Developing a Novel Medulloblastoma Diagnostic with miRNA Biomarkers and Machine Learning

Chloe Wang[a], Valentina Kouznetsova[b-d], Santosh Kesari[f], Igor Tsigelny[b-e]

[a] Mentor Assistance Program, San Diego Supercomputer Center, University of California San Diego, La Jolla, CA, USA

[b]San Diego Supercomputer Center, University of California San Diego, La Jolla, CA, USA

[c] CureScience Institute, San Diego, CA, USA

[d] BiAna, San Diego, CA, USA

[e] Department of Neurosciences, University of California San Diego, La Jolla, CA, USA

[f] Pacific Neuroscience Institute, Santa Monica, CA, USA

Abstract

Background: Medulloblastoma (MB) is the most common malignant brain tumor in children. Current diagnostic methods such as MRI and lumbar punctures are invasive and costly, making early diagnosis challenging. MicroRNAs (miRNAs) have emerged as promising biomarkers for cancer diagnosis due to their dysregulated expression in tumors. This study aims to develop a novel machine learning (ML)-based diagnostic tool for MB using miRNA biomarkers.

Methods: We collected miRNAs associated with MB and random controls, generating sequence- and target gene-based descriptors. We employed the WEKA software to evaluate several ML models, including Logistic Regression, Naive Bayes, and Multilayer Perceptron (MLP). Attribute selection reduced noise by selecting the most significant 24 features. Model performance was evaluated using 10-fold cross-validation and independent test datasets.

Results: Logistic Regression achieved the highest training accuracy (96·2%), while the MLP model was selected for further testing due to its ability to capture complex nonlinear relationships in biological data. The MLP model showed 78·6% accuracy on an independent MB dataset and successfully distinguished MB miRNAs from those associated with chronic myeloid leukemia (CML), further validating its specificity.

Conclusion: The ML-based diagnostic tool using miRNA biomarkers shows promise for improving MB diagnosis, offering a non-invasive and cost-effective alternative to traditional methods. Further validation with larger datasets and diverse control groups is needed to refine the model.

Keywords: Medulloblastoma, microRNA, machine learning, biomarker, diagnostic tool, Multilayer Perceptron.

Introduction

Medulloblastoma (MB) is a malignant central nervous system (CNS) tumor that begins in the cerebellum.[1] It is the most common type of cancerous brain tumor in children and 40% of all posterior fossa tumors are MB.[2] The five-year survival rate for medulloblastoma is around 80%.[3] Being an aggressive brain tumor, earlier detection of MB can significantly increase the survival rate. Additionally, a significant proportion of high-risk patients will relapse from the disease, despite aggressive therapy. Therefore, a convenient follow-up diagnostic tool is desired.

The standard procedure for MB diagnosis is an initial neurological exam, followed by computed tomography (CT) scan and magnetic resonance imaging (MRI) brain scans. CT scans are typically used as the initial imaging method due to their availability and speed; however, non-contrast CT scans often fail to detect medulloblastomas.[4] If a posterior fossa mass is detected, an MRI is used to confirm the diagnosis.[2] After alleviating the brain pressure or removing the tumor, a lumbar puncture, also known as a spinal tap, is performed to check MB cells in the cerebrospinal fluid for confirmation.[4] It takes multiple steps to fully confirm the diagnosis, which can be both expensive and time-consuming. Also, procedures like MRI scans and lumbar punctures are primarily only accessible in high-income countries,[5] but even there, the costs may be prohibitively high for some individuals, deterring them from getting an early diagnosis and treatment, which can lower their survival rate. Therefore, a more cost and time-efficient diagnostic method would help improve the survival rate and benefit MB patients.

MicroRNAs (miRNAs) are short, noncoding RNA segments, measuring 20-22 nucleotides, which are essential for regulating biological processes in multicellular organisms. MiRNAs have demonstrated promise as biomarkers for cancer diagnosis due to differences in their expression levels between normal and cancerous tissues and numerous reports showing that over half of the

miRNA's genes are found within genomic regions associated with cancer or fragile sites.[6] miRNAs can bind to complementary sequences in the 3′ UTR of target genes, modulating gene expression by repressing translation and/or promoting deadenylation followed by mRNA degradation. A single miRNA may target hundreds of mRNAs, while each mRNA can be regulated by multiple miRNAs. This type of modulation allows miRNAs to influence numerous signaling pathways and cellular processes. Changes in miRNA expression are linked to various cancers, including MB, where they impact the expression of tumor suppressor genes, oncogenes, and other signaling molecules.[7] Reverse transcription-quantitative polymerase chain reaction (qRT-PCR) is a widely utilized, efficient, and minimally invasive method for quantifying gene expression.[6,8] In a study comparing miRNA expression in tissues of MB patients versus healthy individuals, Ferretti *et al.* found that 78 of the 248 analyzed miRNAs displayed dysregulated expression, where some of them play a role as tumor suppressors or oncogenes.[9] Furthermore, Dai *et al.* conducted an analysis of a Gene Expression Omnibus (GEO) miRNA expression profiling dataset and discovered that 22 miRNAs were upregulated, while 26 miRNAs were downregulated in MB tissue relative to healthy non-MB tissue.[10] These studies highlight the critical role and potential value of using miRNAs as biomarkers in the diagnosis of medulloblastoma.

There have been many attempts to classify the four subgroups of MB with machine learning (ML) based on radiomics and biological markers. For example, Saju *et al.* utilized radiomic analysis of MRI scans to accurately predict the molecular subgroups of MB.[11] In a separate study, Attallah developed a computer-aided diagnostic system named MB-AI-His by integrating machine learning with various techniques, including "image processing, spatial feature extraction, time-frequency feature extraction, feature fusion and reduction, and classification."[12]

This system was designed for the automated diagnosis of pediatric medulloblastoma and its subtypes using histopathological images.[12] Finally, a study by Gómez *et al.* classified the subtypes of medulloblastoma using machine learning and DNA methylation profiles.[13]

Kang *et al.* developed a new approach utilizing miRNA biomarkers for ML diagnostic of several diseases.[14] Several other studies demonstrated the effectiveness of miRNA as a biomarker in diagnostic ML models achieving 87-96% accuracy for Alzheimer's disease, Parkinson's disease, and laryngeal cancer diagnosis.[15–18] However, to the best of our knowledge, there is no existing research that uses biological attributes of miRNA for MB diagnosis. Thus, we developed a novel ML-based diagnostic method utilizing biomarkers of cancer-associated miRNAs. This approach offers advantages such as low cost, increased efficiency, reduced invasiveness, and the potential for improved diagnostic accuracy either used alone or combined with other methods.

Methods

Figure 1 outlines the methodology used in this study. Initially, miRNAs with known dysregulation in MB were identified and confirmed based on earlier research. Then we selected random miRNAs with no known association with MB. Next, descriptors were generated using both the sequences of these miRNAs and their target genes' names. These descriptors were then analyzed using the Waikato Environment for Knowledge Analysis (WEKA),[19] where several machine learning classifiers were employed to differentiate the miRNAs.

Data Collection

For the training data, we obtained a list of miRNAs associated with MB from a comprehensive study conducted by Bevacqua *et al.* that consolidated recent findings from various sources.[20] Additionally, a set of negative control miRNAs that have no association with MB was randomly

selected from miRBase (https://mirbase.org/).[21] We developed a Python script to randomly choose the control set, as well as find the sequences of all the miRNAs. All included miRNAs are shown in Table 1.

Sequence-based attributes

We adopted Kang *et al.*'s method of generating sequence-based attributes, which involves analyzing the number of nucleotides, frequency, mean mass, hydrogen bonds, and motifs.[14] We analyzed the miRNA sequences because miRNAs identify their target mRNAs through sequence complementarity and subsequently inhibit protein translation by inducing mRNA degradation.[22]

Target gene-based attributes

We opted to use target genes as descriptors since miRNAs influence cancer progression through these genes. When miRNAs become dysregulated, they can influence cancer progression by affecting their target genes, which may function as either oncogenes or tumor suppressors, depending on the specific miRNAs involved.[23] To compile a list of predicted target genes for each miRNA, we utilized miRDB database (https://mirdb.org/mirdb/index.html),[24,25] which provides prediction scores for suggested target genes. We used a Python script to identify target genes with a score above 96 for each miRNA,[17] and these genes were added to our list of attributes. Subsequently, we created a script that iterated through all miRNAs and their target genes.

Attribute selection

With our complete training set of miRNAs and their 1912 attributes, we proceeded to load the data into WEKA. Due to the high number of attributes, many of which could introduce noise to the model, we utilized WEKA's CfsSubsetEval feature to perform attribute selection on the

dataset, retaining only the twenty-four most significant attributes for classification. This feature measures the significance of a subset of attributes by examining both their individual predictive powers and the level of redundancy among them. It prioritizes subsets where the features are strongly correlated with the target class but exhibit minimal correlation with each other.[19]

Training the model

After inputting the set of attributes into WEKA, we used the classification tool and assessed different classification algorithms to see which of them produced the highest classification accuracy. After importing the attribute set into WEKA, we used the classification tool to evaluate various models, aiming to identify the one with the highest classification accuracy. Our objective was to optimize multiple metrics, including accuracy, true-positive rate, false-positive rate, precision, recall, F-measure, Matthew's correlation coefficient (MCC), area under the receiver-operating characteristic curve, and area under the precision-recall curve.

We evaluated multiple models, with the top performers being Logistic Regression at 96·15%, Hoeffding Tree and Naïve Bayes both at 92·31%, Multilayer Perceptron at 88·46%, and Random Forest at 84·62%. Each algorithm was assessed using 10-fold cross-validation, which splits the dataset into ten folds, tests each fold as a validation set, and trains the model on the remaining folds, preventing overfitting in the model. WEKA then averaged the results across these tests to provide an accuracy measure. The five models mentioned performed the best across the different metrics, so we exported all of them from WEKA to use for testing independent data.

Testing models with independent datasets

During the validation phase, we tested the model using a distinct set of miRNAs strongly associated with medulloblastoma (MB), along with another set related to a different condition, chronic myeloid leukemia (CML), a type of bone marrow cancer.

To get an independent dataset that contains miRNA known to be related to MB, we used miR2Disease (http://www.mir2disease.org/), a database that provides an extensive compilation of miRNA deregulation in various human diseases.[26] We identified 14 dysregulated miRNAs for MB in the miR2Disease database that were not part of the training set. To generate the sequence-based attributes of the test set, we followed the same procedure used for the miRNAs in the training set. However, instead of creating new target-gene-based attributes based on the test set, we relied solely on the attributes from the training set. Once the testing set was finalized, we loaded each of the five models that we previously saved into WEKA and tested the independent data on them.

From a systematic review of miRNAs associated with CML, we were able to gather eleven dysregulated miRNAs.[27] We used the same procedure to generate the descriptors and tested the best-performing ML model, which we found to be MLP.

Pathway Analysis

By using the CfsSubsetEval feature in WEKA, we were able to identify the most significant attributes; among those, five were target genes. To further analyze these genes, we queried the GeneFriends database (https://www.genefriends.org/) for co-expressed genes.[28] We extracted genes categorized under the "protein coding" biotype with a p-value less than 10^{-4} using a custom Python script, resulting in a list of 131 co-expressed genes. To find the pathways

involved these genes are involved with, we used Database for Annotation, Visualization and Integrated Discovery (DAVID) (https://david.ncifcrf.gov/home.jsp).[29,30] First, we uploaded the list of significant and co-expressed genes to DAVID's Gene ID Conversion Tool, selecting the ENTREZ_GENE_ID (default) option for Homo sapiens. After conversion, the gene list was analyzed using DAVID's Functional Annotation Tool to identify relevant Reactome pathways.

Results

Model Validation and Justification

During the cross-validation phase, we evaluated the performance of five different ML models to identify the most suitable one for MB diagnosis. We decided to assess the top five models rather than relying on a single model because different models may excel in various performance metrics. Additionally, some models might have high accuracy on the training set but perform poorly on unseen data. Evaluating multiple models helps in identifying those with a lower error rate on test data, ensuring that the chosen model is not just memorizing the training data but also understanding the underlying patterns of descriptors. As shown in Table 4, the Logistic Regression model achieved the highest accuracy during training with a score of 96·2%, followed closely by the Hoeffding Tree and Naive Bayes models, each with an accuracy of 92·3%. The MLP model, with a training accuracy of 88·5%, did not initially emerge as the top performer.

However, despite its lower training accuracy, the MLP model was selected for further testing due to its unique ability to handle complex, nonlinear relationships within the data, which are characteristic of miRNA interactions in MB. The MLP model's architecture allows it to learn intricate patterns that may be missed by more straightforward models, making it a valuable tool in the context of biological data.

Independent new data testing

As seen in Figure 2, among the five models that were chosen after initial dataset cross-validation, the best-performing in the independent new dataset testing is the Multilayer Perceptron (MLP). We input the patient's dataset containing 14 MB-associated miRNAs into our model, and it correctly confirmed the diagnosis of MB with an accuracy of 78.6%. Additionally, it can be seen in Table 4 that its ROC and PRC areas are also high, demonstrating this classifier's ability to discriminate between positive and negative cases and identify true positives.

After evaluating the initial independent dataset, we moved on to test a dataset completely unrelated to MB, to confirm that our model could accurately identify MB without misclassifying other diseases. To test this, we introduced miRNA data from CML to our classifier, which was trained on MB-specific miRNAs. We expected the model to show a low rate of false positives, demonstrating its ability to differentiate between the miRNA profiles of distinct diseases. The results supported this expectation: the model that showed a high accuracy of 78·6% on the MB-specific miRNA dataset showed an accuracy that dropped sharply to 18.2% when applied to the CML dataset. These outcomes suggest that the model is specific to MB, though additional studies with various control groups and more rigorous statistical tests are needed for stronger validation. Thus, the MLP model successfully addresses the clinical requirement for precise MB detection while also reducing the likelihood of false positives.

We confirmed that our MLP model has enough miRNAs used for training and attributes retained after feature selection. These two factors significantly influenced accuracy: the number of miRNAs dictated how much data the model was trained on, while the number of attributes determined which biological characteristics were most crucial for accurate classification.

Figure 3 illustrates the performance of the MLP model with varying numbers of miRNA in the training dataset. The figure confirms that our model is optimized with 52 miRNAs in the training data. Furthermore, it shows improved performance as the number of miRNAs increases, indicating that incorporating additional miRNAs associated with MB into the training dataset could enhance the model's accuracy. Figure 4 displays the accuracy of our model with varying numbers of attributes in the training dataset. The results indicate that using either fewer or more attributes reduce accuracy, affirming that the 24 selected attributes were indeed the most significant contributors to the high performance of our model.

Analysis of model attributes

We also examined the attributes selected. The CfsSubsetEval program assessed the importance of attribute subsets, favoring those where features were highly correlated with the target class but showed minimal correlation with each other. Consequently, the selected attributes highlight the most relevant biological characteristics of miRNAs that link them to MB.

The target genes KCNK10, PRTG, FRS2, NR6A1, and NR2C2, selected by the CfsSubsetEval attribute selection, hold significant potential in advancing our understanding, diagnosis, and treatment of MB.

KCNK10 encodes a potassium channel protein that regulates cell membrane potential and electrical excitability,[31] and its dysregulation may impact the growth and survival of MB tumor cells.

PRTG is involved in cell adhesion and signaling, playing a critical role in neural development. Incorporating information from recent research into PRTG's role in medulloblastoma can enhance the understanding of its relevance in this cancer type. PRTG has been identified as a

crucial marker for Group 3 medulloblastoma, one of the most aggressive and poorly characterized subtypes. Research conducted by Visvanathan and colleagues[32] revealed that PRTG is expressed in the embryonic cells from which these tumors originate. This finding highlights the significance of PRTG in identifying the cell of origin for Group 3 medulloblastoma, which is crucial for developing targeted therapies.

FRS2 acts as an adaptor protein linking fibroblast growth factor receptors (FGFRs) to downstream signaling pathways, including MAPK and PI3K/AKT, which are critical for cell proliferation and survival.[33] Moreover, recent findings indicate that FRS2 plays a role in driving dissemination and tissue invasion in MB, making it a promising candidate for use as a diagnostic biomarker and a therapeutic target.[34]

NR6A1 encodes a nuclear receptor that regulates gene expressions involved in cell differentiation and development.[35] Given that MB arises from undifferentiated neural progenitor cells,[36] NR6A1's role in regulating differentiation could be critical. Dysregulated NR6A1 expression may lead to an undifferentiated, proliferative state in MB cells, promoting tumor growth.

NR2C2 also encodes a nuclear receptor involved in regulating gene expression related to cell growth and differentiation.[37] Thus, similar to NR6A1, NR2C2's regulation of growth and differentiation processes is crucial in preventing tumorigenesis. Its dysregulation in MB could result in aberrant cell cycle control and differentiation, contributing to the formation and maintenance of the tumor.

The selection of these genes for our ML model reflects their potential as key biomarkers for MB. By focusing on attributes with high biological relevance, our model enhances its ability to

accurately classify and predict MB. This approach not only aligns with the current understanding of MB but also validates the model's effectiveness in identifying critical factors that influence tumor characteristics and progression.

Integrative analysis of the co-expressed genes

The attribute selection process, facilitated by the CfsSubsetEval feature in WEKA, identified twenty-four significant target genes including KCNK10, PRTG, FRS2, NR6A1, and NR2C2. These genes play crucial roles in biological processes relevant to MB pathogenesis, contributing to the high performance of our model. Upon further investigation, we discovered that these genes are intricately linked to pathways associated with MB.

Transcriptional regulation of pluripotent stem cells and its proliferative sub-pathway

Many of the co-expressed genes from our model are implicated in the transcriptional regulation of the pluripotent stem cells pathway, a critical regulator of both stem cell maintenance and tumorigenesis. This pathway governs the balance between pluripotency and differentiation, orchestrated by key transcription factors such as POU5F1 (OCT4), SOX2, and NANOG.[38] These factors play a central role in maintaining stem cells in an undifferentiated state, and their dysregulation is associated with unchecked cellular proliferation, a hallmark of aggressive cancers like MB.

Within this broader pathway lies the specific sub-pathway where POU5F1 (OCT4), SOX2, and NANOG activate genes directly related to cellular proliferation. In the context of medulloblastoma, the overexpression of OCT4 has been shown to drive tumor aggressiveness through the activation of the mTOR signaling pathway. In a study that developed humanized models of the Sonic Hedgehog (SHH) subgroup of MB by overexpressing MYCN in different

types of human stem cells, it was reported that mTOR activation, a result of increased OCT4, contributes to tumor aggressiveness and demonstrates the potential of targeting mTOR for treatment.[39] The activation of mTOR signaling due to increased levels of OCT4, a gene linked with stem cell self-renewal and proliferation, underscores a mechanism by which these tumors become more aggressive. This aligns with our findings that genes such as NR6A1 and NANOG, which are linked to these transcriptional circuits, may play pivotal roles in MB progression.

By demonstrating that several genes identified in our model are associated with this proliferative sub-pathway, we further reinforce their significance in SHH medulloblastoma pathogenesis. These findings suggest that targeting the mTOR pathway, driven by OCT4, could offer a promising therapeutic strategy for treating MB. This connection underscores the relevance of the genes we identified and their role in the regulation of pluripotency and tumorigenesis, particularly within the SHH subgroup.

GPCR ligand binding pathway and Secretin family receptors

Many of the co-expressed genes are part of the extensive GPCR ligand-binding pathway, particularly in the sub-pathway of Class B/2 Secretin family receptors, also known as Adhesion G protein-coupled receptors (ADGRs). Notably, in a study comparing gene expressions between normal cerebella and MB samples, it was found that "the ADGRB1 gene, which encodes Brain-specific Angiogenesis Inhibitor 1 (BAI1), is epigenetically silenced in medulloblastoma."[40] This suggests that BAI1 functions as a tumor suppressor, and its silencing contributes to MB progression.[40]

The involvement of the co-expressed genes in critical signaling pathways enhances our understanding of MB progression and highlights new potential targets for therapies. This is

consistent with the genes selected by our model, suggesting that pathways associated with ADGRs may serve as valuable biomarkers for MB diagnosis and treatment. Identifying how these pathways are disrupted in MB provides insight into the molecular mechanisms driving the disease, thereby allowing for more precise therapeutic strategies targeting its root causes.

Discussion

Traditional diagnostic methods for MB, such as MRI scans and lumbar punctures, are the current gold standard for identifying and confirming the presence of tumors. While these methods are effective, they are also invasive, costly, and may not be readily available in all regions, potentially delaying diagnosis and treatment. Additionally, these conventional techniques often lack the molecular precision necessary to identify the specific biological markers associated with MB, which is crucial for personalized treatment strategies.

In response to these limitations, our study presents an ML approach for diagnosing MB using miRNA biomarkers, focusing on the sequence-based and target gene-based attributes of miRNAs. This novel method of diagnosing MB uses miRNA characteristics to improve diagnostic accuracy while maintaining cost-effectiveness and reducing invasiveness compared to conventional diagnostic methods. The cross-validation accuracy of the diagnostics is more than 85%.

We evaluated multiple ML models, with the MLP model demonstrating the highest accuracy of 78·6% on the independent new dataset. The MLP algorithm proved particularly efficient due to its ability to handle complex patterns in the data. The neural network structure allows for capturing intricate relationships between miRNA sequences and their associations with MB. The ROC area indicates the robustness of the model, as it optimizes the classification rate while

minimizing false positives. This highlights the effectiveness of our model to accurately identify critical miRNA attributes linked to MB.

The methodology outlined in our study can be adapted for other diseases by generating attributes for a set of miRNAs known to have associations with the disease. This approach could identify biological processes contributing to various conditions, as miRNAs are linked to many diseases.

Nevertheless, before this method can be fully implemented, further research is required to validate and refine the findings. Future studies should consider incorporating additional attributes, such as age and gender, as these factors are linked to higher incidence rates and could enhance model accuracy.[41] Moreover, integrating this miRNA-based diagnostic method with existing tools like MRI scans could be especially advantageous in regions with limited access to standard diagnostics. In such areas, our study offers a comprehensive, affordable, and less invasive alternative, potentially improving early detection and treatment of MB.

Conclusion

Our research presents a comprehensive analysis and robust application of the MLP classifier for identifying miRNA associations with MB, achieving an 88·5% accuracy, which was further validated with an independent dataset at 78·6% accuracy. Our approach included selecting miRNAs known to be associated with MB and generating both sequence and target gene-based attributes. After employing attribute selection to determine the most significant factors for MB diagnosis, we trained multiple machine-learning models. Among these, the MLP model, utilizing as many miRNAs as possible and twenty-four attributes in the training dataset, proved to be optimal based on various performance metrics. While further research is necessary, our findings

pave the way for more accessible and less invasive diagnostic alternatives, thus advancing the field of cancer diagnostics and patient care.

References

1 Roussel MF, Hatten ME. Chapter 8 - Cerebellum: Development and Medulloblastoma. In: Dyer MA, ed. Current Topics in Developmental Biology. Academic Press, 2011: 235–82.

2 Dhall G. Medulloblastoma. *Journal of Child Neurology* 2009; **24**. DOI:10.1177/0883073809341668.

3 Salloum R, Chen Y, Yasui Y, *et al.* Late Morbidity and Mortality Among Medulloblastoma Survivors Diagnosed Across Three Decades: A Report From the Childhood Cancer Survivor Study. *JCO* 2019; **37**: 731–40.

4 Quinlan A, Rizzolo D. Understanding medulloblastoma. *JAAPA* 2017; **30**: 30.

5 Looking towards the future of MRI in Africa. *Nat Commun* 2024; **15**: 2260.

6 Reddy KB. MicroRNA (miRNA) in cancer. *Cancer Cell International* 2015; **15**: 38.

7 Kumar V, Kumar V, Chaudhary AK, Coulter DW, McGuire T, Mahato RI. Impact of miRNA-mRNA Profiling and Their Correlation on Medulloblastoma Tumorigenesis. *Molecular Therapy - Nucleic Acids* 2018; **12**: 490–503.

8 Le MN, Nguyen TA. Innovative microRNA quantification by qPCR. *Molecular Therapy - Nucleic Acids* 2023; **31**: 628–30.

9 Ferretti E, De Smaele E, Po A, *et al.* MicroRNA profiling in human medulloblastoma. *International Journal of Cancer* 2009; **124**: 568–77.

10 Dai J, Li Q, Bing Z, *et al.* Comprehensive analysis of a microRNA expression profile in pediatric medulloblastoma. *Molecular Medicine Reports* 2017; **15**: 4109–15.

11 Saju AC, Chatterjee A, Sahu A, *et al.* Machine-learning approach to predict molecular subgroups of medulloblastoma using multiparametric MRI-based tumor radiomics. *British Journal of Radiology* 2022; **95**: 20211359.

12 Attallah O. MB-AI-His: Histopathological Diagnosis of Pediatric Medulloblastoma and its Subtypes via AI. *Diagnostics* 2021; **11**: 359.

13 Gómez S, Garrido-Garcia A, Garcia-Gerique L, *et al.* A Novel Method for Rapid Molecular Subgrouping of Medulloblastoma. *Clinical Cancer Research* 2018; **24**: 1355–63.

14 Kang W, Kouznetsova VL, Tsigelny IF. miRNA in Machine-learning-based Diagnostics of Cancers. *Cancer Screening and Prevention* 2022; **1**: 32–8.

15 Aravind VA, Kouznetsova VL, Kesari S, Tsigelny IF. Using Machine Learning and miRNA for the Diagnosis of Esophageal Cancer. *J Appl Lab Med* 2024; **9**: 684–95.

16 Xu A, Kouznetsova VL, Tsigelny IF. Alzheimer's Disease Diagnostics Using miRNA Biomarkers and Machine Learning. *J Alzheimers Dis* 2022; **86**: 841–59.

17 Arora A, Tsigelny IF, Kouznetsova VL. Laryngeal cancer diagnosis via miRNA-based decision tree model. *Eur Arch Otorhinolaryngol* 2024; **281**: 1391–9.

18 Kumar A, Kouznetsova VL, Kesari S, Tsigelny IF. Parkinson's Disease Diagnosis Using miRNA Biomarkers and Deep Learning. *Front Biosci (Landmark Ed)* 2024; **29**: 4.

19 Witten IH, Frank E, Hall MA, Pal CJ. Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques, 4th edn. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2016.

20 Bevacqua E, Farshchi J, Niklison-Chirou MV, Tucci P. Role of MicroRNAs in the Development and Progression of the Four Medulloblastoma Subgroups. *Cancers (Basel)* 2021; **13**: 6323.

21 Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Res* 2019; **47**: D155–62.

22 Ha T-Y. MicroRNAs in Human Diseases: From Cancer to Cardiovascular Disease. *Immune Netw* 2011; **11**: 135–54.

23 Peng Y, Croce CM. The role of MicroRNAs in human cancer. *Sig Transduct Target Ther* 2016; **1**: 1–9.

24 Chen Y, Wang X. miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Research* 2020; **48**: D127–31.

25 Liu W, Wang X. Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data. *Genome Biology* 2019; **20**: 18.

26 Jiang Q, Wang Y, Hao Y, *et al.* miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res* 2009; **37**: D98-104.

27 Elias MH, Syed Mohamad SF, Abdul Hamid N. A Systematic Review of Candidate miRNAs, Its Targeted Genes and Pathways in Chronic Myeloid Leukemia–An Integrated Bioinformatical Analysis. *Front Oncol* 2022; **12**: 848199.

28 Raina P, Guinea R, Chatsirisupachai K, *et al.* GeneFriends: gene co-expression databases and tools for humans and model organisms. *Nucleic Acids Research* 2023; **51**: D145–58.

29 Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009; **4**: 44–57.

30 Sherman BT, Hao M, Qiu J, *et al.* DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res* 2022; **50**: W216–21.

31 Dong YY, Pike ACW, Mackenzie A, *et al.* K2P channel gating mechanisms revealed by structures of TREK-2 and a complex with Prozac. *Science* 2015; **347**: 1256–9.

32 Visvanathan A, Saulnier O, Chen C, *et al.* Early rhombic lip Protogenin+ve stem cells in a human-specific neurovascular niche initiate and maintain group 3 medulloblastoma. *Cell* 2024; **187**: 4733-4750.e26.

33 Santhana Kumar K, Neve A, Guerreiro Stucklin AS, *et al.* TGF-β Determines the Pro-migratory Potential of bFGF Signaling in Medulloblastoma. *Cell Rep* 2018; **23**: 3798-3812.e8.

34 Santhana Kumar K, Brunner C, Schuster M, *et al.* Discovery of a small molecule ligand of FRS2 that inhibits invasion and tumor growth. *Cell Oncol (Dordr)* 2023; **46**: 331–56.

35 Wang Y, Wan X, Hao Y, *et al.* NR6A1 regulates lipid metabolism through mammalian target of rapamycin complex 1 in HepG2 cells. *Cell Commun Signal* 2019; **17**: 77.

36 Faria Assoni A, Giove Mitsugi T, Wardenaar R, *et al.* Neurodegeneration-associated protein VAPB regulates proliferation in medulloblastoma. *Sci Rep* 2023; **13**: 19481.

37 Zhuang W, Qian L, Fei W, *et al.* The role of NR2C2 in the prolactinomas: NR2C2 targeted by miR-129-5p in prolactinomas. *Open Chemistry* 2018; **16**: 817–26.

38 Chen L. A Balanced Network: Transcriptional Regulation in Pluripotent Stem Cells. *J Stem Cell Res Ther* 2012; **01**. DOI:10.4172/2157-7633.S10-004.

39 Čančer M, Hutter S, Holmberg KO, *et al.* Humanized Stem Cell Models of Pediatric Medulloblastoma Reveal an Oct4/mTOR Axis that Promotes Malignancy. *Cell Stem Cell* 2019; **25**: 855-870.e11.

40 Zhu D, Osuka S, Zhang Z, *et al.* BAI1 Suppresses Medulloblastoma Formation by Protecting p53 from Mdm2-mediated Degradation. *Cancer Cell* 2018; **33**: 1004-1016.e5.

41 Soon WC, Goacher E, Solanki S, *et al.* The role of sex genotype in paediatric CNS tumour incidence and survival. *Childs Nerv Syst* 2021; **37**: 2177–86.

Figures and Tables

Legends

Figure 1. Flowchart of methodology used in this study

Figure 2. Comparison of the accuracies of five different ML classifiers on the independent dataset

Figure 3. Performance of MLP model with different numbers of miRNA in the training dataset

Figure 4. Performance of MLP model with different numbers of attributes in the training dataset

Table 1. MB and Negative Control miRNA Training Set

Table 2. Independent MB set of miRNAs

Table 3. Independent CML set of miRNAs

Table 4. Performance metric scores of top five ML models during 10-fold cross validation; TP Rate = rate of true positives, FP Rate = rate of false positives, Precision = proportion of instances that are truly of a class divided by the total instances classified as that class, Recall = equivalent to TP rate, F-Measure = A combined measure for precision and recall, MCC = A combined measure for precision and recall, ROC Area = Receiver Operating Characteristic Area, PRC Area = Precision-Recall Curve Area

Figure 1

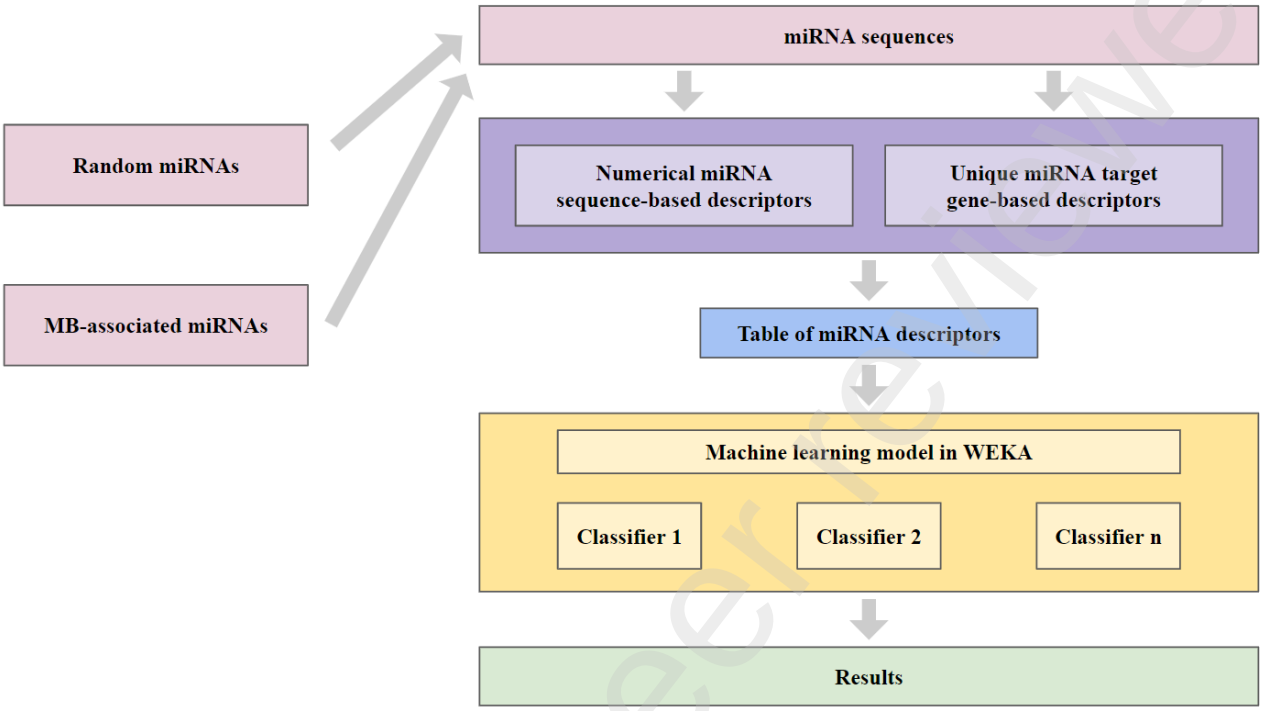

Figure 2

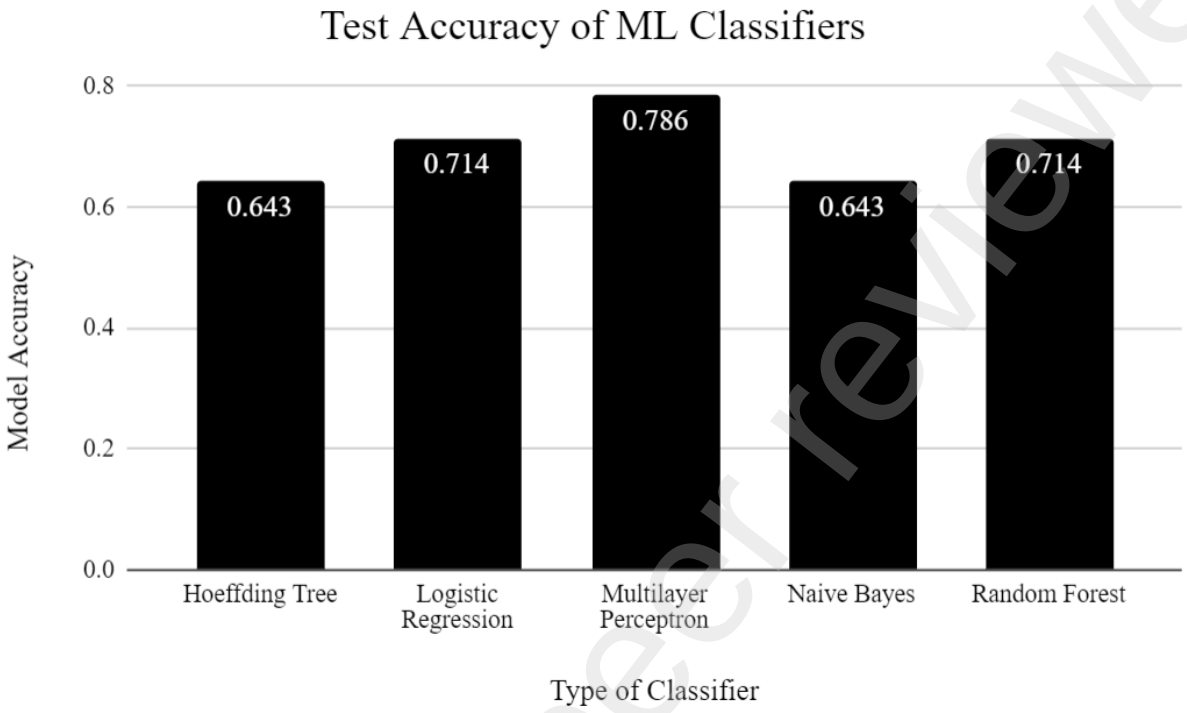

Figure 3

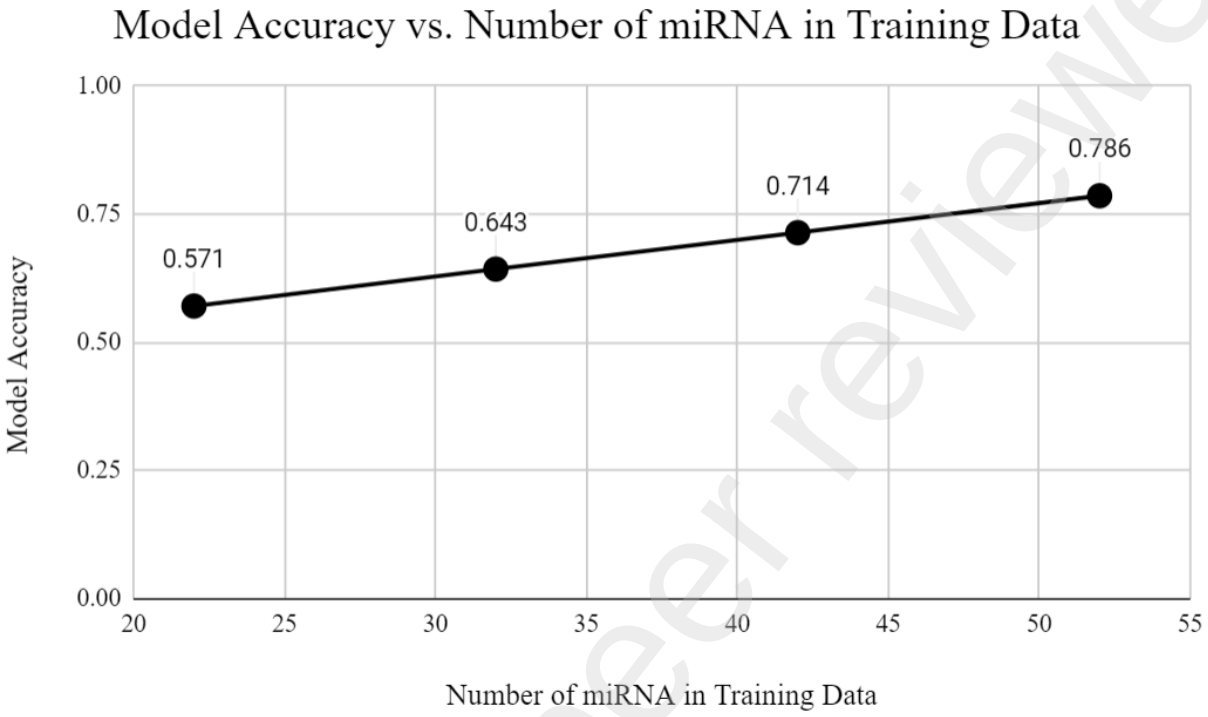

Figure 4

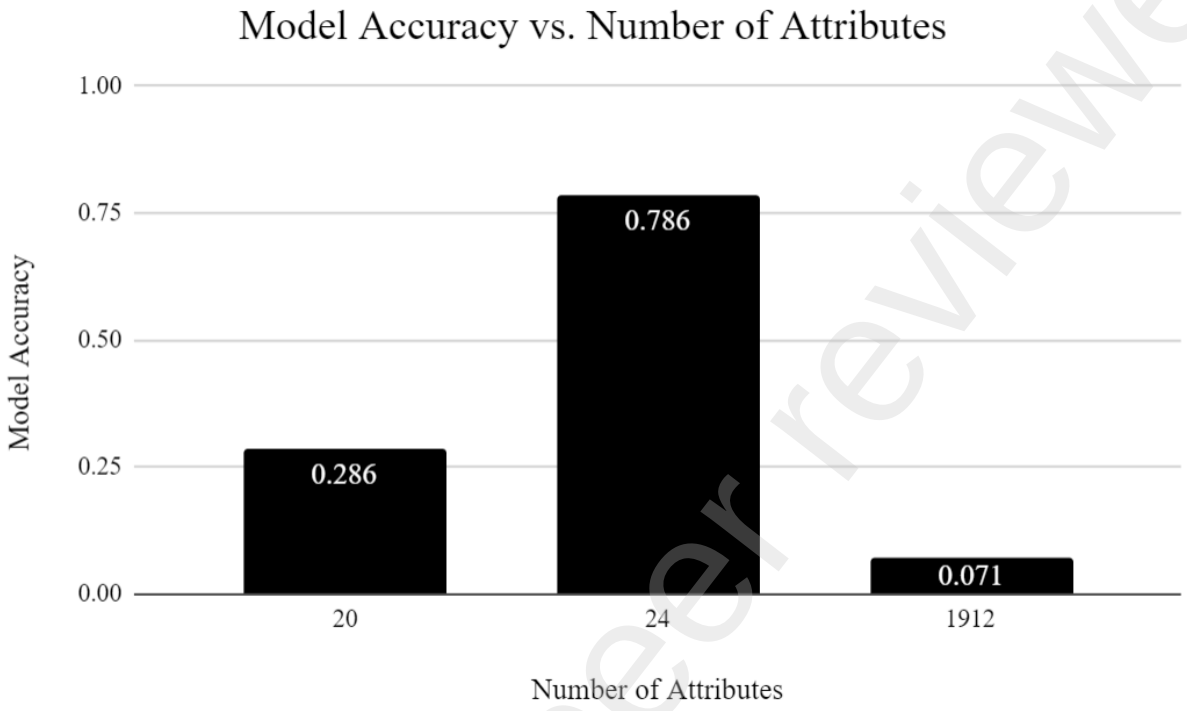

Table 1

MB-associated	Negative control
hsa-miR-30b-5p, hsa-miR-30d-5p, hsa-miR-193a-3p, hsa-miR-148a-3p, hsa-miR-224-3p, hsa-miR-183-5p, hsa-miR-96-5p, hsa-miR-182-5p, hsa-miR-196b-5p, hsa-miR-200b-3p, hsa-miR-135a-5p, hsa-miR-219a-5p, hsa-miR-326, hsa-miR-204-5p, hsa-miR-495-3p, hsa-miR-1253, hsa-miR-21-5p, hsa-miR-106a-5p, hsa-miR-363-3p, hsa-miR-106b-5p, hsa-miR-367-3p, hsa-miR-31-5p, hsa-miR-206, hsa-miR-378a-3p, hsa-miR-383-3p, hsa-let-7g-5p	hsa-miR-216b-3p, hsa-miR-921, hsa-miR-610, hsa-miR-520e-5p, hsa-miR-6764-5p, hsa-miR-7162-5p, hsa-miR-924, hsa-miR-3140-3p, hsa-miR-1322, hsa-miR-1914-5p, hsa-miR-4689, hsa-miR-30a-5p, hsa-miR-559, hsa-miR-151a-3p, hsa-miR-5090, hsa-miR-4714-3p, hsa-miR-1323, hsa-miR-4743-3p, hsa-miR-6834-5p, hsa-miR-103a-3p, hsa-miR-3650, hsa-miR-3689d, hsa-miR-3677-3p, hsa-miR-3976, hsa-miR-3945, hsa-miR-3126-5p

Table 2

Independent MB-associated
hsa-miR-152-3p, hsa-miR-362-5p, hsa-miR-140-5p, hsa-miR-409-5p, hsa-miR-570-3p, hsa-miR-320a-3p, hsa-miR-424-5p, hsa-miR-23a-3p, hsa-miR-29b-3p, hsa-miR-223-5p, hsa-miR-10a-5p

Table 3

Independent CML-associated
hsa-miR-124-3p, hsa-miR-125b-5p, hsa-miR-324-5p, hsa-miR-326, hsa-miR-125a-5p, hsa-miR-9-3p, hsa-miR-19a-3p, hsa-miR-20a-5p, hsa-miR-92a-3p, hsa-miR-199b-5p, hsa-miR-17-5p, hsa-miR-18a-5p, hsa-miR-19b-3p, hsa-miR-34a-5p

Table 4

Model	Accuracy	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Hoeffding Tree	0·923	0·923	0·077	0·923	0·923	0·923	0·846	0·975	0·977
Logistic Regression	0·962	0·962	0·038	0·962	0·962	0·962	0·923	0·987	0·987
Multilayer Perceptron	0·885	0·885	0·115	0·885	0·885	0·885	0·769	0·970	0·972
Naive Bayes	0·923	0·923	0·077	0·923	0·923	0·923	0·846	0·975	0·977
Random Forest	0·846	0·846	0·154	0·848	0·848	0·848	0·694	0·946	0·950