HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 3 (2024)

Machine Learning Based Early Prediction of Glioblastoma Using Gene Expression

Arunraj Jeyaprakash[1] and Ed Gerstin[#]

[1]Canyon Crest Academy, USA
[#]Advisor

ABSTRACT

Glioblastoma multiforme (GBM), is an invasive highly malignant fast-growing tumor in the central nervous system. GBM ranks among the highest mortality rate cancers globally, with a survival rate of 5-10% even with combination therapies. The introduction of molecular prognostic-diagnostic biomarkers for central nervous system tumors, including microRNAs in exosomes plays a significant role in the early detection of Glioblastoma. The diagnosis by experts using clinical biomarkers is time consuming and variable. A Prediction Diagnostic model was developed for GBM based on miRNA molecular biomarkers from exosomes using Machine Learning algorithms. It was hypothesized to accurately distinguish glioblastoma patients from healthy individuals using the gene expression, offering a promising diagnostic tool for early detection of the disease with the help of Machine Learning. Data was collected from the NCBI [1] Gene Expression Omnibus site. Eight machine learning models were trained and compared using accuracy, precision, recall, confusion matrices, and AUC ROC curves. Logistic Regression was found to be the best model based on the comparison and matching the expert diagnosis. The study matches the six overexpressed microRNAs in GBM (hsa-miR-4443, hsa-miR-422a, hsa-miR-494-3p, hsa-miR-5025p, hsa-miR-520f-3p, and hsa-miR-549a). Primarily the following two expressed microRNAs hsa-miR-549a and hsa-miR-502-5p plays a significant role in the prediction of prognosis in patients with tumors of glial origin. A group of genetically expressed miRNAs that may serve as reliable biomarkers for brain cancer were identified using machine learning, which represents a powerful tool in biomarker identification.

Introduction

Glioblastoma is the most aggressive and common type of primary brain tumor, meaning it starts in the brain itself.

It's classified as a grade 4 astrocytoma, the highest grade, due to its fast-growing and aggressive nature.

Glioblastomas as shown in Figure 1 arise from star-shaped glial cells called astrocytes, which support and protect neurons.

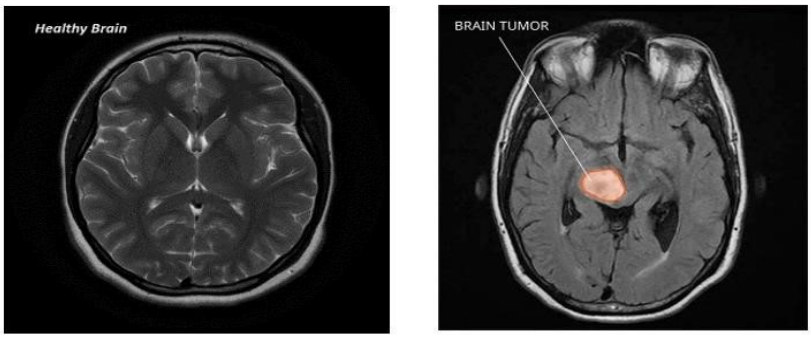

Figure 1. Healthy brain and brain with tumor Credit: NCI-CONNECT

Glioblastomata are most commonly found in the frontal or temporal lobes of the brain, and while they can invade nearby tissue, they typically don't spread to other organs. Despite ongoing research and treatment advancements, the prognosis for glioblastoma remains poor, with a median survival rate of only 12-15 months after diagnosis as depicted in Figure 2. However, there are treatment options available that can help improve quality of life and manage symptoms.

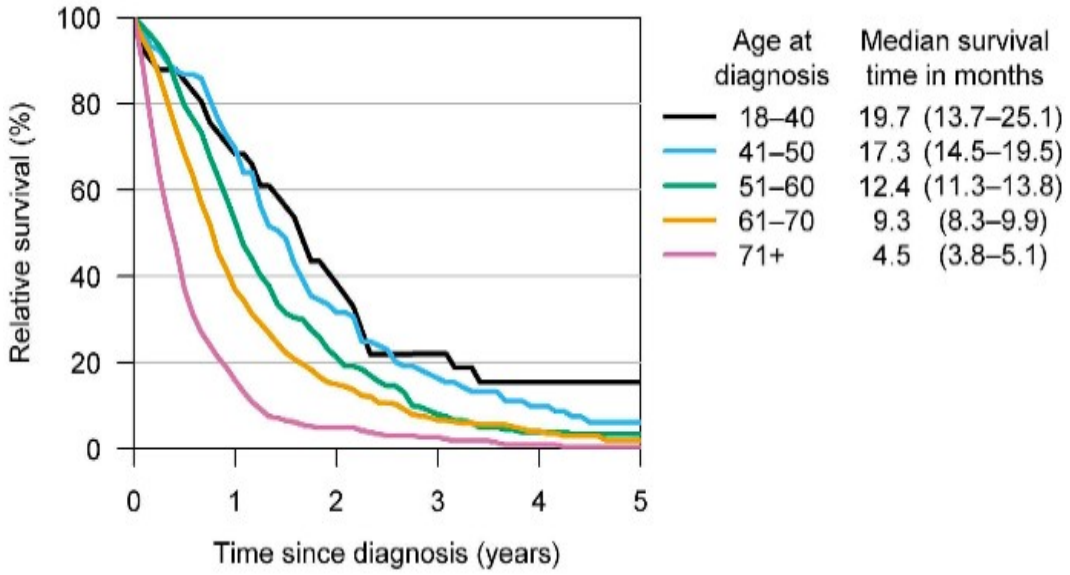

Figure 2. Median survival times and relative survival rates for different glioblastoma patient age groups [10].

Exosomes, meticulously crafted nanoscale messengers, originate from the internal machinery of most eukaryotic cells. These membranous envelopes, ranging from 30-150 nanometers in diameter, ferry a diverse cargo of biomolecules including proteins, lipids, and nucleic acids (mRNA, miRNA, DNA). Acting as intercellular communicators, exosomes shuttle these molecules between cells, impacting various physiological processes in both health and disease. Their involvement ranges from orchestrating tissue repair and immune responses to potentially contributing to the spread of pathological signals in neurodegenerative disorders and cancer.

microRNA (miRNA) is a type of small RNA that has been mainly overlooked in the past. miRNAs are endogenous, non-coding RNA molecules, typically 19-25 nucleotides in length, that have emerged as key

regulators of gene expression across a wide range of biological processes in eukaryotes. Despite their small size, miRNAs exert a profound impact on cellular function by binding to the messenger RNA (mRNA) transcripts of protein-coding genes. This interaction can lead to either the degradation of the mRNA or the inhibition of its translation into protein, effectively fine-tuning protein production within the cell. miRNA can be considered a part of the central dogma due to its vital role in gene expression.

The intricate connection between glioblastoma, exosomes, and microRNA reveals a complex interplay at the molecular level. In glioblastoma, the aggressive nature of these tumors, originating from astrocytes, underscores the urgency for understanding the underlying molecular mechanisms. Exosomes, as nanoscale messengers, contribute to intercellular communication by shuttling biomolecules, including microRNAs, between cells[6]. This dynamic interaction becomes particularly significant in the context of glioblastoma, where microRNAs, despite their small size, emerge as pivotal regulators of gene expression. The fine-tuning of protein production by miRNAs within cells may play a role in the pathophysiology of glioblastoma, impacting its growth and aggressiveness. The exploration of these connections not only deepens our understanding of glioblastoma but also highlights the potential involvement of exosomes and microRNAs in the intricate molecular landscape of this formidable brain tumor[7].

Review of Literature

Gliomas are a common type of tumor originating in the brain. Gliomas are called intra-axial brain tumors because they grow within the substance of the brain and often mix with normal brain tissue[11].

Oligodendroglioma is a growth of oligodendrocyte cells that starts in the brain and found in the cerebrum. A substance was made by these cells protects nerve cells and helps with the flow of electrical signals of central nervous system in the brain and spinal cord.

Astrocytoma is a growth of astrocyte cells that starts in the brain or spinal cord. Astrocytes support and connect nerve cells in the brain and spinal cord. They are most often found in the cerebrum (the large, outer part of the brain), but also in the cerebellum (located at the base of the brain).

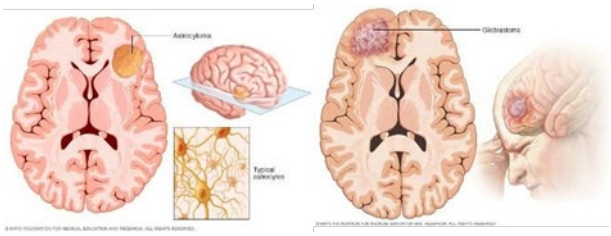

Figure 3. Glioblastoma is a type of cancer that starts in cells called astrocytes that support nerve cells[11].

Glioblastoma (GBM) is a type of cancer that starts as a growth of cells in the brain or spinal cord as shown in Figure 3. It grows quickly and can invade and destroy healthy tissue. Glioblastoma forms from cells called astrocytes that support nerve cells. Glioblastoma is the most aggressive and challenging type of brain cancer.

Gene Expression

Gene expression begins with DNA and results in a protein as illustrated in Figure 5. Gene expression is the process by which genetic instructions are used to synthesize protein products. The process by which a gene's

information is converted into the structures and functions of a cell by a process of producing a biologically functional molecule of either protein or RNA.

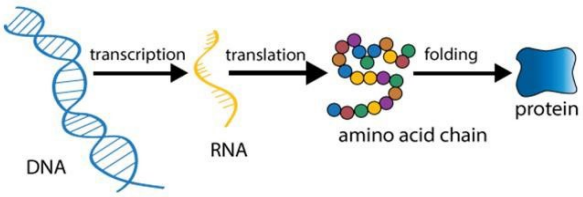

Figure 5. Role of RNA in Protein synthesis [12] .

Gene expression controlled at various points in the process leading to protein synthesis. Gene expression be thought of as an "on/off switch" to control when and where RNA molecules and proteins are made and as a "volume control" to determine how much of those products are made.

Exosomes

Exosome are extracellular vesicles secreted by most eukaryotic cells and participate in intercellular communication as shown in Figure 6. Exosomes serve as vehicle to transfer bioactive cargos.

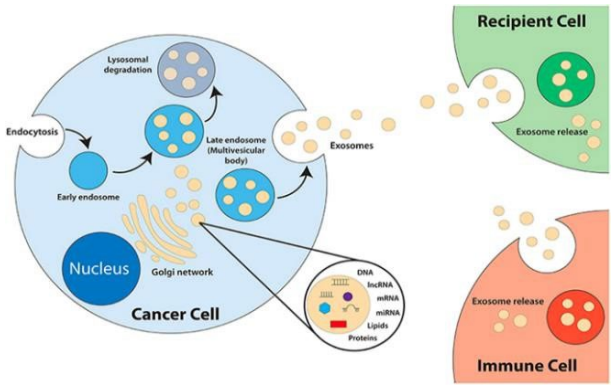

Figure 6. Exosomes serve as vehicle to transfer bioactive cargos [13] .

The components of exosomes including proteins, DNA, mRNA, microRNA etc., which play a crucial role in regulating tumor growth, metastasis, and angiogenesis in the process of cancer development[6].

Messenger RNA

Messenger RNA (mRNA) molecules in cells carry codes from DNA in the nucleus to the sites of protein synthesis in the cytoplasm (the ribosomes).

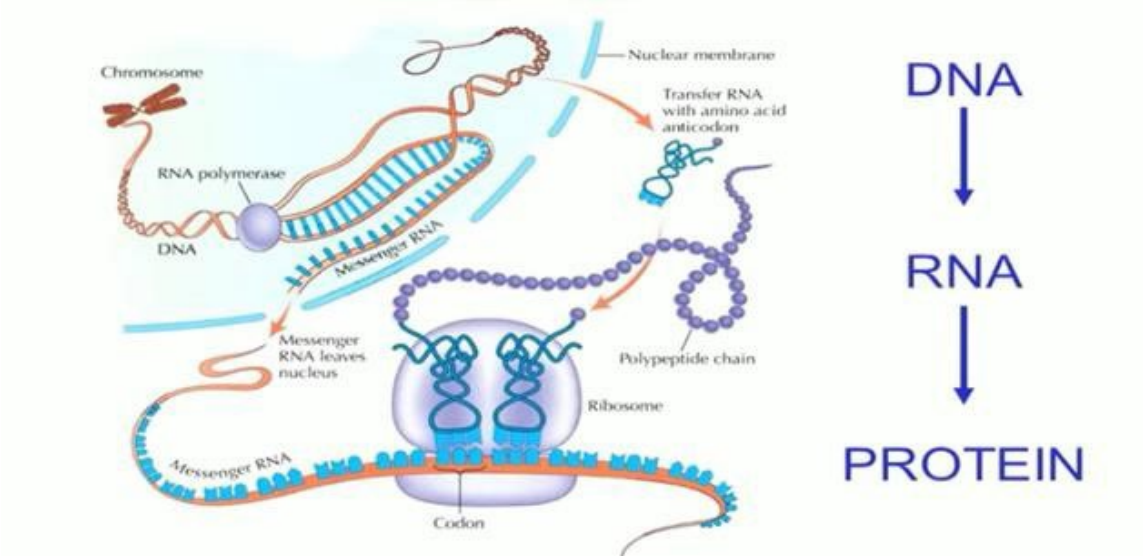

Figure 7. Role of messenger RNA (mRNA) in Protein synthesis [14].

mRNA is the translated form of DNA that the machinery can recognize and use to assemble amino acids and proteins as illustrated in Figure 7.

MicroRNA

MicroRNAs (miRNAs) as shown in Figure 8 are regulatory non-coding RNAs that negatively regulate protein coding genes and other non-coding transcript expressions. microRNAs play a significant role in brain tumor initiation, progression and metastasis. miRNAs control gene expression mainly by binding with mRNAs in the cell cytoplasm instead of allowing it to be translated quickly into a protein [4] .

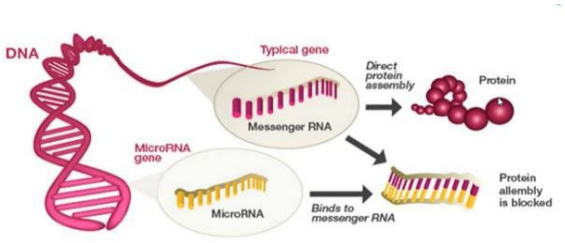

Figure 8. Role of micro-RNA (miRNA) in Protein synthesis regulation [15].

Glioma Development

The central nervous system glioma cells transfer information to other cells *via* exosomal miRNAs in the tumor microenvironment. Exosomes derived from glioma cells alter the behavior of normal cells, including promoting glioma carcinogenesis, angiogenesis, drug resistance, and/or helping cancer cells escape from the host's immune system. They can be useful diagnostic and/or prognostic biomarkers.

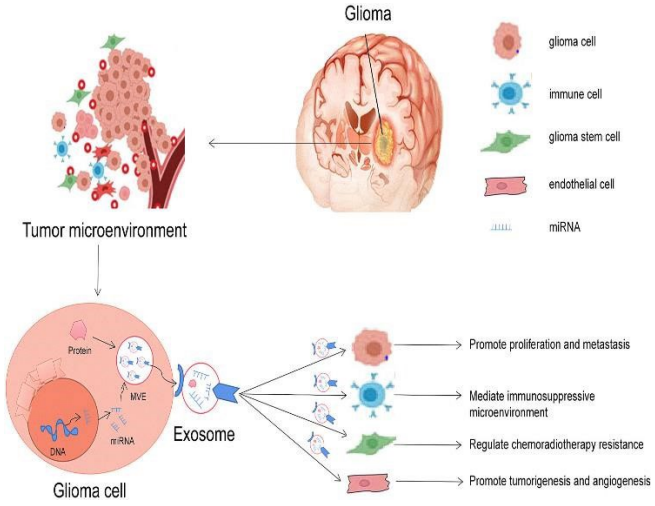

Figure 9. Roles of glioma-derived exosomes in glioma development [16].

A mutation in a microRNA gene can leave the cell without that particular microRNA or reduce it a low level in the cell as displayed in Figure 9. Abnormally low levels of a microRNA can lead to overexpression of genes that microRNA regulates[2], and that can lead to cancer development and progression.

Glioblastoma Prediction

Methodology

The Glioblastoma prediction methodology can be explained in four simple steps as in Figure 10. The methodology: Curation, Learning, Modelling and Prediction. Curation involves the collection and isolation of RNA and miRNA from the sample. The machine learning involves the training of the model using the genetically expressed miRNA data. The best model is identified based on the key evaluation statistics and deployed for the early prediction of Glioblastoma.

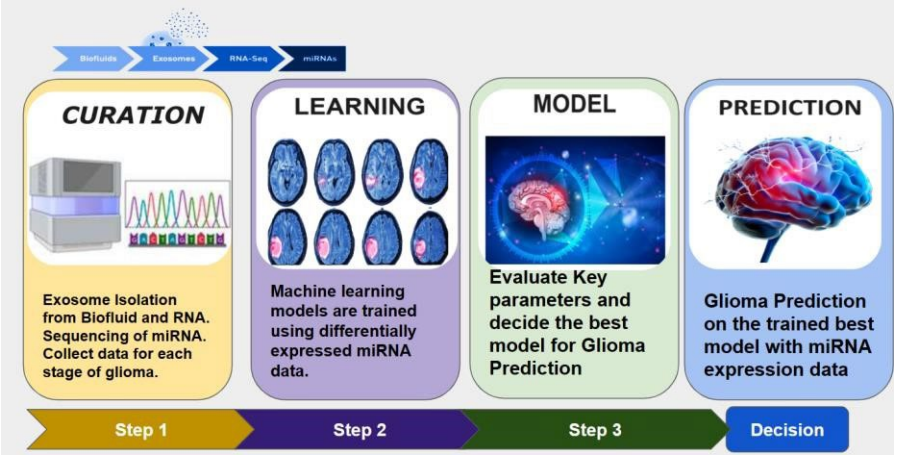

Figure 10. Machine learning based Glioblastoma prediction methodology

Essentially, we procure and clean data, train machine learning models, determine the best model, and then use that model for Glioma predictions.

Procedure

The Glioblastoma prediction procedure is described in detail as follows,

Genome Data Dataset

Raw data from clinical information and sequencing were downloaded from the NCBI [1] database, an available dataset that catalogs brain cancer-causing genomic changes

Preprocessing and Extraction of Differentially Expressed miRNA

Processing and analysis of raw data were accomplished using R software LIMMA package (Linear Models for Microarray Data). The differentially expressed miRNAs (demiRs) were extracted with the following threshold, adj. $p < 0.05$ and $\log|\text{Fold Change}| > 2.0$.

Detecting Predictive Biomarkers

- Machine learning was used to find miRNAs with diagnostic and predictive qualities. For this purpose, in the processing step, all important/relevant miRNAs were identified using heatmap analysis (miRPathDB v2.0).
  Use Pearson Linear Correlation Coefficient Maps to compare the patient samples to find similarities and differences between stages.

Machine Learning

- Implement machine learning model and fine-tune the model on the glioblastoma dataset to adapt the knowledge for improved detection accuracy.
- Validate the machine learning model on independent datasets not used during the training phase. Ensure the model's ability to generalize and maintain diagnostic accuracy across diverse datasets.
- Compare with other Machine Learning Models; selected ones are SVM, RF, KNN, Decision Tree, Logistic Regression, Gradient Boosting, Adaboost, and Naive Bayes.
- The machine learning algorithms were compared with key metrics (accuracy, precision, ROC_curve, and confusion matrix) to find the most accurate algorithm. The algorithms were used in this study to measure the diagnostic ability of differentially expressed miRNA in Glioma.
- Compare the performance of the best model with other existing diagnostic methods for glioblastoma. Analyze the effectiveness of transferring knowledge from related miRNA datasets for enhanced diagnostic accuracy.

Implementation

Integrate the optimized model into a diagnostic pipeline for automated exosome-based glioblastoma diagnosis.

Materials and Tools

The materials and various software tools used in the Glioblastoma prediction is detailed in the Figure 11.

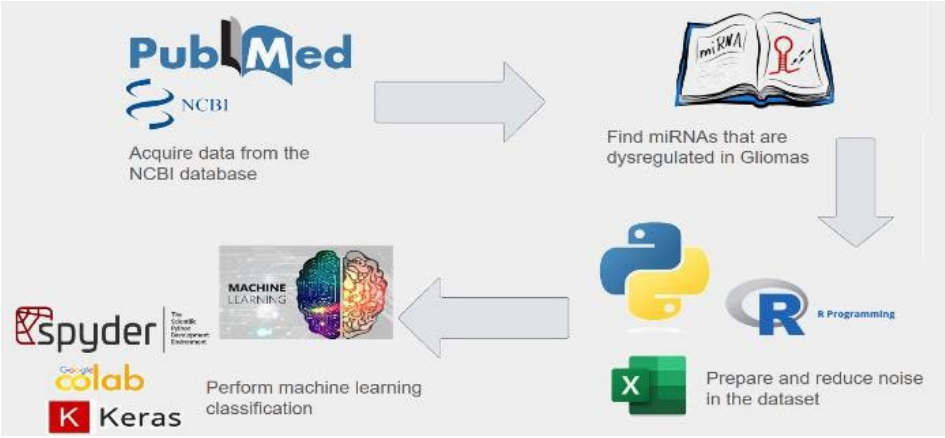

Figure 11. Materials and tools

Machine Learning Model

The Glioblastoma prediction Machine learning model in Figure 12 illustrates the different blocks involved in the training and evaluation.

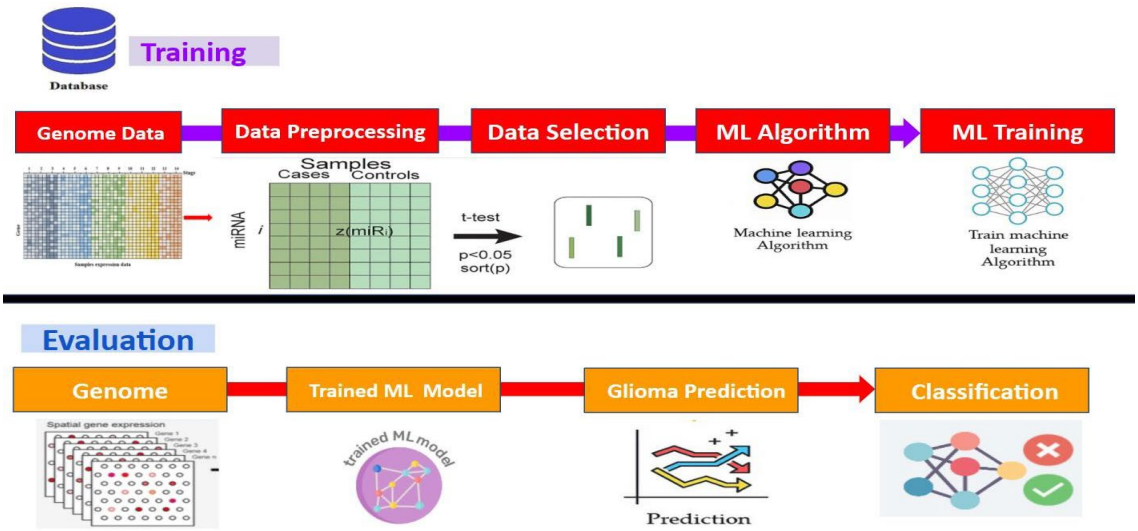

Figure 12. Machine learning Model Training and Evaluation

Results

This heat map shows highly validated miRNAs and the pathways are involved in using miRPathDB v2.0. The circled Glioma column in Figure 13 shows 8 different miRNAs that are involved in the progression and regression of glioma.

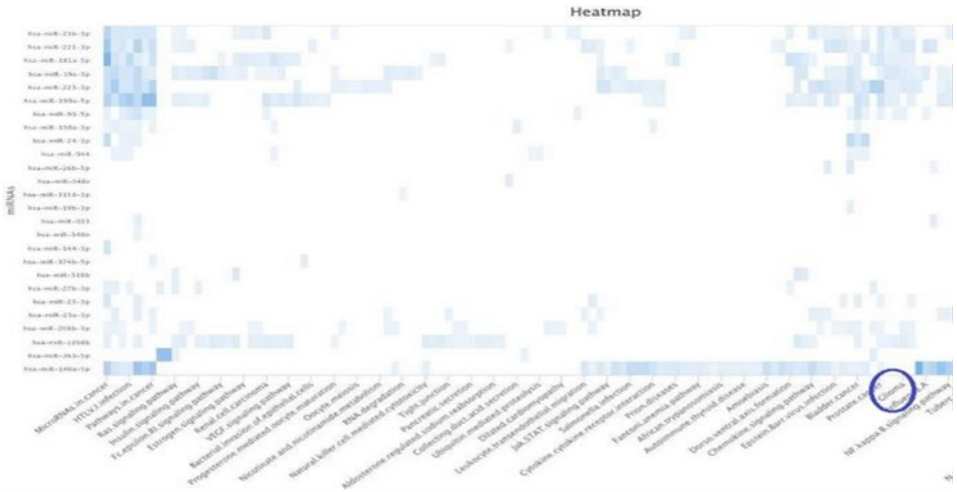

Figure 13. Heat Map of microRNA

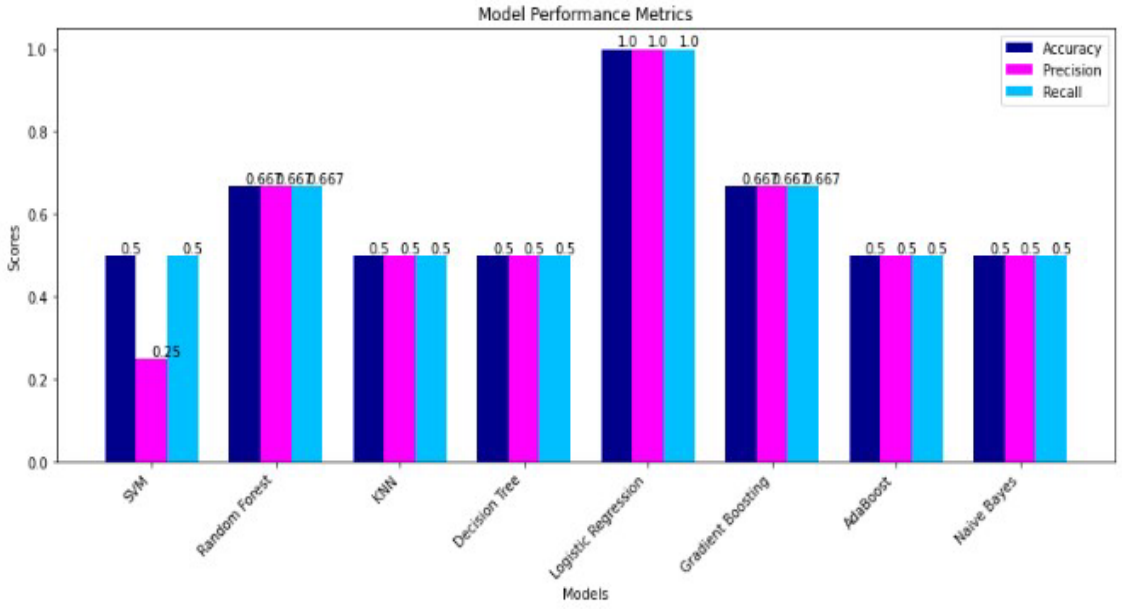

Figure 14. Machine Learning Models comparison for Glioblastoma detection

The bar chart in Figure 14, shows evidently that the Logistic Regression model performed the best with accuracy, precision, and recall scores are near 100%. The KNN, Decision Tree, AdaBoost and Naïve Bayes shows only 50% score. SVM model under performs out of all the models. Random forest and Gradient Boosting perform better than the rest of the models but lesser than linear regression.

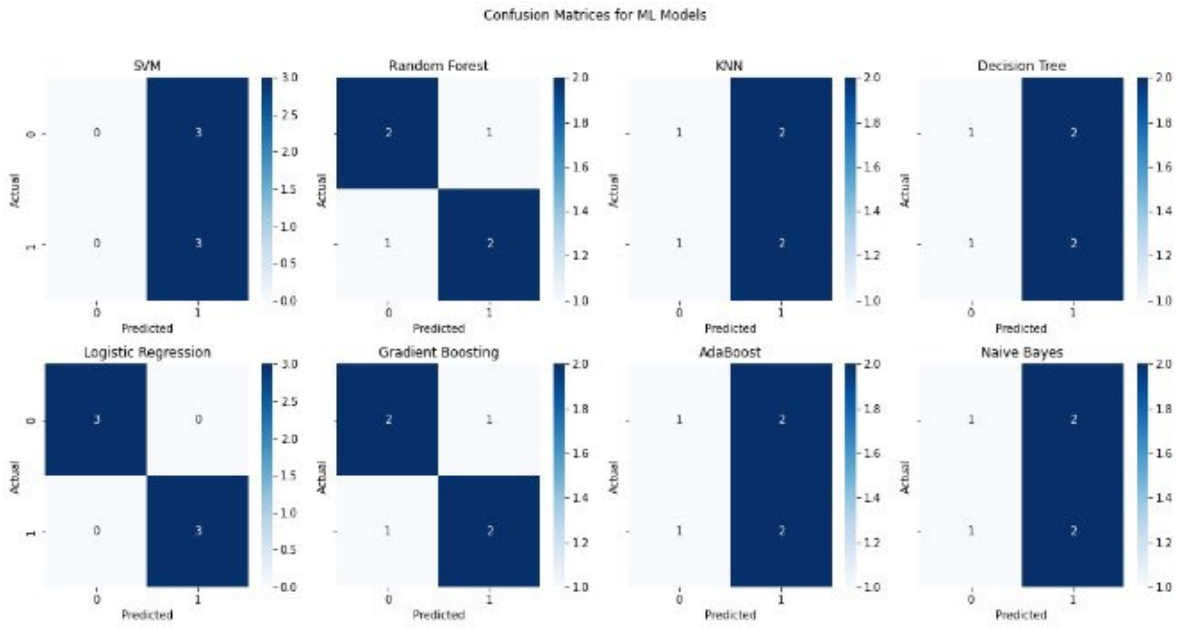

Figure 15. Confusion Matrix for Machine Learning Models of Glioblastoma detection

The confusion matrices shown in Figure 15 demonstrate that the Logistic Regression model works the best as there are three true positives and three true negative values. The others have some false positive and negative values whereas the Logistic Regression Model has none. In a diagnostic perspective, false positives and false negatives are very dangerous.

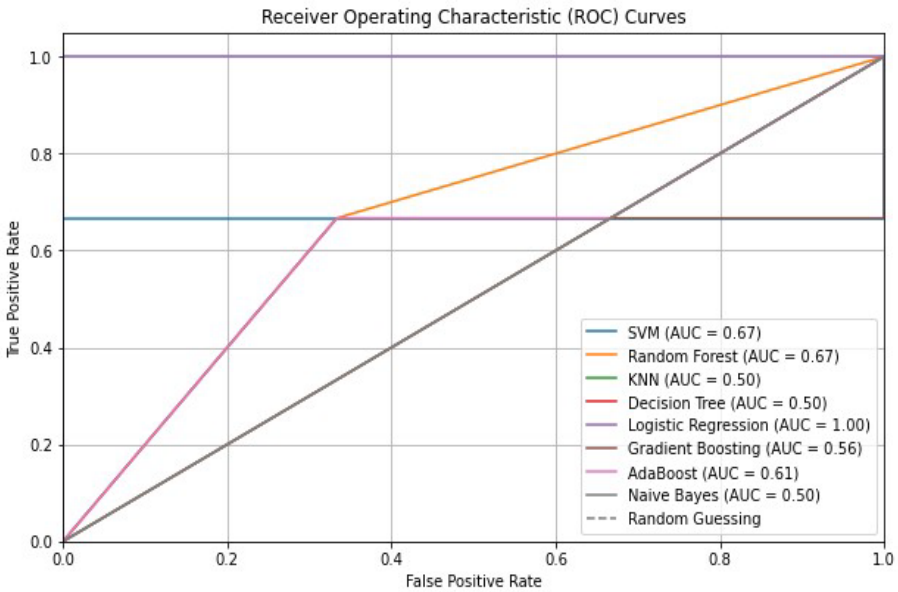

Figure 16. Receiver Operating characteristics (ROC) curves for different Machine Learning Models

The AUC is widely used to measure the accuracy of diagnostic tests. In Figure 16 the logistic regression model in the graph is closer to the top left corner of the ROC curve which evidently shows the highest accuracy of the test. The logistic regression model turns out to be the best model for the glioma prediction.

Discussion

Glioblastoma is a heterogeneous tumor with a very poor prognosis. The overall survival rate is 10- 15 months even though there are many advances in treatment including the use of secondary therapies such as chemotherapy, radiotherapy, surgery and targeted therapy. This forecasts the significance of identifying the biomarkers which helps to detect, identify and likely prognosis of the tumor in its early stages. miRNAs inside the exosomes can control the different post-transcriptional processes and play an important role in carcinogenesis including angiogenesis, cancer metastasis and drug resistance. In addition, the importance of miRNA-mRNA networks in activating or inhibiting many cancer-related molecular signaling pathways has recently been observed [17] and approved by the world health organization for central nervous system tumors. The recent advancement in the development of bioinformatics and machine learning algorithms, used to detect biomarkers plays a vital role in early detection, treatment, and prognosis of cancer.

The role of machine learning plays a significant role in the feature selection in miRNAs even without the prior knowledge (unsupervised). In the present study, NIH data analysis of brain cancer six genetically expressed exosomal miRNAs were identified using the machine learning algorithms.

Conclusion

In conclusion, six miRNAs were identified using machine learning algorithms, subsequent analyses showed a panel of six miRNAs including hsa-miR-4443, hsa-miR-422a, hsa-miR-494-3p, hsa-miR-502-5p, hsa-miR-520f-3p, and hsa-miR-549a with high diagnostic and prognostic power, which was validated by several datasets. hsa-miR-549a and hsa-miR-502-5p expression predicted prognosis in patients with tumors of glial origin matching the expert analysis. This study emphasizes the importance of machine learning as an alternative option for predicting biomarkers in brain cancer.

The Significance of the project is a group of genetically expressed miRNAs that may serve as reliable biomarkers for brain cancer were identified using machine learning, which represents a powerful tool in biomarker identification.

Future Steps include to build a more robust machine learning model utilizing the dataset from various geography and cultures. A whole-body level miRNA-based disease classifier as it is definitely possible that miRNA expression is part of most types of anomalies.

Acknowledgments

I would like to thank Dr. Ed Gerstin for his tremendous support throughout my research journey. Dr. Ed Gerstin's expertise in biomedical sciences and experience in research helped me piece together my research findings into this research paper. Last but not least, I would like to thank my parents for always encouraging and empowering me in my work.

References

1. GEO Accession viewer. (n.d.). Retrieved 14 February 2024, from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE112462

2. Durocs, A., Fadda, P., Nigita, G., Fassan, M., Bottoni, A., Gardiman, M. P., … Croce, C. M. (2018). Circulating microRNAs predict survival of patients with tumors of glial origin. *EBioMedicine*, *30*, 105–112. https://doi.org/10.1016/j.ebiom.2018.03.022
3. Azari, H., Nazari, E., Mohit, R., Asadnia, A., Maftooh, M., Nassiri, M., … Avan, A. (2023). Machine learning algorithms reveal potential miRNAs biomarkers in gastric cancer. *Scientific Reports*, *13*(1), 6147. https://doi.org/10.1038/s41598-023-32332-x
4. Xiao, C., & Rajewsky, K. (2009). MicroRNA control in the immune system: basic principles. *Cell*, *136*(1), 26– 36. https://doi.org/10.1016/j.cell.2008.12.027
5. Garcia-Garcia, F., Panadero, J., Dopazo, J., & Montaner, D. (2016). Integrated gene set analysis for microRNA studies. *Bioinformatics (Oxford, England)*, *32*(18), 2809–2816. https://doi.org/10.1093/bioinformatics/btw334 6. Chen, M., Medarova, Z., & Moore, A. (2021). Role of microRNAs in glioblastoma. *Oncotarget*, *12*(17), 1707– 1723. https://doi.org/10.18632/oncotarget.28039
7. 9Aili, Y., Maimaitiming, N., Mahemuti, Y., Qin, H., Wang, Y., & Wang, Z. (2021). The role of exosomal miRNAs in glioma: Biological function and clinical application. *Frontiers in Oncology*, *11*, 686369. https://doi.org/10.3389/fonc.2021.686369
8. Marangon, D., & Lecca, D. (2023). Exosomal non-coding RNAs in glioma progression: insights into tumor microenvironment dynamics and therapeutic implications. *Frontiers in Cell and Developmental Biology*, *11*, 1275755. https://doi.org/10.3389/fcell.2023.1275755
9. Lee, E., Yong, R. L., Paddison, P., & Zhu, J. (2018). Comparison of glioblastoma (GBM) molecular classification methods. *Seminars in Cancer Biology*, *53*, 201–211. https://doi.org/10.1016/j.semcancer.2018.07.006
10. Lehtinen, P. (2018, October 16). Brain cancer survival has improved – but not much for elderly. Retrieved 29 May 2024, from University of Helsinki website: https://www.helsinki.fi/en/news/healthier-world/brain-cancersurvival-has-improved-not-much-elderly
11. Glioblastoma - Overview - mayo clinic. (2024, March 7). Retrieved 15 April 2024, from https://www.mayoclinic.org/diseases-conditions/glioblastoma/cdc-20350148
12. Epigenetics tutorial. (n.d.). Retrieved 29 May 2024, from https://biosocialmethods.isr.umich.edu/researchsupport/videos-tutorials/epigenetics-tutorial/
13. Othman, N., Jamal, R., & Abu, N. (2019). Cancer-derived exosomes as effectors of key inflammation-related players. *Frontiers in Immunology*, *10*, 2103. https://doi.org/10.3389/fimmu.2019.02103
14. (N.d.). Retrieved 29 May 2024, from http://www.evolutiontextbook.org/content/free/figures/02_EVOW_Art/27_EVOW_CH02.pdf
15. Leitch, C. (2016, June 14). Disease progression in brain tumors may be predictable with. Retrieved 29 May 2024, from Labroots website: https://www.labroots.com/trending/cell-and-molecular-biology/3339/diseaseprogression-brain-tumors-predictable-micrornas
16. Aili, Y., Maimaitiming, N., Mahemuti, Y., Qin, H., Wang, Y., & Wang, Z. (2021). The role of exosomal miRNAs in glioma: Biological function and clinical application. *Frontiers in Oncology*, *11*, 686369. https://doi.org/10.3389/fonc.2021.686369
17. Publication of the WHO classification of tumours, 5th edition, volume 6: Central nervous system tumours. (n.d.). Retrieved 29 May 2024, from https://www.iarc.who.int/news-events/publication-of-the-whoclassification-of-tumours-5th-edition-volume-6-central-nervous-system-tumours/