NeuroXNet: Creating a Novel Deep Learning Model Architecture that Diagnoses Neurological Disorders and Finds New Blood-Based Biomarkers with a miRNA Drug Discovery Pipeline Using Medical Imaging and Genomic Data

Vaibhav Mishra[1]

[1]West Laurens High School, 3692 GA-257, Dexter, GA, USA, 31019

Personal Section

My research journey started from when I was a volunteer at a memory and rehabilitation center at my local city. There, I saw the drastic effects of neurodegenerative diseases on individuals making it incredibly hard to continue daily life. Motivated by my interest in neuroscience, I decided to start a research project in computational neuroscience.

Due to COVID-19 restrictions, I performed all parts of my research at home. Due to the computational aspects, I first had to learn advanced calculus, linear algebra, differential equations, machine learning, genomics, and several programming languages. I spend hours upon hours learning from a plethora of online courses, tutorials, articles, and books, until I had learned enough to start researching. My next step was to identify a project that would help alleviate conditions for neurodegenerative disease affected patients. After months of hard work, I succeeded in created NeuroXNet, a genomics and AI powered application with diagnostic and treatment capabilities for neurodegenerative diseases.

Through my research journey, I learned not only unique skills in computer science, mathematics, and neuroscience but also the power of science and mathematics in solving real world problem faced by millions throughout the world.

My advice for any high school student who would like to undertake a research project in the sciences is to not be daunted by the terminology and research methodology required to succeed in research. Research can be intense at times where there seems to be no way to improve a machine learning model or to prove a specific theorem. In these times, I would suggest trying new techniques and methods to succeed. I would also suggest focusing on fields that are interesting, so that a student is willing to spend months or even years researching a topic in the area without losing motivation.

Research Section

I. Introduction

1.1 Background

Neurological disorders continue to affect millions of people worldwide, with diseases leading to loss of cognitive function, a decline in memory, and even death. These diseases contribute to nearly a trillion dollars of healthcare spending and drastically change the lives of those affected. With the advent of new medical imaging and computational techniques, it has become possible to use large amounts of imaging data to build and train deep learning models that can diagnose many diseases with high accuracy rates using clinical tests and medical imaging tests like MRI. Some of the most common neurodegenerative disorders include Alzheimer’s disease, Parkinson’s disease, and Mild Cognitive Impairment.

1.2 Previous Literature

Through the usefulness of large amounts of data and advancements in medical imaging research, it has become possible to diagnose and even treat patients with diseases in various fields of medicine. A large amount of medical imaging research using MRI data has been used to classify AD from normal patients. For instance, Al- Khuzaie et al. [1] developed a new model, AlzNet, which achieved an accuracy of 99.30% in diagnosing AD from normal using 2D MRI slices. One of the highest accuracies was achieved in AD diagnosis using machine learning models.

However, relatively few machine learning models have performed multiclass diagnosis in neurological disorders. A multiclass approach to classifying multiple neurodegenerative diseases was studied in very few papers, notably in [2], where AD, PD, and CN were classified with an accuracy of

90% for AD, 90% for control from ADNI, 89% for control from PPMI, and 90% for PD using transfer learning on the VGG19 model which performed the best out of the ResNet 50, Inception Net, and VGG16 model which were also tested in the paper and in another article by Tong et al. [3], where a five-class model was proposed that achieved a 75.2% that classified AD and other dementia-like diseases using the RUSBoost algorithm. Even in these studies, the model was limited to classifying only neurodegenerative diseases in lesser than five classes. Therefore, this study aims to solve the problem of multiclass diagnosis and treatment of neurological disorders.

1.3 Research Problem

Current diagnosis of neurodegenerative diseases and cancerous brain tumors can be inaccurate, costly, time-inefficient, and invasive creating risk for patients. Moreover, only part of the patient data is used for diagnosis. To offer a more robust diagnosis, blood-based biomarkers and MRI imaging data offer a better diagnosis. Furthermore, diagnosing these medical conditions often have limited integration of utilizing patient data to offer the best treatment approaches. Moreover, current therapeutic treatments for diseases involve billions of dollars, a 90% failure rate, and 12-years to develop a successfully follow the drug discovery process. In addition, current treatment procedures fail to accurately predict the highest survival percentages for patients to suggest the best treatment plans. Patient data is also not used accurately to help integrate multiple components like a patient’s genomic profile to find overexpressed genes and their corresponding miRNA regulatory pathways for faster drug discovery.

1.4 Proposed Solution

This study proposes a novel deep learning architecture, NeuroXNet, which performs multiclass diagnosis of AD, PD, MCI, glioma, meningioma, pituitary, and normal patients. NeuroXNet is the first model in published literature that diagnoses neurological diseases in seven classes using MRI images. This is also the first model in published literature which creates a novel architecture to classify neurodegenerative disorders instead of relying on previously built models like ResNet50 or VGG16.

Furthermore, novel blood-based biomarkers and their corresponding miRNA regulatory pathways are identified with potential to aid in clinical drug discovery research through target identification, having the potential to drastically fasten the drug discovery process and reduce costs for in vitro experiments. In addition, NeuroXNet generates recommendations for treatment based on classification of disease from its convolutional neural network (CNN) model combined with the patient's genomic data and clinical data. These recommendations include treatment plans for surgery, radiation, or drug therapy. Moreover, this model is the first that combines diagnosis with treatment plans and a miRNA drug discovery pipeline for neurological disorders. Therefore, this model has great potential to be used in neurological medicine and provide a low-cost, efficient, and quick solution to patients worldwide.

II. Materials and Methods

2.1 Data Acquisition and Description for MRI Classifier

Data of AD and MCI patients used in the paper was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI), a free open source database (adni.loni.usc.edu), and data for the PD and normal patients was obtained from the Parkinson's Progression Markers Initiative (PPMI) database (ppmi-info.org). Data for MRI of patients with glioma, meningioma, and pituitary tumors were acquired from Kaggle's Brain Tumor Dataset [4]. For each of the seven AD, PD, MCI, Pituitary Tumor, Glioma Tumor, Meningioma Tumors, and Normal Patients, 1000 Brain MRI images were collected NIFTI format from the respective databases for a total of 7000 images. All the MRIs were T1-weighted and equally distributed to reduce model bias. These images were then converted to png format using the MRIcro tool. The images were split into training/validation/test folders using a 70%/30%/10% ratio. Table 1 shows the number of patient MRIs that were split into Training, Validation, and Test sets for each disease class.

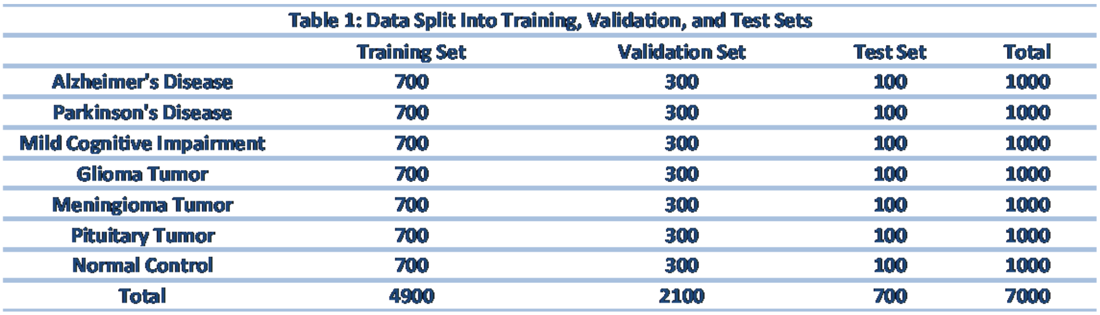


Table 1: Data Split Into Training, Validation, and Test Sets				
	Training Set	Validation Set	Test Set	Total
Alzheimer's Disease	700	300	100	1000
Parkinson's Disease	700	300	100	1000
Mild Cognitive Impairment	700	300	100	1000
Glioma Tumor	700	300	100	1000
Meningioma Tumor	700	300	100	1000
Pituitary Tumor	700	300	100	1000
Normal Control	700	300	100	1000
Total	4900	2100	700	7000

Table 1: Number of Patients per Data Set Folder

The demographic data including age and gender for the neurodegenerative classes is shown below in Table 2:

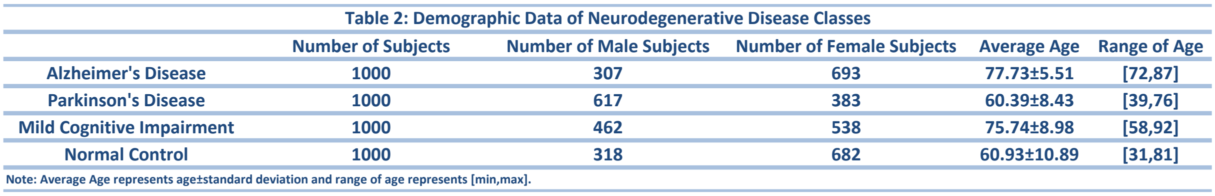


Table 2: Demographic Data of Neurodegenerative Disease Classes					
	Number of Subjects	Number of Male Subjects	Number of Female Subjects	Average Age	Range of Age
Alzheimer's Disease	1000	307	693	77.73±5.51	[72,87]
Parkinson's Disease	1000	617	383	60.39±8.43	[39,76]
Mild Cognitive Impairment	1000	462	538	75.74±8.98	[58,92]
Normal Control	1000	318	682	60.93±10.89	[31,81]
Note: Average Age represents age±standard deviation and range of age represents [min,max].					

Table 2: Demographic Data for Neurodegenerative Diseases

2.2 Data Preprocessing for Diagnosis

After being split into their respective folder, each MRI image was normalized by rescaling the size of each image. Then, data augmentation was applied, including shearing, zooming, and flipping the data to decrease overfitting and improve the model's overall accuracy. The data was preprocessed to reduce the amount of data bias the model gains and helps make the model regularize to fit all ranges of MRI images.

2.3 Identification of Novel Biomarkers

Identifying biomarkers from blood, tissue, or another type of body cell for early detection of neurological diseases was one of the main focuses of this study. Blood biomarkers are analyzed with microarrays to gather data on gene expression for thousands of genes and subsequently used as a non-invasive neurological test for diagnosis and research in finding new treatments. This study utilized genomic datasets from Gene Expression Omnibus. The data used in this study consisted of the series GSE74385 for meningioma tumor classification, GSE31095 for glioma tumors, GSE4488 for pituitary tumors, GSE63063 for MCI classification, GSE6613 PD classification, and GSE4226 for AD classification. Each gene expression study was used to find differentially expressed genes that would serve as blood biomarkers for diagnosing the specific neurological disorder. The NeuroXNet model used to classify patients by genomic data.

Using GEO2R, the patients were split into two groups for each series number, with one group consisting of the specific disease class and the other group with normal patients. Then, all differentially expressed genes with a p-value of less than 0.05 were used to generate a protein interaction network using the STRING tool. The protein interaction network was then used to identify hub genes with a high number of nodes using the Cytoscape tool. The hub genes are the differentially expressed genes that are over-expressed in patients with certain neurological conditions and are potential biomarkers for the disease. These hub genes were sorted by degree (number of nodes connected to) and used for gene ontology analysis using the PANTHER tool, which gave the false positive rate, fold enrichment, and p-values for the specific genes present in biological, molecular, and cellular processes.

The fold change represents the ratio of the average gene expression in the experimental group vs. the control group. This study focused on the over-expressed genes with fold change values greater than 1.

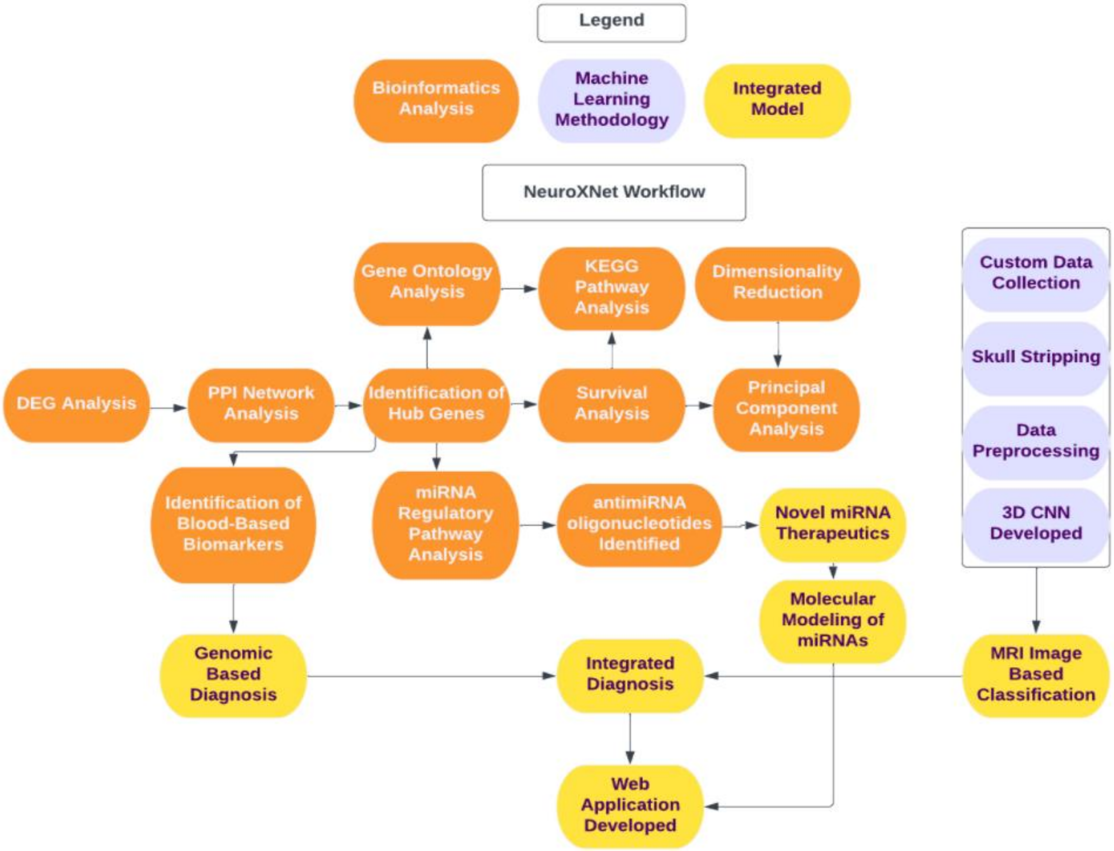


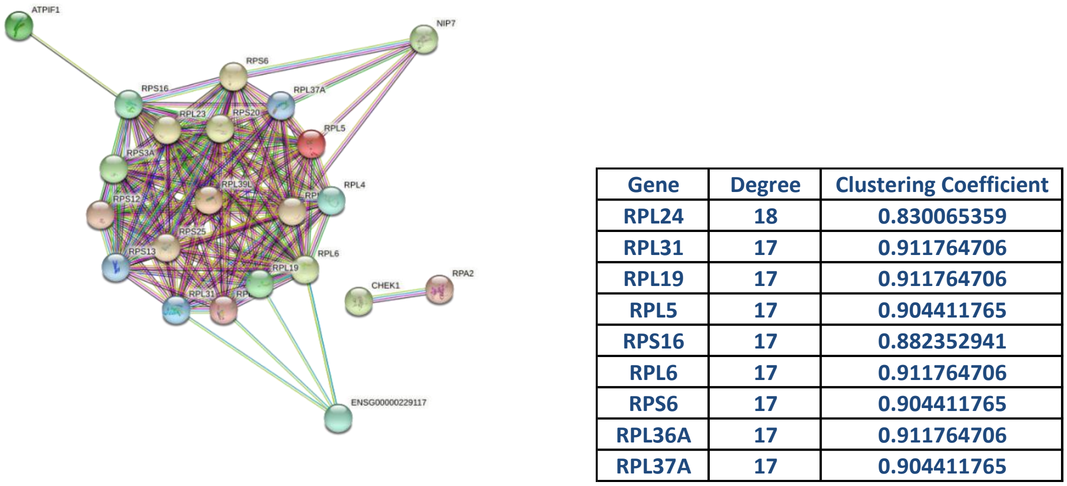
Figure 1: NeuroXNet Model Workflow Analysis

III. Results and Analysis

3.1 Results for Identification of Blood Biomarkers

Five hub genes were found for PD patients with a degree of 7. These genes were CENPF, DLGAP5, KLAA0101, TOP2A, and BUB1B. CENPF or Centromere Protein F is related to the centromere-kinetochore complex and is responsible for chromosome separation during mitosis. The gene is also differentially expressed in cancer patients. DLG Associated Protein 5 codes for protein and is also found in cancerous patients. DNA Topoisomerase II Alpha (TOP2A) is crucial in the transcription process, helping create enzymes for chromosome segregation and condensation. This gene also counters drug resistance in patients with ataxia-telangiectasia [5].

For the biological function of neuron-glial cell signaling, a fold enrichment value of 58.29 was observed in 3 over-expressed PD genes. In molecular functions, the alpha1-adrenergic receptor activity process had three genes with a fold enrichment of 97.15. This shows that the over-expressed genes are essential biomarkers of PD in patients, and the extremely low false discovery rate supports that this observance is not by random chance.



Gene	Degree	Clustering Coefficient
RPL24	18	0.830065359
RPL31	17	0.911764706
RPL19	17	0.911764706
RPL5	17	0.904411765
RPS16	17	0.882352941
RPL6	17	0.911764706
RPS6	17	0.904411765
RPL36A	17	0.911764706
RPL37A	17	0.904411765

Figure 2: Protein Interaction Network of Biomarkers for AD

For AD blood biomarkers, nine hub genes were over-expressed in the patients with the condition. RPL24 had a degree of 18, and the other eight genes each had a degree of 8, showing that these genes interacted closely and were signals of AD in patients. Ribosomal Protein L24 (RPL24) codes for protein synthesis and is part of the ribosomal proteins family of L24E. The RPL series and RPS series of genes which are the hub genes, are associated with certain cancers. The model shows that the ubiquitin ligase inhibitor activity and ubiquitin-protein transferase inhibitor activity have extremely large fold enrichments for molecular processes. These hub genes are efficient signals for AD diagnosis and can be studied for possible treatment of the disease.

For MCI patients, three hub genes were found, namely: FYN, SNRNP70, and CHD4. FYN is responsible for cell growth control in the tyrosine kinase protein family. SNRNP70 is responsible for many types of diseases involving body tissue. These hub genes are essential in classifying MCI through blood samples and can serve as a way for neurologists to differentiate between MCI and AD patients because of the different hub genes for each disease.

Five hub genes were found for patients with glioma tumors, and two biological processes of calcium ion export and regulation of postsynaptic cytosolic calcium ion concentration had high enrichment values of 38.5 and 35, respectively. PRKCA is associated with other types of cancer.

For patients with pituitary tumors, seven hub genes were found by the model. MPDZ proteins are responsible for HTR2C genes clot in the cell. Together, these hub genes can be used as non-invasive approaches for helping identify pituitary tumors in patients.

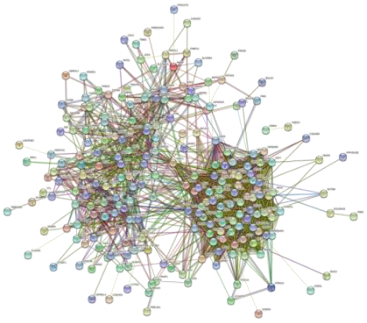


Figure 3: Protein Interaction Network for Meningioma Over Expressed Genes

One of the most significant results of the NeuroXNet model was the biomarkers it found for meningioma tumors. The model found 37 hub genes that were overly expressed in patients with meningioma tumors, and all had degrees greater than 64. The model shows that the biological process of DNA replication preinitiation complex assembly has a high enrichment value of 78.31 among the associated hub genes. The cellular function of the cyclin B1-CDK1 complex also has a high enrichment value of 78.31. Some of the hub genes with the highest interactions included CDK1, CCNA2, AURKA, and BUB1 (also associated with PD patients). Cyclin-Dependent Kinase 1 (CDK1) has been associated with breast cancer and is involved in M-phase promoting factors. Cyclin A2 (CCNA2) is also in the same family of genes and is responsible for protein transition. The gene is also found in patients with other types of cancer.

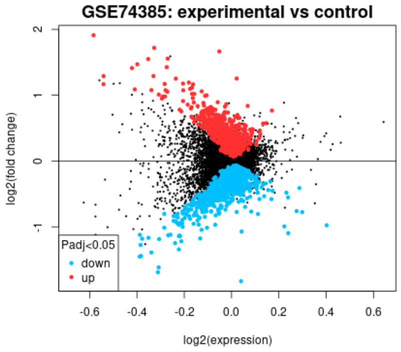


Figure 4: Mean Difference Plot for Meningioma Related Genes

3.2 NeuroXNet Classification Results

For my sequential model, NeuroXNet, the model starts with a convolutional layer with 32 filters, a stride of 2, the same padding, and a kernel size of 3. The layer takes in the input mages using an activation ReLU function. Next, the images are passed into a max-pooling layer with a pool size of 2 and a stride of 2. Then, the images pass through another convolutional layer with filter size 128. Next, the images pass into another max-pooling layer followed by a convolutional layer with filter size 128. Next, the images are passed into a batch normalization layer with a momentum of 0.8. Finally, the images are flattened through a flatten layer and are passed into the dense layer, which uses the softmax function to classify and diagnose the MRI images.

The model had an input of 6300 images from the Training and Validation folders, which is used to train itself. The model was run for 30 epochs and had a training accuracy of 99.79% with a loss of 0.0067 and a validation accuracy of 100% with a 0.0010 loss at the end of the 30 epochs. The training and validation accuracy and loss graphs over the epochs are shown below:

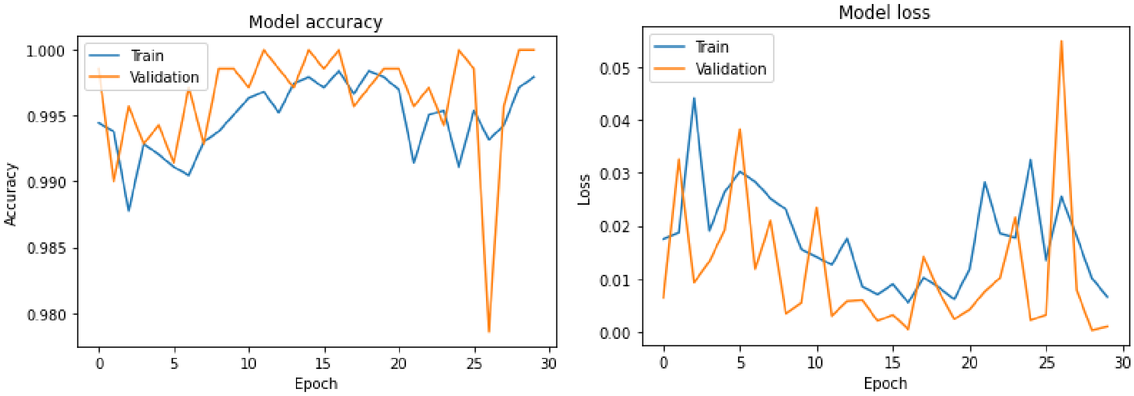
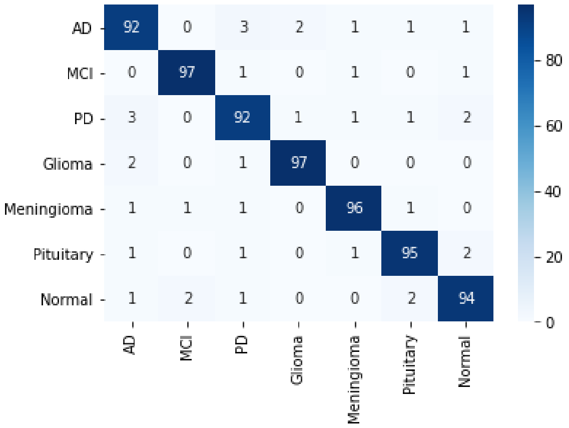


Figure 5: Training and Validation Model Accuracy and Loss

Then, the model ran the images on the test set comprising 700 images evenly split into seven classes. The confusion matrix for the testing data is shown below:



	AD	MCI	PD	Glioma	Meningioma	Pituitary	Normal
AD	92	0	3	2	1	1	1
MCI	0	97	1	0	1	0	1
PD	3	0	92	1	1	1	2
Glioma	2	0	1	97	0	0	0
Meningioma	1	1	1	0	96	1	0
Pituitary	1	0	1	0	1	95	2
Normal	1	2	1	0	0	2	94

Figure 6: Model Confusion Matrix and Heatmap

The confusion matrix shows the number of images in each class and their respective predicted values compared to the actual values from the test set. The confusion matrix also helps visualize the model's performance on the test set through the heatmap colors. The diagonal values of 92, 97, 92, 97, 96, 95, and 94 represent the number of images that NeuroXNet correctly classified in each of the seven classes of AD, MCI, PD, Glioma, Meningioma, Pituitary, and Normal respectively. Out of the 700 images in the test set, NeuroXNet correctly classified 663 pictures, and the rest of the 37 were incorrectly classified. The few incorrect classifications could be caused by the model perceiving features of certain patient MRIs to be similar to multiple diseases making the model rely on a close probability value from the softmax function to diagnose the particular incorrect disease over the correct class. Otherwise, the model performed well, achieving an accuracy of 94.71% for multiclass diagnosis.

Other than the confusion matrix, another way to help see the model's overall performance on the test set is the classification report which gives the precision, recall, f1-score, and support for each of the classes with macro averaging as well as the weighted averaging. The classification report for the model is shown below:

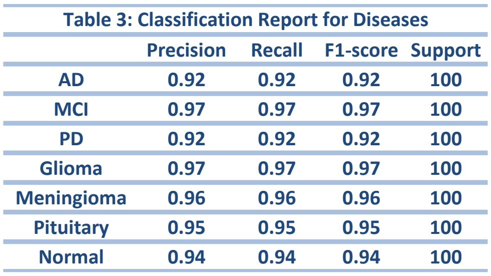


Table 3: Classification Report for Diseases

	Precision	Recall	F1-score	Support
AD	0.92	0.92	0.92	100
MCI	0.97	0.97	0.97	100
PD	0.92	0.92	0.92	100
Glioma	0.97	0.97	0.97	100
Meningioma	0.96	0.96	0.96	100
Pituitary	0.95	0.95	0.95	100
Normal	0.94	0.94	0.94	100

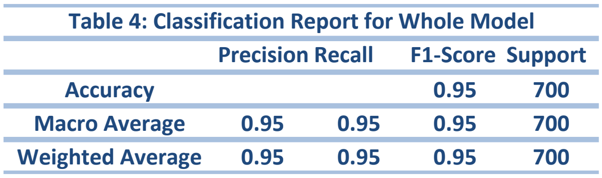


Table 4: Classification Report for Whole Model

	Precision	Recall	F1-Score	Support
Accuracy			0.95	700
Macro Average	0.95	0.95	0.95	700
Weighted Average	0.95	0.95	0.95	700

Table 3&4: Disease Classification Report

Below are the formulas used in the Classification Report:

$$Precision = \frac{TP}{TP + FP} \ (Equation\ 1)$$

$$Recall = \frac{TP}{TP + FN} \ (Equation\ 2)$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \ (Equation\ 3)$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \ (Equation\ 4)$$

The precision, recall, and f1-score values for the model are 92%, 97%, 92%, 97%, 96%, 95%, and 94% for AD, MCI, PD, Glioma, Meningioma, Pituitary, and Normal respectively. The accuracy, macro average, and weighted average of all the different methods of calculating performance are 95%, with a support of 700.

In addition, the Cohen's Kappa score for the model was calculated to be 0.9383 with the equation for the score shown below (p_0 represents relating observed agreement and p_e represents the probability of chance agreement):

$$Kappa\,Score = 1 - \frac{1 - p_0}{1 - p_e}\,(Equation\ 5)$$

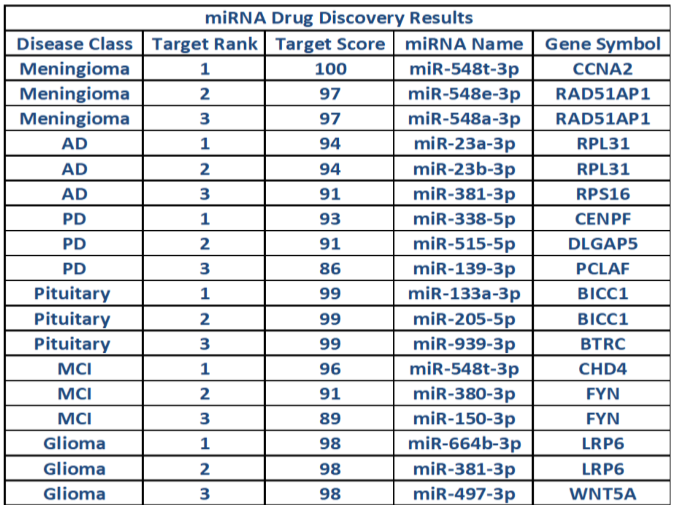
The Matthews correlation coefficient was calculated to be 0.9383 with the equation for the coefficient calculation shown below:

$$Coefficent = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}\,(Equation\ 6)$$

3.3 Identification of Novel miRNA Therapeutics

The blood-based biomarkers identified from the previous section for each disease was utilized and the hub genes were shortlisted. These hub genes were fed into the miRDB database to find target predictions for the corresponding miRNAs. The miRNAs which were identified had target prediction scores of greater than 60. These miRNA targets were used to create miRNA regulatory pathways using Cytoscape. The miRNAs identified with the greatest target prediction score were miR-664b for glioma tumors, miR-548t for meningioma tumors, miR-133a for pituitary tumors, miR-338 for PD, miR-23a for AD, and miR-548t for MCI. Thus, the identified miRNAs act as targets for drug inhibition and better treatment of these diseases. To treat the disease, the doctors would have to inject the corresponding anti miRNA oligonucleotides to downregulate the overexpressed miRNA and overexpressed genes which are the underlying mechanisms causing the particular disease.

The data tables and miRNA regulatory networks generated during the research for each of the six neurological conditions are shown below:



miRNA Drug Discovery Results				
Disease Class	Target Rank	Target Score	miRNA Name	Gene Symbol
Meningioma	1	100	miR-548t-3p	CCNA2
Meningioma	2	97	miR-548e-3p	RAD51AP1
Meningioma	3	97	miR-548a-3p	RAD51AP1
AD	1	94	miR-23a-3p	RPL31
AD	2	94	miR-23b-3p	RPL31
AD	3	91	miR-381-3p	RPS16
PD	1	93	miR-338-5p	CENPF
PD	2	91	miR-515-5p	DLGAP5
PD	3	86	miR-139-3p	PCLAF
Pituitary	1	99	miR-133a-3p	BICC1
Pituitary	2	99	miR-205-5p	BICC1
Pituitary	3	99	miR-939-3p	BTRC
MCI	1	96	miR-548t-3p	CHD4
MCI	2	91	miR-380-3p	FYN
MCI	3	89	miR-150-3p	FYN
Glioma	1	98	miR-664b-3p	LRP6
Glioma	2	98	miR-381-3p	LRP6
Glioma	3	98	miR-497-3p	WNT5A

Figure 7: New miRNA Therapeutics Discovered for Each Class of Disease

IV. Discussion and Conclusions

This paper helped describe the layers and characteristics of the NeuroXNet model, which achieved a test accuracy of 94.71%. This model is the first CNN model which approaches the diagnosis of neurodegenerative diseases, primarily Alzheimer's disease, Parkinson's disease, and Mild Cognitive Impairment and brain tumors (glioma, meningioma, and pituitary) through a novel deep learning architecture (NeuroXNet). The model helps find new blood biomarkers for the six diseases, through which a robust miRNA drug discovery pipeline is also developed. Furthermore, NeuroXNet is the first to integrate a treatment component into the model and uses genomic data with the MRI images to diagnose patients. Through this model, doctors and radiologists can diagnose neurological diseases at an earlier stage and use the diagnosis in treating patients with the proper medications and treatment procedures, helping prevent the disease from progressing onto a deadlier stage that could affect the patient's health drastically. Consequently, this model has great potential to be used clinically and improve the lives of numerous patients. Many results including the machine learning image classifier and novel miRNA therapeutics discovered have never been reported or published in scientific literature before. Through this paper, a new model is proposed and seen to attain a high accuracy that has many practical applications to radiology, neuroscience, and medicine, helping make a breakthrough in the diagnosis and treatment of neurological diseases using computational and biological approaches.

V. References

[1] F. E. Al-Khuzaie, O. Bayat, and A. D. Duru, “Diagnosis of Alzheimer disease using 2D MRI slices by convolutional neural network,” *Applied Bionics and Biomechanics*, vol. 2021.

[2] Bhatele K. R, Bhadauria S. S. Classification of Neurodegenerative Diseases Based on VGG 19 Deep Transfer Learning Architecture: A Deep Learning Approach. Biosc.Biotech.Res.Comm. 2020;13(4)

[3] Tong T, Ledig C, Guerrero R, Schuh A, Koikkalainen J, Tolonen A, Rhodius H, Barkhof F, Tijms B, Lemstra AW, Soininen H, Remes AM, Waldemar G, Hasselbalch S, Mecocci P, Baroni M, Lötjönen J, Flier WV, Rueckert D. Five-class differential diagnostics of neurodegenerative diseases using random undersampling boosting. Neuroimage Clin. 2017 Jun 12;15:613-624. doi: 10.1016/j.nicl.2017.06.012. PMID: 28664032; PMCID: PMC5479966.

[4] Sartaj Bhuvaji, Ankita Kadam, Prajakta Bhumkar, Sameer Dedge, and Swati Kanchan, “Brain Tumor Classification (MRI).” Kaggle, 2020, doi: 10.34740/KAGGLE/DSV/1183165.

[5] Friedman HS, Kerby T, Calvert H. Temozolomide and treatment of malignant glioma. Clin Cancer Res. 2000 Jul;6(7):2585-97. PMID: 10914698.