA sequencing (DNA-Seq), RNA sequencing (RNA-Seq), miRNA sequencing (miRNA-Seq), DNA methylation, and protein expression data.

[13] Iván Salgado, Ernesto Prado Montes de Oca, Isaac Chairez, Luis Figueroa-Yáñez, Alejandro Pereira-Santana, Andrés Rivera Chávez, Jesús Bernardino Velázquez-Fernandez, Teresa Alvarado Parra, and Adriana Vallejo. Deep learning techniques to characterize the rps28p7 pseudogene and the metazoa-srp gene as drug potential targets in pancreatic cancer patients. *Biomedicines*, 12(02), 2024

[14] GDC Data Portal. Genomic data commons data portal, 2025. URL https://portal.gdc.cancer.gov/

[15] NHGRI. The cancer genome atlas program, 2025. URL https://www.genome.gov/Funded-Programs-Projects/Cancer-Genome-Atlas

Given that the relative importance of information contained in each data modality had not been previously established for this specific classification task, a preliminary screening process identified nine specific components as input features for the neural network, corresponding to the main data categories retrieved from the GDC portal, as summarized in Table 3.2. While the full pipeline incorporated five data modalities (multiomic data), the genomic and transcriptomic features, specifically the DNA-Seq copy number metrics and the RNA-Seq upper-quartile FPKM normalization strategy, constitute the primary methodological reference for the present work, given that gene expression profiling via RNA-Seq represents the foundational data type which can be translated to microarray data for the classification pipeline developed herein.

Table 3.1: Data types and features retrieved from the GDC repository [Salgado et al., 2024].

Data type	**Description**
ASCAT DNA-Seq	• Weighted median of strand copy numbers. • Greater strand copy number of the two DNA strands (copy number segment files only). • Smaller strand copy number of the two DNA strands (copy number segment files only).
RNA-Seq	• Upper-quartile FPKM (FPKM-UQ): a modified FPKM in which the protein-coding gene at the 75th percentile substitutes for the sequencing quantity.
miRNA sequences	• Read count and normalized count in reads-per-million. • Isoform information: coordinates and region type within the full miRNA transcript.
Methylation	• Ratio between methylated array intensity and total array intensity.
Peptide/protein counts	• Unique ID for the antigen-binding target site. • Relative protein expression levels via interpolation of each dilution curve to the slide supercurve (antibody).

The nine identified components represent the complete input layer of the ANN, intended to model the relationship between the selected multiomics data and the vital status of pancreatic cancer patients. No normalization was applied to the input data; instead, files were processed through an automated import routine that detected the class of information in each file and associated it with the corresponding input node in the DL architecture. To address missing values, a three stage preprocessing strategy was implemented: in the first stage, biological features (genes, miRNAs, and analogous elements) were removed if zero values appeared in more than 20% of patients, and incomplete samples were eliminated if missing values exceeded 20% of features; in the second stage, the input package in R was used to fill

previously missing values; in the third stage, input features with zero values across all samples were removed. This preprocessing pipeline yielded a curated dataset suitable for DL analysis [Salgado et al., 2024].

3.2.1 *Deep Learning Architecture: LSTM Networks*

The study employed Long Short-Term Memory (LSTM) networks as the core component of the DL approach, justified by their capacity to extract complex relationships between gene expression, protein distribution, miRNA, and methylation profiles, data modalities whose molecular layers exhibit complex interdependencies, enabling the identification of survival status in pancreatic cancer patients, as illustrated in Figure 3.1.

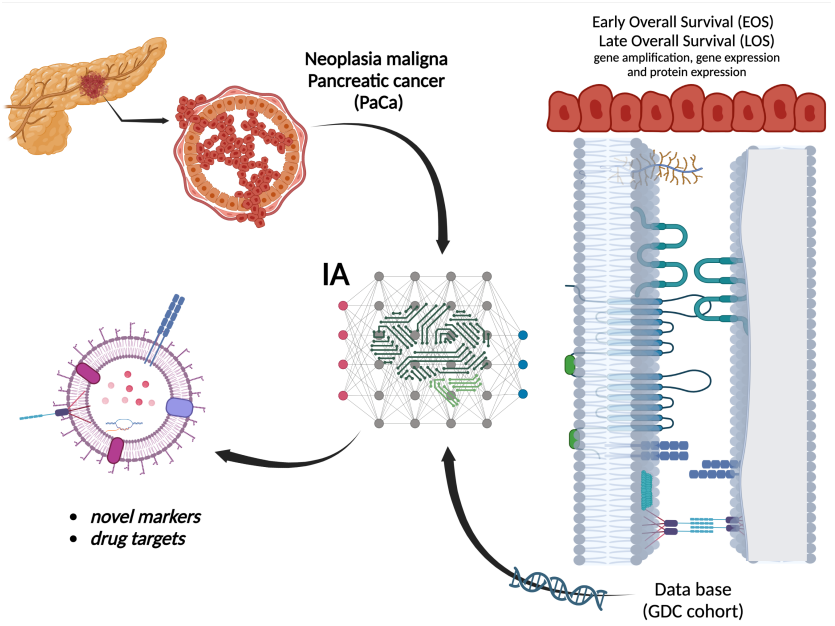

Figure 3.1: Procedure following a general methodology towards identifying the survival status in patients with pancreatic cancer based on DL focused on detecting possible markers or therapeutic targets [Salgado et al., 2024].

The selected LSTM topology operated as a classifier comprising five sequential components: a LSTM layer, a dropout layer, a fully connected feedforward layer, a softmax layer, and a classification output, its general topology can be observed in Figure 3.2. The LSTM contained 128 hidden units, determined through progressive structural adaptation using classification accuracy as a stoppage criterion. Sigmoidal activation functions were distributed across the LSTM inputs according to a defined partition scheme, with zero-padding applied to homogenize input vectors across omics modalities of varying dimensionality. A switching on/off layer between the LSTM and dropout sections enabled evaluation of the relative contribution of each omics subset to the classification outcome. The dropout layer applied a probability of 0.5 to mitigate overfitting, and the feedforward

section comprised two hidden layers with 96 and 76 activation functions respectively, culminating in a two-output softmax layer corresponding to vital status [Salgado et al., 2024], as summarized in Table 3.2.

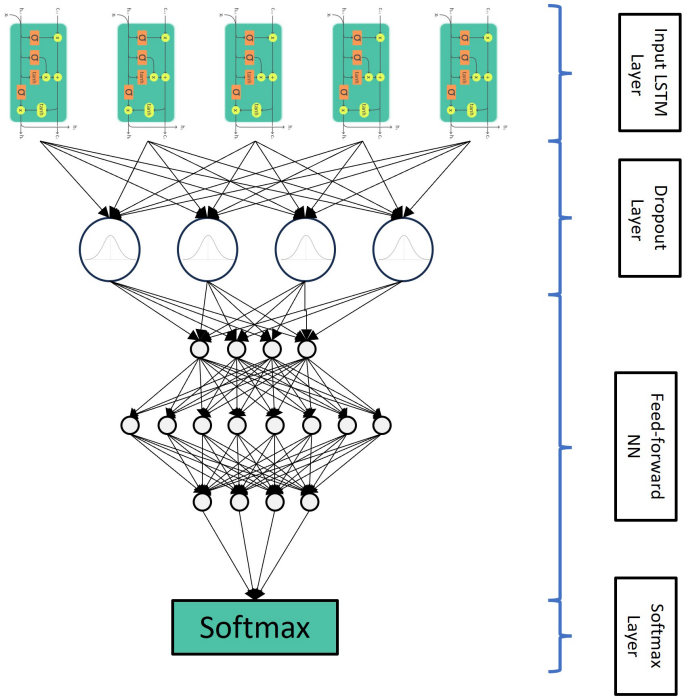

Figure 3.2: LSTM general topology used in the design of classification status across a multiomics approach [Salgado et al., 2024].

Layer	Key Parameters	Function
LSTM	128 hidden units, sigmoidal activation	Captures long- and short-term dependencies across multi-omics input modalities
Dropout	Probability = 0.5	Randomly deactivates neurons during training to mitigate overfitting
Fully connected feed-forward	2 hidden layers: 96 and 76 units	Models complex relationships between dropped-out representations and target class
Softmax	2 output nodes	Produces probability distribution over binary vital status classification
Classification output	—	Final predicted vital status (deceased vs. alive)

Table 3.2: LSTM-based classifier architecture employed by Salgado et al. [Salgado et al., 2024].

The LSTM network was trained under a supervised learning scheme, with known vital status labels enabling a transfer learning strategy

in which the pre-trained model structure served as a starting point, accelerating weight adjustment across the preprocessed architecture. Training incorporated 5-fold cross-validation (k=5), with the dataset partitioned randomly into five subsets; across five rounds, four folds served for training and one for validation, yielding five independently trained models whose performance metrics were averaged to obtain final estimates. The asymmetric lengths of the the multiomics input vectors necessitated the construction of asymmetric inputs sequences across folds. Each complete training cycle required approximately 78 hours on an Intel i7 10th generation processor with 8 GB RAM, reflecting the computational complexity of LSTM networks processing high-dimensional multiomics data. To assess model robustness, a sensitivity analysis was performed over all network weights, computing the sensitivity absolute averages as the sum of the average temporal distribution of absolute values of partial derivatives across input-output pairs. A threshold criterion ($\varepsilon > 0.01$) was applied such that the classification task was restarted until all sensitivity values exceeded this threshold, requiring twelve sequential runs before convergence. Additionally, the feedforward section of this architecture operated as an autoencoder, enabling extraction of genes, RNAs, proteins, and other compounds with strongly propagated influence on the reduced-dimensional internal encoding, thereby providing a bias for evaluating the relative importance of each input component [Salgado et al., 2024].

Model performance was evaluated using three standard classification metrics, formally defined in Table 3.3. To assess the stability of these estimates, each input subset was partitioned according yo the 5-fold cross-validation scheme, with the LSTM topology combining individual fold sets across omics inputs to yield an unbiased performance evaluation. Input bias was further controlled through early stopping and the sensitivity analysis threshold criterion described previously.

Metric	**Formula**	**Interpretation**
Sensitivity	$\frac{TP}{TP+FN}$	Proportion of patients with pancreatic cancer correctly identified as deceased
Specificity	$\frac{TN}{TN+FP}$	Proportion of alive patients correctly classified as such
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Overall proportion of samples correctly classified

Table 3.3: Performance metrics used to evaluate the LSTM-based classifier [Salgado et al., 2024]. TP: true positives; TN: true negatives; FP: false positives; FN: false negatives.

From multiple projects in the GDC database including TCGA-PAAD and CPTAC-3, 1,666 files were retrieved as of January 2023, from which

sample subsets were selected for input into the LSTM architecture. Table 3.4 presents the accuracy and computational cost obtained for each omics subset independently and in combination.

Subset	Survived	Deceased	Accuracy	Processing Time (h)
DNA-Seq	5,064	3,199	0.92	56
RNA-Seq	1,818	1,165	0.92	49
Methylation	774	504	0.81	51
miRNA-Seq	848	662	0.88	67
Protein	70	50	0.80	43
All subsets	—	—	0.96	78

Table 3.4: Classification accuracy and processing time per omics subset for the proposed LSTM classifier [Salgado et al., 2024].

Individual omics subsets yielded accuracies ranging from 0.80 (protein) to 0.92 (DNA-Seq and RNA-Seq), while the integration of all subsets achieved the highest classification accuracy of 0.96, confirming the added predictive value of the multiomics approach. The confusion matrix obtained for all subsets configuration is presented in Table 3.5. These results confirm that the proposed network effectively predicts the relationship between actual and predicted vital status, with the all subset configuration demonstrating robust discriminative capacity between deceased and alive pancreatic cancer patients.

While the LSTM structure was determined through a standard partition scheme, the authors acknowledge that adaptive parameter adjustment methods could further simplify structural selection and improve classification performance. A key limitation of the study was the absence of quantified time to death from diagnosis; notably deceased patients presented temporal disparity that warrants considerations in further survival analyses [Salgado et al., 2024].

	Predicted Positive	Predicted Negative
Actual Positive	4,061	2,030
Actual Negative	128	3,071

Table 3.5: Confusion matrix for the all-subsets LSTM classification configuration [Salgado et al., 2024].

3.3 *ANN-Based Molecular Classification of Medulloblastoma Subgroups*

The study by [Reveles-Espinoza et al., 2025] introduced a structured methodology to identify molecular targets capable of classifying the four MB subgroups (WNT, SHH, Group 3, and Group 4) using an Artificial Neural Network (ANN) model trained on microarray gene expression data. The primary objective was to determine minimal gene combinations for each subgroup that could achieve high classification

accuracy while maintaining biological interpretability. Unlike the multiomics LSTM approach employed in pancreatic cancer, this study focused exclusively on transcriptomic profiling through microarray technology, prioritizing methodological parsimony and experimental validation by identifying key markers on novel PCR experimental data, the whole framework is illustrated in Figure 3.3.

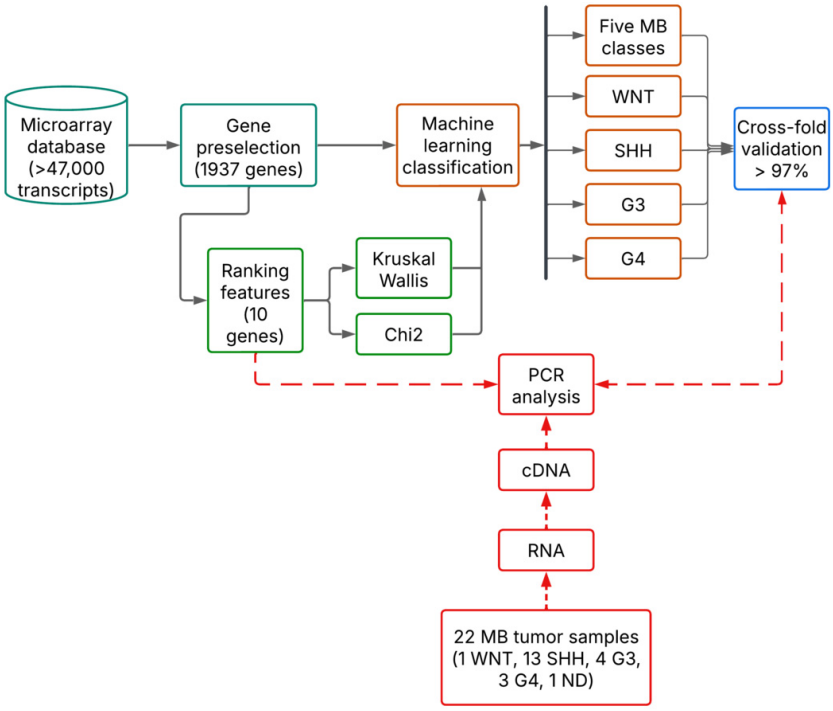

Figure 3.3: Diagram of the general methodology to identify key genes in MB subgroup classification [Reveles-Espinoza et al., 2025].

Data acquisition followed a multi cohort strategy, with the Affymetrix microarray U133 Plus 2.0 Array as the main platform (GPL570) retrieved from the Gene Expression Omnibus (GEO) database, comprising MB tumor samples Kool et al. [2008] and healthy control cerebellar tissue (CT) across five independent studies. Two datasets contributed healthy cerebellar tissue (CT) as reference controls: GSE4036, comprising cerebellar tissue from schizophrenia patients and matched healthy controls, and GSE44971 Lambert et al. [2013], which provided normal cerebellum samples from a pilocytic astrocytoma methylation and transcriptome profiling study. The remaining three datasets provided labeled MB tumor cohorts: GSE10327 Pöschl et al. [2014] profiled 62 MB tumors identifying five distinct molecular subtypes through unsupervised hierarchical clustering; GSE37418 Robinson et al. [2012] contributed whole-genome sequencing and expression profiles from 93 MB tumors with subgroup specific annotations across all four subtypes; and GSE49243 Lambert et al. [2013] enriched the SHH subgroup

representation through transcriptomic profiling of SHH-MB across pediatric and adult age groups.

Each dataset contained comprehensive sample metadata including tissue type, disease status, RNA purification method, RNA quality metrics, and microarray protocol specifications. A preliminary differential analysis based on prior data mining of pediatric medulloblastoma microarray datasets was conducted to obtain the gene expression profile associated with MB, with the CT group serving as the genomic reference for comparative analysis against the four subgroups. Differential gene expression was calculated using Partek Genomics Suite v7.18, applying stringent statistical thresholds of $p < 0.005$ and absolute fold-change > 5. These conservative parameters were selected to prioritize high confidence differentially expressed genes and minimize false positive discoveries, yielding a curated fingerprint of 1,937 genes identified as a potential molecular signature for MB subgroup classification[16].

[16] Rosa Angélica Castillo-Rodríguez, Víctor Manuel Dávila-Borja, and Sergio Juárez-Méndez. Data mining of pediatric medulloblastoma microarray expression reveals a novel potential subdivision of the group 4 molecular subgroup. *Oncology letters*, 15(5):6241–6250, 2018

The final integrated dataset comprised 124 samples distributed across the four MB subgroups and the CT reference group, as summarized in Table 3.6. The notable class imbalance, with G4 comprising 34% of samples while WNT and CT each represented only 14%, reflects an inherent characteristic of MB epidemiology rather than a sampling artifact, and was accounted for during model evaluation and interpretation [Reveles-Espinoza et al., 2025]. Beyond overall multiclass classification of the four MB subgroups alongside the CT reference, the pipeline was designed to perform a binary one-versus-all classification across each individual subgroup and to rank features by their discriminative contribution to subgroup classification. Feature ranking was implemented through Kruskal-Wallis and χ^2 tests, further reducing the gene set to a minimal signature of 10 genes, which were subsequently validated through an independent classification task. Experimental validation was performed on 22 MB tumor samples through RNA extraction, cDNA synthesis, and digital PCR (dPCR) analysis with subgroup-specific primers, complemented by a literature review of the functional and oncological relevance of the selected genes.

Subgroup	**Number of Samples**
Control Tissue (CT)	17
WNT	17
SHH	28
Group 3 (G3)	20
Group 4 (G4)	42
Total	**124**

Table 3.6: Sample distribution across MB subgroups and control

3.3.1 *Artificial Neural Network Architecture*

The study employed feedforward Artificial Neural Networks (ANNs) to differentiate among MB subgroups. Unlike the LSTM approach used for pancreatic cancer, this architecture prioritized computational efficiency and biological interpretability. The classification strategy consisted of two stages: a five-category multiclass classifier and a set of binary one-versus-all subgroup specific classifiers, both operating on the 1,937 genes which were differentially expressed fingerprint as input. The first stage implemented a five category classifier evaluating the feasibility of simultaneously differentiating all four MB subgroups alongside the CT reference group. The second stage established four independent binary classifiers, each distinguishing one MB subgroup from all remaining cases including CT, enabling the isolation of subgroup-specific molecular signatures, improved handling of class imbalance through focused binary comparisons, and enhanced interpretability of feature importance per subgroup. Each classifier was optimized independently with subgroup-specific hyperparameters, as summarized in Table 3.7.

Component	**WNT**	**SHH**	**G3**	**G4**	**5-Class**
Input layer (genes)	1937	1937	1937	1937	1937
Hidden layer 1 (neurons)	600	750	600	700	700
Hidden layer 2 (neurons)	600	700	600	700	700
Hidden layer 3 (neurons)	600	750	600	700	700
Output layer (classes)	2	2	2	2	5
Activation function	ReLU	ReLU	Tanh	ReLU	ReLU

Table 3.7: ANN configuration for binary classifiers [Reveles-Espinoza et al., 2025].

Activation functions and layer sizes were determined through progressive structural adaptation, incrementally adjusting network capacity until classification accuracy plateaued. As discussed in Chapter 2, the Tanh activation function was selected specifically for the G3 classifier given the elevated molecular homogeneity between G3 and G4 subgroups. All configurations were implemented in MATLAB, ensuring tractability on standard hardware.

The ANN training protocol incorporated 10-fold cross-validation (k=10), dividing the dataset into ten subsets with each fold serving once as validation while the remaining nine comprised the training set. This approach follows the cross-validation paradigm described by Kuijjer and colleagues[17], whereby a single dataset in divided into training and test sets n times repeating the training and evaluation process across each fold. As noted by the authors, there is no absolute standard for the number of folds or dataset division, however a common split of 90/10 between training and test sets with at least 10 folds is recommended when working with relatively small datasets, as is

[17] Marieke L. Kuijjer, Joseph N. Paulson, and John Quackenbush. Expression analysis. In Andreas D. Baxevanis, Gary D. Bader, and David S. Wishart, editors, *Bioinformatics*, chapter 10. John Wiley & Sons, 4th edition, 2020

frequently the case in rare disease research or cancer. Given the limited sample size inherent to the MB cohort, a characteristic of rare pediatric tumor samples, the selection of k=10 was methodologically consistent with this recommendation, ensuring robust performance estimation while minimizing dependence on any particular data partition.

Two distinct training procedures were evaluated: absolute expression, using raw normalized gene expression values directly from microarray data, and differential expression, using fold-change values relative to the CT reference group. Each complete training cycle required approximately 5 minutes on a standard Intel i7 processor with 8 GB RAM, reflecting the computational advantages of feedforward architectures over sequential models such as LSTM for high-dimensional non-sequential data. Model convergence was monitored through validation accuracy with early stopping applied when performance plateaued or began degrading.

The five category classifier achieved an average accuracy of 96% under absolute expression training with subgroup-specific performance ranging from 90% for G3, consistent with its known molecular heterogeneity, to 98-99% for SHH. WNT and CT both exceeded 97% accuracy with minimal classifications. Under differential expression training, overall accuracy reached 97.2%, with only 3 misclassifications out of 107 samples, two involving confusion between G3 and G4 or SHH, and one WNT sample misclassified as G4, demonstrating robustness despite challenging molecular overlap between subgroups.

The four binary one-versus-all classifiers achieved consistently high validation accuracy across both training modalities, as summarized in Table 3.8 .

Subgroup	Absolute Expression	Differential Expression
WNT vs. All	>98%	>97%
SHH vs. All	99%	98%
G3 vs. All	98%	93.5%
G4 vs. All	98%	95%

Table 3.8: Binary classifier validation accuracy

The notable accuracy decrease for G3 under differential expression (93.5%) likely reflects the biological complexity of this subgroup and potential loss of discriminative features when expression is normalized against CT. ROC curve analysis confirmed excellent discriminative ability across all subgroups; the ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity) across varying classification thresholds, with the Area Under the Curve (AUC) providing an aggregated performance measure across all possible thresholds on a scale from 0 to 1[18]. AUC values ranging from 0.9874 to 0.9985 were obtained across all subgroup classifiers, with Model Operating Points aligned closely with the ROC curve, confirming robust

[18] MathWorks. Performance curves. https://la.mathworks.com/help/stats/performance-curves.html, 2024b. Accessed: 2025

threshold selection and consistent discriminative performance [Reveles-Espinoza et al., 2025].

To identify a minimal gene signature retaining maximum discriminative power, feature ranking was performed using two non-parametric statistical methods. The Kruskal-Wallis test evaluated statistically significant expression differences across multiple independent groups without assuming normal distribution, defined as:

$$H = \frac{12}{n(n+1)} \sum_i n_i^{-1} \left(R_i - \overline{R}\right)^2 \tag{3.1}$$

where n is the total number of observations, R_i is the sum of ranks for group i, and $\overline{R}$ is the mean rank. The χ^2 test evaluated associations between gene expression categories and subgroup classification, defined as:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} E_{ij}^{-1} \left(O_{ij} - E_{ij}\right)^2 \tag{3.2}$$

where O_{ij} is the observed frequency and $E_{ij} = N^{-1} R_i C_j$ is the expected frequency. Both methods converged on similar sets of highly discriminative genes, with ANN classifiers retained on the top 10 identified genes retaining robust performance, with accuracy above 90% under absolute expression and approximately 80% under differential expression, demonstrating that a minimal gene subset contains sufficient information for reliable subgroup classification.

The study acknowledged several limitations inherent to its design: the restricted sample size of 124 cases with uneven subgroup distribution, particularly for G3 (n=20) and WNT (n=17), constrains generalization to larger cohorts; G3-G4 discrimination remained the most challenging classification task (90-93% accuracy), reflecting the known molecular overlap between these subgroups; experimental validation was limited to 22 fresh tumor samples with particularly small representation for G3 (n=4) and G4 (n=3); and the exclusive focus on transcriptomic data, while methodologically parsimonious, may omit complementary information available through multiomic profiling. Notwithstanding these limitations, the study established a reproducible and computationally efficient pipeline for MB molecular classification, demonstrating that feedforward ANNs can achieve accuracy comparable to more complex DL architectures while identifying minimal gene signatures of 10-16 genes retaining high discriminative power through the integration of statistical feature selection and experimental dPCR validation. The authors proposed as future directions the expansion of validation to larger independent cohorts, the incorporation of additional omic data layers such as methylation and proteomics, and the exploration of clinical utility

of the identified gene signatures for risk stratification and therapeutic decision making [Reveles-Espinoza et al., 2025].

As highlighted by [Kuijjer et al., 2020], a common application of gene expression profiling in clinical and translational context is the development of classification models capable of assigning new samples to predefined phenotypic or genotypic groups. Constructing a robust and reproducible classifier requires, first, the selection of a discriminative feature set, in the present context, differentially expressed genes, and second, the fitting of model parameters through a training process that enables accurate sample classification. While numerous statistical and ML methods are available for this purpose, no clear consensus exist in the field regarding optimal methodology, reinforcing the value of the comparative framework examined across the two anchor studies reviewed in this chapter. Key criteria for successful classifier development include an adequate balance across classification groups, the separation of training and independent test population, and, when working with small or rare disease cohorts, the application of cross-validation in which both feature selection and algorithm training are re-performed at each fold using only the training subset before evaluation on the independent test subset. Both the multiomics LSTM pipeline and the feedforward ANN approach adhere to these foundational principles, collectively establishing the methodological basis upon which the present work is developed.

4 Methodology

Contents

4.1 Experimental Design and Data Collection . . . **85**
4.1.1 Experimental Design Considerations . 87
4.1.2 Data Mining and Repository Search . 89
4.2 Metadata Preprocessing and Sample Identification **90**
4.3 Gene Expression Data Processing and Differential Expression Analysis **95**
4.3.1 Microarray Data Acquisition and Preprocessing 97
4.3.2 Quality Control and Normalization . 98
4.3.3 Differential Expression Analysis . . . 101
4.4 Deep Learning Classification **109**
4.4.1 Network Architectures 110
4.4.2 Layer Construction and Regularization 110
4.4.3 Training Configuration 111
4.4.4 Cross-Validation Protocol 111
4.4.5 Evaluation Metrics and Outputs . . . 112

4.1 Experimental Design and Data Collection

This study develops an integrated computational pipeline that combines bioinformatics preprocessing, statistical transcriptomic analysis, and deep learning classification to address the central research question of whether a minimal, clinically actionable biomarker set can discriminate among the four molecular subgroups of medulloblastoma (WNT, SHH, Group 3, and Group 4) and cerebellar control tissue. The methodological framework is grounded on the foundational principles established by Reveles-Espinoza et al. [2025] and Salgado et al. [2024], extending their respective approaches toward a unified diagnostic strategy that operates on a larger single-platform cohort, incorporates a homogeneous biological reference, and preserves full analytical traceability from raw microarray intensities through classifier output. Although the primary disease model is medulloblastoma, the pipeline

is explicitly architected as a transferable protocol suitable for extension to other transcriptomic classification problems in pediatric and adult oncology, particularly in resource-limited research contexts where gold-standard molecular techniques such as DNA methylation profiling and RNA sequencing remain inaccessible.

The pipeline is organized into four sequential yet interdependent stages, each designed to feed structured, quality-controlled outputs into the next. The first stage comprises experimental design and data acquisition, including statistical power analysis for multi-class classification, systematic mining of the Gene Expression Omnibus (GEO) repository, and selection of the GSE85217 medulloblastoma cohort [Cavalli et al., 2017] together with anatomically matched cerebellar controls drawn from GSE45878 [National Center for Biotechnology Information, 2024], yielding a unified dataset of 794 samples profiled on compatible Affymetrix Human Gene 1.1 ST Array platforms. The second stage addresses metadata preprocessing and sample harmonization, resolving inconsistencies across annotation versions (GPL22286 and GPL16977), standardizing age and gender encodings, and integrating cancer and control cohorts into a single curated metadata table suitable for joint downstream analysis. The third stage implements the bioinformatics core of the pipeline: joint processing of all raw CEL files under a common annotation framework, RMA normalization, probe-to-Ensembl harmonization across the 20,372 genes shared by both platforms, quality control through distributional and dimensionality-reduction diagnostics, and differential expression analysis via empirical Bayes moderated linear modeling with `limma`. The fourth stage employs the resulting differentially expressed genes as the input feature space for a hierarchical feedforward neural network classifier, comprising one five-class model and four one-versus-rest binary models trained under stratified ten-fold cross-validation.

The rationale underlying this modular architecture is that each stage constitutes both a self-contained analytical block and a transfer point at which intermediate outputs are validated before propagation. Raw microarray intensities are converted into quality-controlled expression matrices; expression matrices yield a statistically principled set of differentially expressed genes; and these genes constitute a compact, biologically interpretable feature space that feeds the deep learning stage. By enforcing this layered structure, the pipeline avoids conflating preprocessing artifacts with biological signal, isolates the sources of variance at each step, and preserves full reproducibility through script-level documentation of every transformation. Figure 4.1 summarizes the complete workflow, highlighting the logical dependencies between stages and indicating where the principal analytical decisions are made. Each subsequent section of this chapter develops one stage in full

methodological detail.

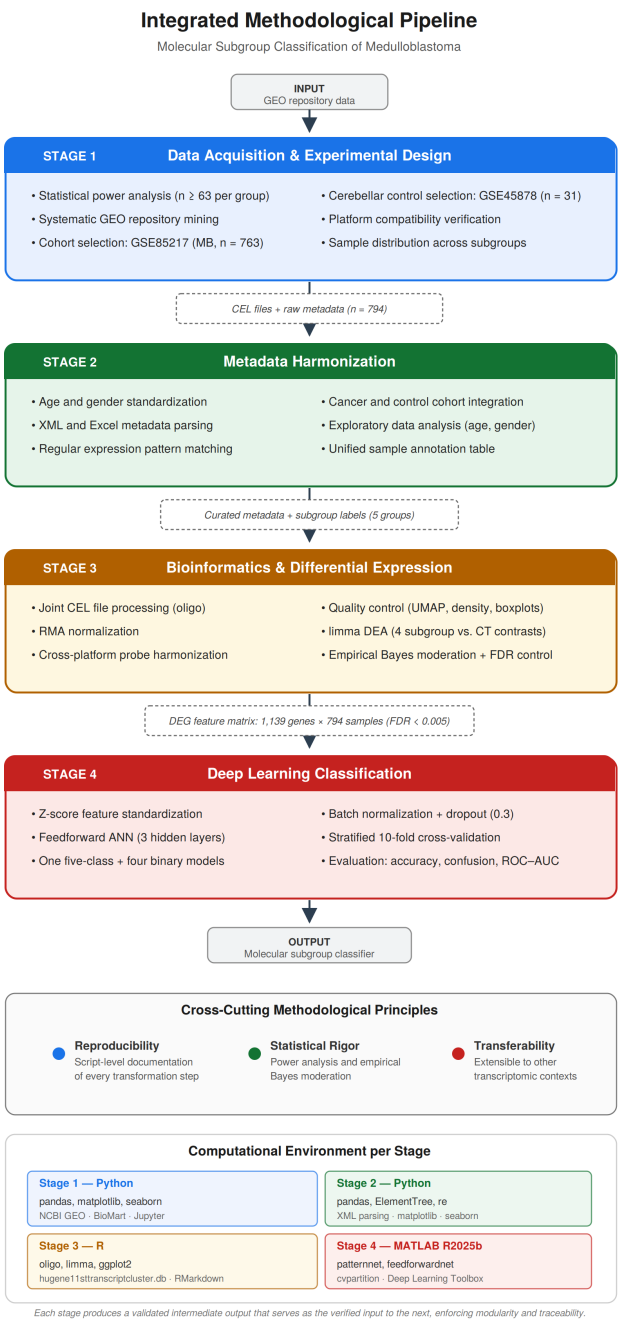

Figure 4.1: Overview of the integrated methodological pipeline. The four sequential stages — data acquisition and experimental design, metadata harmonization, bioinformatics preprocessing and differential expression analysis, and deep learning classification — are connected by well-defined intermediate outputs. Each stage is developed in full in the corresponding section of this chapter.

4.1.1 *Experimental Design Considerations*

The experimental design began with a comprehensive literature review to identify larger cohorts and maximize sample size, thereby addressing potential limitations of previous models that may have suffered

from insufficient statistical power. This approach also facilitated proper power calculations and enabled batch effect identification and correction, as it encompassed a single cohort instead of multiple dataset as reported by [Reveles-Espinoza et al., 2025]. Dataset discovery was conducted through systematic searches of the Gene Expression Omnibus (GEO) database, augmented by AI-assisted search tools for efficient cohort identification, mainly based on properly identify GEO datasets IDs, with subsequent human validation.

Power analysis for multi-class classification was conducted to ensure adequate sample representation across molecular subgroups. Given the four MB subgroups (WNT, SHH, G3, and G4) with historically documented prevalence rates of approximately 10%, 30%, 25%, and 35% respectively[1], a minimum required sample size per group was calculated to ensure robust differential gene expression detection and stable deep learning model training. Given the class imbalance inherent to MB subgroup distribution, this calculation was applied to the smallest subgroup (WNT, ∼10% prevalence) as the most conservative bounding estimate.

[1] Abhinav Dey, Anshu Malhotra, and Neha Garg. Medullo-blastoma; and National Cancer Institute (NCI). Tratamiento del meduloblastoma y otros tumores embrionarios del sistema nervioso central infantil, 2025. URL https://www.cancer.gov/espanol/tipos/cerebro/pro/tratamiento-embrionarios-snc-infantil-pdq

The sample size per group was estimated using the standard two-group detection formula[2]:

[2] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013

$$n = \frac{2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{d^2}$$

where n is the required sample size per group and the remaining parameters are defined in Table 4.1.

Parameter	Value	Definition
α	0.05	Significance level (Type I error rate)
$1-\beta$	0.80	Statistical power (probability of detecting true differences)
d	0.5	Effect size (Cohen's d; medium biological effect)
$z_{1-\alpha/2}$	1.96	Critical value for two-tailed test at $\alpha = 0.05$
$z_{1-\beta}$	0.84	Critical value for power $= 0.80$

Table 4.1: Power analysis parameters used for minimum sample size estimation.

Substituting these values:

$$n = \frac{2\left(1.96 + 0.84\right)^2}{(0.5)^2} = \frac{2(2.80)^2}{0.25} = \frac{15.68}{0.25} = 62.72 \approx 63$$

This establishes a minimum threshold of 63 samples per group to achieve adequate statistical power for detecting medium effect size differences in gene expression between molecular subgroups and for training robust classification models.

4.1.2 Data Mining and Repository Search

This comprehensive search identified one of the largest curated gene expression datasets encompassing cerebellar and medulloblastoma samples GSE124814[3]. This metadataset was originally compiled to improve cure rates and reduce treatment side effects by integrating 23 transcription datasets spanning 1,350 MB and 291 normal brain samples. The integration approach enabled removal of the majority of batch effects, which represent one of the most significant challenges when combining datasets from different microarray platforms due to inherent differences in data structures and distributions[4]. However, GSE124814 encompasses four different reading platforms, introducing residual batch effects that would require correction beyond the current scope of study. Consequently, the decision was made to extract the largest single-platform dataset embedded within this metacohort GSE85217[5], containing 763 primary medulloblastoma samples derived from RNA extraction and hybridization of Affymetrix Human Gene 1.1 ST Array microarrays (GPL22286, ENSG v19.0.0).

The sample distribution across molecular subgroups in GSE85217 is as follows: WNT ($n = 70$), SHH ($n = 223$), Group 3 ($n = 144$), and Group 4 ($n = 326$). All four subgroups exceeded the calculated minimum threshold of 63 samples per group as it can be seen in Figure 4.2, ensuring adequate statistical power for differential gene expression analysis, deep learning model training, and biomarker identification. Although a uniform distribution across subgroups would be theoretically preferable for balanced model training, as it was previously established to be a key criteria towards developing a classifier[6], downsampling to equalize group sizes would introduce selection bias and discard valuable information. Therefore, the entire cohort was retained to maximize statistical power, preserve the natural epidemiological distribution of MB subgroups [Dey et al., National Cancer Institute (NCI), 2025], and ensure that the classification pipeline generalizes to real-world clinical proportions.

Normal cerebellar tissue samples ($n = 31$) were incorporated as controls to provide a biological baseline for differential gene expression analysis. These samples were sourced from the GSE45878 dataset[7], selecting only those labeled as "Brain - Cerebellum" ($n = 16$) and "Brain - Cerebellar Hemisphere" ($n = 15$), yielding 31 anatomically relevant cohorts based on the retrieved metadata. Although this control group falls below the calculated threshold of 63 samples, this is methodologically acceptable because controls serves as a reference baseline rather than a primary classification category, and their purpose is to characterize normal cerebellar gene expression patterns rather than to achieve balanced representation in multi-class classification task.

[3] Holger Weishaupt, Patrik Johansson, Anders Sundström, Zelmina Lubovac-Pilav, Björn Olsson, Sven Nelander, and Fredrik J Swartling. Batch-normalization of cerebellar and medulloblastoma gene expression datasets utilizing empirically defined negative control genes. *Bioinformatics*, 35(18):3357–3364, 2019

[4] Steven M Foltz, Casey S Greene, and Jaclyn N Taroni. Cross-platform normalization enables machine learning model training on microarray and rna-seq data simultaneously. *Communications Biology*, 6(1):222, 2023

[5] Florence MG Cavalli, Marc Remke, Ladislav Rampasek, John Peacock, David JH Shih, Betty Luu, Livia Garzia, Jonathon Torchia, Carolina Nor, A Sorana Morrissy, et al. Intertumoral heterogeneity within medulloblastoma subgroups. *Cancer cell*, 31(6):737–754, 2017

[6] Marieke L. Kuijjer, Joseph N. Paulson, and John Quackenbush. Expression analysis. In Andreas D. Baxevanis, Gary D. Bader, and David S. Wishart, editors, *Bioinformatics*, chapter 10. John Wiley & Sons, 4th edition, 2020

[7] National Center for Biotechnology Information. GEO accession GSE85217, 2024. URL https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85217. Gene Expression Omnibus (GEO), NCBI

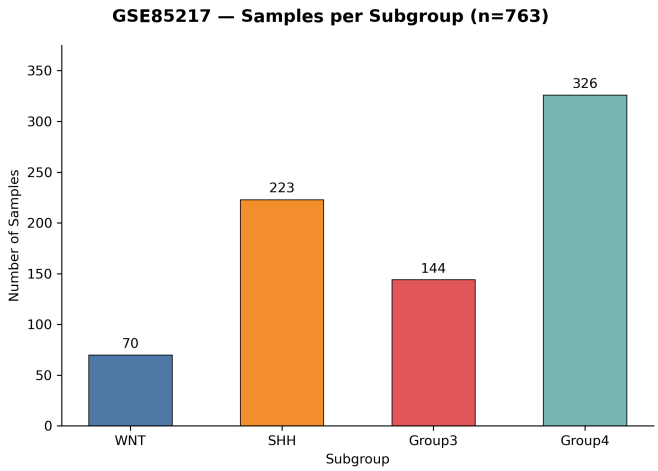

Figure 4.2: Distribution of medulloblastoma molecular subgroups in the GSE85217 dataset. All molecular subgroups exceed the minimum calculated sample size ($n = 63$) required for adequate statistical power (shown as dashed line).

A critical technical consideration was ensuring platform compatibility between cancer samples and control. GSE85217 employed GPL22286 (ENSG v19.0.0), while GSE45878 utilized GPL16977 (ENSG v14.1.0). Both platforms are derived from the same physical Affymetrix Human Gene 1.1 ST Array chip, differing only in the annotation file version used to map probes to genes via ENSEMBL database (Table 4.2). The distribution of samples across annotation versions is shown in Figure 4.3. To quantify the extent of identifiers overlap, a comparative analysis of gene identifiers was performed, revealing 20,372 common genes shared between platforms as illustrated in Figure 4.4, representing 94.13% of GPL22286 identifiers and 89.72% of GPL16977 identifiers. This high degree of overlap confirms that observed expression differences reflect true biological variation rather than technical artifacts, and that the majority of genes can be directly compared across cancer and control samples without loss of biological information. The annotation version discrepancy was resolved during preprocessing by harmonizing gene identifiers to a common reference, as described in the following section.

GPL ID	Platform	Anot.	Comentario
GPL22286	HuGene-1_1-st	ENSG v19.0.0	Most recent
GPL16977	HuGene-1_1-st	ENSG v14.1.0	Older, same chip technology

Table 4.2: GPL platform annotation comparison

4.2 *Metadata Preprocessing and Sample Identification*

Prior to gene expression analysis, comprehensive metadata preprocessing was conducted to ensure data quality and enable proper integration of cancer samples with anatomically appropriate controls. The GSE124814 metadataset served as the initial repository for identifying medulloblastoma samples, with metadata

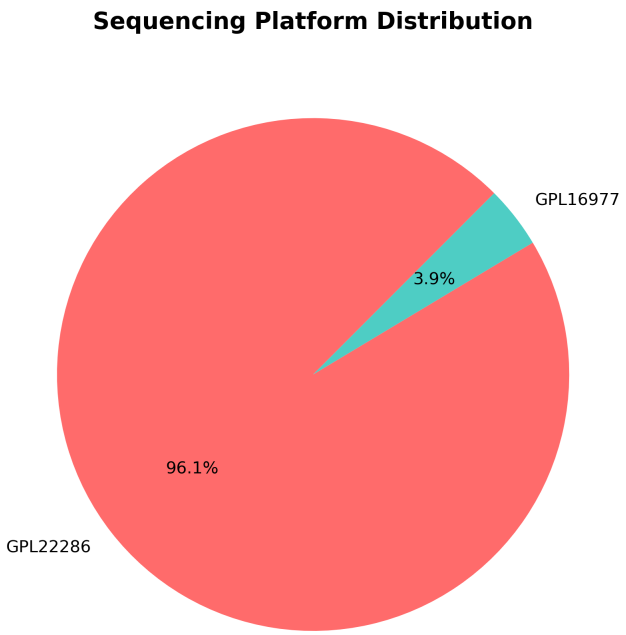

Figure 4.3: Proportional distribution of samples across annotation versions. GPL22286 (ENSG v19.0.0) accounts for the medulloblastoma samples ($n = 763$) and GPL16977 (ENSG v14.1.0) for the cerebellar controls ($n = 31$), both derived from the same Affymetrix Human Gene 1.1 ST Array chip.

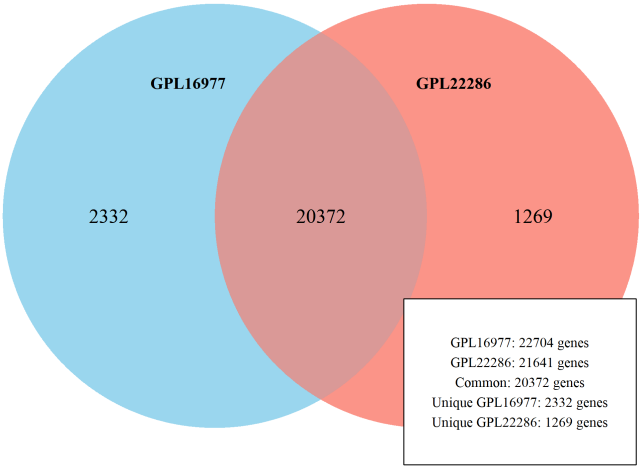

Figure 4.4: Venn diagram showing the overlap of gene identifiers between GPL22286 (ENSG v19.0.0, used for medulloblastoma samples) and GPL16977 (ENSG v14.1.0, used for cerebellar controls). Both annotation versions are derived from the same physical Affymetrix Human Gene 1.1 ST Array chip, with 20,372 genes (>89%) shared between platforms.

obtained from the Series Matrix File and the supplementary file GSE124814_sample_descriptions.xlsx (available at: https://ftp.ncbi.nlm.nih.gov/geo/series/GSE124nnn/GSE124814/suppl/, last modified 2019-01-08). Substantial data cleansing was required to

standardize metadata formatting and ensure compatibility across datasets prior to downstream analysis.

Several preprocessing steps were implemented to harmonize the GSE124814 metadata structure. Characteristic columns were systematically renamed to provide unambiguous identifiers for each data field. Gender labels, which exhibited inconsistent formatting (e.g., "M", "male", "F", "female", mixed case variations, and missing values), were standardized to three unified categories: Male, Female, and Unknown. Age normalization represented a particularly critical preprocessing challenge, as the metadataset contained age information in multiple incompatible formats including days, weeks, months, years, post-conceptional weeks (PCW), and various abbreviated notations (e.g., "6y3m", "19 PCW", "2.04 years"). Given that medulloblastoma exhibits strong age-dependent epidemiological patterns with peak incidence in pediatric populations as mentioned by [Dey et al., National Cancer Institute (NCI), 2025], precise age representation was essential. Days were selected as the fundamental unit for normalization, capturing clinically relevant distinctions while accommodating the wide age range present in the dataset. The normalization algorithm implemented pattern-matching regular expressions to convert diverse formats to a standardized scale using appropriate conversion factors (365.25 days/year, 30.44 days/month, 7 days/week). Age values were subsequently converted back to years for interpretability, with both representations retained in the final metadata structure to facilitate precise computational analysis and intuitive clinical interpretation.

Following these preprocessing steps, samples belonging to the GSE85217 series were filtered from the larger GSE124814 metacohort based on their Series ID annotation. This operation successfully identified the 763 primary medulloblastoma samples previously validated for single-platform consistency, retaining all standardized metadata fields including molecular subgroup assignment (WNT, SHH, G3, G4), demographic information, and technical specifications.

Normal cerebellar tissue samples required independent preprocessing due to differences in metadata structure. Unlike GSE124814, the GSE45878 dataset[8] lacked a pre-formatted metadata file, necessitating direct extraction from the GSE45878_family.xml file (available at: https://ftp.ncbi.nlm.nih.gov/geo/series/GSE45nnn/GSE45878/miniml/, last modified 2013-05-30). An XML parser was implemented to extract relevant metadata fields for all samples, including GEO accession numbers, tissue classification, anatomical details, age ranges, gender, and platform information. Samples were first filtered to retain only those originating from brain tissue, followed by a second filter selecting exclusively cerebellar samples annotated as "Brain - Cerebellum" ($n = 16$) or "Brain - Cerebellar Hemisphere" ($n = 15$), yielding the final control

[8] dbGaP. Medulloblastoma genomics. dbGaP Study Accession: phs000424.v2.p1, 2024. URL https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v2.p1

cohort of 31 samples. This anatomical specificity was critical because medulloblastoma originates in the cerebellum, making cerebellar tissue the most appropriate biological reference for differential expression analysis.

Age information in GSE45878 was provided as ranges (e.g., "60-69 years", "50-59 years") reflecting the identification protocols of the GTEx consortium. These ranges were converted to single numerical values by calculating the midpoint of each interval (e.g., "60-69 years" $\rightarrow$ 64.5 years), representing the most statistically unbiased estimate when converting from categorical to continuous age data. The same conversion factor applied to cancer samples were used, ensuring consistency across the integrated dataset.

Following independent preprocessing of both cohorts, the datasets were merged into a unified metadata structure containing 794 total samples (763 medulloblastoma + 31 cerebellar controls), as shown in Figure 4.5. Prior to merging, column names were harmonized and a "Subgroup" label of "CT" was assigned to control samples to distinguish them from molecular subgroup categories. The final integrated metadata file retained all relevant fields including Sample ID, Series ID, Platform ID, tissue source, molecular subgroup, age, gender, organism, and molecular type.

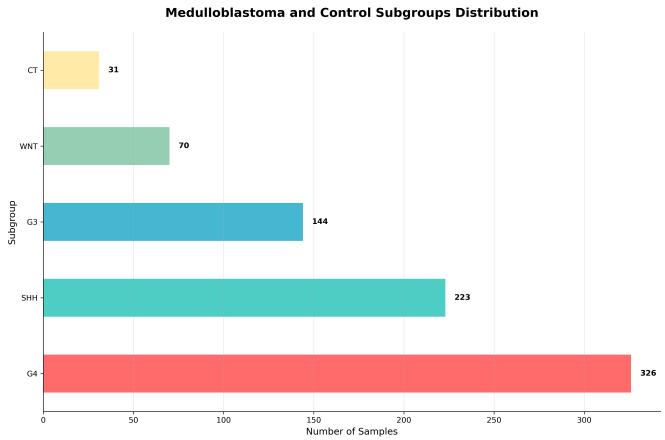

Figure 4.5: Sample distribution across molecular subgroups and cerebellar controls in the integrated cohort (n = 794). Tumor subgroups (WNT, SHH, G3, G4) are derived from GSE85217 and controls (CT) from GSE45878.

Exploratory data analysis (EDA) was performed on the integrated metadata to verify data quality, assess sample distributions, and identify potential confounding variables. Age distribution analysis revealed patterns consistent with established MB epidemiology[9], with cancer samples concentrated in pediatric age ranges (median: 8.0 years) (Figure 4.6), while control samples reflected the adult-biased composition of the GTEx cohort (median: ~55.0 years), as evidenced by the separation between conditions (Figure 4.7). Gender distribution was relatively balanced within each molecular subgroup,

[9] National Cancer Institute (NCI). Tratamiento del meduloblastoma y otros tumores embrionarios del sistema nervioso central infantil, 2025. URL https://www.cancer.gov/espanol/tipos/cerebro/pro/tratamiento-embrionarios-snc-infantil-pdq

with males representing 50-70% of cases, consistent with the known male predominance in medulloblastoma (Figure 4.8). The proportional gender heatmap across all conditions (Figure 4.9) further confirms this distribution pattern and provides a reference baseline for potential stratified analyses in future works. These demographic characteristics confirm that the integrated dataset reflects the expected clinical profile of the disease and do not introduce systematic confounds that would compromise subsequent gene expression analyses.

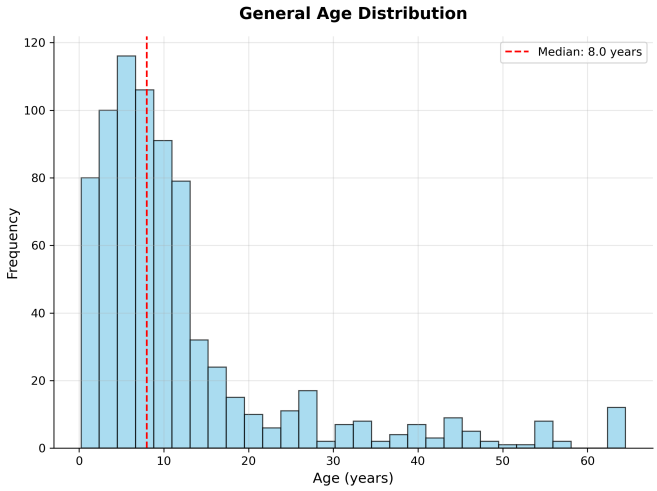

Figure 4.6: Age distribution across the integrated cohort. Median age of 8.0 years is consistent with the known pediatric predominance of medulloblastoma [National Cancer Institute (NCI), 2025].

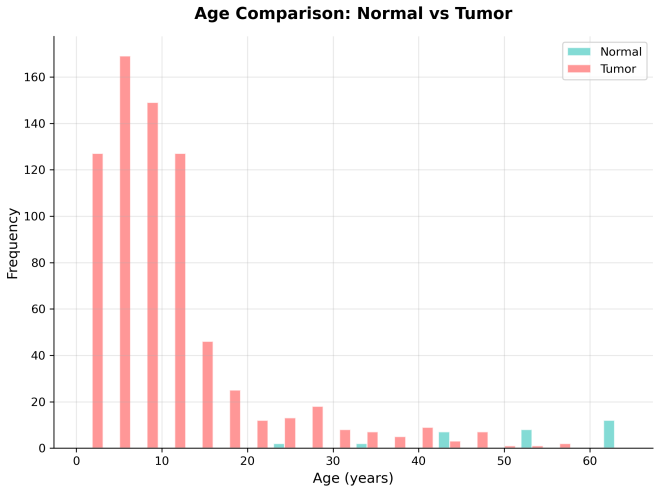

Figure 4.7: Age distribution stratified by condition (tumor vs. normal). The separation between pediatric cancer samples and adult cerebellar controls reflects the distinct cohort origins of GSE85217 and GSE45878.

The complete preprocessed metadata, containing all 794 samples with standardized annotations, was exported as a comma-separated values (CSV) file for integration with gene expression data in subsequent analysis steps as illustrated in Figure 4.10. All preprocessing steps were implemented in Python using the `pandas` library for data manipulation, the `re` module for pattern-matching age normalization,

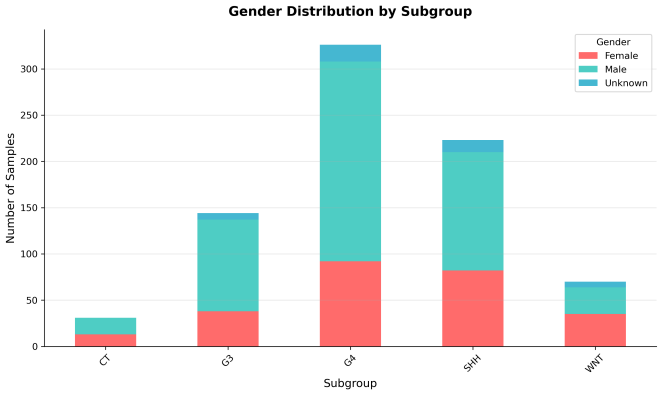

Figure 4.8: Gender distribution across molecular subgroups and cerebellar controls. Male predominance of 50-70% per subgroup is consistent with established medulloblastoma epidemiology.

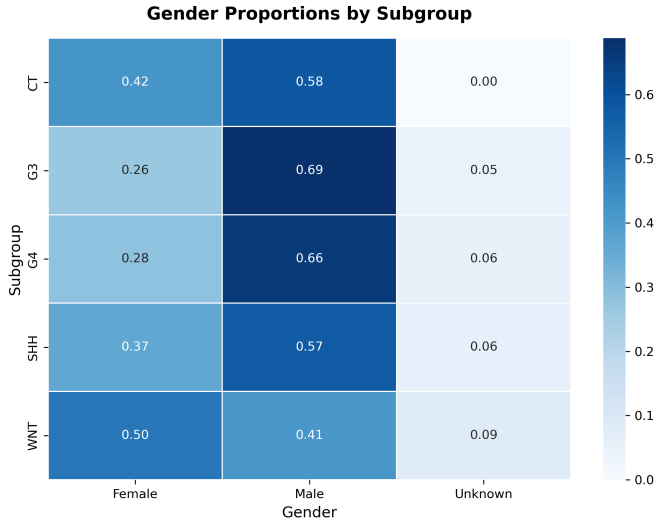

Figure 4.9: Proportional gender heatmap across all conditions. Values represent the relative frequency of male and female samples within each molecular subgroup and control group.

the `ElementTree` library for XML parsing, and `matplotlib/seaborn` for visualization. The complete pipeline is fully documented and publicly available, ensuring reproducibility and facilitating potential extension to other cancer datasets requiring similar metadata harmonization prior to bioinformatics analysis. [10]

[10] Pipeline scripts are available at: https://github.com/AndresRivera99.

4.3 *Gene Expression Data Processing and Differential Expression Analysis*

Raw microarray gene expression data were subjected to a systematic bioinformatics pipeline encompassing quality control, normalization, and statistical analysis to identify differentially expressed genes (DEGs) distinguishing medulloblastoma molecular subgroups from normal cerebellar tissue. Following the GeneChip microarray protocol, fluoresce intensity measurements for each probe are stored at the sample level in CEL format[11], yielding one file per sample. These

[11] Debojyoti Dhar and Gopala Kallapura. Analysis of microarray data from medulloblastoma tissue samples. In *Medulloblastoma: Methods and Protocols*, pages 59–64. Springer, 2022

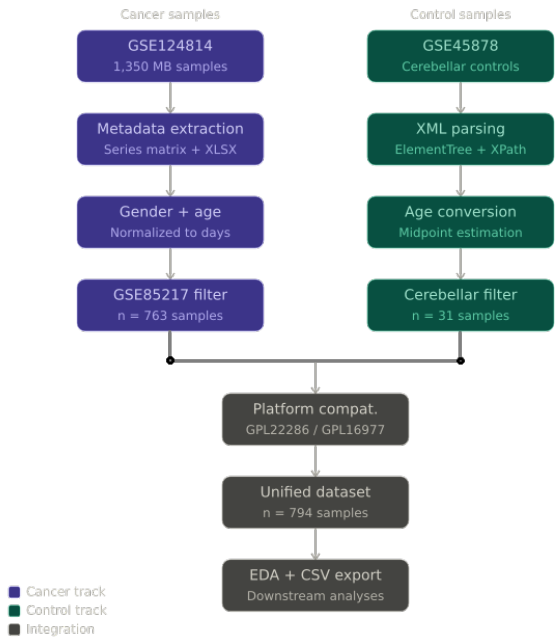

Figure 4.10: Metadata preprocessing pipeline for the integrated medulloblastoma cohort ($n = 794$). Cancer samples (GSE85217) and cerebellar controls (GSE45878) are processed through independent tracks before harmonization into a unified dataset for downstream gene expression analysis.

files contain probe-level intensity values prior to any normalization or summarization, representing the most granular form of microarray data available and affording full experimental control over all downstream transformation steps.

Prior to any expression analysis, a pattern recognition algorithm was implemented to validate the correspondence between the hand-curated GEO sample identifiers (GSM IDs) and the CEL files present in each group-specific directory. This step confirmed that all 794 samples across five groups were correctly assigned, with zero mismatches or unexpected files detected, establishing a verified and reproducible data foundation before downstream processing.

Two different hyperparameters played a major role to define differential expression analysis, reflecting statistical frameworks reported in literature. The primary configuration applied a false discovery rate threshold of FDR < 0.005 and a fold change threshold of $|\text{FC}| > 5$ ($|\log_2 \text{FC}| > 2.32$), as established by Castillo-Rodríguez et al. [2018] and subsequently adopted by Reveles-Espinoza et al. [2025] in the methodological precedent most closely aligned with the present study in terms of platform, disease, and subgroup structure yielding a conservative gene set suitable as a first feature selection for downstream classification. An alternative configuration grounded in complementary literature considers a more permissive threshold of FDR < 0.05 and $|\log_2 \text{FC}| > 0.3$, following the criteria reported by Arora et al. [2026] for transcriptomic analysis of medulloblastoma, with the FDR threshold further supported by Dhar and Kallapura [2022]. This second approach

broadens the detectable gene set and serves as a complementary reference for evaluating the robustness of biological signals identified under more stringent conditions. The complete bioinformatics pipeline, from raw CEL file validation through differential expression analysis and result export, is illustrated in Figure 4.11.

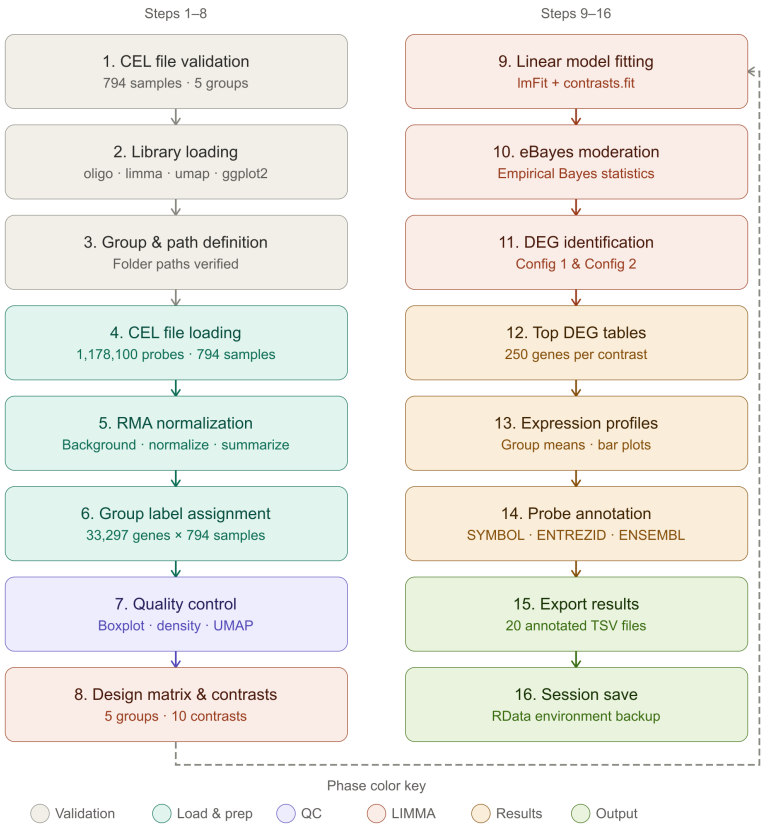

Figure 4.11: Bioinformatics pipeline for gene expression data processing and differential expression analysis. The pipeline comprises 16 sequential steps organized across six phases: data validation, loading and preprocessing, quality control, LIMMA-based statistical modeling, results extraction and annotation, and output generation. Two hyperparameter configurations were applied at the differential expression stage, corresponding to distinct statistical thresholds established in the literature, yet the main data downstream analysis considered for the present work follows the hyperparameters reported by [Castillo-Rodríguez et al., 2018, Reveles-Espinoza et al., 2025].

4.3.1 *Microarray Data Acquisition and Preprocessing*

To resolve the annotation version discrepancy between GPL22286 and GPL16977, all 794 CEL files were processed jointly from raw intensities using a single standardized annotation framework. Rather than harmonizing pre-normalized data across annotation versions, which risks introducing systematic biases, the `pd.hugene.1.1.st.v1` platform definition definition package was applied uniformly across all samples, ensuring that probe-to-gene mapping, background correction, and normalization were performed under identical parameters regardless of the original annotation version reported in GEO. All CEL files were read into R using the `oligo` package[12], which provides comprehensive support for Affymetrix Gene ST arrays

[12] Benilton S Carvalho and Rafael A Irizarry. A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, 26(19):2363–2367, 2010

and implements modern processing algorithms for exon-level and gene-level expression estimation. CEL files were organized into five group-specific directories corresponding to the four molecular subgroups and the cerebellar control cohort, and loaded jointly into a single `GeneFeatureSet` object via the `read.celfiles()` function. Successful data acquisition was confirmed by verification of the resulting object, which contained 1,178,100 probe-level features across all 794 samples with `pd.hugene.1.1.st.v1` confirmed as the active annotation consistent with the expected probe count for the HuGene-1.1-st array architecture.

4.3.2 *Quality Control and Normalization*

Gene expression normalization was performed using the Robust Multi-array Average (RMA) algorithm[13], implemented via the `rma()` function in the `oligo` package. RMA encompasses three sequential steps: background correction, which adjust for non-specific hybridization by separating true signal from background noise; quantile normalization, which forces the distribution of probe intensities to be identical across all arrays, removing systematic technical variation while preserving biological differences; and probe summarization, which aggregates multiple probes targeting the same transcript into a single expression value using median polish. RMA was applied to all 794 samples jointly in a single call, producing a normalized expression matrix of 33,297 probe-level features across 794 samples, with expression values on a log2 scale ranging from 1.42 to 14.25, consistent with the expected output from the HuGene-1.1-st array architecture. No manual log2 transformation was applied, as RMA outputs log2-scaled values by default.

[13] Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003

Following normalization, group membership was assigned to each sample by iterating over five predefined group directories, constructing a sample metadata frame containing the full CEL file path, a cleaned sample identifier, and the corresponding molecular subgroup label for each of the 794 samples. The subgroup variable was encoded as an order factor with CT specified as the reference level (`levels = c("CT", "WNT", "SHH", "G3", "G4")`), yielding the following distribution: CT ($n = 31$), WNT ($n = 70$), SHH ($n = 223$), G3 ($n = 144$), G4 ($n = 326$), subsequently stored in the `phenoData` slot of the expression object for downstream modeling.

Comprehensive quality control was performed on the full RMA-normalized expression matrix of 33,297 features through three complementary visualizations. For visualization clarity, sample-level expression distributions were summarized as group means across all 33,297 features, yielding a compact five-group representation of

the normalized expression profiles as shown in Figure 4.12. All five groups exhibited comparable median log2 expression levels around 7.0, with similar interquartile ranges spanning approximately 5.6 to 8.0, confirming successful quantile normalization across the 794 samples regardless of molecular subgroup or tissue origin. Expression density curves per group, as illustrated in Figure 4.13, revealed closely overlapping unimodal distributions centered around an intensity of 5–6 consistent with well-normalized microarray data. Notably, the cerebellar control samples exhibited a sharper, more pronounced density peak relative to the tumor subgroups, reflecting the greater transcriptional homogeneity of normal cerebellar tissue compared to the heterogeneous tumor populations. Uniform Manifold Approximation and Projection (UMAP)[14] was applied to the normalized expression matrix with parameters `n_neighbors = 15` and `random_state = 123` for reproducibility, projecting the 794 samples into a two-dimensional embedding, as observed in Figure 4.14. The projection revealed the presence of potential outlier samples, likely attributed to the heterogeneous classification protocols across the source GEO datasets, while confirming the expected molecular subgroup structure: WNT samples formed a tight, well-separated cluster in the upper left region; SHH samples occupied a distinct cluster in the lower central region; cerebellar control samples formed an isolated compact cluster in the lower right; and Group 3 and Group 4 samples exhibited partial spatial overlap in the upper right region, consistent with their known molecular similarity [Cavalli et al., 2017]. These results validate the integrity of subgroup assignments and confirm that transcriptional differences between molecular subgroups are preserved following joint preprocessing of the two source datasets.

[14] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018

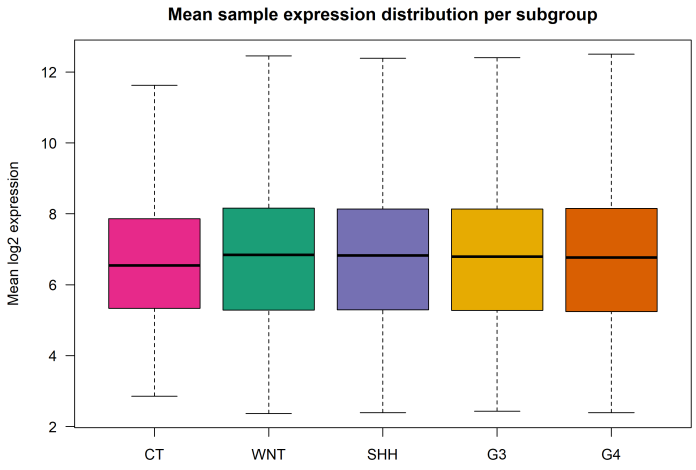

Figure 4.12: Box-and-whisker plot of mean log2 expression per molecular subgroup and cerebellar control (CT) following RMA normalization. Each box summarizes the distribution of group-averaged expression values across 33,297 gene-level features. Comparable median levels and interquartile ranges across all five groups confirm successful quantile normalization.

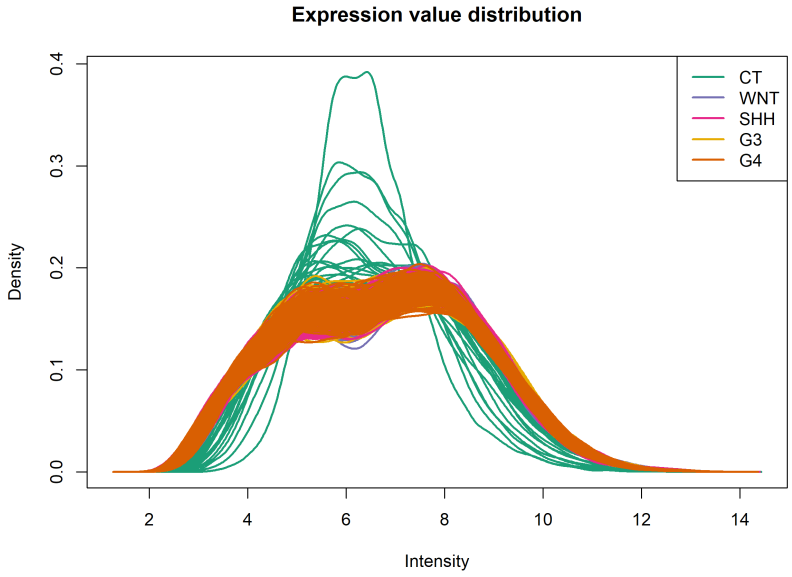

Figure 4.13: Expression value density curves per sample following RMA normalization. Each line represents one of the 794 samples, colored by molecular subgroup. Overlapping unimodal distributions confirm technical consistency across samples. Cerebellar control samples exhibit a sharper density peak, reflecting greater transcriptional homogeneity relative to tumor subgroups.

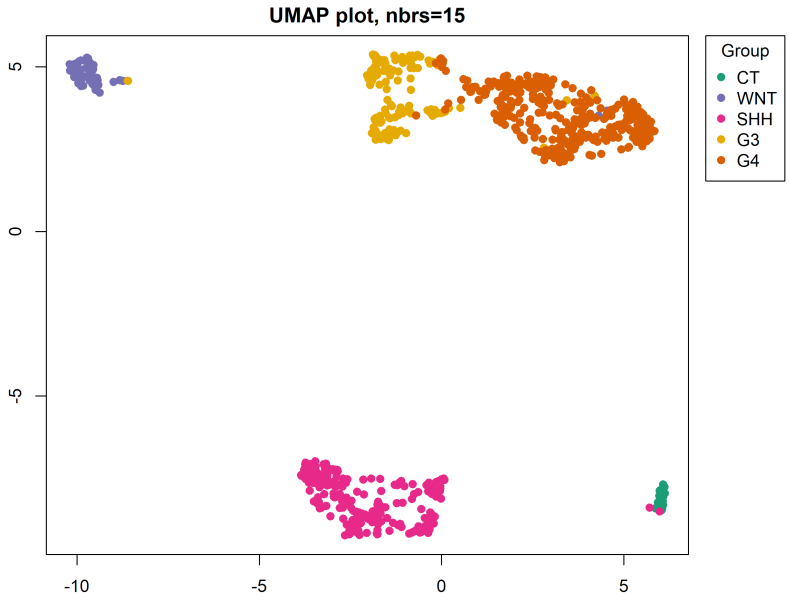

Figure 4.14: UMAP projection of 794 samples following RMA normalization (`n_neighbors = 15, random_state = 123`). Each point represents one sample, colored by molecular subgroup. WNT and SHH subgroups from tight well-separated clusters, cerebellar controls occupy a distinct isolated region, and Group 3 and Group 4 exhibit partial overlap consistent with their molecular similarity.

Following quality control, cross-platform probe harmonization was performed to restrict the analytical feature space to genes represented in both source datasets (as mentioned previously in the metadata preprocessing section). Annotation files for both platforms were parsed from the GEO soft format using a custom helper function that identifies

where the platform table begins and where it ends, as each platform arrange and structures data and metadata all together varying across different platforms. The custom helper functions marks and read the intervening data block, bypassing metadata headers that would otherwise corrupt the parse. Each platform file was loaded into a three-column annotation table containing probe identifier, Ensembl gene identifier (SPOT_ID), and probe description; the two tables were merged via `bind_rows()` and deduplicated by probe identifier, retaining the GPL22286 entry when a probe appeared in both, yielding a joint annotation of 23,973 unique probe entries. Cross-platform overlap analysis identified 20,372 probes shared by both platforms, 1,269 probes exclusive to GPL22286, and 2,332 probes exclusive to GPL16977. To avoid confounding downstream differential expression results with platform-specific features, only the 20,372 shared probes were retained for all subsequent analyses.

Numeric probe identifiers from the RMA-normalized matrix were mapped to Ensembl gene identifiers using the `hugene11sttranscriptcluster.db` package, which returned 47,153 probe-Ensembl mappings across 33,297 probes, of which 35,445 carried a valid Ensembl identifier and 11,708 were unannotated. The expression data frame was subsequently filtered to retain only those probes with a valid Ensemble identifier present among the 20,372 shared Ensembl IDs, yielding 18,931 rows after the inner join. Since multiple probes may target the same gene, probe-level features were collapsed to a single representative per Ensembl gene identifier using a highest-mean-expression criterion: the probe with the greatest mean expression across all 794 samples was retained, as probes with higher average hybridization signal are more likely to be on target and less susceptible to background noise. This strategy is consistent with standard practice for Affymetrix microarray analysis. Probe collapse produced a final expression matrix of 17,032 unique Ensembl gene identifiers across 794 samples, which was exported as a comma separated values file (CSV) together with the corresponding sample metadata table for downstream differential expression analysis and classifier training.

4.3.3 *Differential Expression Analysis*

Differential gene expression analysis was performed using the `limma` package[15] (linear models for microarray data), which implements empirical Bayes moderated t-statistics specifically designed for high-dimensional gene expression data with moderate sample sizes. The `limma` framework offers several advantages over traditional t-tests: information is borrowed across genes to stabilize variance estimates, improving statistical power particularly for genes with low expression;

[15] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015

moderated test statistics follow a t-distribution with increased degrees of freedom, reducing false positives caused by unreliable variance estimates; and the linear modeling framework accommodates complex experimental designs and multiple contrast simultaneously.

A design matrix was constructed without an intercept term (`model.matrix(~ 0 + subgroup)`), where `subgroup` represents the five study categories: CT, WNT, SHH, G3, and G4,with CT specified as the reference level. This parametrization facilitates direct comparisons between each molecular subgroup and normal cerebellar tissue through explicit contrast specification. Following Reveles-Espinoza et al. [2025] and Castillo-Rodríguez et al. [2018], the analytical focus was placed on four subgroup versus cerebellar control contrasts (WNT–CT, SHH–CT, G3–CT, G4–CT), as these directly capture the transcriptional programs distinguishing each molecular subgroup from normal tissue and are most relevant for downstream classification and biomarker identification. The contrast matrix is specified in Table 4.3.

Levels	G3–CT	G4–CT	SHH–CT	WNT–CT
CT	-1	-1	-1	-1
WNT	0	0	0	1
SHH	0	0	1	0
G3	1	0	0	0
G4	0	1	0	0

Table 4.3: Contrast matrix specifying the four subgroup versus cerebellar control comparisons defined for differential expression analysis. Values of $+1$ and -1 indicate the subgroup and reference group respectively; 0 indicates groups not involved in the comparison.

Linear models were fit to the 17,032 x 794 gene level expression matrix using `lmFit()`, contrasts were applied via `contrasts.fit()`, and empirical Bayes moderation was performed using `eBayes()`, producing moderated t-statistics , B-statistics (log-odds of differential expression), and adjusted p-values for each gene across all four contrasts. The mean-variance trend plot (Figure 4.15) confirmed that residual variance is not strongly dependent on expression level, as evidenced by a relatively flat trend line across the expression range, validating the appropriateness of the standard `limma` approach for this dataset.

Two additional diagnostic visualizations were generated to assess model behavior prior to DEG extraction. The adjusted p-value distribution histogram (Figure 4.16), confirmed strong enrichment of adjusted p-values near zero with the vast majority of the 17,032 genes concentrated below 0.05 and negligible counts across the remainder of the distribution, indicating a pervasive differential signal consistent with the comparison of tumor subgroups against normal cerebellar tissue. The Q-Q plot of moderated t-statistic (Figure 4.17, generated via `qqt()`, revealed near-perfect alignment between observed and theoretical quantiles across the central range with pronounced divergence at bot tails, sample quantiles extending to approximately

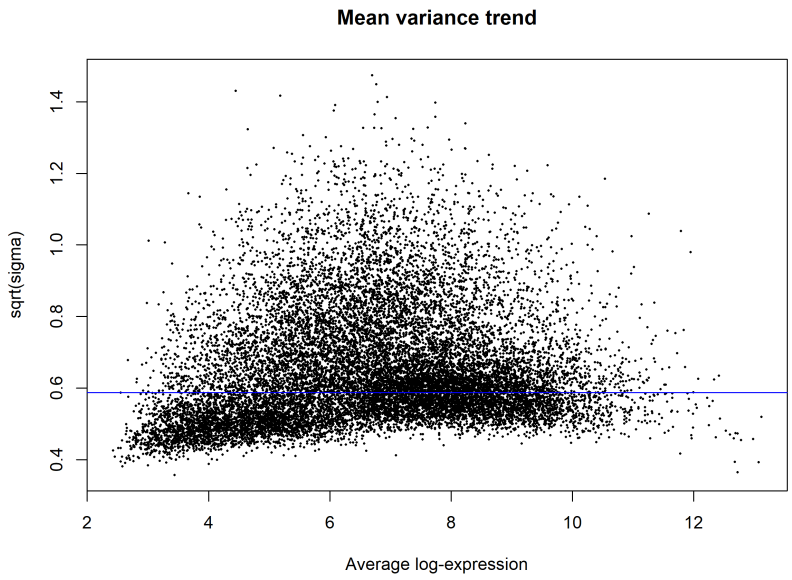

Figure 4.15: Mean-variance trend plot following empirical Bayes moderation. The relatively flat blue trend line cross the expression range confirms that residual variance is not dependent on expression level, validating the standard `limma` modeling approach for this microarray dataset.

±45 against theoretical limits of ±4, consistent with the presence of a substantial number of strongly differentially expressed genes and confirming appropriate model calibration for this dataset. Mean-difference (MD) plots were additionally generated for each contrast via `plotMD()`, displaying the log2 fold change against the average log2 expression for all genes, with significantly differentially expressed genes highlighted to facilitate visual inspection of the effect size distribution relative to expression level.

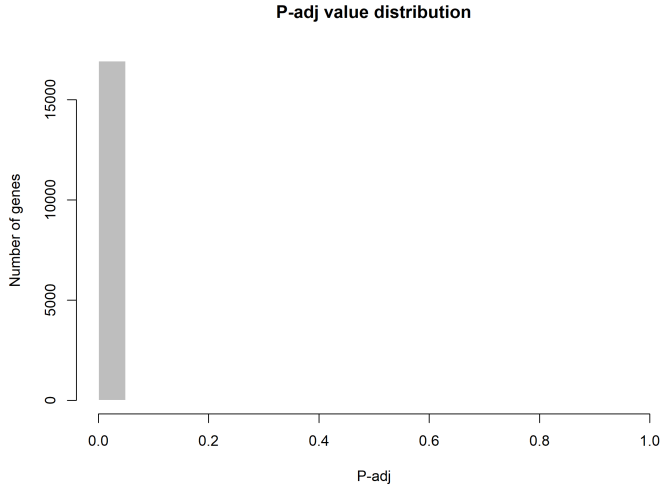

Figure 4.16: Adjusted p-value distribution across all 17,032 genes following empirical Bayes moderation. Concentration of values below 0.05 with negligible counts across the remainder of the distribution indicates a pervasive differential expression signal across all four subgroup versus cerebellar control contrasts.

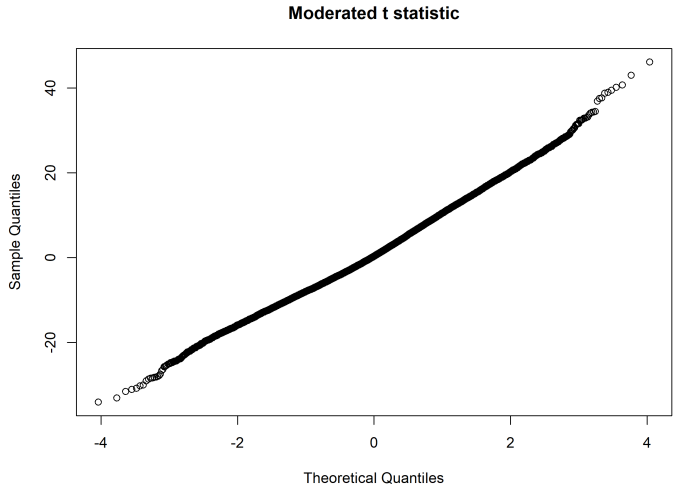

Figure 4.17: Q-Q plot of moderated t-statistics against the theoretical t-distribution. Near-perfect alignment across the central quantile range with pronounced tail divergence — sample quantiles reaching approximately ±45 against theoretical limits of ±4 — confirms appropriate model calibration and reflects the presence of strongly differentially expressed genes.

Adjusted p-values were calculated using the Benjamini-Hochberg procedure[16], which controls the expected proportion of false positives among genes called significant. Differentially expressed genes were identified applying the primary configuration of FDR < 0.005 and $|FC| > 5$ ($|\log_2 FC| > 2.32$), as established by Castillo-Rodríguez et al. [2018] and Reveles-Espinoza et al. [2025]. These stringent criteria prioritize high-confidence candidates for downstream biological interpretation and feature selection for the classification model. The complete ranked list for each contrasts was extracted via `topTable()` with `p.value = 1`, retaining all genes regardless of significance to preserve flexibility for downstream pathway enrichment and ANN input construction; threshold filtering was applied at the point of use rather than at extraction

[16] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995

Application of the primary configuration thresholds across the four subgroup versus CT contrasts yielded the DEG counts summarized in Table 4.4.

Contrast	Upregulated	Downregulated	Total DEGs
WNT vs. CT	404	304	708
SHH vs. CT	268	209	477
G3 vs. CT	260	274	534
G4 vs. CT	251	216	467

Table 4.4: Summary of differentially expressed genes per molecular subgroup versus cerebellar control contrast under the primary configuration (FDR < 0.005, $|\log_2 FC| > 2.32$). Up: upregulated genes; Down: downregulated genes.

As an alternative reference, the permissive configuration (FDR $<$ 0.05, $|\log_2 FC| > 0.3$) following Arora et al. [2026] and Dhar and Kallapura [2022] yielded substantially broader gene sets, with subgroup versus CT contrasts ranging from 22,112 (G4–CT) to 23,067 (WNT–CT) DEGs. The marked difference in DEG counts between configurations

reflects the distinct stringency levels and confirms that the primary configuration effectively restricts the feature space to biologically robust signals suitable for downstream classification. All subsequent analyses and visualizations reported in this section are based on the primary configuration.

Volcano plots for the four primary contrasts visualize the relationship between fold-change magnitude and statistical significance for each molecular subgroup versus cerebellar control (Figures 4.18, 4.19, 4.20, and 4.21). Plots were generated using `ggplot2`, displaying the $-\log_{10}$(FDR-adjusted p-value) on the vertical axis against the $\log_2$ fold change on the horizontal axis, with dashed lines indicating the applied significance thresholds (FDR < 0.005, $|\log_2 \text{FC}| > 2.32$). Genes meeting both criteria are highlighted in red (upregulated) or blue (downregulated), while non-significant genes are shown in gray. This dual threshold coloring is preferred over highlighting based on the B-statistic alone, as the latter can designate genes near zero fold change as significant solely on statistical grounds, without requiring biological effect size. The WNT–CT contrast exhibited the greatest number of significant DEGs (708 total), with a broad fold-change and several genes reaching $-\log_{10}$, reflecting the high transcriptional distinctiveness of the WNT subgroup. The G3–CT and G4–CT contrasts displayed comparable distributions with moderate fold-change spreads, consistent with the intermediate transcriptional divergence of these subgroups from control tissue. The SHH–CT contrast yielded 477 significant DEGs within a slightly more compact fold-change distribution. Across al four contrasts, the characteristic volcano shape confirms appropriate statistical behavior.

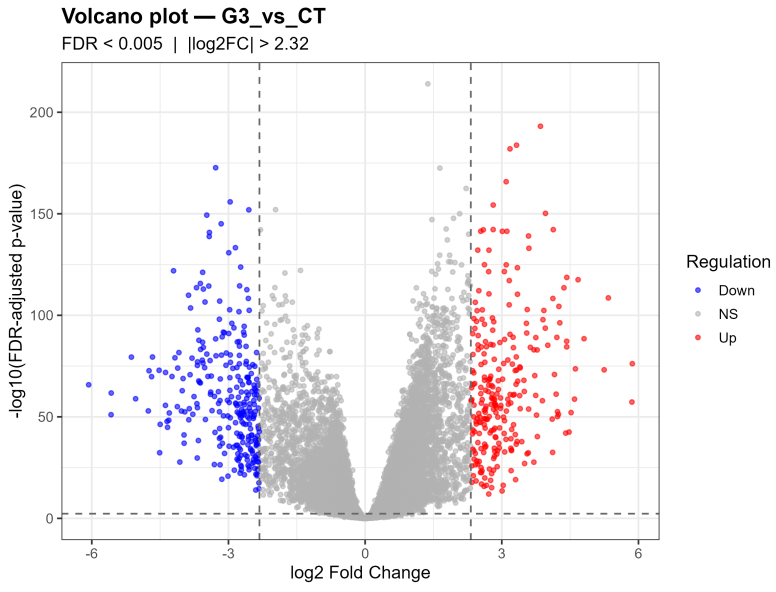

Figure 4.18: Volcano plot for G3 vs. CT contrast (FDR < 0.005, $|\log_2 \text{FC}| > 2.32$). A total of 534 DEGs were identified (260 upregulated, 274 downregulated). Red points: significantly upregulated genes; blue points: significantly downregulated genes; gray points: non-significant genes. Dashed lines indicate the applied significance thresholds.

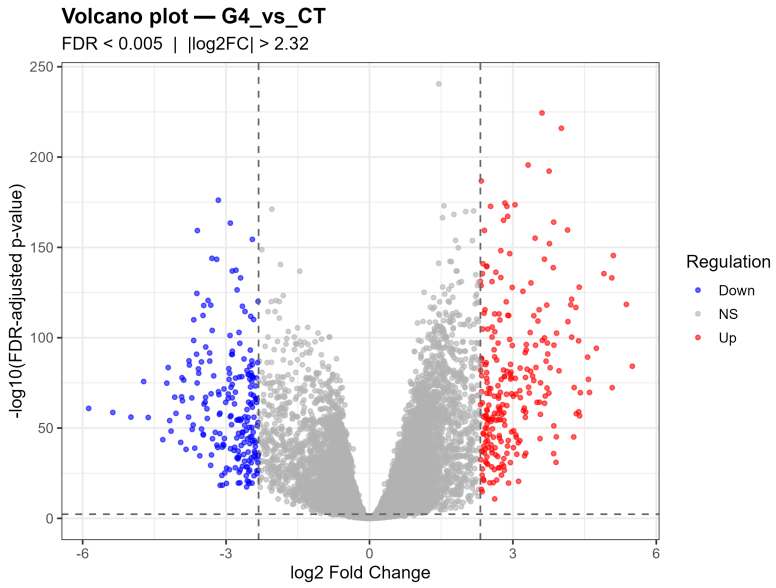

Figure 4.19: Volcano plot for G4 vs. CT contrast (FDR < 0.005, $|\log_2 \mathrm{FC}| > 2.32$). A total of 467 DEGs were identified (251 upregulated, 216 downregulated). Red points: significantly upregulated genes; blue points: significantly downregulated genes; gray points: non-significant genes. Dashed lines indicate the applied significance thresholds.

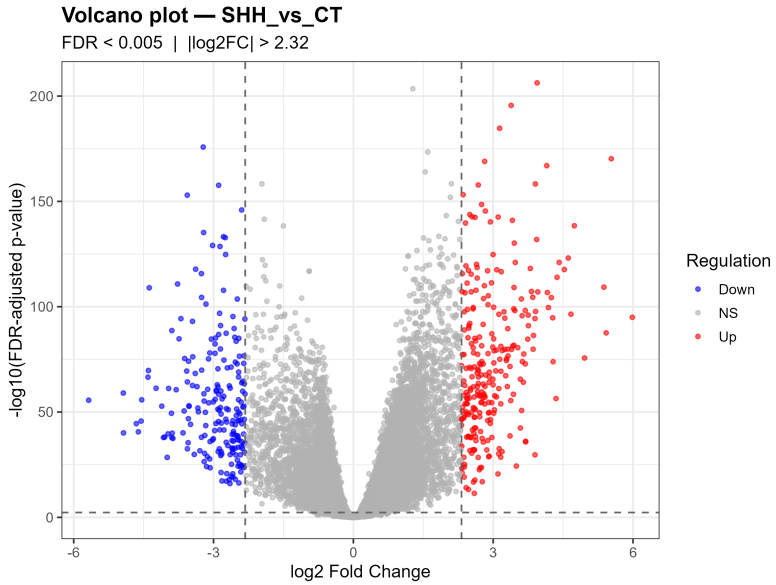

Figure 4.20: Volcano plot for SHH vs. CT contrast (FDR < 0.005, $|\log_2 \mathrm{FC}| > 2.32$). A total of 477 DEGs were identified (268 upregulated, 209 downregulated). Red points: significantly upregulated genes; blue points: significantly downregulated genes; gray points: non-significant genes. Dashed lines indicate the applied significance thresholds.

Venn diagram analysis of the four subgroup versus cerebellar control contrasts (Figure 4.22) revealed 174 genes commonly differentially expressed across all four molecular subgroups, suggesting a core transcriptional signature consistently altered in medulloblastoma regardless of molecular subtype. Pairwise overlaps between adjacent contrasts were also identified: G3–CT and G4–CT shared 78 genes exclusively, G4–CT and SHH–CT shared 13 genes, SHH–CT and WNT–CT shared 69 genes, and three-way intersections ranged from 16 to 60 genes across combinations. Subgroup-exclusive DEGs (those

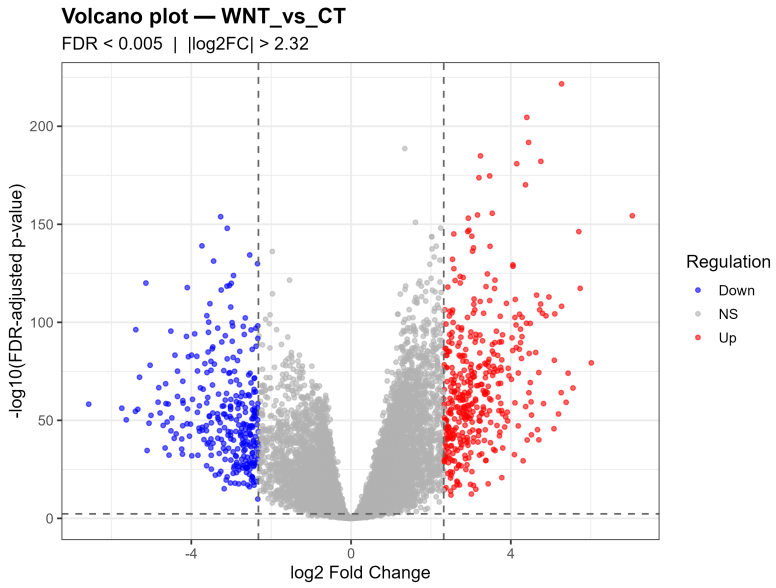

Figure 4.21: Volcano plot for WNT vs. CT contrast (FDR < 0.005, $|\log_2 \mathrm{FC}| > 2.32$). A total of 708 DEGs were identified (404 upregulated, 304 downregulated), the highest among all four contrasts. Red points: significantly upregulated genes; blue points: significantly downregulated genes; gray points: non-significant genes. Dashed lines indicate the applied significance thresholds.

differentially expressed in only one contrasts) ranged from 76 (G4–CT) to 271 (WNT–CT), with WNT exhibiting the largest subgroup-specific transcriptional program, consistent with its recognized molecular distinctiveness [Cavalli et al., 2017]. The 15,893 genes outside all four contrasts represent non-significant features under the primary configuration thresholds.

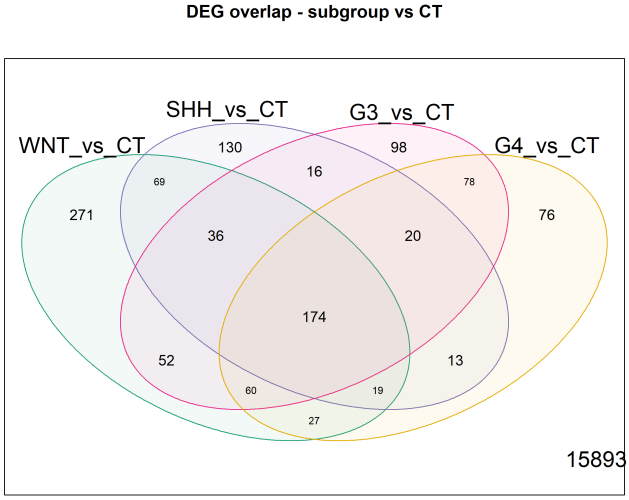

Figure 4.22: Venn diagram showing DEG overlap across the four subgroup versus cerebellar control contrasts (FDR < 0.005, $|\log_2 \mathrm{FC}| > 2.32$). Numbers indicate genes meeting both significance criteria in each region. The central intersection of 174 genes represents a core transcriptional signature common to all four molecular subgroups. Subgroup-exclusive DEGs range from 76 (G4–CT) to 271 (WNT–CT). The value 15,893 represents non-significant features outside all four contrasts.

Following differential expression analysis, results were systematically annotated and exported to structured file formats to support downstream analyses and ensure full reproducibility. The top 250 differentially expressed genes per contrast were extracted via

`topTable()` for both hyperparameter configurations, retaining the complete statistical output including probe identifiers, $\log_2$ fold change, average expression (`AveExpr`), moderated t-statistic, raw and FDR-adjusted p-values, and B-statistic. Group-averaged expression values across all five molecular subgroups were additionally computed for the top 50 genes under each configuration, yielding compact summary matrices that replicate the sample-level expression tables available in GEO2R. Both configurations returned the same top 50 genes, confirming that the most statistically robust signals are consistent regardless of threshold stringency.

Since CEL files were processed directly from raw intensities, probe-level annotation metadata available in GEO2R (including gene symbols, descriptions, and database identifiers) was not automatically attached to the expression object. Probe annotation was therefore performed using the `hugene11sttranscriptcluster.db` package via `AnnotationDbi::select()`, mapping all 33,297 probe identifiers to ENSEMBL gene identifiers, HGNC gene symbols, full gene names, and Entrez IDs. This mapping yielded 49,370 initial rows due to one-to-many probe-to-gene relationships, which were resolved by retaining the first available annotation per probe identifier, producing a unique annotation table of 33,297 entries. The resulting annotation enables downstream Gene Ontology enrichment analysis across biological process, molecular function, and cellular component categories; KEGG pathway enrichment; and STRING protein-protein interaction analysis, as applied in comparable medulloblastoma transcriptomic studies[17].

[17] Debojyoti Dhar and Gopala Kallapura. Analysis of microarray data from medulloblastoma tissue samples. In *Medulloblastoma: Methods and Protocols*, pages 59–64. Springer, 2022

Since probe annotation to Ensembl gene identifiers was performed during the preprocessing stage, the differential expression results carry Ensembl IDs as row identifiers throughout without requiring a secondary annotation step. Full ranked gene lists for all four contrasts were exported as comma-separated values via `write.csv()`, retaining all 17,032 genes with complete statistical output including $\log_2$ fold change, average expression, moderated t-statistic, raw and FDR-adjusted p-values, and B-statistic per gene. The complete R environment, including all fitted models, expression matrices, and group factor assignments, was preserved as an `.RData` binary object to ensure full session reproducibility without requiring reprocessing of the raw CEL files.

Following DEG extraction, the ANN input feature matrix was constructed by taking the union of significant DEGs across all four subgroups versus CT contrasts under the primary configuration (FDR < 0.005, $|\log_2 \mathrm{FC}| > 2.32$). This yielded 1,139 unique Ensembl gene identifiers, spanning genes that are differentially expressed in at least one molecular subgroup relative to cerebellar control. The expression matrix was filtered to these 1,139 features across all 794

samples and exported in transposed format (samples as rows, genes as columns), which is the standard input convention for machine learning frameworks. A companion label file mapping each sample identifier to its corresponding molecular subgroup was exported separately. The ratio of 794 samples to 1,139 features is within the acceptable range for feedforward neural network training, and residual dimensionality concerns are addressed through regularization during model construction.

4.4 *Deep Learning Classification*

The differentially expressed genes identified under the primary analytical configuration (FDR < 0.005, $|\log_2 \mathrm{FC}| > 2.32$) constituted the molecular feature set for the ANN classification pipeline. Following the two-stage hierarchical framework established by Reveles-Espinoza et al. [2025], the classification approach comprised a simultaneous five class classifier distinguishing all molecular subgroups and cerebellar controls, each trained to isolate one subgroup from the remaining cases (WNT-vs-rest, SHH-vs-rest, G3-vs-rest, G4-vs-rest).

The ANN input matrix and corresponding label file, exported from the bioinformatics pipeline as described in the previous section, were loaded into MATLAB 2025b via `readtable()`. The expression matrix comprised 794 samples (rows) and 1,139 DEG-derived features (columns), with values on the RMA $\log_2$ scale and Ensembl gene identifiers as column names. This 1,139-gene feature vector constitutes the input layer of all five architectures in the present study, replacing the 1,937-gene vector reported by Reveles-Espinoza et al. [2025], whose composition reflected a smalled dataset processed through a different normalization pipeline.

Although RMA normalization renders samples mutually comparable, the 1,139 genes retain heterogeneous absolute expression ranges. Prior to network training, each gene was therefore standardize by z-score normalization across samples (`zscore()`), yielding a zero-mean, unit-variance distribution per feature. This operates on a different axis than RMA: it prevents high-expression genes from numerically dominating the input weights during gradient descent, a convergence requirement regardless of whether log-scale normalization has been applied upstream.

Class labels were retained as MATLAB categorical variables rather than integer or one-hot-encoded vectors, allowing the `classificationLayer` to manage internal encoding. The five-class label vector followed the distribution CT ($n = 31$), WNT ($n = 70$), SHH ($n = 223$), G3 ($n = 144$), and G4 ($n = 326$). For each binary classifier, labels were recoded by collapsing all non-target subgroups into a single

rest category while preserving the target class name, producing two-level categorical vectors suitable for one-vs-rest discrimination. Class imbalance, particularly pronounced in the CT vs the others and WNT vs the rest settings, was addressed through stratified partitioning in the cross-validation stage.

4.4.1 Network Architectures

All five networks share a common structural template inherited from Reveles-Espinoza et al. [2025]: a three hidden layer fully connected feedforward architecture with softmax output and cross-entropy loss. Per-network topology, activation function, and output dimensionality are summarized in Table 4.5. Hidden layer widths were retained from the original specification; the input dimensionality was adjusted from 1,937 to 1,139 to reflect the DEG set obtained under the present analytical configuration.

Network	Hidden layers	Output units	Activation
5-class	700–700–700	5	ReLU
WNT vs. rest	600–600–600	2	ReLU
SHH vs. rest	750–700–750	2	ReLU
G3 vs. rest	600–600–600	2	Tanh
G4 vs. rest	700–700–700	2	ReLU

Table 4.5: Feedforward network topologies adapted from Reveles-Espinoza et al. [2025]. Input dimensionality corresponds to the 1,139 DEGs identified in the present pipeline. Hidden-layer widths are given as neurons per layer.

The G3-vs-rest network retains the hyperbolic tangent activation, motivated by the well documented transcriptomic overlap between Group 3 and Group 4 of medulloblastoma. The bounded, zero centered output of tanh was reported to improve discrimination along this ambiguous subgroup boundary. All remaining networks use ReLU, which is favored in modern feedforward architectures for its non saturating gradient behavior and its compatibility with batch normalization.

4.4.2 Layer Construction and Regularization

Each network was assembled programmatically via the helper function `build_layers()`, which returns a layer array composed of an input layer, three repeated hidden blocks, and an output head. The `featureInputLayer` accepts the 1,139 dimensional standardized expression vector with internal normalization disabled (`'Normalization', 'none'`), since z-score scaling as applied at the preprocessing stage. Each hidden block comprises four sequential layers: a `fullyConnectedLayer` of the prescribed width, the corresponding activation (`reluLayer` or `tanhLayer`), a `batchNormalizationLayer`, and a `dropoutLayer` with rate 0.3. The

topology reported in Reveles-Espinoza et al. [2025] included neither batch normalization nor dropout; both were introduced in the present study to counteract overfitting in the moderately high-dimensional regime arising from 794 training samples and 1,139 input features. Batch normalization stabilizes the distribution of pre-activations across mini-batches and accelerates convergence, whereas dropout provides stochastic regularization by deactivating a random 30% of units during each forward pass. The output head consists of a final `fullyConnectedLayer` projecting onto the class dimension, followed by a `softmaxLayer` and a `classificationLayer`, the latter internally computing the categorical cross-entropy loss used for backpropagation.

4.4.3 *Training Configuration*

All networks were trained using the Adam optimizer under a uniform set of hyperparameters, summarized in Table 4.6. The initial learning rate was set to 10^{-3} and reduced by a factor of 0.5 every 20 epochs under a piecewise schedule, allowing rapid early convergence while enabling fine-grained adjustment in later stages. An L_2 weight decay of 10^{-4} provided additional regularization in parameter space, complementing the dropout-based regularization applied at the activation level. Mini-batch stochastic optimization was conducted with batch size 32, and training data were reshuffled at every epoch to decorrelate successive gradient updates.

Parameter	Value
Optimizer	Adam
Maximum epochs	100
Mini-batch size	32
Initial learning rate	1×10^{-3}
Learning rate schedule	Piecewise
Learning rate drop factor	0.5
Learning rate drop period	20 epochs
L_2 regularization	1×10^{-4}
Shuffle	Every epoch
Validation frequency	10 iterations

Table 4.6: Training hyperparameters applied uniformly across all five networks. Validation data correspond to the held-out fold within each cross-validation iteration.

4.4.4 *Cross-Validation Protocol*

Given the pronounced class imbalance in the dataset, most acute in the contrast between G4 ($n = 326$) and CT ($n = 31$), stratified 10-fold cross-validation was employed to ensure that every fold preserved the global class distribution. Fold assignments were generated via `cvpartition()` with `'KFold', 10` and `'Stratify', true`, producing

ten training/validation splits in which each class was represented in approximate proportion to its overall frequency. For each of the five network an identical training routine was executed on each fold: layers were instantiated anew via `build_layers()`, the held-out partition was attached as validation data, and the network was fitted through `trainNetwork()`. Predicted labels and posterior class scores were obtained by applying `classify()` to the validation fold, and both were accumulated across all ten iterations for aggregated evaluation. For the four binary classifiers the label recording described in the classification section was performed prior to partitioning, so that stratification operated on the collapsed two-level distribution and preserved the position-class proportion within every fold.

4.4.5 *Evaluation Metrics and Outputs*

Network performance was quantified along three complementary dimensions: classification accuracy, confusion structure, and probabilistic discrimination. Per-fold accuracy was computed as the proportion of correctly classified validation samples and summarized across the ten folds as mean $\pm$ standard deviation. Aggregate confusion matrices were constructed from the concatenated predictions across all folds and rendered via `confusionchart()` with both row-normalized (recall per true class) and column-normalized (precision per predicted class) annotations.

Receiver operating characteristic (ROC) curves were computed for each class under a one-vs-rest formulation using `perfcurve()`, with per-class areas under the curve (AUC) and mean AUC reported per network. Wall-clock training time was recorded at both the per-fold and per-network level via `tic/toc` timers, providing a pragmatic measure of computational cost for comparison across architectures on the workstation specified as Intel Core i9-14900HX, 32 GB RAM, MATLAB 2025b. Per-network graphical outputs (confusion matrix, fold-accuracy bar plot, and ROC curve panel) were exported as PNG figures, which will be discussed in the next chapter, whereas aggregate summary statistics across all five networks (mean accuracy, standard deviation, mean AUC, total training time) were consolidated into a CSV for latter reference.

5 Results

Contents

5.1 Overview of Classification Performance 113
5.2 Five-class Classifier: CT vs. WNT vs. SHH vs. G3 vs. G4 114
5.3 Binary Classifiers 116
5.4 Computational Cost Analysis 121
5.5 Comparison with the State of the Art 124

5.1 Overview of Classification Performance

The deep learning pipeline described in the previous chapter was executed on the feature matrix of 794 samples and 1,139 differentially expressed genes. Five networks were trained under the stratified 10-fold cross-validation protocol: one five-class classifier discriminating all molecular subgroups and cerebellar controls simultaneously, and four binary classifiers implementing the hierarchical one-vs-rest decomposition inherited from Reveles-Espinoza et al. [2025]. Aggregate performance metrics across the five architectures are consolidated in Table 5.1.

Network	Mean accuracy (%)	Std. dev. (%)	Mean AUC	Time (min)
5-class	97.99	0.65	0.999	17.5
WNT vs. rest	99.75	0.53	1.000	11.6
SHH vs. rest	99.87	0.40	0.996	16.6
G3 vs. rest	97.99	1.98	0.988	34.1
G4 vs. rest	99.12	1.04	0.999	11.9

Table 5.1: Summary of 10-fold cross-validated performance across all five ANN architectures. Accuracy is reported as mean ± standard deviation across folds; mean AUC corresponds to the average one-vs-rest area under the ROC curve across classes within each network. Total time refers to cumulative wall-clock training across all ten folds on an Intel Core i9-14900HX workstation with 32 GB RAM.

All five networks achieved mean cross-validated accuracy above 97.99%, with mean AUC values between 0.988 and 1.000. Performance was remarkably stable across folds: standard deviations remained below 1.1% for four of the five networks, with the G3-vs-rest network exhibiting the highest fold-to-fold variability (1.98%), a behavior that is biologically consistent with the well-documented transcriptomic

overlap between Group 3 and Group 4 medulloblastoma. Training time ranged from 11.6 minutes for the WNT-vs-rest network to 34.1 minutes for the G3-vs-rest network, reflecting both architectural differences (tanh activation in G3 vs. ReLU elsewhere) and convergence dynamics on the most ambiguous binary boundary.

5.2 *Five-class Classifier: CT vs. WNT vs. SHH vs. G3 vs. G4*

The five-class network, which simultaneously discriminates the four molecular subgroups and cerebellar controls, achieved a mean cross-validated accuracy of 97.99% ± 0.65% with a mean one-vs-rest AUC of 0.999. The aggregate confusion matrix, per-fold accuracy distribution, and ROC curves are shown in Figures 5.1, 5.2, and 5.3, respectively.

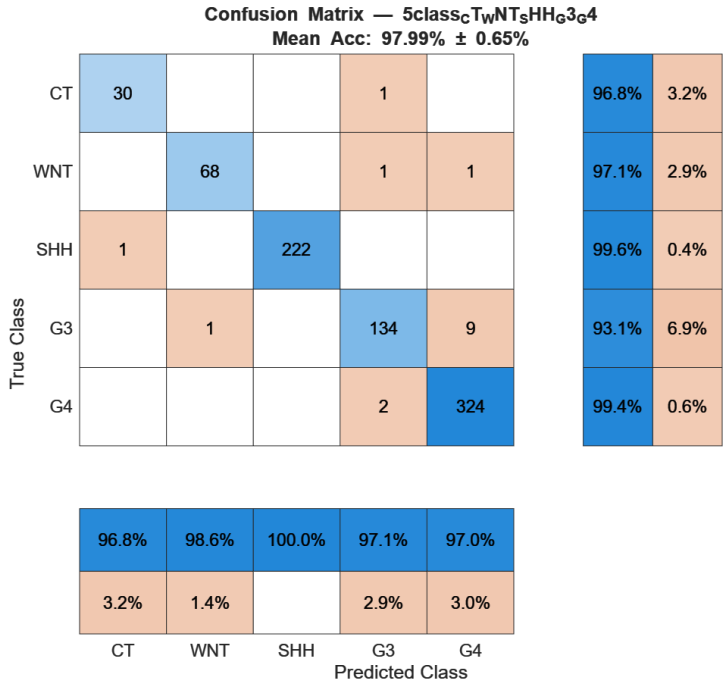

Figure 5.1: Aggregate confusion matrix for the five-class classifier across all ten cross-validation folds. Row and column summaries report class-wise recall and precision, respectively.

Examination of the confusion matrix in Figure 5.1 reveals that misclassification is concentrated along the Group 3 / Group 4 boundary: 9 G3 samples were predicted as G4 and 2 G4 samples as G3, accounting for the majority of classification errors. SHH samples were classified with the highest recall (99.6%, 222 of 223), consistent with the molecular distinctness of the SHH subgroup. CT and WNT each exhibited a single misclassification routed towards G3, producing recall values of 96.8% and 97.1% respectively, numerically modest deviations that are largely driven by the small sample sizes of these two classes ($n = 31$

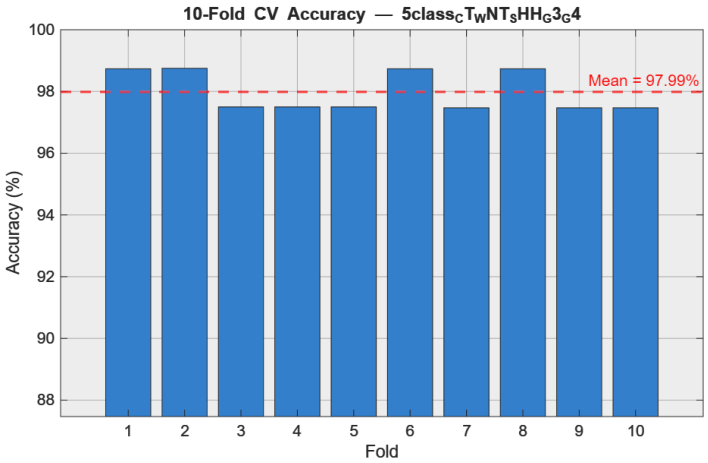

Figure 5.2: Per-fold classification accuracy for the five-class network. The dashed horizontal line indicates the mean across folds (97.99%).

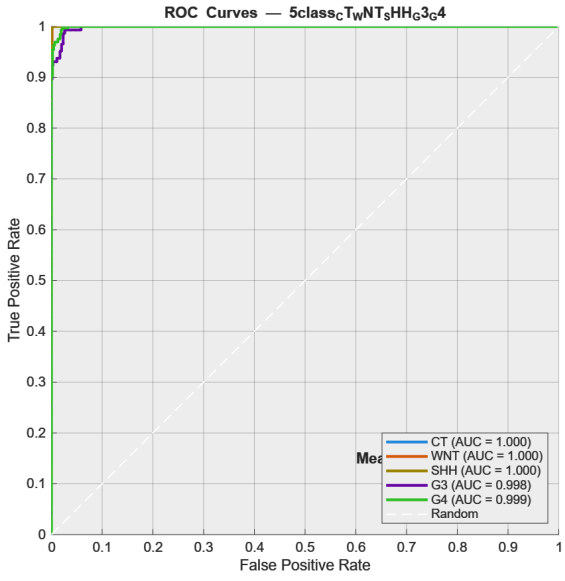

Figure 5.3: One-vs-rest ROC curves for the five-class classifier. Per-class AUC values are reported in the legend; the mean AUC across classes is 0.999.

and $n = 70$). Per-fold accuracy (Figure 5.2) remained tightly clustered between 97.47% and 98.75%, indicating that network performance is not artefactually inflated by a favorable fold partition.

5.3 *Binary Classifiers*

The WNT-vs-rest network achieved a mean cross-validated accuracy of 99.75% ± 0.53% and a mean AUC of 1.000, the highest discriminative performance among all five networks. Results are shown in Figures 5.4–5.6.

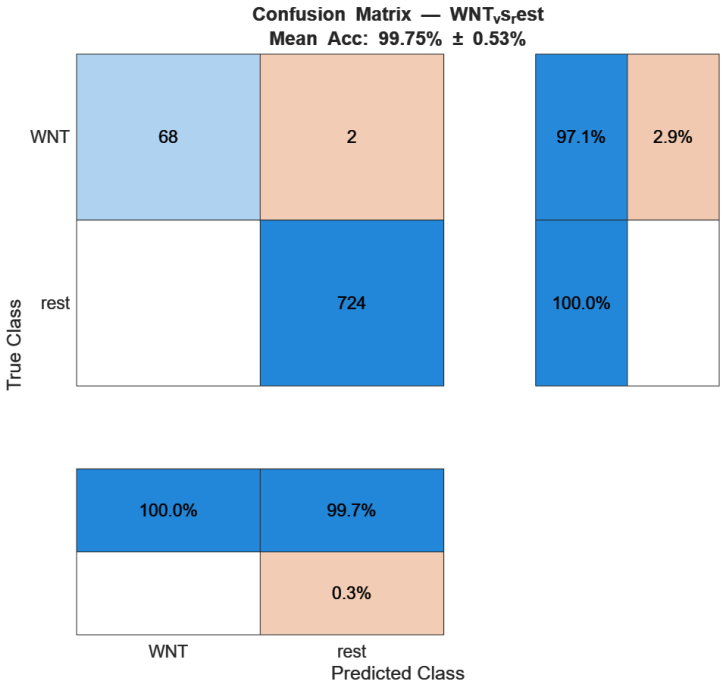

Figure 5.4: Confusion matrix for the WNT-vs-rest binary classifier aggregated across all ten cross-validation folds.

Only 2 of 70 WNT samples were misclassified across the entire cross-validation procedure, while all 724 non-WNT samples were correctly rejected. This performance is biologically expected: WNT medulloblastoma is characterized by a transcriptionally distinctive signature, which collectively produce a strongly separable expression profile. The attainment of AUC = 1.000 under stratified cross-validation confirms that the 1,139 DEG feature set retains the discriminative information required to isolate this subgroup without ambiguity.

The SHH-vs-rest network achieved the highest point accuracy of the binary classifiers, with mean 99.87% ± 0.40% and mean AUC of 0.996. Results are presented in Figures 5.7–5.9.

Only one SHH sample was misclassified and non-SHH samples was

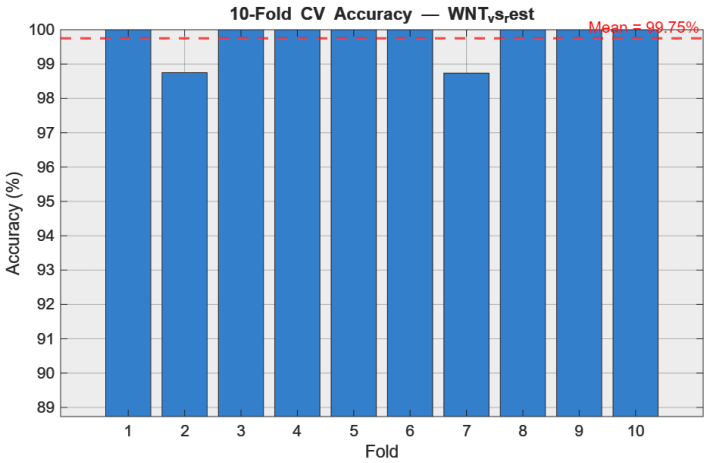

Figure 5.5: Per-fold accuracy for the WNT-vs-rest classifier. Eight of ten folds achieved 100% accuracy.

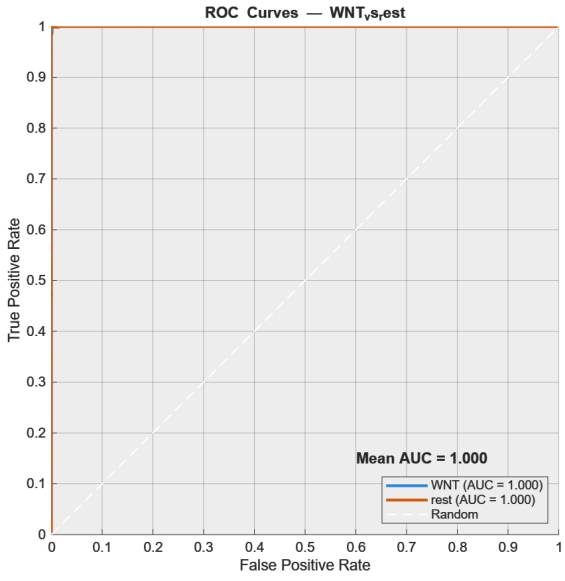

Figure 5.6: ROC curves for the WNT-vs-rest classifier. Both classes achieved AUC = 1.000, indicating perfect probabilistic separability across the cross-validation procedure.

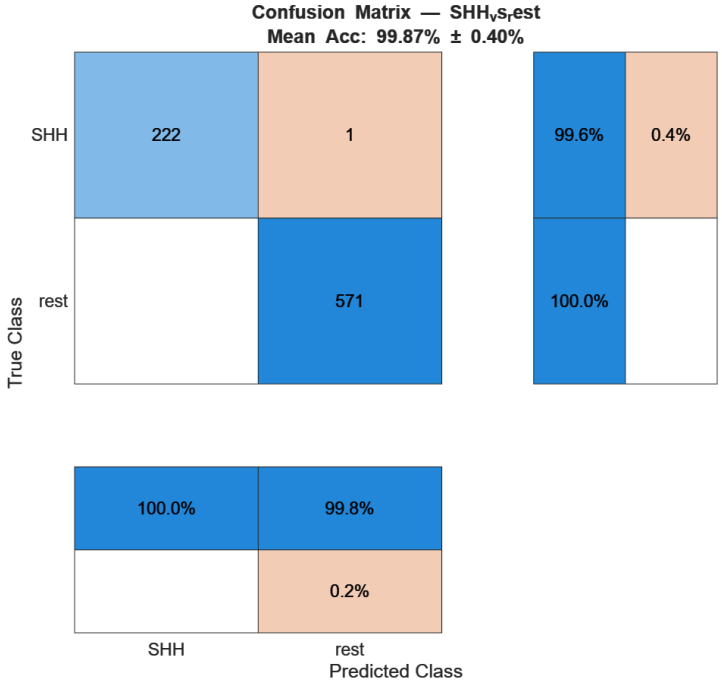

Figure 5.7: Confusion matrix for the SHH-vs-rest classifier. A single misclassification occurred across all 794 samples evaluated during cross-validation.

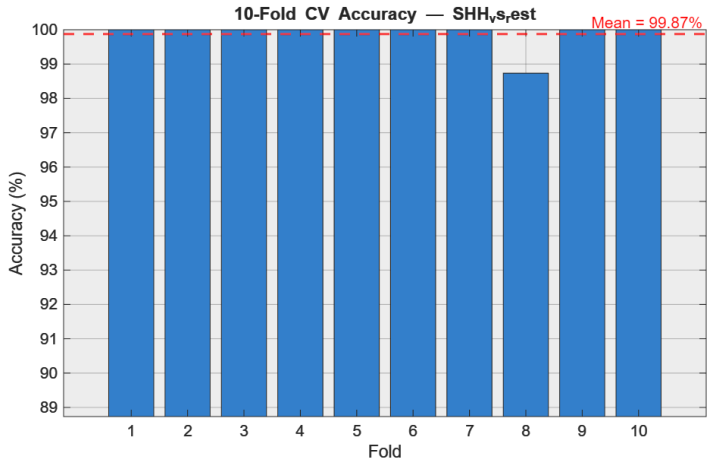

Figure 5.8: Per-fold accuracy for the SHH-vs-rest classifier. Nine of ten folds achieved perfect classification.

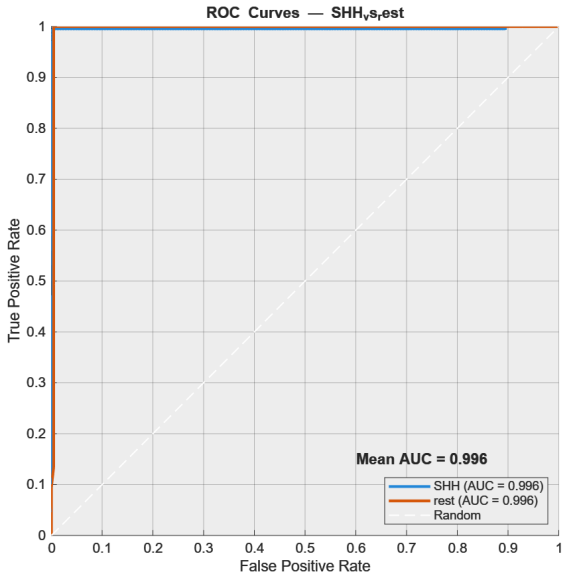

Figure 5.9: ROC curves for the SHH-vs-rest classifier, with AUC = 0.996 for both the positive and rest classes.

incorrectly assigned to the SHH class. The unambiguous transcriptomic identity identity of SHH-driven medulloblastoma, underpinned by activation of the Sonic Hedgehog signaling pathway and associated downstream effectors, accounts for the near-perfect separability observed. The narrow per-fold standard deviation (0.40%) further indicates that performance was not contingent on the specific cross-validation partition.

The G3-vs-rest network obtained a mean accuracy of 97.99% $\pm$ 1.98% and a mean AUC of 0.988, still a strong result in absolute terms, but noticeably below the performance of the other three binary classifiers and with approximately four times the fold-to-fold variability of the WNT and SHH networks. Results are shown in Figures 5.10–5.12.

The reduced performance is biologically justified and, importantly, concordant with the pattern reported by Reveles-Espinoza et al. [2025] and by earlier consensus work on medulloblastoma molecular subtyping. Group 3 and Group 4 share a substantial portion of their transcriptomic landscape, which can be seen in the partial overlap of the UMAP showing their closeness in a 2D dimensional space in the previous chapter. Consequently, a fraction of G3 tumors lie near the G3/G4 decision boundary at the transcriptomic level, and network performance on this contrast provides a lower bound for what a purely expression-based classifier can achieve in the absence of orthogonal information such as copy-number variation or methylation status. The retention of the hyperbolic tangent activation function for this network,

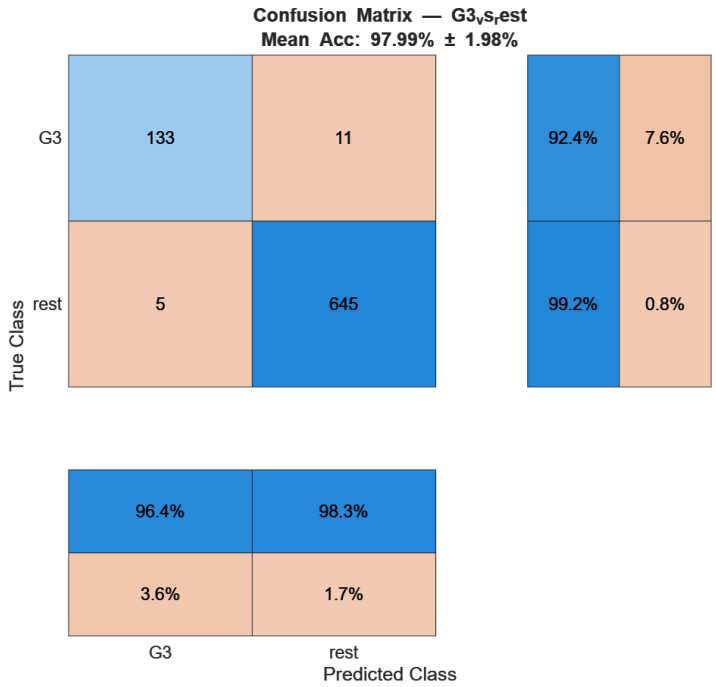

Figure 5.10: Confusion matrix for the G3-vs-rest classifier. Eleven of 144 G3 samples were misclassified as non-G3, and 5 non-G3 samples were predicted as G3.

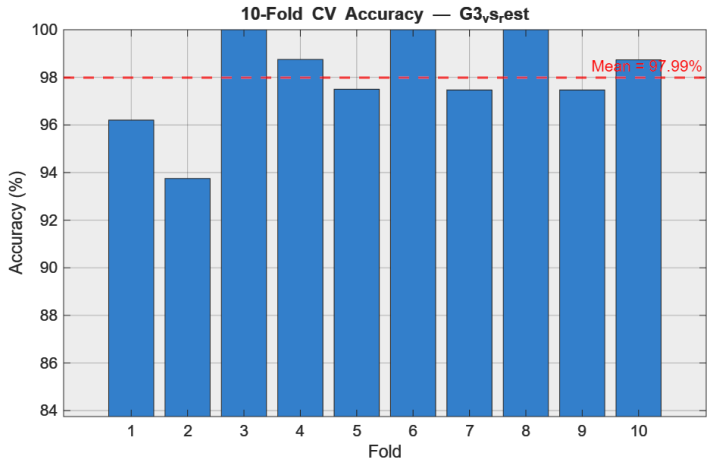

Figure 5.11: Per-fold accuracy for the G3-vs-rest classifier. The highest fold-to-fold variability among all five networks (range 93.75%–100%) reflects the transcriptomic overlap between Group 3 and Group 4.

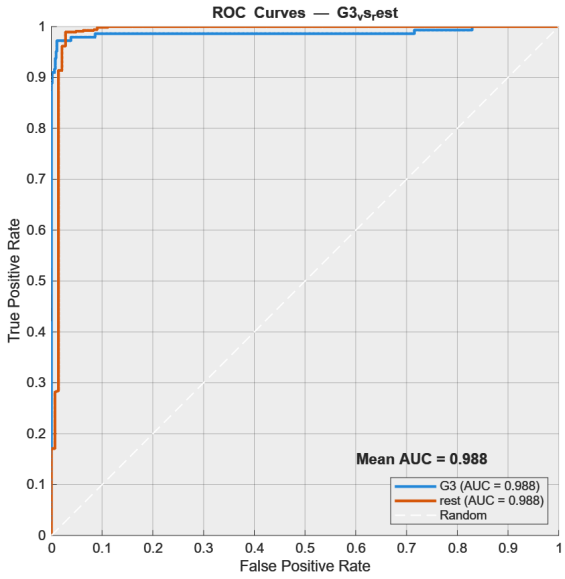

Figure 5.12: ROC curves for the G3-vs-rest classifier, with AUC = 0.988 for both classes.

rather than the ReLU used elsewhere, is motivated by exactly this ambiguity.

The G4-vs-rest network achieved a mean accuracy of 99.12% ± 1.04% and a mean AUC of 0.999. Results are presented in Figures 5.13–5.15.

The G4-vs-rest classifier recovers substantially better discriminative performance than its G3 counterpart despite operating on the same ambiguous subgroup boundary. This asymmetry is attributed to the class-size imbalance of the contrasts: G4 constitutes the largest single class in the dataset ($n = 326$), providing the network with a richer positive-class manifold during training and increasing the effective statistical power of the ReLU-based architecture. Misclassification in both directions, three G4 samples predicted as non-G4 and four non-G4 predicted as G4, remain concentrated along the G3/G4 boundary, mirroring the pattern observed in the five-class confusion matrix (Figure 5.1).

5.4 *Computational Cost Analysis*

Training time was recorded at both per-fold and per-network levels to assess the computational cost of each architecture. The relationship between cumulative training time and mean cross-validated accuracy is shown in Figure 5.16.

The G3-vs-rest network required 34.1 minutes of cumulative training time, approximately threefold the cost of the WNT-vs-rest network (11.6

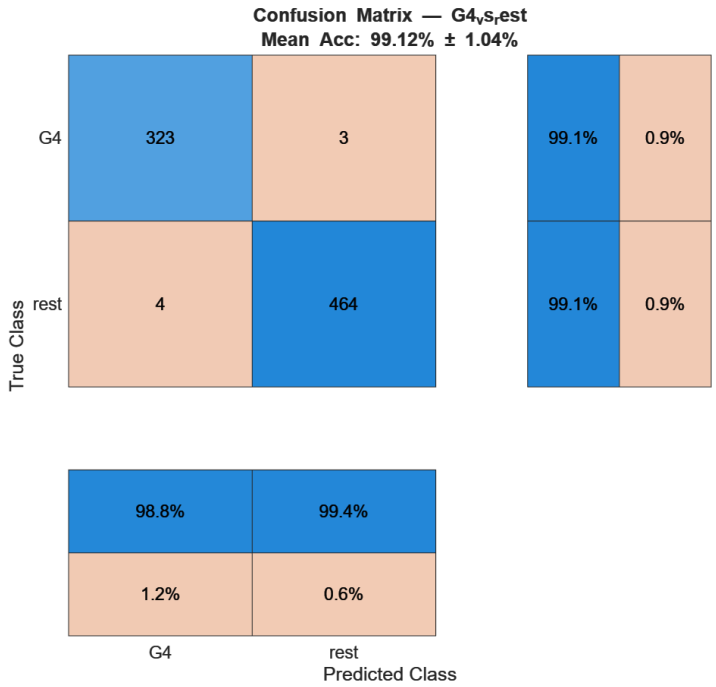

Figure 5.13: Confusion matrix for the G4-vs-rest classifier. Three G4 samples and four non-G4 samples were misclassified across the ten folds.

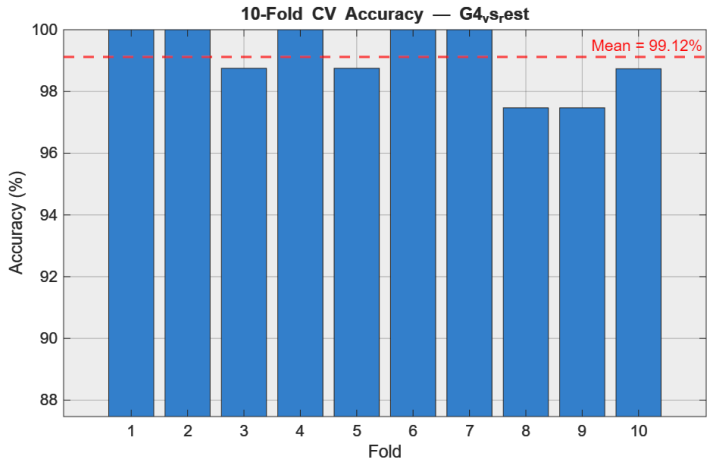

Figure 5.14: Per-fold accuracy for the G4-vs-rest classifier, ranging from 97.47% to 100% across folds.

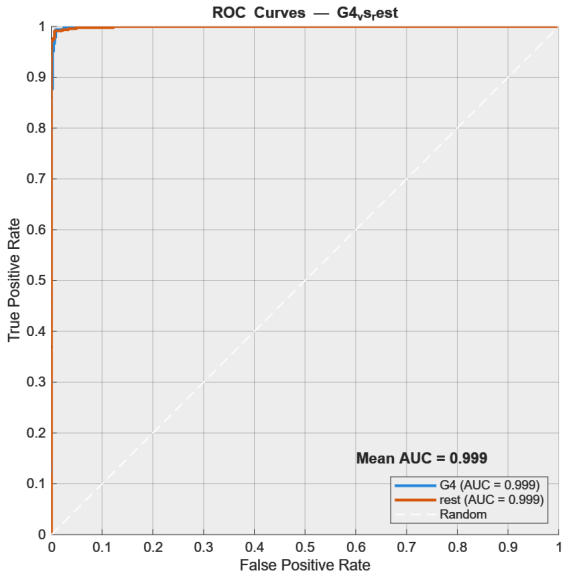

Figure 5.15: ROC curves for the G4-vs-rest classifier, with AUC = 0.999 for both classes.

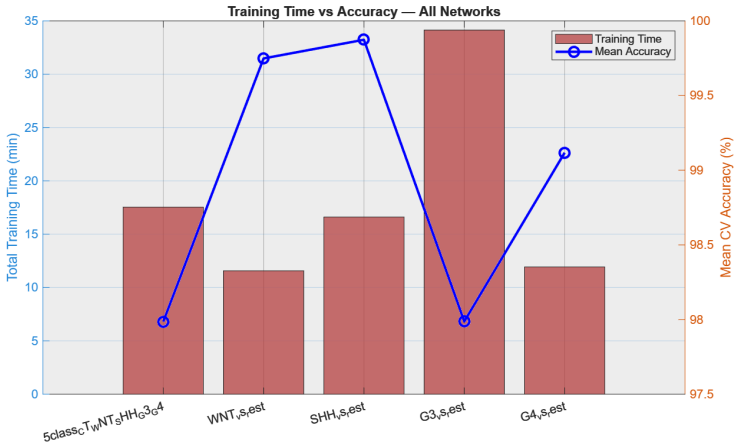

Figure 5.16: Total training time (red bars, left axis) and mean cross-validated accuracy (blue line, right axis) for each of the five networks.

minutes) and twice the cost of the five-class network (17.5 minutes). This elevated cost does not correspond to higher accuracy; on the contrary, G3-vs-rest yielded the lowest mean accuracy and the highest fol-to-fold variability. Two factors contributed to this pattern. First, the hyperbolic tangent activation, while biologically motivated for the G3/G4 boundary, saturated at its extremes and produces smaller gradients than ReLU during backpropagation, slowing convergence. Second, the transcriptomic ambiguity between Group 3 and Group 4 induces a flatter loss surface near the decision boundary, extending the number of epochs required for the network to settle on a stable solution. The remaining four networks trained between 11.6 and 17.5 minutes, a range consistent with their uniform ReLU activations and the clearer separability of their respective target classes.

5.5 *Comparison with the State of the Art*

Performance across all five architectures was benchmarked against the reference results reported Reveles-Espinoza et al. [2025], which represent the immediate methodological predecessors to the present work and share the hierarchical two-stage classification design. Table 5.2 summarizes the comparison.

Network	Reveles et al. (2025)	Present work
5-class	97.6%	97.99%
WNT vs. rest	98.4%	99.75%
SHH vs. rest	99.2%	99.87%
G3 vs. rest	98.4%	97.99%
G4 vs. rest	98.4%	99.12%

Table 5.2: Comparison of cross-validated accuracy obtained in the present work against the reference values reported by Reveles-Espinoza et al. [2025]. Both studies employ the same hierarchical ANN framework with one multi-class network and four one-vs-rest binary networks.

While the comparison above is restricted to studies sharing both the disease and the transcriptomic microarray modality, a broader point of reference is offered by Salgado et al. [2024], who evaluated an analogous ANN-based classification strategy across multiple omic data types (DNA-seq, RNA-seq, DNA methylation, miRNA-seq, and protein expression, individually and combined) for pancreatic cancer classification. Although the disease context and input modalities differ from the present work, this comparison is informative because it benchmarks classification performance against computational cost, a dimension not captured in Table 5.2. Table 5.3 summarizes these results alongside the corresponding metrics from the present work.

Despite relying on a single transcriptomic modality rather than the five-modality integration evaluated by Salgado et al. [2024], the present pipeline achieves comparable or superior accuracy (97.99% versus 92% for the closest single modality analogue, RNA-seq, and versus 96% for

Input configuration	Accuracy	Process Time (h)	FLOPs
DNA-seq	0.92	56	1.5×10^7
RNA-seq	0.92	49	4.6×10^7
Methylation	0.81	51	7.1×10^7
miRNA-seq	0.88	67	9.4×10^7
Protein	0.80	43	2.3×10^7
All subsets (combined)	0.96	78	9.5×10^8
Present work (5-class)	0.9799	-	-

Table 5.3: Comparison of accuracy, processing time, and computational cost (FLOPs) between the multi-omic ANN classifiers reported by Salgado et al. [2024] for pancreatic cancer and the present medulloblastoma classification pipeline.

the full multiomic integration at a substantially lower computational cost. This suggests that a carefully that a carefully curated, disease-specific transcriptomic feature set, selected via stringent differential expression criteria rather than broad multiomic aggregation, can recover much of the classification performance attributed to multiomic integration without its associated computational burden, an outcome of particular relevance to the resource constrained diagnostic setting motivating this work.

The present pipeline matches or exceeds the reference benchmark across four of the five networks, the G3-vs-rest network is the sole exception, underperforming the reference benchmark by 0.4 percentage points. This margin is consistent with the elevated fold-to-fold variability reported in Section 5.3 and with the intrinsic G3/G4 transcriptomic ambiguity, and it remains within the cross-fold standard deviation of both studies. Despite the molecular data being drawn from a different dataset constructed specifically for the present *in silico* design. The observed improvement is attributable to the following modifications relative to the original methodology: the introduction of batch normalization between each hidden layer, which stabilized pre-activation distributions and accelerated convergence; the restriction of tumor samples to a single dataset (GSE85217), which eliminated the batch-effect heterogeneity present in the original study, where samples were pooled from five independent datasets; the addition of dropout with rate 0.3, which mitigated overfitting in the $p > n$ regime inherent to microarray-scale feature sets; the refinement of the input feature vector, determined by the DEG analysis under the stringent configuration (FDR < 0.005, $|\log_2 \mathrm{FC}| > 2.32$) applied to the harmonized GSE85217 + GSE124814 dataset; and the adoption of z-score normalization in place of the min-max normalization used originally, as z-score standardization is the established convention for microarray-derived expression data prior to ANN training and prevents high-expression genes from numerically dominating the input weights. The combination of a larger training cohort (794 samples versus the smaller cohort used in the original study) with a more stringent DEG

selection criterion yielded a feature set that is both more focused and better supported statistically, producing the observed gains in cross-validated accuracy relative to the reference benchmark.

6 Discussions and Conclusions

Contents

6.1 Summary of Findings **129**
6.2 Limitations **132**
6.2.1 Subtyping of Medulloblastoma 132
6.2.2 Normalization Strategies 134
6.2.3 Single Platform and Dataset Validation 135
6.2.4 Mono-omic Experimental Design . . . 136
6.2.5 Wet Laboratory Confirmation 137
6.3 Future Work **137**
6.3.1 Molecular Subtype of Medulloblastoma´s Subgroups 138
6.3.2 Normalization Strategies Benchmark 139
6.3.3 Molecular Linear Separability Hypothesis 140
6.3.4 Gene Ontology and Gene Set Enrichment Analysis 140
6.3.5 Migration to Open-Source Pipelines . 141
6.3.6 Multiomics Data Integration 142
6.4 Final Considerations **142**

The work presented in the preceding chapters established a reproducible computational pipeline for the molecular subgroup classification of medulloblastoma from microarray transcriptomic data, extending the methodological framework introduced by Reveles-Espinoza et al. [2025] to a substantially larger and independently constructed dataset. By integrating cross-platform annotation harmonization, robust differential expression analysis, and hierarchical feedforward artificial neural network architecture, the pipeline addresses several recurring challenges of microarray-based classification studies: platform heterogeneity and batch effects, the curse of dimensionality inherent to high-throughput transcriptomic feature spaces, and the class imbalance commonly observed across molecular subgroups in publicly available cohorts. Beyond classification accuracy, the practical value of the pipeline lies in its selection steps: by reducing the full transcriptome

to a compact, subgroup-discriminative set of differentially expressed genes, the same preprocessing and feature selection protocol that drives the classifier also constitutes a systematic route to candidate biomarker identification. Because the approach operates on a single transcriptomic (mono-omic) data source rather than requiring multi-omic integration, the resulting gene panels are directly translatable to lower-cost, targeted assays and could in principle guide the design oligonucleotide based reagents, custom probes or primer sets, for subgroup assignment in resource constrained settings and even across multiple types of cancer considering the same or similar type of data and experimental design considerations.

The classification results reported in the previous chapter support three principal conclusions, which frame the synthesis presented in the remainder of this chapter. First, the feedforward ANN pipeline achieves cross-validated accuracy at or above 97.99% across all five classification tasks, meeting or exceeding the current state of the art for transcriptomic medulloblastoma subgroup classification. Second, the residual classification errors are biologically interpretable rather than stochastic, clustering systematically along Group 3 / Group 4 boundary in agreement with the well documented transcriptomic overlap between these subgroups and consistent with the dimensional structure observed in the UMAP projection (Figure 2.10). Third, the combined regularization strategy, batch normalization, dropout, L_2 weight decay, and stratified cross-validation, successfully controls overfitting in the high-dimensional regime where feature count exceeds sample count, supporting the validity of this architectural configuration for transcriptomic classification tasks in other cancer types.

Methodologically, the present work supports the position that careful upstream design, rather than algorithmic complexity, is the principal determinant of classification performance. The combination of a shared-genome intersection between annotation platforms, stringent DEG selection governed by log_2FC and false discovery rate thresholds, UMAP-based assessment of subgroup separability, and regularization strategy combining batch normalization, dropout, and L_2 weight decay produces a pipeline that matches or exceeds state of the art benchmarks while remaining computationally tractable on standard workstation hardware. This stands in deliberate contrast to the multi-omic deep learning approach of Salgado et al. [2024], in which broader data integration was offset by substantially greater computational demand without commensurate gains in classification fidelity for a comparably scoped classification task.

6.1 Summary of Findings

The first principal finding is that the proposed pipeline achieves cross-validated accuracy at or above 97.99% across all five classification tasks, meeting or exceeding the current state of the art for transcriptomic medulloblastoma subgroup classification. The five-class network attained a mean cross-validated accuracy of 97.99%, while the four binary one-vs-rest networks reached 99.75% (WNT), 99.87% (SHH), 97.99% (G3), and 99.12% (G4), with mean AUC values ranging from 0.988 to 1.000. These figures should be read in the context of the validation strategy under which they were obtained: stratified ten-fold cross-validation provides k independent estimates of out-of-sample prediction error by partitioning the data into ten complementary folds of approximately equal size, training the network on $k-1$ folds and evaluating on the remaining fold, and reporting the arithmetic mean of the resulting estimates [1]. By construction, this procedure ensures that the evaluation set disjoint from the training set at every iteration, and by stratifying the partition on the class label, the procedure further preserves the empirical class-frequency distribution within each fold, a non-trivial requirement given the substantial class imbalance of the GSE85217 cohort. The narrow per-fold dispersion observed across all networks (standard deviation between 0.53% and 1.93% for all networks) confirming that the reported accuracies are not artifacts of favorable single train/test split but stable estimate of generalization performance across reshuffled partitions of the data.

The second principal finding is that the residual classification errors are biologically interpretable rather than stochastic. Examination of the aggregated confusion matrix for the five-class network reveal that nine Group 3 samples were predicted as Group 4 and two Group 4 samples as Group 3, accounting for the majority of classification errors, while SHH samples were classified with the highest recall (99.6%, 222 of 223), consistent with the molecular distinctness of the SHH subgroup. This systematic concentration of errors along the Group 3/Group 4 boundary reproduces a pattern extensively documented in the molecular subtyping literature [2] and reflects genuine transcriptomic overlap between the two non-WNT, non-SHH subgroups rather than a deficiency of the classifier. The same boundary ambiguity is independently observed in the unsupervised UMAP projection presented in Figure 4.14, in which Group 3 and Group 4 occupy adjacent regions of the embedding with partial inter-cluster contact, whereas WNT, SHH, and the cerebellar control class form well-separated islands. The choice of Uniform Manifold Approximation and Projection [3] as the unsupervised diagnostic is itself methodologically motivated: linear projections such as principal

[1] Osval Antonio Montesinos López, Abelardo Montesinos López, and Jose Crossa. Fundamentals of artificial neural networks and deep learning. In *Multivariate statistical machine learning methods for genomic prediction*, pages 379–425. Springer, 2022

[2] Abhinav Dey, Anshu Malhotra, and Neha Garg. Medulloblastoma methods and protocols. Springer Nature, 2022. ISBN 978-1-0716-1952-0

[3] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018

component analysis (PCA) identify directions of maximal variance and consequently obscure finer-scale structure that is not aligned with the dominant variance axes, a limitation that has driven the development of nonlinear neighbor graphs algorithms over the past decade [4]. UMAP, in particular, constructs a weighted nearest-neighbor graph in the original feature space and then optimizes a low-dimensional representation that preserves local neighborhood topology rather than absolute distances. This neighborhood preserving property allows data dense regions to be expanded in the embedding while retaining the local relational structure of the original space, which is precisely the property required to reveal subtle subgroup boundaries in transcriptomic cohorts where global variance is dominated by a small number of strongly differentiated classes. The same property has driven the widespread adoption of UMAP in single-cell genomics, bulk transcriptomics, and population genetics, where it has been shown to recover demographic structure, ancestry composition, and co-variation between genetics, geography, and phenotype that linear methods systematically miss [5]. In the present context, the embedding therefore functions as an unsupervised counterpart to the confusion matrix: WNT, SHH, and the cerebellar controls form visually distinct neighborhoods because their transcriptomic profiles are locally distinct from those of every other class, while the partial overlap between Group 3 and Group 4 neighborhoods reflects the genuine local proximity of their gene expression profiles in the original 1,139-dimensional feature space. The convergence of these two independent diagnostics, supervised classification error and unsupervised dimensional structure, provides cross-method evidence that the residual classifier uncertainty corresponds to a real biological signal rather than a methodological artifact.

[4] Alex Diaz-Papkovich, Luke Anderson-Trocmé, and Simon Gravel. A review of umap in population genetics. *Journal of Human Genetics*, 66(1):85–91, 2021

[5] Howard M Cann, Claudia De Toma, Lucien Cazes, Marie-Fernande Legrand, Valerie Morel, Laurence Piouffre, Julia Bodmer, Walter F Bodmer, Batsheva Bonne-Tamir, Anne Cambon-Thomsen, et al. A human genome diversity cell line panel. *Science*, 296(5566):261–262, 2002

The third principal finding is that the regularization strategy adopted in this work successfully controlled overfitting in the high-dimensional regime where feature count exceeds sample count. The combination of three architectural and optimization-level mechanisms proved decisive in this regard. At the architectural level, each hidden block in the networks coupled a `fullyConnectedLayer` with a `batchNormalizationLayer` and a `dropoutLayer` with rate 0.3, deactivating a random 30% of units at every forward pass and stabilizing the distribution of pre-activation across mini-batches. At the optimization level, the Adam optimizer was configured with an initial learning rate of 10^{-3} reduced by factor of 0.5 every 20 epochs under a piecewise schedule, an L_2 weight decay of 10^{-4}, mini-batch size of 32, and per-epoch reshuffling of the training data; these settings allowed rapid early convergence while progressively constraining parameter updates as training advanced. The resulting per-fold accuracies

remained tightly clustered between 97.47% and 98.75% for the five-class network, with no systematic divergence between training and validation curves across folds, indicating that the network learned a generalizable mapping from the 1,139-dimensional differentially expressed gene fingerprint to the five-class label space rather than memorizing the training partitions. This configuration therefore stands as a validated template for transcriptomic classification tasks operating under similar high-dimensional, moderate-sample-size conditions in other cancer types.

A second principal finding concerns the validation of the stringent differential expression thresholds adopted in the present pipeline. The primary configuration, following the criteria of Castillo-Rodríguez et al. [2018] and Reveles-Espinoza et al. [2025] (FDR < 0.005, $|\log_2 \mathrm{FC}| > 2.32$), produced a discriminative gene set of 1,139 unique Ensembl identifiers across the ten pairwise contrasts. The same `limma/voom` fit, re-evaluated under the more permissive configuration of Arora et al. [2026] and Dhar and Kallapura [2022] (FDR < 0.05, $|\log_2 \mathrm{FC}| > 0.3$), yielded substantially broader gene sets: subgroup versus control contrasts ranged from 22,112 (G4–CT) to 23,067 (WNT–CT) DEGs, and pairwise subgroup contrasts from 6,722 (G3–G4) to 12,122 (G4–WNT). Two contrasts illustrate the magnitude of the gap with particular clarity: the G3–G4 contrast dropped from 6,722 DEGs under the permissive thresholds to 41 DEGs under the stringent thresholds (24 up-regulated, 17 down-regulated), while the WNT–CT contrast contracted from 23,067 to 930 DEGs. These differences are not incremental but operate at an order of magnitude scale, and they determine the dimensionality and the biological character of the feature space ultimately presented to the classifier.

The finding, therefore, is that the stringent thresholds did not merely reduce the feature count for computational convenience: they produced a feature space that retained sufficient discriminative information to support cross-validated accuracy at or above 97.99% across all five networks, while remaining tractable in a moderate sample size regime ($n = 794$, $p = 1{,}139$) where the more permissive thresholds would have inverted the sample to feature ratio and forced the classifier to learn from a substantially noisier signal. The 1,139-feature configuration therefore represents an empirically validated operating point in the bias–variance trade off inherent to high-dimensional transcriptomic classification: stringent enough to exclude weakly differentiated probes whose expression differences fall within the noise envelope of microarray quantification, yet permissive enough to retain the molecular diversity required to separate five biologically distinct classes. This validates the position adopted in the methodology, that threshold selection should be governed by downstream classification performance

and biological interpretability rather than by the maximization of detected DEGs, and supports the use of the [Castillo-Rodríguez et al., 2018] and [Reveles-Espinoza et al., 2025] configuration as a defensible default for microarray-based subgroup classification studies operating under comparable cohort sizes.

A third principal finding, emerging from the integration of the unsupervised and supervised analyses in this work, is that UMAP separability of the differentially expressed feature space constitutes a useful pre-selection diagnostic for whether a given transcriptomic cohort will support accurate downstream subgroup classification. In the medulloblastoma cohort, the UMAP projection of the 1,139-feature DEG space resolved WNT, SHH, and the cerebellar control class into well-separated, approximately linearly separable neighborhoods on the UMAP plane, while Group 3 and Group 4 formed adjacent regions with partial inter-cluster contact (Figure 4.14); the supervised classification performance subsequently obtained perfect or near perfect discrimination of WNT, SHH, and CT, with residual error concentrated along the G3/G4 boundary, mirrored the embedding structure with high fidelity. Preliminary exploratory analysis conducted on colorectal transcriptomic cohort under the same preprocessing pipeline yielded a qualitatively different picture: the UMAP projection produced overlapping, non-separable clusters across the candidate molecular classes (based on normal tissue, metastatic dead cohort, and metastatic alive cohorts), and the corresponding ANN classifier trained on that feature space failed to converge to the accuracy range observed for medulloblastoma (around 60%). While that exploratory result is based on a single, unpublished side-experiment and cannot be generalized on its own, the contrast between the two cohorts suggests that the linear separability of class clusters in the UMAP embedding may serve as a pre-training diagnostic of the intrinsic discriminability of the transcriptomic data, allowing investigators to assess whether a cohort carries sufficient class-specific signal to justify the computational cost of full ANN training before that training is undertaken. This positions UMAP not solely as a visualization tool but as an empirical sample and feature space suitability check for transcriptomic classification pipelines.

6.2 *Limitations*

6.2.1 *Subtyping of Medulloblastoma*

The first limitation of the present work concerns the granularity of the classification target. The pipeline was developed and validated exclusively for the four canonical molecular subgroups of

medulloblastoma (WNT, SHH, Group 3, and Group 4) established by the 2010 consensus and reaffirmed in the WHO Classification of CNS Tumors. The finer subtype stratification reported in the GSE85217 metadata, in which Group 3 and Group 4 are subdivided into α, β, and γ subtypes and the SHH subgroup into four subtypes (α, β, γ, δ), was not addressed. This finer schema emerged from a second generation of molecular characterization studies in the latter half of the 2010s, when independent groups applied advanced clustering techniques to larger cohorts in order to resolve the clinical and biological heterogeneity of the original four-subgroup framework. Northcott et al. [2017] applied t-SNE and DBSCAN to the methylation profiles of 740 Group 3 and Group 4 samples and reported eight subclasses (I–VIII); Schwalbe et al. [2017] stratified Group 3 and Group 4 into high- and low-risk categories using methylation consensus non-negative matrix factorization on 273 tumors; and Cavalli et al. [2017] described six intertumoral subtypes (α, β, γ within both Group 3 and Group 4) using similarity network fusion on combined microarray and methylation data from 470 specimens. The Cavalli subtype labels are the schema provided in the GSE85217 metadata and are increasingly cited as a more clinically actionable framework, as they capture risk–relevant heterogeneity that are the four subgroup classification collapses together [6].

Extending the present pipeline to the twelve subtype schema was not pursued because the per-subtype sample sizes available within GSE85217 are substantially smaller than those of the four canonical subgroups and fall close to or below the minimum statistical power threshold of 63 samples er group derived in the methodological chapter. As shown in Figure 6.2.1, only six of the twelve subtypes (SHH_alpha, SHH_delta, Group3_alpha, Group4_alpha, Group4_beta, Group4_gamma) exceed the 63-sample threshold, while the remaining seven subtypes range from 21 (WNT_beta) to 49 (WNT_alpha) samples, sample sizes at which DEG selection becomes unstable and ANN training is highly susceptible to overfitting and high variance per-fold performance under the same ten fold cross-validation regime used in this work. The four subgroup classification therefore represents an appropriate level of granularity given the available cohort, the chosen validation methodology, and the scope of the present thesis. The extension of the pipeline to the finer subtype schema, which would require either larger cohorts, alternative validation strategies, or hierarchical classifier designs in which the four subgroup network is followed by within subgroup subtype classifiers, is reported as a direction for future works.

[6] Yujin Suk, William D Gwynne, Ian Burns, Chitra Venugopal, and Sheila K Singh. Childhood medulloblastoma: an overview. *Medulloblastoma: Methods and Protocols*, pages 1–12, 2022

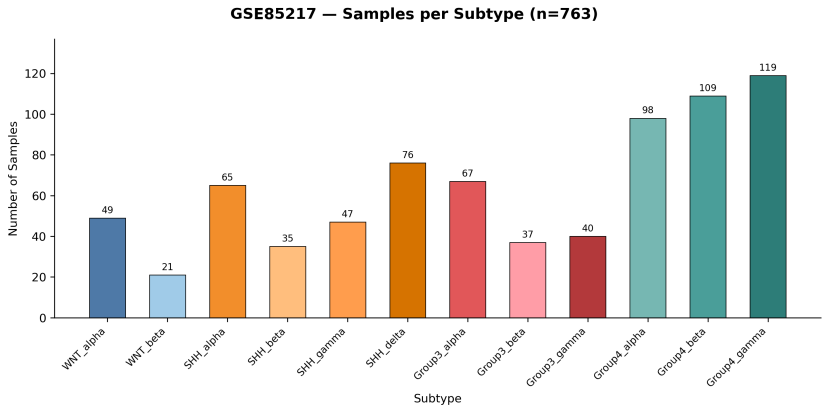

Figure 6.1: Sample distribution across the twelve molecular subtypes reported in the GSE85217 metadata, following the schema of Cavalli et al. [2017]. Total $n = 763$. Several subtypes fall below the 63-sample minimum threshold derived from the power analysis, which motivated the restriction of the present work to the four canonical subgroups.

6.2.2 *Normalization Strategies*

A second limitation concerns the choice of input-feature normalization strategy and the absence of a formal benchmark against alternatives. The 1,139-feature input vector entering the network is constrained in dimensionality by the differentially expressed gene set produced upstream, a fixed property of the pipeline given the stringent selection thresholds adopted in the DEG analysis, but the numerical values entering the input layer depend critically on the normalization applied to that feature vector immediately prior to training. The reference methodology of Reveles-Espinoza et al. [2025] applied min–max rescaling, mapping each feature to the closed interval $[-1, 1]$. The present work departed from that choice and adopted z-score standardization via `zscore()`, on the methodological grounds that z-score preserves the shape of the gene-wise expression distribution while centering each feature at zero mean and unit variance, which prevents high-expression genes from numerically dominating the input weights during gradient descent and is robust to the presence of extreme `logFC` values that would otherwise compress the rescaled range under min–max. While this rationale supports the choice, the present work does not include a controlled benchmark in which the five network architectures are retrained under alternative normalization strategies and the cross-validated accuracies compared under matched hyperparameters and seeds.

The space normalization strategies that would warrant inclusion in such benchmark is well-defined and directly supported by the MATLAB `normalize()` function [7], which exposes z-score (`"zscore"`, default), robust z-score (`"zscore","robust"`, centering on the median and scaling by the median absolute deviation), min–max range rescaling (`"range"`), unit-norm scaling (`"norm"`), median–interquartile-

[7] MathWorks. normalize - normalize data, 2026. URL https://www.mathworks.com/help/matlab/ref/double.normalize.html

range standardization (`"medianiqr"`), and centering or scaling alone (`"center"`, `"scale"`) as separate options. Each of these strategies operates on the 1,139-feature input vector identically in dimensionality but produces a numerically distinct input distribution, and there is no a apriori theoretical guarantee that z-score is the optimal choice for the present architecture, the present feature set, or the present cohort. Consequently, the present work cannot claim that z-score is the optimal normalization choice for transcriptomic ANN classification under the conditions established in this thesis; it can only claim that z-score, applied in conjunction with batch normalization, dropout, and L_2 weight decay, was sufficient to achieve cross-validated accuracy at or above 97.99% across all five networks.

6.2.3 *Single Platform and Dataset Validation*

A third limitation concerns the breadth of the validation cohort. The classification pipeline was developed, trained, and validated on a single transcriptomic dataset, GSE85217, hybridized on a single microarray platform, the Affymetrix Human Gene 1.1 ST array under GPL annotation GPL22286, complemented by cerebellar control samples from GSE124814 hybridized on the same physical chip but annotated under GPL16977. While the cross-platform harmonization step resolved the annotation version discrepancy between the two GPL identifiers and the resulting feature space is internally consistent, no independent external validation cohort, hybridized on a different platform or acquired by a different consortium, was processed through the pipeline to confirm the generalization of the trained networks beyond the GSE85217 / GPL22286 substrate. Cross-validated accuracy within a single cohort, however stable across folds, is a weaker form of evidence than performance on an unseen external cohort, and the present work cannot rule ut the possibility that some fraction of the reported classification accuracy is attributable to cohort-specific covariate structure or platform specific probe behavior rather than to the underlying biological signal of the four molecular subgroups.

This limitation is partially mitigated by an aspect of the pipeline design that distinguishes it from its immediate methodological predecessor. Reveles-Espinoza et al. [2025] trained their classifier on samples drawn from multiple datasets, but those datasets were unified under a single physical platform, the Affymetrix Human Genome U133 Plus 2.0 Array (GPL570), meaning that platform level sources of variation were homogeneous across the entire training corpus. The present work inverts that configuration: a single dataset is used, but the cerebellar controls are drawn from a GPL16977-annotated subset that, while hybridized on the same physical Affymetrix HuGene-1.1 ST

chip as the cancer samples, was annotated under a distinct Ensembl release. The pipeline was therefore obliged to demonstrate cross-annotation robustness even within its single-cohort training regime, and the 89% gene level overlap between the two GPL annotations was preserved as the working feature space. This is not equivalent to true external multi-cohort validation, but it does establish that the pipeline is not silently dependent on a single annotation version. A formal external validation, in which the trained networks are applied to a medulloblastoma cohort hybridized on a structurally different platform, for example, the Illumina HumanHT-12 BeadChip used in several methylation-paired transcriptomic cohorts, or RNA-Seq derived expression matrices from independent consortia, would be the natural following step in future works.

6.2.4 *Mono-omic Experimental Design*

A fourth limitation concerns the dimensional scope of the molecular data underlying the classifier. The pipeline operates exclusively on transcriptomic measurements derived from a single quantification modality, gene level RMA normalized expression intensities from Affymetrix microarray hybridization, and does not integrate any complementary molecular layer. The four canonical medulloblastoma subgroups, however are biologically defined not solely by their gene-expression signatures but by the convergence of multiple molecular phenomena: DNA methylation patterns, somatic mutation and copy-number landscapes, micro-RNA regulatory programs, and downstream proteomic activity each contribute distinct and non-redundant information about the underlying tumor biology. By restricting itself to the transcriptomic axis, the present classifier necessarily projects this multi-dimensional biology onto a single information channel and treats the resulting expression fingerprint as a sufficient statistic for subgroup membership. The high cross-validated accuracy obtained suggest that, for the four subgroup classification target, the transcriptomic projection does in act retain enough discriminative information to resolve the classes, but the present work cannot speak to whether multi-omic integration would further improve the residual G3/G4 boundary, support the finer subtype stratification discussed previously, or generalize more robust external cohorts.

The contrasts with Salgado et al. [2024] is instructive in this regard. That work assembled a multi-omic input space drawn from The Cancer Genome Atlas via the Genomic Data Commons, integrating ASCAT-derived copy number profiles, bulk RNA-Seq expression, micro-RNA-Seq, DNA methylation arrays, and peptide level proteomic

quantification into a single deep learning classification framework. The breadth of that input space allows the classifier to draw evidence from molecular layers that operate on distinct biological timescales, genomic alterations as long term causes, transcriptomic and micro-RNA signals as intermediate regulatory states, methylation as an epigenetic record, and proteomic abundance as a near functional readout, and is in principle better positioned to capture the full biological identity of a tumor subgroup than any single modality approach. The present work cannot make that claim. Its scope was deliberately restricted to a single modality in order to isolate and validate the microarray to ANN classification pipeline as a self contained methodological unit, and to enable the direct benchmarking against Reveles-Espinoza et al. [2025] that defined the comparative framework of this thesis. Multiomic extension of the pipeline, whether through early fusion of harmonized molecular matrices, late fusion via ensembled subgroup-specific classifiers, or attention-based architectures designed for heterogeneous biological inputs, represents a substantial next stage of development and is reported as a direction for future work.

6.2.5 *Wet Laboratory Confirmation*

A final limitation, is that the discriminative gene set identified through the differential expression analysis was validated exclusively in silico. The 1,139 features retained under the stringent FDR and $\log_2$ fold-change thresholds carry strong statistical evidence of subgroup specific disregulation and demonstrate substantial discriminative power for separating the four molecular subgroups from cerebellar controls, as confirmed by the cross-validated classification performance reported previously. However, no wet-lab confirmation, whether through quantitative PCR validation of the representative DEGs, immunohistochemical staining of the corresponding protein products, or functional perturbation experiments, was conducted within the scope of this thesis. The present work therefore establishes the discriminative gene set as a computationally validated molecular fingerprint, but its biological significance and translational potential remains to be confirmed experimentally.

6.3 *Future Work*

The most immediate extension of this work is the validation of the proposed methodology beyond medulloblastoma, applying the same preprocessing, differential expression, and classification protocol to other transcriptomically characterized cancers, and other similar

types of data such as RNA-seq or other omic format to complement the works of the state of the art. The objective is not merely to reproduce the present accuracy in a new disease, but to consolidate the individual stages into a single, disease-agnostic pipeline whose parameters and decision points are fixed in advance rather than tuned per dataset. Homogenizing the methodology in this way is what makes results comparable across diseases: a consistent protocol ensures that performance differences reflect genuine biological separability between molecular subtypes rather than incidental analytical choices. This consolidation effort underpins the more specific directions that follow, and finds its concrete realization in the open-source pipeline described below, which packages the homogenized workflow into a reusable form for systematic multi-disease analysis.

6.3.1 *Molecular Subtype of Medulloblastoma´s Subgroups*

A first natural extension of the present pipeline is its application to the finer subtype stratification of Group 3 and Group 4 reported by Suk et al. [2022] and adopted in the GSE85217 metadata, in which each of the two non-WNT, non-SHH subgroups its subdivided into α, β, and γ subtypes. As discussed previously, this finer schema is increasingly cited as more clinically actionable framework than the four-subgroup classification, but the per-subtype sample sizes available within GSE85217 fall close to or below the 63 sample power threshold for several classes. The central methodological question, therefore, is whether the same five network architecture validated the present work scales directly to the twelve-class problem, or whether the smaller and more imbalanced per-subtype cohorts require a redesigned classification topology.

Two architecture strategies are worth evaluating in this regard. The first is a flat extension in which the existing five-class network is reconfigured for twelve-class output and trained end to end on the full subtype label space, with the regularization strategy validated (batch normalization, dropout, L_2 weight decay, stratified k-fold cross-validation) carried over unchanged. The second is a hierarchical design more closely matched to the underlying biological taxonomy, in which the four subgroup classifier established in this work serves as a first stage routing network and a second tier of within subgroup classifier, trained specifically on the SHH, Group 3, and Group 4 subtype labels, performs the finer stratification on samples already assigned to the relevant parent subgroup. The hierarchical configuration has the advantage of confining the small sample subtype problem to its relevant parent subgroup, where the available DEGs are drawn from within subgroup contrasts rather than the broader pan-subgroup feature space, and is consistent with the two-stage methodology inherited

from Reveles-Espinoza et al. [2025]. Comparative evaluation of the two strategies under matched cross-validation conditions would establish whether the hierarchical design recovers classification accuracy where the flat extension fails, or whether both configurations remain limited by the underlying sample size constraint and require external cohort augmentation to resolve.

6.3.2 *Normalization Strategies Benchmark*

A second extension addresses the absence of a controlled normalization benchmark. The present work adopted z-score standardization as the preprocessing step immediately upstream of the ANN input layer, departing from the min–max rescaling of the state of the art, but not comparative evaluation was performed against the broader space of normalization strategies. A systematic benchmark covering the five most relevant alternatives (min–max range rescaling to $[-1, 1]$, standard z-score, robust z-score (median centering with median absolute deviation scaling), interquartile-range standardization, and gene-wise quantile normalization) would determine empirically which preprocessing choice maximizes cross-validated classification performance for the present architecture, feature set, and cohort, and would establish whether the z-score result reported in this work represents an optimal or defensible point on the preprocessing landscape.

For the comparison to be methodologically meaningful, the benchmark must hold all other factors constant: the same 1,139-feature DEG matrix, the same five-network architecture (one five-class plus four binary one-vs-rest), identical Adam optimizer configuration, L_2 weight decay, dropout rate, batch normalization placement, and stratified ten-fold cross-validation partitions, with only the upstream normalization function varying across runs. Reporting per-strategy mean and standard deviation of cross-validated accuracy across all five networks, paired with confusion matrices and ROC–AUC values for the residual G3/G4 boundary, would allow the identification of normalization strategies that improve overall accuracy, those that specifically alleviate the G3/G4 confusion, and those that achieve neither, a three way distinction that is biologically and methodologically more informative than a single accuracy ranking. Such benchmark would close the residual methodological uncertainty around the normalization choice and provide a defensible empirical default for transcriptomic ANN classification studies under comparable feature-set and cohort-size conditions.

6.3.3 Molecular Linear Separability Hypothesis

A third direction concerns the systematic application of the pipeline to additional tumor types with established molecular heterogeneity, specifically a colorectal cancer cohort that came from GSE17538 dataset with preliminary exploratory work on this colorectal cohort (which followed the previously established pipeline) yielded UMAP projections in which the candidate molecular classes based on survival status failed to resolve into linearly separable clusters, and the corresponding ANN classifier did not approach the accuracy range observed for medulloblastoma by almost 30% accuracy difference. This contrast suggest a testable hypothesis, that linear separability of class clusters in the UMAP embedding of the DEG-filtered feature space functions as an empirical baseline indicator of whether the downstream ANN pipeline will achieve strong classification performance, prior to any training. Formalizing this hypothesis through cross-tumor application of the pipeline, and evaluating, in cohorts where baseline cluster separability is weak, whether targeted interventions such as more stringent DEG filtering, alternative feature-selection schemes, or supervised dimensionality reduction can recover separability and downstream classification accuracy, would establish UMAP separability as a principled pre-screening diagnostic for transcriptomic classification studies and clarifying the operational boundary between cohorts where the pipeline generalizes and cohorts where additional preprocessing investment is required.

6.3.4 Gene Ontology and Gene Set Enrichment Analysis

A fourth extension addresses the gap between the classifier´s statistical decision and the biological mechanisms underlying medulloblastoma subgroup identity. The 1,139 differentially expressed genes that constitute the input feature space where selected on purely statistical grounds (FDR adjusted significance and effect size magnitude) and their collective predictive power has been demonstrated empirically by the cross-validated classification accuracy reported in the present work. Differential expression analysis at the gene level, however, does not capture the functional implications of expression disregulation: a list of significantly altered transcripts cannot, on its own, explain which cellular processes are coordinately disrupted across the subgroups, nor whether the classifier is leveraging coherent biological programs or merely a statistically convenient collection of discriminative probes. Pathway analysis offers the appropriate analytical bridge; by aggregating individual genes into biologically meaningful units (collections of actors connected through functional relationships such as protein–protein interactions, signaling cascades, or shared metabolic

substrates) it moves the interpretation from the gene level to the level of disrupted cellular processes that sustain each subgroup´s phenotype.

A staged enrichment workflow applied to the present DEG sets would clarify this biological substrate. Gene Set Enrichment Analysis (GSEA) [8], using the Kolmogorov–Smirnov statistic [9] against the ranked DEG lists derived from each pairwise subgroup contrast, would identify the Gene Ontology biological process, molecular function, and cellular component categories most strongly enriched in each subgroup-specific signature. A subsequent extension to topological pathway analysis, in which the position, direction, and signal type of each gene within curated biological pathways is incorporated into the enrichment statistic, would sharpen the interpretation by distinguishing genes operating as upstream regulators from those acting as downstream effectors within the same disrupted program. Two specific outcomes would substantially strengthen the present work: first, the identification of subgroup-specific functional signatures that correspond to the established molecular biology of WNT, SHH, G3, and G4 medulloblastoma, for example, recovery of WNT signaling enrichment in the WNT subgroup, Hedgehog pathway enrichment in SHH, and MYC-related transcriptional programs in G3, which would constitute a form of biological face validity for the discriminative gene set; and second, the identification of pathway-level features that distinguish G3 from G4 along the boundary where the classifier accumulates its residual errors, potentially clarifying whether the G3/G4 confusion identifier reflects genuinely shared pathway disregulation or distinct pathway signatures that the gene level features fail to fully resolve.

[8] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the national academy of sciences*, 102(43): 15545–15550, 2005

[9] AN Kolmogorov. Sulla determinzione empírica di una legge di distribuzione, ist i tal. *Atti. Giorn*, 4:88–91, 1933

6.3.5 *Migration to Open-Source Pipelines*

A fifth direction concerns the migration of the ANN training stage from MATLAB to an open-source deep learning framework. The present work was implemented in MATLAB 2025b for direct architectural parity with Reveles-Espinoza et al. [2025], but porting the five-network configuration to Python would deliver three practical benefits without altering the underlying methodology. First, it would remove the licensing constraint that currently limits reproducibility of the training stage to institutions with active MATLAB and Deep Learning Toolbox licenses, allowing the pipeline to be re-executed on standard scientific Python infrastructure available without cost. Second, it would unify the tool chain end-to-end: the upstream R-based bioinformatics preprocessing and the Python based UMAP diagnostic stages already operate with open source environments, and consolidating the ANN stage within the same ecosystem would eliminate the inter-language

data handoffs (CSV exports, manual matrix orientation alignments) that introduce the most common implementation errors in the current pipeline. Third, it would provide direct access to the broader open source ecosystem of pre-trained transcriptomic models, attention based architectures, and multiomic integration libraries that are developed predominantly outside the MATLAB environment , positioning the pipeline for the multiomic extension proposed and for collaborative reuse by groups without commercial software infrastructure.

6.3.6 *Multiomics Data Integration*

A sixth and final direction frames the longer term horizon of the research program. The multiomic and external validation limitations identified converge on a single methodological horizon: the application of the pipeline to molecular cohorts that are simultaneously broader in modality and independent of the GSE85217 / GPL22286 substrate (GEO environment). The Genomic Data Commons (GDC) provides the natural infrastructure for this extension, hosting harmonized multiomic data (RNA-seq, DNA methylation, somatic mutation, copy-number variation, micro RNA-seq, and proteomic quantification) for medulloblastoma and a wide range of other cancer types within a single curated repository. Migration of the input layer from microarray intensities to GDC-derived RNA-Seq expression would itself constitute a form of cross-platform external validation, while the parallel availability of methylation and mutation layers within the same cohort would enable the early or late fusion multiomic architectures discussed previously to be tested on biologically matched samples rather than on independently constructed cohorts. The combination of these two extensions, external validation through RNA-Seq and multiomic integration through paired modalities, would together address the two principal scope limitations of the present work and constitute the next major iteration of the classification pipeline.

6.4 *Final Considerations*

The methodological choices that define this thesis carry, beyond their immediate technical justification, a broader commitment to the values of reproducibility, transparency, and accessibility in computational oncology research. The decision to implement the bioinformatics preprocessing and differential expression analysis stages in R and Python (rather than in commercial platforms such as PARTEK Genomic Suite, used by both Reveles-Espinoza et al. [2025] and Salgado et al. [2024] as the methodological antecedents to the present work) was not a matter of convenience but of principle. Commercial environments

such as PARTEK provide polished graphical interfaces and curated parameter defaults that lower the barrier to entry for non-specialist users, but they do so at a methodological cost: the analytical decisions made during preprocessing, normalization, and differential expression modeling are obscured behind menu selections and dialog boxes whose state cannot be reconstructed from the published results alone. An R or Python code base, by contrast, encodes every analytical decision explicitly (every threshold, every parameter, every intermediate matrix transformation) as text that can be read, audited, version-controlled, and re-executed by any researcher with a standard computing environment, irrespective of institutional licensing arrangements. This is not an incidental property of the implementation but the substantive condition under which the pipeline can claim to be reproducible in any meaningful sense.

The same reasoning extends to the ANN training stage and motivates the migration proposed to free software analysis. The MATLAB implementation that produced the classification results reported in this thesis was the appropriate choice for direct architectural parity with the reference methodology, but it leaves the pipeline in a methodologically asymmetric stage: an open and freely reproducible upstream half coupled to a closed and license dependent downstream half. Completing the migration to Python is therefore not merely a practical optimization; it is the methodological act that transforms the pipeline from a partially open implementation into a fully open and reproducible tool chain, end-to-end auditable from the raw CEL files to the trained classifier weights. Until that migration is undertaken, the reproducibility commitment articulated in this section applies only to the bioinformatics half of the pipeline, and its completion is regarded here as an outstanding methodological obligation rather than a discretionary refinement.

The broader implication of this commitment becomes clearest when the pipeline is considered in the context of the research environment from which it emerges. Computational oncology in Mexico, and in Latin America more broadly, operates under structural constraints that differ substantively from those of the high-resource research centers where many of the antecedent methodologies were developed: institutional access to commercial bioinformatics software is intermittent and often contingent on individual project budgets; dedicated wet lab validation infrastructure is similarly variable; and the scale if locally generated multiomic cohorts remains modest relative to the consortia from which large public datasets such as TCGA and the Genomic Data Commons were assembled. A computationally lightweight, reproducible, and license independent classification pipeline is therefore not only a methodological preference but an operational fit for the research

realities of the region. The present work is offered, in this sense not solely as a technical contribution to medulloblastoma molecular classification, but as a methodological template, a worked example of how carefully designed open source pipeline can deliver state of the art classification performance under the operational constraints typical of Latin America computational oncology research, and how such pipelines can be extended, audited, and reused by groups operating under comparable conditions.

The arc of this thesis began with the proposition that medulloblastoma molecular classification is most productively approached as the convergence of three analytical domains: the molecular data that characterizes the disease at the transcriptomic level, the statistical rigor required to extract reliable signal from high-dimensional and heterogeneously distributed expression measurements, and the deep learning methodology that translates that signal into discriminative classification of clinically meaningful subgroups. The pipeline developed and validated across the preceding chapters demonstrates that this convergence, when implemented under a deliberate commitment to reproducibility and open source methodology, produces classification performance at or above the current state of the art while remaining tractable on standard computing infrastructure and accessible to any researcher prepared to read the code. It is this combination (methodological rigor without computational extravagance, classification accuracy without licensing dependence, biological interpretability without analytical opacity) that the present work seeks to advance as a viable model for accessible, reproducible computational oncology research in the future.

Bibliography

Boletín epidemiológico semanal. URL https://www.gob.mx/salud/acciones-y-programas/historico-boletin-epidemiologico. Semana Epidemiológica [52].

Giuseppe Agapito. Preface. In Giuseppe Agapito, editor, *Microarray Data Analysis*. Springer, 2022.

Alexander Aliper, Sergey Plis, Artem Artemov, Alvaro Ulloa, Polina Mamoshina, and Alex Zhavoronkov. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Molecular pharmaceutics*, 13(7):2524–2530, 2016.

Sonali Arora, Nicholas Nuechterlein, Matt Jensen, Gregory Glatzer, Philipp Sievers, Srinidhi Varadharajan, Andrey Korshunov, Felix Sahm, Stephen C Mack, Michael D Taylor, et al. Integrated transcriptomic landscape of medulloblastoma and ependymoma reveals novel tumor subtype-specific biology. *Neuro-oncology*, 28(2):505–519, 2026.

Awais Athar, Anja Füllgrabe, Nancy George, Haider Iqbal, Laura Huerta, Ahmed Ali, Catherine Snow, Nuno A Fonseca, Robert Petryszak, Irene Papatheodorou, et al. Arrayexpress update–from bulk to single-cell expression data. *Nucleic acids research*, 47(D1):D711–D715, 2019.

Noam Auslander, Ayal B. Gussow, and Eugene V. Koonin. Incorporating machine learning into established bioinformatics frameworks. *International Journal of Molecular Sciences*, 22(6):2903, 2021. DOI: 10.3390/ijms22062903.

Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F Kim, Alexandra Soboleva, Maxim Tomashevsky, and Ron Edgar. Ncbi geo: mining tens of millions of expression profiles—database and tools update. *Nucleic acids research*, 35(suppl_1):D760–D765, 2007.

Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H

Phillippy, Patti M Sherman, Michelle Holko, et al. Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research*, 41 (D1):D991–D995, 2012.

Andreas D. Baxevanis. Biological sequence databases. In Andreas D. Baxevanis, Gary D. Bader, and David S. Wishart, editors, *Bioinformatics*, chapter 1. John Wiley & Sons, 4th edition, 2020.

Andreas D. Baxevanis, Gary D. Bader, and David S. Wishart. *Bioinformatics*. Wiley, 2020.

Robert G. Beiko. Metagenomics and microbial community analysis. In Andreas D. Baxevanis, Gary D. Bader, and David S. Wishart, editors, *Bioinformatics*, chapter 16. John Wiley & Sons, 4th edition, 2020.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Anna Bernasconi and Silvia Cascianelli. Scenarios for the integration of microarray gene expression profiles in covid-19–related studies. In Giuseppe Agapito, editor, *Microarray Data Analysis*, chapter 19. Springer, 2022.

Eduardo Francisco Caicedo Bravo. *Una aproximación práctica a las redes neuronales artificiales*. Universidad del Valle, 2009.

Alvis Brazma, Pascal Hingamp, John Quackenbush, Gavin Sherlock, Paul Spellman, Chris Stoeckert, John Aach, Wilhelm Ansorge, Catherine A Ball, Helen C Causton, et al. Minimum information about a microarray experiment (miame)—toward standards for microarray data. *Nature genetics*, 29(4):365–371, 2001.

Vince Buffalo. *Bioinformatics data skills: Reproducible and robust research with open source tools*. " O'Reilly Media, Inc.", 2015.

Barbara Calabrese. Web and cloud computing to analyze microarray data. In *Microarray Data Analysis*, pages 29–38. Springer, 2021.

Barbara Calabrese. Web and cloud computing to analyze microarray data. In Giuseppe Agapito, editor, *Microarray Data Analysis*, chapter 3. Springer, 2022.

Howard M Cann, Claudia De Toma, Lucien Cazes, Marie-Fernande Legrand, Valerie Morel, Laurence Piouffre, Julia Bodmer, Walter F Bodmer, Batsheva Bonne-Tamir, Anne Cambon-Thomsen, et al. A human genome diversity cell line panel. *Science*, 296(5566):261–262, 2002.

M Carlson, S Falcon, N Li, et al. Annotationdbi: Manipulation of sqlite-based annotations in bioconductor. *R Package Version 1.54. 1*, 2021.

Benilton S Carvalho and Rafael A Irizarry. A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, 26(19):2363–2367, 2010.

Rosa Angélica Castillo-Rodríguez, Víctor Manuel Dávila-Borja, and Sergio Juárez-Méndez. Data mining of pediatric medulloblastoma microarray expression reveals a novel potential subdivision of the group 4 molecular subgroup. *Oncology letters*, 15(5):6241–6250, 2018.

Francesco Cauteruccio. Alignment of microarray data. In Giuseppe Agapito, editor, *Microarray Data Analysis*, chapter 14. Springer, 2022.

Florence MG Cavalli, Marc Remke, Ladislav Rampasek, John Peacock, David JH Shih, Betty Luu, Livia Garzia, Jonathon Torchia, Carolina Nor, A Sorana Morrissy, et al. Intertumoral heterogeneity within medulloblastoma subgroups. *Cancer cell*, 31(6):737–754, 2017.

Chris Cheadle, Marquis P Vawter, William J Freed, and Kevin G Becker. Analysis of microarray data using z score transformation. *The Journal of molecular diagnostics*, 5(2):73–81, 2003.

Davide Chicco. geneexpressionfromgeo: An r package to facilitate data reading from gene expression omnibus (geo). In Giuseppe Agapito, editor, *Microarray Data Analysis*, chapter 12. Springer, 2022.

F Chollet and JJ Allaire. Deep learning with r. manning publications, manning early access program (mea), 2017.

Emily Clough and Tanya Barrett. The gene expression omnibus database. In *Statistical Genomics: Methods and Protocols*, pages 93–110. Springer, 2016.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013.

Mikołaj Danielewski, Marlena Szalata, Jan Krzysztof Nowak, Jarosław Walkowiak, Ryszard Słomski, and Karolina Wielgus. History of biological databases, their importance, and existence in modern scientific and policy context. *Genes*, 16(1):100, 2025.

dbGaP. Medulloblastoma genomics. dbGaP Study Accession: phs000424.v2.p1, 2024. URL https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v2.p1.

Ruth-Mary DeSouza, Benjamin RT Jones, Stephen P Lowis, and Kathreena M Kurian. Pediatric medulloblastoma–update on molecular classification driving targeted therapies. *Frontiers in oncology*, 4:176, 2014.

Abhinav Dey, Anshu Malhotra, and Neha Garg. Medullo-blastoma.

Abhinav Dey, Anshu Malhotra, and Neha Garg. Medulloblastoma methods and protocols. Springer Nature, 2022. ISBN 978-1-0716-1952-0.

Debojyoti Dhar and Gopala Kallapura. Analysis of microarray data from medulloblastoma tissue samples. In *Medulloblastoma: Methods and Protocols*, pages 59–64. Springer, 2022.

Alex Diaz-Papkovich, Luke Anderson-Trocmé, and Simon Gravel. A review of umap in population genetics. *Journal of Human Genetics*, 66 (1):85–91, 2021.

Lars MT Eijssen, Magali Jaillard, Michiel E Adriaens, Stan Gaj, Philip J de Groot, Michael Müller, and Chris T Evelo. User-friendly solutions for microarray quality control and pre-processing on arrayanalysis. org. *Nucleic acids research*, 41(W1):W71–W76, 2013.

EMBL-EBI. Biostudies arrayexpress, 2025. URL https://www.ebi.ac.uk/biostudies/arrayexpress.

Shahla Faisal and Gerhard Tutz. Missing value imputation for gene expression data by tailored nearest neighbors. *Statistical applications in genetics and molecular biology*, 16(2):95–106, 2017.

Antonio Federico, Laura Aliisa Saarimäki, Angela Serra, Giusy Del Giudice, Pia Anneli Sofia Kinaret, Giovanni Scala, and Dario Greco. Microarray data preprocessing: from experimental design to differential analysis. In *Microarray Data Analysis*, pages 79–100. Springer, 2021.

Steven M Foltz, Casey S Greene, and Jaclyn N Taroni. Cross-platform normalization enables machine learning model training on microarray and rna-seq data simultaneously. *Communications Biology*, 6(1):222, 2023.

Agostino Forestiero, Giuseppe Papuzzo, Rosaria De Simone, and Rosa Varchera. A microarray analysis technique using a self-organizing multiagent approach. In *Microarray Data Analysis*, pages 39–50. Springer, 2021.

Rocío Elizabeth García-Dávila, Sergio Díaz-Bello, Raúl Villanueva-Rodríguez, León René López, Jorge Luis Valencia-Vázquez, and Belén

Rivera-Bravo. Utilidad de la tomografía por emisión de positrones/ tomografía computada (pet/ct) en pacientes con diagnóstico de meduloblastoma. 63(1). DOI: 10.22201/fm.24484865e.2020.63.1.06. URL https://doi.org/10.22201/fm.24484865e.2020.63.1.06.

GDC Data Portal. Genomic data commons data portal, 2025. URL https://portal.gdc.cancer.gov/.

Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004.

Ángel Herráez. Hibridación de ácidos nucleicos. In *Biología molecular e ingeniería genética*, chapter 12, pages 163–190. Elsevier Health Sciences, 2012.

Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

Michael Iv, Michelle Zhou, Katie Shpanskaya, Sebastian Perreault, Zhewei Wang, Evan Tranvinh, Brent Lanzman, Sridhar Vajapeyam, Nicholas A. Vitanza, Paul G. Fisher, et al. MR imaging–based radiomic signatures of distinct molecular subgroups of medulloblastoma. *American Journal of Neuroradiology*, 40:154–161, 2019. DOI: 10.3174/ajnr.A5899.

W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

Audrey Kauffmann, Robert Gentleman, and Wolfgang Huber. arrayqualitymetrics—a bioconductor package for quality assessment of microarray data. *Bioinformatics*, 25(3):415–416, 2009.

AN Kolmogorov. Sulla determinzione empírica di una legge di distribuzione, ist i tal. *Atti. Giorn*, 4:88–91, 1933.

Marcel Kool, Jan Koster, Jens Bunt, Nancy E Hasselt, Arjan Lakeman, Peter Van Sluis, Dirk Troost, Netteke Schouten-van Meeteren, Huib N Caron, Jacqueline Cloos, et al. Integrated genomics identifies five medulloblastoma subtypes with distinct genetic profiles, pathway signatures and clinicopathological features. *PloS one*, 3(8):e3088, 2008.

Max Kotlyar, Serene WH Wong, Chiara Pastrello, and Igor Jurisica. Improving analysis and annotation of microarray data with protein interactions. *Microarray Data Analysis*, pages 51–68, 2021.

Max Kotlyar, W. H. Wong, Chiara Serene, Pastrello, and Igor Jurisica. Improving analysis and annotation of microarray data with protein interactions. In Giuseppe Agapito, editor, *Microarray Data Analysis*, chapter 5. Springer, 2022.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Marieke L. Kuijjer, Joseph N. Paulson, and John Quackenbush. Expression analysis. In Andreas D. Baxevanis, Gary D. Bader, and David S. Wishart, editors, *Bioinformatics*, chapter 10. John Wiley & Sons, 4th edition, 2020.

Sally R Lambert, Hendrik Witt, Volker Hovestadt, Manuela Zucknick, Marcel Kool, Danita M Pearson, Andrey Korshunov, Marina Ryzhova, Koichi Ichimura, Nada Jabado, et al. Differential expression and methylation of brain developmental genes define location-specific subsets of pilocytic astrocytoma. *Acta neuropathologica*, 126(2):291–301, 2013.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10): 733–739, 2010.

Jeffrey T Leek, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, 2012.

Pietro Di Lena, Claudia Sala, Andrea Prodi, and Christine Nardini. Methylation data imputation performances under different representations and missingness patterns. *BMC bioinformatics*, 21(1): 268, 2020.

Boyu Lyu and Anamul Haque. Deep learning based tumor type classification using gene expression data. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 89–96, 2018.

Shiyong Ma and Zhen Zhang. Omicsmapnet: Transforming omics data to take advantage of deep convolutional neural network for discovery. *arXiv preprint arXiv:1804.05283*, 2018.

Ganiraju Manyam, Aybike Birerdinc, and Ancha Baranova. Kpp: Kegg pathway painter. *BMC Systems Biology*, 9(Suppl 2):S3, 2015.

Fabrizio Marozzo and Loris Belcastro. High-performance framework to analyze microarray data. In *Microarray Data Analysis*, pages 13–27. Springer, 2021.

Veer Singh Marwah, Giovanni Scala, Pia Anneli Sofia Kinaret, Angela Serra, Harri Alenius, Vittorio Fortino, and Dario Greco. eutopia: solution for omics data preprocessing and analysis. *Source code for biology and medicine*, 14(1):1, 2019.

MathWorks. traingdm - gradient descent with momentum backpropagation. https://la.mathworks.com/help/deeplearning/ref/traingdm.html, 2024a. MATLAB Deep Learning Toolbox Documentation.

MathWorks. Performance curves. https://la.mathworks.com/help/stats/performance-curves.html, 2024b. Accessed: 2025.

MathWorks. normalize - normalize data, 2026. URL https://www.mathworks.com/help/matlab/ref/double.normalize.html.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Memorial Sloan Kettering Cancer Center. cbioportal for cancer genomics. https://about.cbioportal.org/, 2025.

Christian von Mering, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. String: a database of predicted functional associations between proteins. *Nucleic acids research*, 31(1):258–261, 2003.

Osval Antonio Montesinos López, Abelardo Montesinos López, and Jose Crossa. Fundamentals of artificial neural networks and deep learning. In *Multivariate statistical machine learning methods for genomic prediction*, pages 379–425. Springer, 2022.

National Cancer Institute. Medulloblastoma, 2025. URL https://www.cancer.gov/rare-brain-spine-tumor/tumors/medulloblastoma.

National Cancer Institute, Center for Cancer Genomics. The cancer genome atlas program. https://www.cancer.gov/ccg/research/genome-sequencing/tcga, 2024.

National Cancer Institute (NCI). Tratamiento del meduloblastoma y otros tumores embrionarios del sistema nervioso central infantil, 2025. URL https://www.cancer.gov/espanol/tipos/cerebro/pro/tratamiento-embrionarios-snc-infantil-pdq.

National Center for Biotechnology Information. GEO accession GSE85217, 2024. URL https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85217. Gene Expression Omnibus (GEO), NCBI.

National Human Genome Research Institute. The human genome project. https://www.genome.gov/human-genome-project, 2024.

NCBI. Miniml (miame notation in markup language), 2024a. URL https://www.ncbi.nlm.nih.gov/geo/info/MINiML.html.

NCBI. Miame and minseqe guidelines, 2024b. URL https://www.ncbi.nlm.nih.gov/geo/info/MIAME.html.

NCBI. Geo overview, 2025. URL https://www.ncbi.nlm.nih.gov/geo/info/overview.html.

NHGRI. The cancer genome atlas program, 2025. URL https://www.genome.gov/Funded-Programs-Projects/Cancer-Genome-Atlas.

Paul A Northcott, Adrian M Dubuc, Stefan Pfister, and Michael D Taylor. Molecular subgroups of medulloblastoma. *Expert review of neurotherapeutics*, 12(7):871–884, 2012.

Paul A Northcott, Ivo Buchhalter, A Sorana Morrissy, Volker Hovestadt, Joachim Weischenfeldt, Tobias Ehrenberger, Susanne Gröbner, Maia Segura-Wang, Thomas Zichner, Vasilisa A Rudneva, et al. The whole-genome landscape of medulloblastoma subtypes. *Nature*, 547(7663):311–317, 2017.

Josh Patterson and Adam Gibson. *Deep learning: A practitioner's approach*. " O'Reilly Media, Inc.", 2017.

Alisa Pavel, Angela Serra, Luca Cattelani, Antonio Federico, and Dario Greco. Network analysis of microarray data. In *Microarray Data Analysis*, pages 161–186. Springer, 2021.

Belinda Phipson, Stanley Lee, Ian J Majewski, Warren S Alexander, and Gordon K Smyth. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *The annals of applied statistics*, 10(2):946, 2016.

Scott L Pomeroy, Pablo Tamayo, Michelle Gaasenbeek, Lisa M Sturla, Michael Angelo, Margaret E McLaughlin, John YH Kim, Liliana C Goumnerova, Peter M Black, Ching Lau, et al. Prediction of central

nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–442, 2002.

Julia Pöschl, Sebastian Stark, Philipp Neumann, Susanne Gröbner, Daisuke Kawauchi, David TW Jones, Paul A Northcott, Peter Lichter, Stefan M Pfister, Marcel Kool, et al. Genomic and transcriptomic analyses match medulloblastoma mouse models to their human counterparts. *Acta neuropathologica*, 128(1):123–136, 2014.

Xing Qiu, Hulin Wu, and Rui Hu. The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. *BMC bioinformatics*, 14(1):124, 2013.

John Quackenbush. Microarray data normalization and transformation. *Nature genetics*, 32(4):496–501, 2002.

Alicia Reveles-Espinoza, Ulises Villela, Edgar Hernandez-Martinez, Isaac Chairez, Sergio Juárez-Méndez, J Casanova-Moreno, Ma del Pilar Eguía-Aguilar, Luis Figueroa-Yáñez, Adriana Vallejo-Cardona, and Iván Salgado. Machine learning identification of molecular targets for medulloblastoma subgroups using microarray gene fingerprint analysis. *Computational and Structural Biotechnology Journal*, 2025.

Matthew E Ritchie, Jeremy Silver, Alicia Oshlack, Melissa Holmes, Dileepa Diyagama, Andrew Holloway, and Gordon K Smyth. A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, 23(20):2700–2707, 2007.

Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

Giles Robinson, Matthew Parker, Tanya A Kranenburg, Charles Lu, Xiang Chen, Li Ding, Timothy N Phoenix, Erin Hedlund, Lei Wei, Xiaoyan Zhu, et al. Novel mutations target distinct subgroups of medulloblastoma. *Nature*, 488(7409):43–48, 2012.

James T Rutka and Harold J Hoffman. Medulloblastoma: a historical perspective and overview. *Journal of neuro-oncology*, 29(1):1–7, 1996.

I. Salgado, E. Prado Montes de Oca, I. Chairez, L. Figueroa-Yáñez, A. Pereira-Santana, A. Rivera Chávez, et al. Deep learning techniques to characterize the RPS28P7 pseudogene and the Metazoa-SRP gene as drug potential targets in pancreatic cancer patients. 12(02).

Iván Salgado, Ernesto Prado Montes de Oca, Isaac Chairez, Luis Figueroa-Yáñez, Alejandro Pereira-Santana, Andrés Rivera Chávez,

Jesús Bernardino Velázquez-Fernandez, Teresa Alvarado Parra, and Adriana Vallejo. Deep learning techniques to characterize the rps28p7 pseudogene and the metazoa-srp gene as drug potential targets in pancreatic cancer patients. *Biomedicines*, 12(02), 2024.

Edward C Schwalbe, Janet C Lindsey, Sirintra Nakjang, Stephen Crosier, Amanda J Smith, Debbie Hicks, Gholamreza Rafiee, Rebecca M Hill, Alice Iliasova, Thomas Stone, et al. Novel molecular subgroups for clinical classification and outcome prediction in childhood medulloblastoma: a cohort study. *The Lancet Oncology*, 18(7):958–971, 2017.

Francesca Scionti, Mariamena Arbitrio, Daniele Caracciolo, Licia Pensabene, Pierfrancesco Tassone, Pierosandro Tagliaferri, and Maria Teresa Di Martino. Integration of dna microarray with clinical and genomic data. In *Microarray Data Analysis*, pages 239–248. Springer, 2021.

Angela Serra, Michele Fratello, Luca Cattelani, Irene Liampa, Georgia Melagraki, Pekka Kohonen, Penny Nymark, Antonio Federico, Pia Anneli Sofia Kinaret, Karolina Jagiello, et al. Transcriptomics in toxicogenomics, part iii: data modelling for risk assessment. *Nanomaterials*, 10(4):708, 2020.

Angela Serra, Luca Cattelani, Michele Fratello, Vittorio Fortino, Pia Anneli Sofia Kinaret, and Dario Greco. Supervised methods for biomarker detection from microarray experiments. In *Microarray Data Analysis*, pages 101–120. Springer, 2021.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Tanvi Sharma, Edward C Schwalbe, Daniel Williamson, Martin Sill, Volker Hovestadt, Martin Mynarek, Stefan Rutkowski, Giles W Robinson, Amar Gajjar, Florence Cavalli, et al. Second-generation molecular subgrouping of medulloblastoma: an international meta-analysis of group 3 and group 4 subtypes. *Acta neuropathologica*, 138 (2):309–326, 2019.

Gautam B Singh. Fundamentals of bioinformatics and computational biology. *Cham: Springer International Publishing*, pages 159–170, 2015a.

Gautam B Singh. Microarrays. In *Fundamentals of Bioinformatics and Computational Biology*, chapter 17, pages 159–170. Springer International Publishing, Cham, 2015b.

Gordon K Smyth, Joëlle Michaud, and Hamish S Scott. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, 21(9):2067–2075, 2005.

Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the national academy of sciences*, 102(43):15545–15550, 2005.

Yujin Suk, William D Gwynne, Ian Burns, Chitra Venugopal, and Sheila K Singh. Childhood medulloblastoma: an overview. *Medulloblastoma: Methods and Protocols*, pages 1–12, 2022.

Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, et al. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, 47(D1):D607–D613, 2019.

Thermo Fisher Scientific. Supporting multi-omics approaches, n.d. URL https://www.thermofisher.com/mx/es/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/spotlight-articles/supporting-multi-omics-approaches.html.

Daniel Urda, Julio Montes-Torres, Fernando Moreno, Leonardo Franco, and José M Jerez. Deep learning to analyze rna-seq gene expression data. In *International work-conference on artificial neural networks*, pages 50–59. Springer, 2017.

Hao Wang, Ruifeng Liu, Patric Schyman, and Anders Wallqvist. Deep neural network models for predicting chemically induced liver toxicity endpoints from transcriptomic responses. *Frontiers in pharmacology*, 10: 42, 2019.

Holger Weishaupt, Patrik Johansson, Anders Sundström, Zelmina Lubovac-Pilav, Björn Olsson, Sven Nelander, and Fredrik J Swartling. Batch-normalization of cerebellar and medulloblastoma gene expression datasets utilizing empirically defined negative control genes. *Bioinformatics*, 35(18):3357–3364, 2019.

Joshua Fredrick Wiley. *R deep learning essentials*. Packt, 2016.

M. Yousef and J. Allmer. Deep learning in bioinformatics. 47(06).

Ye Yuan and Ziv Bar-Joseph. Gcng: Graph convolutional networks for inferring cell-cell interactions. *bioRxiv*, pages 2019–12, 2019.

Paolo Zaffino and Maria Francesca Spadea. Algorithms to preprocess microarray image data. In *Microarray Data Analysis*, pages 69–78. Springer, 2021.

Index

Activation Functions, 66
ANN-Based Molecular Classification of Medulloblastoma Subgroups, 78
Artificial Neural Network Architecture, 81
Artificial Neural Networks, 63

Binary Classifiers, 116
Bioinformatics and Computational Biology, 41

Comparison with the State of the Art, 124
Computational Cost Analysis, 121
Cross-Validation Protocol, 111

Data Mining and Repository Search, 89
Deep Learning Architecture: LSTM Network, 75
Deep Learning Classification, 109
Deep Learning Methods and Applications on Molecular Data, 71
Deep Learning Techniques for Survival Prediction in Pancreatic Cancer, 73
Differential Expression Analysis, 101
Discussions and Conclusions, 127

Evaluation Metrics and Outputs, 112
Experimental Design and Data Collection, 85
Experimental Design Considerations, 87

Feedforward Neural Network Topology, 67
Final Considerations, 142
Five-class Classifier: CT vs. WNT vs. SHH vs. G3 vs. G4, 114
Fundamentals of Machine Learning and Deep Learning, 62
Future Work, 137

Gene Expression Data Processing and Differential Expression Analysis, 95
Gene Expression Omnibus (GEO), 48
Gene Ontology and Gene Set Enrichment Analysis, 140

Hypothesis, 34

Introduction, 25

Justification, 29

Layer Construction and Regularization, 110
Limitations, 132
Loss Function, 68

Medulloblastoma, 39
Metadata Preprocessing and Sample Identification, 90
Methodology, 85
Microarray Data Acquisition and Preprocessing, 97
Microarray Data Analysis, 51
Microarray Data Preprocessing, 53
Microarray Technology and Gene Expression Analysis, 44
Migration to Open-Source Pipelines, 141
Molecular Linear Separability Hypothesis, 140
Molecular Subtype of Medulloblastoma´s Subgroups, 138
Mono-omic Experimental Design, 136
Multiomics Data Integration, 42, 142

Network Architecture, 110
Normalization Strategies, 134
Normalization Strategies Benchmark, 139
Novel Contributions, 36

Objectives, 34
Overview of Classification Performance, 113

Problem Statement, 32
Public Biological Databases and Data Repositories, 48

Quality Control and Normalization, 98

Research Question, 33
Results, 113

Single Platform and Dataset Validation, 135
Specific Objectives, 35
State of the Art, 71
Subtyping of Medulloblastoma, 132
Summary of Findings, 129

Theoretical and conceptual framework, 39
Thesis Proposal, 35
Training Configuration, 111

Wet Laboratory Confirmation, 137