PIONEER
ACADEMIC PUBLISHING LIMITED

Journal of Research in Social Science and Humanities
ISSN 2709-1910
www.pioneerpublisher.com/jrssh
Volume 3 Number 12 December 2024

Leveraging Random Forest Machine Learning for Subgroup Classification of Medulloblastoma

Bo Lan[1]

[1] United World College of South East Asia, Singapore
Correspondence: Bo Lan, United World College of South East Asia, Singapore.

doi:10.56397/JRSSH.2024.12.06

Abstract

Medulloblastoma, the most prevalent malignant brain tumour in children, necessitates precise diagnostic methods due to its heterogeneous molecular subgroups. This study leverages Random Forest machine learning algorithms to classify medulloblastoma subgroups by analysing DNA methylation and gene expression data. Utilising the Gene Expression Omnibus dataset GSE85218 — comprising 763 primary MB samples — the study implements variance threshold feature selection for preprocessing. Models were evaluated based on precision, recall, F1 score, and accuracy — with the highest performance observed in models utilising Top 1% varied combined DNA methylation and gene expression data. Models performed similarly however, meaning only targeted gene expression and DNA methylation data are required for an accurate diagnosis. Gene Set Enrichment Analysis (GSEA) identified significant pathways related to neural processes, underscoring the tumour's impact on neural development and function. Biomarkers were identified from the most important features identified by the ML model, with possible new biomarkers for subgroup diagnosis being discovered.

1. Introduction

Among children from the ages of 5-9 within the US, cancer is the second leading cause of death (CDC, 2024). Cancer is a disease stemming from mutations within the DNA of the cell, leading to abnormal cell proliferation. As such, they buildup in masses of tissue known as tumours, disrupting organ activity and sapping nutrients from healthy cells, ultimately leading to death. As the most common malignant brain tumour amongst children, accounting for around 20% of all brain tumours (Rossi et al., 2008), medulloblastoma (MB) is specifically problematic due to its wide array of molecular subgroups, all having varying survival rates. The four main molecular subgroups identified are Wingless Activated (WNT), Sonic Hedgehog (SHH), Group 3, and Group 4 (Taylor et al., 2011). The five-year survival rates for WNT are over 90%, while the survival rates for other subgroups are much lower by comparison, with Group 3 MB patients only being around 50% (Tenley et al., 2017). As the genetic and epigenetic makeups of each subgroup of MB are different, distinct clinical procedures must be followed based on the diagnosis (Ray et al., 2021). Hence, accurate prognosis is crucial for personalised therapy and improved treatment.

DNA methylation sequencing and gene expression profiling are indispensable in providing insight towards the molecular subgroup. However, the aforementioned

processes are quite expensive (Orr, B.A., 2020). By targeting specific regions rather than sequencing the entire genome, costs could be reduced. Recent advancements in technology have allowed for generation of DNA methylation and gene expression profiles in publicly accessible datasets. These datasets provide a valuable resource for identifying biomarkers associated with MB subtypes. To properly utilise the aforementioned datasets, appropriate computational methods must be utilised.

Machine learning (ML) has become an essential tool in the analysis of large-scale biological data, enabling researchers to uncover complex patterns and relationships that are not immediately apparent. By leveraging ML techniques, scientists can efficiently process and analyse vast amounts of data, identifying key features and biomarkers for classification purposes (Larrañaga et al., 2006). These computational methods can also enhance the accuracy and speed of data interpretation, facilitating the development of targeted therapies and personalized medicine (Libbrecht & Noble, 2015). An appropriate ML model for cancer classification is Random Forest.

Random Forest is a supervised machine learning method for classification (Ho, 1995). The algorithm constructs multiple decision trees during training, outputting the mode of resultant classes for improved reliability (Breimen, 2001). The classification method is commonly used in bioinformatics, and related research has shown promising levels of accuracy (Minnoor & Baths, 2023; Cavalli et al., 2017).

This study aims to utilise machine learning algorithms to predict medulloblastoma subgroups based on biomarker data; understanding the scope of data required for an accurate classification, as well as identifying biological pathways and biomarkers after obtaining results. Through analysis of DNA methylation and gene expression data in primary MB samples, the study hopes to compare the performance of the Random Forest classifier with varied amounts of data. The study will also investigate the change in effectiveness when training models with gene expression and DNA methylation data together rather than separately. The most important pathways and biomarkers identified through the classifier will also be analysed, finding new further research directions to better understand the disease.

2. Results

For each variance threshold (Top 1%, 0.5%, 100), 3 models were developed; one utilising both methylation and gene expression data, one using gene expression data exclusively, and one inputting methylation data. Table 1 displays the results from each model:

Table 1. Random Forest model comparison

Input	Accuracy (%)	Mean CVA (%)
Top 1% DNA Methylation profiling + gene expression data	98.68	98.73
Top 1% DNA methylation profiling only	98.68	98.65
Top 1% gene expression data only	96.71	98.81
Top 0.5% DNA Methylation profiling + gene expression data	98.68	98.54
Top 0.5% DNA methylation profiling only	98.68	98.35
Top 0.5% gene expression data only	98.03	97.81
Top 100 DNA Methylation profiling + gene expression data	98.03	98.44
Top 100 DNA methylation profiling only	98.03	96.86
Top 100 gene expression data only	98.68	97.78

Overall, the Random Forest classifier demonstrated high accuracy and robustness across different data inputs, maintaining accuracies and mean 5-fold cross validation accuracies above 95%. While the model with the highest combined accuracy and mean cross validation is the model which inputted the top 1% of both datasets, the difference in average performance compared to all other models is negligible. This signifies an accurate subgroup diagnosis does not require vast amounts of data, merely the most relevant genes/methylation

markers.

The WNT subgroup displayed higher levels of recall when sorted using exclusively gene expression data, specifically the models beyond top 1% (Table 1). This could insinuate some noise within the cutoff from top 1% to top 0.5%, hindering the performance of WNT classification.

Gene Set Enrichment Analysis

Results taken from Enrichr for genes associated with the top 1% variance genes and methylation data includes the three most enriched pathways identified from each database.

Table 2. Pathways identified

Database	Pathway	Odds Ratio	Combined score	P-value	Adjusted P-value
KEGG 2021 Human	Glutamatergic Synapses	4.30	84.31	$3.088e^{-9}$	$8.862e^{-7}$
	Nicotine Addiction Axon Guidance	6.74 2.69	97.07 33.10	$5.596e^{-7}$ $4.458e^{-6}$	$8.031e^{-5}$ $4.265e^{-}$
Reactome 2022	Neuronal System R-HSA-112316	2.84	75.62	$2.688e^{-12}$	$3.378e^{-9}$
	Transmission Across Chemical Synapses R-SHA-112315	2.92	55.44	$5.542e^{-9}$	$3.483e^{-6}$
	Netrin-1 Signalling R-HSA-112316	5.01	58.36	$8.619e^{-6}$	$3.611e^{-3}$
GO Biological Process 2023	Nervous System Development	2.97	96.08	$8.483e^{-15}$	$3.095e^{-11}$
	Axonogenesis	3.65	84.44	$8.883e^{-11}$	$1.620e^{-7}$
	Regulation of Transcription by RNA Polymerase II	1.66	37.30	$1.695e^{-10}$	$2.061e^{-7}$
DisGeNet	Autism Disorder	2.38	68.00	$3.792e^{-13}$	$2.734e^{-9}$
	Medulloblastoma	2.35	57.80	$2.032e^{-11}$	$6.732e^{-8}$
	Alcohol Intoxication (Chronic)	2.62	63.26	$3.249e^{-11}$	$6.732e^{-8}$

Reactome 2022 analysis identified Neuronal System, Transmission Across Chemical Synapses, and Netrin-1 Signalling, all neural processes. This reinforces the association of such genes to medulloblastoma, hinting at its potential to interrupt neuronal functions, disrupt synaptic transmissions, and contribute to abnormal cell migration differently based on the subgroup. The low P-values also indicates an extremely strong association.

KEGG Human 2021 analysis identified Glutamatergic Synapse, Nicotine Addiction, and Axon Guidance as the primary pathways. Enrichment of Glutamatergic Synapses indicates dysregulation of the pathway may lead to differences in medulloblastoma pathophysiology. There is a possibility that targeting glutamate receptors and transporters may offer therapeutic benefits. Nicotine Addiction interacts with dopamine signalling and reward pathways, insinuating possible different levels of dopamine signalling based on the MB subgroup. Axon Guidance refers to the navigation of growing axons to targeted locations for the establishment of neural networks. Its enrichment suggests aberrant axon guidance signalling in MB can lead to disorganised neural tissue. This finding is consistent with the Netrin-1 Signalling pathway identified from the Reactome database. Again, the minimal P- values provides legitimacy to the findings.

GO Biological Process 2023 analysis identified Nervous System Development, Axonogenesis, and Regulation of Transcription by RNA

Polymerase II. Nervous System Development involves the formation and growth of the nervous system, consistent with pathways found previously. Enriched Axonogenesis suggests medulloblastoma's involvement in disrupting the ability for axons to form neural networks, again consistent with other pathways identified. Enrichment of Regulation of Transcription by RNA Polymerase II suggests dysregulation of transcriptional processes due to MB, possibly leading to overexpression of oncogenes or under-expression of tumour suppressors, leading to malignant growth. Low P-values ensure reliability of pathways identified.

Finally, DisGeNet identified pathways correlating with autism disorder, medulloblastoma, and alcohol intoxication (Chronic). The three diseases are all related to neurological processes, and could suggest common genetic or epigenetic factors between the diseases.

Biomarker Identification

The genes in Table 3 were associated with the top 20 most important classification features identified by the Random Forest model for the input of Top 100 mixed.

Downloading all genes most relevant to medulloblastoma from Gene Cards, only the following three were within the list: SGK1, CAMTA1, OTX2.

Table 3. Biomarkers identified not within Gene Cards

Biomarker	Role	Significance	Previous studies
EXOC5	Component of exocyst complex, involved in cellular transport and signalling	Mutation may cause disruptions in cell signalling, which are common within cancers (including MB)	Shin et al. (2019) found EXOC5 overexpression leads to breast cancer progression due to increased cell invasion and proliferation.
TXNDC15	Protein coding gene and positive regulator of Ciliary Hedgehog (HH) signalling	SHH pathway is a component of the HH pathway, therefore maybe critical to identifying SHH subgroup MB	Mutations in HH pathway are directly affiliated with MB (Fujia et al., 2017)
LINC02058	Long non-coding RNA (lncRNA), regulating processes such as cell proliferation or apoptosis (Liu et al., 2024)	lncRNAs are generally implicated in cancers through regulation of oncogenes and tumour suppressors.	Previous study has confirmed potential for lncRNAs as biomarkers for MB (Kesherwani et al., 2020)
H3K27AC	Not a genebut a histone modification associated with regulation of gene transcription	Methylation may lead to epigenetic dysregulations, affecting cell processes leading to cancer.	Lin et al. (2016) revealed H3K27AC methylation could affect gene expression patterns particularly in SHH and Group 3 MB
SASH1	Scaffolding protein involved in signal transduction and tumour suppression. Involved in TLR4 signalling pathway	Shown associations with cancer, so maybe case for MB as well. Mutation may lead to reduced tumour suppression capabilities	Liu et al. (2015) found SASH1 expression to directly correlate with Glioma's (another Brain tumour) survival rates
RREB1	Zinc finger transcription factor, regulating gene expression	Dysregulations of transcriptional factors may lead to overexpression and uncontrolled cell proliferation	RREB1 was found to be directly correlated with Group 3 MB (Masihi et al., 2020)
CYBRD1	Involved in iron	Could impact MB cancer cell	CYBRD1 upregulation is

	metabolism	metabolism and proliferation. Shown association with other cancers.	shown to be linked to Ovarian cancer (Chen et al., 2021)
ZSWIM5	Protein coding gene. Zinc finger SWIM-type containing 5 proteins. Role in axon guidance as well as brain development	Connected to brain development as well as axon guidance, may affect how MB cells affect surrounding nervous tissue.	Xu et al. (2018) found ZSWIM5 expression linked to tumour suppression in Lung cancer
CDHR1	Protein coding gene encoding a photoreceptor- specific cadherin. Mainly connected to the eye	Photoreceptor encoding genes have been shown to be expressed in MB cases. Could be due to common progenitor cells between neural and optical cells in early development. (Gabriel et al., 2021)	Castillo-Rodriguez et al. (2018) found CDHR1 to be linked to Group 4 MB subtype classification Photoreceptor genes are shown to be expressed in specific MB classifications (Kool et al., 2008)
RPGRIP1	Protein coding gene encoding photoreceptor protein	Photoreceptor encoding genes have been shown to be expressed in MB cases. Could be due to common progenitor cells between neural and optical cells in early development. (Gabriel et al., 2021)	RPGRIP1 is shown to be expressed in specific MB classifications (Kool et al., 2008)
FBXL21	Pseudogene heavily involved in regulation of circadian rhythm	Mutation may lead to disruption of circadian rhythms, significantly associated with cancers	Circadian rhythms were found to impact Glioma (Brain tumour) physiopathology (Wang et al., 2022)
RIPOR2	Inhibitor of small G protein RhoA which regulates myoblast and hair cell differentiation	RhoA has been associated with oncological processes as well as the WNT pathway	Rodrigues et al. (2014) found RhoA expression to reduce WNT signalling as well as suppressing Colorectal cancer
LEMD1	Protein coding gene connected to Pancreatic cancer subtypes and Colorectal cancer	Connection to other cancers, dysregulation may impact cell signalling pathways	Kool et al. (2008) found LEMD1 to be associated with non WNT and SHH subgroup MB
PLPPR4	Protein coding gene, involved in glutamatergic neuron pathway	Glutamatergic neuron pathways are strongly associated with MB	Hooper et al. (2014) found SHH, Group 3, and Group 4 MB to all Glutamatergic neuron pathways

This could indicate the discovery of new medulloblastoma biomarkers linked to previously identified significant pathways, along with corroboration of previously identified biomarkers. Overall, the GSEA confirmed the pathways of the gene expression and methylation data's relation to medulloblastoma, as well as introducing possible new biomarkers for medulloblastoma diagnosis and subgroup identification.

3. Discussion

This study utilised the Random Forest machine learning algorithm to analyse DNA methylation and gene expression data from primary

medulloblastoma samples, aiming to identify significant biological pathways associated with different medulloblastoma subgroups. The results from the Gene Set Enrichment Analysis (GSEA) and subsequent pathway analysis provide credibility to the data's association with medulloblastoma and important insights into the molecular mechanisms underlying medulloblastoma.

The classification accuracy using the Random Forest model averages at around 98%, similar to Cavalli et al. (2017). Utilising the same dataset, their model achieved a higher classification accuracy for WNT subgroups generally, while being generally worse in Group 3 classification compared to this study's models. Compared to Gershanov et al. (2021) which input exclusively gene expression data from the same dataset which achieved an accuracy of 97.8%, this study's gene expression exclusive models overall performed similarly.

The increased recall exclusively stemming from gene expression data lends credibility to exclusivities in patterns based on if the data used is gene expression or DNA methylation mentioned in subtype classification also being applicable to subgroup classification (Cavalli, Florence M G et al., 2017). The sustained similarities throughout methylation data model results and mixed model results could indicate a favouring of methylation data within the training process, possibly stemming from the higher number of variables (probes).

The recall of WNT being higher from models trained exclusively using gene expression data compared to models including DNA methylation profiling could lend slight credibility to the claim that DNA methylation and gene expression contribute differently to the molecular differences between subgroups, expanding to the claim by Florence et al. (2017) that "both DNA methylation and gene expression contributing to the heterogeneity observed within each subgroup."

Pathways identified through the Reactome 2022 Database are crucial for neuronal function and communication. These findings align with Ray et al. (2021), which emphasised the role of neural development pathways in different medulloblastoma subtypes, suggesting that disruptions in these pathways contribute to tumorigenesis and progression.

Enriched pathways highlighted by KEGG human 2021 include functions of synaptic transmission and glutamatergic synapse signalling. The finding that medulloblastoma may lead to an upregulation in the aforementioned pathways is consistent with Korshunovva et al. (2019). Roussel et al. (2011) also noted the significance of neural development pathways in medulloblastoma, reinforcing this study's findings.

The DisGeNet analysis identified significant associations with diseases such as autism disorder, medulloblastoma, and chronic alcohol intoxication. The association with autism spectrum disorder (ASD) suggests that common genetic or epigenetic factors may underlie both medulloblastoma and ASD. This is supported by previous studies indicating shared molecular mechanisms between neurodevelopmental disorders and brain tumours Ray et al. (2021).

The genes from the most significant variables identified by the Random Forest classifier included multiple biomarkers previously associated with MB subgroup classification. The newly discovered biomarkers are the following: EXOC5, LINC02058, SASH1, CYBRD1, ZSWIM5, FBXL21, PLPPR4.

4. Limitations and Future Research Directions

The reason the WNT subgroup received suboptimal results compared to other subgroups within methylation and mixed trials could also be due to a lower number of samples, and thus data within the dataset. Results from methylation input models and mixed models are extremely similar, insinuating a tendency for the Random Forest classifier to favour methylation data, most likely as there are more probes.

Experimental testing to solidify associations of newly identified pathways and biomarkers related to medulloblastoma subgroup classification is required. Furthermore, understanding their roles in MB development and propagation maybe crucial to clinical diagnoses and treatments to improve survival rates.

5. Methodology

Figure 1. Flowchart

Figure 1 represents the general workflow utilised for this research. Data is downloaded from GEO, then pre-processed. Features selection limits variables accounted for in the Random Forest classifier to only the most relevant, reducing computing power required as well as reducing noise. This splits the data into three sets for DNA methylation profiling and gene expression data each. The data are split into training and testing sets, and used to train the Random Forest algorithm. During training, 5-fold cross-validation is used to ensure reliability, and hyperparameter tuning is used to improve accuracy. Then the models are tested with the testing set. Gene set enrichment analysis is subsequently undertaken to understand the biological pathways related the classification in detail. Biomarkers identified from the model are also investigated.

Dataset Description

The dataset "DNA methylation and gene expression profiling of primary medulloblastoma samples" was downloaded from the online database Gene Expression Omnibus (GEO) under the GEO accession code GSE85218. The dataset includes 763 primary medulloblastoma samples, each with a DNA methylation profile and a gene expression profile. Each sample is also classified as one of the 4 medulloblastoma subgroups; 70 WNT samples, 223 SHH samples, 144 Group 3 samples, and 326 Group 4 samples respectively.

Each sample has methylation profiling on 321174 probes, as well as gene expression data on 21641 genes. This amount of data is unnecessary and leads to longer processing times and inaccurate results due to the noise. Hence, feature selection is necessary to root out the most relevant probes and genes.

Data Preprocessing

The files are in txt format, with the samples as the columns and the data as the rows. They are transposed and converted to csv to be used for machine learning. The gene expression data also had irrelevant labels for gene names, which were removed within the preprocessing process.

Feature Selection

As mentioned previously, having excess amounts of data leads to inefficiencies and possible inaccuracies. Therefore, variance th19reshold feature selection was employed to identify the most informative features from the gene expression and DNA methylation data. This method prioritises features which show the greatest variability across samples under the assumption that these features are more likely to be biologically significant.

First, the variances across the gene expression and methylation datasets are calculated individually. The features which displayed the highest (top 1%/0.5%/100) variances are then saved into separate csv files, while the rest are discarded, reducing the dimensionality of the data. Three separate pairs of gene expression/methylation data resulted from this process (top 1%/0.5%/100 variance), leaving methylation profiling with (3212/1606/100) probes and gene expression data with (216/108/100) genes respectively.

Model Training

The 763 medulloblastoma samples are split into a training set and testing set randomly; the training set contains 80% of every subgroup's sample while the testing set contains 20% of every subgroup's samples. The processed training sets are then inputted into the Random Forest classifier from Sci-kit Learn in python. The model parameters were optimised using a grid search with cross-validation to achieve the best possible performance. Below are the modified parameters:

Table 4. Hyperparameters utilised

Hyperparameter	Value	Description
n_estimators	100, 200, 300	Number of trees in the forest

max_depth	10, 20, 30, None	Maximum depth of the tree
min_samples_split	2, 5, 10	Number of samples required to split an internal node
min_samples_leaf	1, 2, 4	Minimum number of samples required to be a leaf node

While the Random Forests taking in exclusively DNA methylation data or exclusively gene expression data could be directly inputted, Random Forests which take in both DNA methylation data and gene expression data required standardisation. The data were standardised to have a mean of 0 and a standard deviation of 1, ensuring equal weighting of the data.

In essence, 5-fold cross validation splits the training data into five equal folds and engages in five iterations with one fold as the testing set and the rest as training sets, then shuffling the folds until all folds have been used as the testing set. By doing so, it prevents overfitting and leads to higher general reliability (Berrar, 2019).

To evaluate the results of the Random Forest classifier, appropriate metrics must be applied. All models trained were evaluated based on the metrics in Table 5. Overall, each feature selected dataset underwent 10 trials, with the highest mean 5-fold cross validation accuracy determining the most successful model used to test the testing set. The final model's performance was evaluated based on its accuracy in predicting the medulloblastoma subgroups. The evaluation metrics included overall accuracy, precision, recall, and F1 score. These metrics provided a comprehensive assessment of the model's predictive capabilities.

Table 5. Definitions of metrics used

Metric	Definition	Symbol/Formula
True positives	Number of correctly predicted positives	TP
True negatives	Number of correctly predicted negatives	TN
False positives	Number of incorrectly predicted positives	FP
False negatives	Number of incorrectly predicted negatives	FN
Accuracy	Ratio of correctly predicted observations against total predicted observations	$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$
Precision	Ratio of correctly predicted positive observations against total predicted positive observations for each classification	$\frac{TP}{\mathrm{TP}+\mathrm{FP}}$
Recall	Ratio of correctly predicted positive observations against total observations which fit the classification	$\frac{TP}{TP+FN}$
Macro Average (MA)	Unweighted mean of precision, recall, and F1 score over all classifications, not taking into account sample size for each classification	Eg: $\frac{Acc(1)+\cdots Acc(N)}{Number\ of\ Classes}$

Weighted Average (WA)	Weighted mean of precision, recall, and F1 score over all classifications, taking into account sample size for each classification	Eg: $\frac{Ass(1) \cdot Sample(1) + \cdots}{2aNumber\ of\ Samples}$
F1 Score	WA of precision and recall for the specific classification	$\frac{2 \cdot precision \cdot recall}{precision + recall}$
Mean Cross Validation Accuracy (Mean CVA)	Average 5 fold cross validation accuracy over 10 trials; ameasure of reliability	$\frac{Meancv(1) + \cdots Meancv(N)}{N}$

Gene set enrichment analysis (GSEA)

The top 1% of methylation data and gene expression data (Table 2) are analysed through GSEA, to better understand the biological pathways regarding medulloblastoma subgrouping. Methylation probes are mapped to their respective genes via the Illumina HumanMethylation450 Beadchip's annotation file, while the Ensembl ids are mapped to their genes via the Affymetrix Human Gene 1.1 ST Array's annotation file. Then, the genes are placed into Ma'ayan laboratory's online GSEA program Enrichr for analysis. The metrics for significance of each pathway within Enrichr (aside from adjusted P-value) is calculated using Fisher's exact test, a statistical significance test used to determine if there are non-random associations between two categorical variables in a contingency table. (Fisher, 1925)

After this, the 20 most significant features identified by the Random Forest Classifier from the top 100 mixed model are analysed and compared to previously identified biomarkers for medulloblastoma.

Table 6. Enrichr pathway databases used

Database	Category	Description
Kegg 2021 Human	Pathway	A comprehensive encyclopaedia for understanding biological functions
Reactome 2022	Pathway	Database of pathways and reactions in the human biology
GO Biological Process 2023	Ontologies	Assists with helping understanding roles of genes in biological processes
DisGeNet	Disease	Information on common and rare gene-disease associations

Table 7. Pathway metrics

Metric	Definition
P-value	Probably of obtaining observed results assuming there is no relation between the pathway and the data
Adjusted P-value	P-value adjusted to account for multiple tests, reducing the chance of false positives.
Odds Ratio	Likelihood of an event occurring in the presence of a particular condition compared to its absence. Returns z-score to assess deviation from expected rank
Combined score	Log of P-value multiplied by Odds ratio.

References

Archer, T. C., Mahoney, E. L., & Pomeroy, S. L. (2017). Medulloblastoma: Molecular classification-based personal therapeutics. *Neurotherapeutics, 14*(2), 265-273. https://doi.org/10.1007/s13311-017-0526-y

Berrar, D. (2019). Cross-validation. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of Bioinformatics and Computational Biology* (pp. 542-545). Academic Press. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Brandes, A. A., et al. (2014). Shedding light on adult medulloblastoma: Current management and opportunities for advances. *ASCO Educational Book, 34*, e82-e87. https://doi.org/10.14694/EdBook_AM.2014.34.e82

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5-32. https://doi.org/10.1023/A:1010933404324

Castillo-Rodríguez, R. A., Dávila-Borja, V. M., & Juárez-Méndez, S. (2018). Data mining of pediatric medulloblastoma microarray expression reveals a novel potential subdivision of the Group 4 molecular subgroup. *Oncology letters, 15*(5), 6241-6250. https://doi.org/10.3892/ol.2018.8094

Cavalli, F. M. G., Remke, M., Rampasek, L., Peacock, J., Shih, D. J. H., Luu, B., Garzia, L., Torchia, J., Nor, C., Morrissy, A. S., Agnihotri, S., Thompson, Y. Y., Kuzan-Fischer, C. M., Farooq, H., Isaev, K., Daniels, C., Cho, B. K., Kim, S. K., Wang, K. C., Lee, J. Y., … Taylor, M. D. (2017). Intertumoral Heterogeneity within Medulloblastoma Subgroups. *Cancer cell, 31*(6), 737–754.e6. https://doi.org/10.1016/j.ccell.2017.05.005

Centers for Disease Control and Prevention, National Center for Health Statistics. (2024). *National Vital Statistics System, Mortality 2018-2022 on CDC WONDER Online Database* [Data set]. Multiple Cause of Death Files, 2018-2022. https://wonder.cdc.gov/ucd-icd10-expanded.html (Accessed: July 6, 2024).

Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics; 128*(14).

Chen, R., Cao, J., Jiang, W., Wang, S., & Cheng, J. (2021). Upregulated Expression of CYBRD1 Predicts Poor Prognosis of Patients with Ovarian Cancer. *Journal of oncology*, 7548406. https://doi.org/10.1155/2021/7548406

Fisher, R.A. (1925). *Statistical methods for research workers* (11th ed. rev.). Oliver and Boyd: Edinburgh.

Gabriel, E., Albanna, W., Pasquini, G., Ramani, A., Josipovic, N., Mariappan, A., Schinzel, F., Karch, C. M., Bao, G., Gottardo, M., Suren, A. A., Hescheler, J., Nagel- Wolfrum, K., Persico, V., Rizzoli, S. O., Altmüller, J., Riparbelli, M. G., Callaini, G., Goureau, O., Papantonis, A., Busskamp, V., Schneider, T., & Gopalakrishnan, J. (2021). Human brain organoids assemble functionally integrated bilateral optic vesicles. *Cell Stem Cell, 28*(10), 1740-1757.e8. https://doi.org/10.1016/j.stem.2021.07.010

Gershanov, S., Madiwale, S., Feinberg-Gorenshtein, G., Vainer, I., Nehushtan, T., Michowiz, S., Goldenberg-Cohen, N., Birger, Y., Toledano, H., & Salmon-Divon, M. (2021). Classifying Medulloblastoma Subgroups Based on Small, Clinically Achievable Gene Sets. *Frontiers in oncology, 11*, 637482. https://doi.org/10.3389/fonc.2021.637482

Ho, T. K. (1995). Random decision forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition* (Vol. 1, pp. 278-282). IEEE. https://doi.org/10.1109/ICDAR.1995.598994

Hooper, C. M., Hawes, S. M., Kees, U. R., Gottardo, N. G., & Dallas, P. B. (2014). Gene expression analyses of the spatio-temporal relationships of human medulloblastoma subgroups during early human neurogenesis. *PloS one, 9*(11), e112909. https://doi.org/10.1371/journal.pone.0112909

Kesherwani, V., Shukla, M., Coulter, D.W. et al. (2020). Long non-coding RNA profiling of pediatric Medulloblastoma. *BMC Med Genomics 13*, 87. https://doi.org/10.1186/s12920-020-00744-7

Kool, M., Koster, J., Bunt, J., Hasselt, N. E., Lakeman, A., van Sluis, P., Troost, D., Meeteren,N. S., Caron, H. N., Cloos, J., Mrsić, A., Ylstra, B., Grajkowska, W., Hartmann, W., Pietsch, T., Ellison, D., Clifford, S. C., & Versteeg, R. (2008). Integrated genomics identifies five medulloblastoma subtypes with distinct genetic profiles, pathway signatures and clinicopathological features. *PloS one, 3*(8), e3088. https://doi.org/10.1371/journal.pone.0003088

Korshunov, A., Sahm, F., Okonechnikov, K.,

Ryzhova, M., Stichel, D., Schrimpf, D., Casalini, B., Sievers, P., Meyer, J., Zheludkova, O., Golanov, A., Lichter, P., Jones, D. T. W., Pfister, S. M., Kool, M., & von Deimling, A. (2019). Desmoplastic/nodular medulloblastomas (DNMB) and medulloblastomas with extensive nodularity (MBEN) disclose similar epigenetic signatures but different transcriptional profiles. *Actaneuropathologica, 137*(6), 1003-1015. https://doi.org/10.1007/s00401-019-01981-6

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, gkw377.

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armañanzas, R., Santafé, G., Pérez, A., & Robles, V. (2006). Machine learning in bioinformatics. *Briefings in bioinformatics, 7*(1), 86–112. https://doi.org/10.1093/bib/bbk007

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. Nature reviews. *Genetics, 16*(6), 321-332. https://doi.org/10.1038/nrg3920

Lin, C. Y., Erkek, S., Tong, Y., Yin, L., Federation, A. J., Zapatka, M., Haldipur, P., Kawauchi, D., Risch, T., Warnatz, H. J., Worst, B. C., Ju, B., Orr, B. A., Zeid, R., Polaski, D. R., Segura-Wang, M., Waszak, S. M., Jones, D. T., Kool, M., Hovestadt, V., … Northcott, P. A. (2016). Active medulloblastoma enhancers reveal subgroup-specific cellular origins. *Nature, 530*(7588), 57-62. https://doi.org/10.1038/nature16546

Liu, H., Tu, M., Yin, Z., Zhang, D., Ma, J., & He, F. (2024). Unraveling the complexity of polycystic ovary syndrome with animal models. *Journal of Genetics and Genomics, 51*(2), 144-158. https://doi.org/10.1016/j.jgg.2023.09.012

Masihi, M. B., Lee, C., Furnari, G. A., Garancher, A., & Wechsler-Reya, R. J. (2020). MBRS-12. A TRANSPOSON MUTAGENESIS SCREEN IDENTIFIES Rreb1 AS A DRIVER FOR GROUP 3 MEDULLOBLASTOMA. *Neuro-Oncology, 22*(Suppl 3), iii400. https://doi.org/10.1093/neuonc/noaa222.529

Minnoor, M., & Baths, V. (2023). Diagnosis of breast cancer using random forests. *Procedia Computer Science, 218*, 429-437. https://doi.org/10.1016/j.procs.2023.01.025

Orr, B. A. (2020). Pathology, diagnostics, and classification of medulloblastoma. *Brain Pathology, 30*, 664-678. https://doi.org/10.1111/bpa.12837

Pedregosa, Fabian, et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*(85), pp. 2825-2830.

Ray, S., Chaturvedi, N. K., Bhakat, K. K., Rizzino, A., & Mahapatra, S. (2021). Subgroup-specific diagnostic, prognostic, and predictive markers influencing pediatric medulloblastoma treatment. *Diagnostics, 12*(1), 61. https://doi.org/10.3390/diagnostics12010061

Rodrigues, P., Macaya, I., Bazzocco, S. et al. (2014). RHOA inactivation enhances Wnt signalling and promotes colorectal cancer. *Nat Commun, 5*, 5458. https://doi.org/10.1038/ncomms6458

Rossi, A., Caracciolo, V., Russo, G., Reiss, K., & Giordano, A. (2008). Medulloblastoma: from molecular pathology to therapy. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 14*(4), 971-976. https://doi.org/10.1158/1078-0432.CCR-07-2072

Roussel, M. F., & Hatten, M. E. (2011). Cerebellum development and medulloblastoma. *Current topics in developmental biology, 94*, 235-282. https://doi.org/10.1016/B978-0-12-380916-2.00008-5

Shin, J., Choi, J. H., Jung, S., Jeong, S., Oh, J., Yoon, D. Y., Rhee, M. H., Ahn, J., Kim, S. H., & Oh, J. W. (2019). MUDENG Expression Profiling in Cohorts and Brain Tumor Biospecimens to Evaluate Its Role in Cancer. *Frontiers in genetics, 10*, 884. https://doi.org/10.3389/fgene.2019.00884

Taylor, M.D., Northcott, P.A., Korshunov, A. et al. Molecular subgroups of medulloblastoma: the current consensus. *Acta Neuropathol 123*, 465-472 (2012). https://doi.org/10.1007/s00401-011-0922-z

Wu, F., Zhang, Y., Sun, B., McMahon, A. P., &

Wang, Y. (2017). Hedgehog Signaling: From Basic Biology to Cancer Therapy. *Cell chemical biology, 24*(3), 252-280. https://doi.org/10.1016/j.chembiol.2017.02.010

Xie Z, Bailey A, Kuleshov MV, Clarke DJB., Evangelista JE, Jenkins SL, Lachmann A, Wojciechowicz ML, Kropiwnicki E, Jagodnik KM, Jeon M, & Ma'ayan A. (2021). Gene set knowledge discovery with Enrichr. *Current Protocols, 1*, e90. doi: 10.1002/cpz1.90

Xu, K., Liu, B., Ma, Y., Xu, B., & Xing, X. (2018). A novel SWIM domain protein ZSWIM5 inhibits the malignant progression of non-small-cell lung cancer. *Cancer management and research, 10*, 3245-3254. https://doi.org/10.2147/CMAR.S174355

Yang, L., Zhang, H., Yao, Q., Yan, Y., Wu, R., & Liu, M. (2015). Clinical Significance of SASH1 Expression in Glioma. *Disease markers*, 383046. https://doi.org/10.1155/2015/383046