Graph Deep Learning:

Weighing the Benefits of Multimodal Machine Learning in Medical Diagnosis

Gaurav Sushanth Gajavelli

Pioneer Academics

Graph Analytics: Algorithms and Machine Learning

Dr. Julian Shun

September 2, 2021

Abstract

Today's medical professionals are trained to diagnose patients' illnesses by interpreting multiple modalities of patient data (such as imaging techniques and medical tests). Prohibitive costs, however, have sparked the use of machine learning (ML) which has recently shown huge promise with the task of medical diagnosis. To achieve usable accuracy at diagnosis tasks, the ML community has found themselves at a cross-roads: use a multimodal approach to mimic medical professionals or unimodal to avoid high model complexity. In this paper, we focus on the task of using functional Magnetic Resonance Imaging (fMRI) scans to diagnose Parkinson's Disease. We propose a unimodal graph convolutional network with mean pooling trained with cross-validation. Our experiments show that with a limited dataset of 260 fMRI scans, our model is able not only to outperform a state-of-the-art deep multimodal brain network, but ensures correctness through interpretability by identifying a key Parkinson's Disease biomarker used by professionals.

1 - Introduction

1.1 Describing and Motivating the Problem

Significant cost of healthcare and increased access to computing power have led to an increase in the applications of machine learning in medical diagnosis (Olveres et al., 2021). The versatility of machine learning allows it to be applied to a wide range of problems in the field, and many frameworks and models have been created and applied towards various diseases (Ahmedt-Aristizabal et al., 2021). A recent trend in the field has been to focus on the relatively novel topic of graphs neural networks (e.g., graph convolutional networks), as prior machine learning models were not usually able to capture the interdependencies of the many varieties of networks such as those in the brain (Zhang et al., 2020).

Another topic that has been gaining popularity is multimodal machine learning, in which multiple modalities (such as both text and images) are used to learn how to solve a given problem (Baltrušaitis et al., 2017). Multimodal learning can be used in many scenarios and on average performs machine learning tasks more accurately than learning on a single modality (D'Mello & Kory, 2012a). The ability to leverage information from multiple modalities has made multimodal machine learning a top choice for the field of automated medical diagnosis (Zhang et al., 2020). Additionally, on a high level, several modalities of medical tests and statistics (e.g., age and BMI) can be factored into diagnoses by doctors, indicating that machine learning accounting for multiple modalities is likely well-suited for medical diagnosis (Cleveland et al., 2007). As such, there has been a corresponding surge in medical diagnosis papers utilizing multimodal learning (Ahmedt-Aristizabal et al., 2021), leading to many new state-of-the-art results. There is, however, still much room for improvement in the field (Gao et al., 2020). There is value in determining how well-suited the rising multimodal learning approaches are in medical diagnosis, a field which could be benefitted by increased computational efficiency and explainability.

There has been growing interest in applications of automated medical diagnosis on mobile devices, where computing power is limited, in order to massively increase accessibility by leveraging the prevalence of smartphone ownership (Olveres et al., 2021). As such, an automated diagnosis with increased computational efficiency has been sought after. Additionally, explainability, consisting of interpretability and replicability (*IEEE SPM Special Issue on Explainability in Data Science: Interpretability, Reproducibility, and Replicability*, 2020), has been an emerging priority for automated medical diagnosis. In life-threatening situations, it is necessary for medical professionals to be able to trust machine learning models via interpretable decision-making (Joshi et al., 2021). The ability to identify biomarkers, an example of which would be traits indicative of a disease, would also require increased model interpretability (the degree to which a human can understand the cause of a decision and can consistently predict the model's results). Finally, the ability to generalize models to multiple diseases via replication is important to the field of diagnosis to fully utilize the potential of diagnostic frameworks. Although there are a variety of medical databases, the data tends to come in smaller sets than those used in other domains of machine learning (Chen et al., 2019). Recent studies by Algan and Bakbak review how medical datasets are also usually noisier than the average dataset; due to their complexity there is often disagreement even among human experts as to the correct labels (Algan & Bakbak, 2021). Since they tend to be smaller, the noise has the potential to significantly reduce the predictive performance of machine learning models (Algan & Bakbak, 2021).

The merits of multimodal machine learning in the context of medical diagnosis can be debated due to the relatively high computational costs and a lack of explainability. In this paper we compare these merits to those of unimodal approaches in the context of automated medical diagnosis, identify the characteristics of the most effective diagnosis models, and test these characteristics by leveraging them to create a unimodal deep learning model for Parkinson's diagnosis via graph classification.

1.2 Related Work

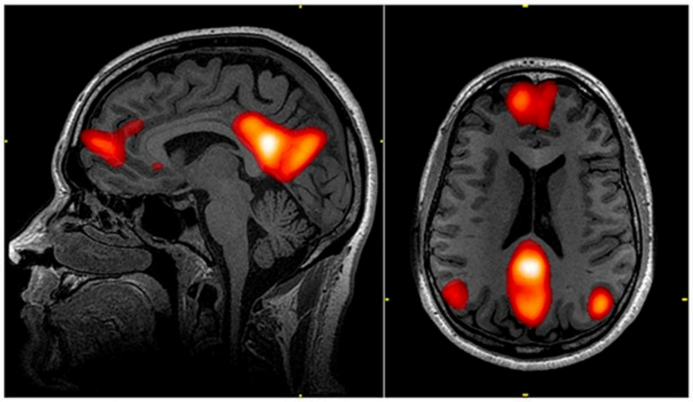


Figure 1. *(Left) An fMRI scan from a sagittal view. (Right) A transverse view of the same scan.* (Sahakian, 2017)

Magnetic Resonance Imaging (MRI) data has been a popular method of training machine learning algorithms to diagnose Parkinson's Disease (PD) patients. Different types of MRI have been developed and utilized in diagnoses (Zhang et al., 2020; Solana-Lavalle & Rosas-Romero, 2021), including functional MRI, or fMRI. Some approaches to diagnosis include ensemble learning, image classification, graph classification, voxel-based morphometry, and multimodal data fusion (Solana-Lavalle & Rosas-Romero, 2021). A difficulty in the field that research is forced to face is caused by limitations on the amount of data; even Parkinson's Progression Markers Institute (PPMI), one of the most extensive datasets for MRI and fMRI data, only contains a few hundred fMRI sessions. Despite this, several studies have been conducted on creating models for PD diagnosis based on the dataset.

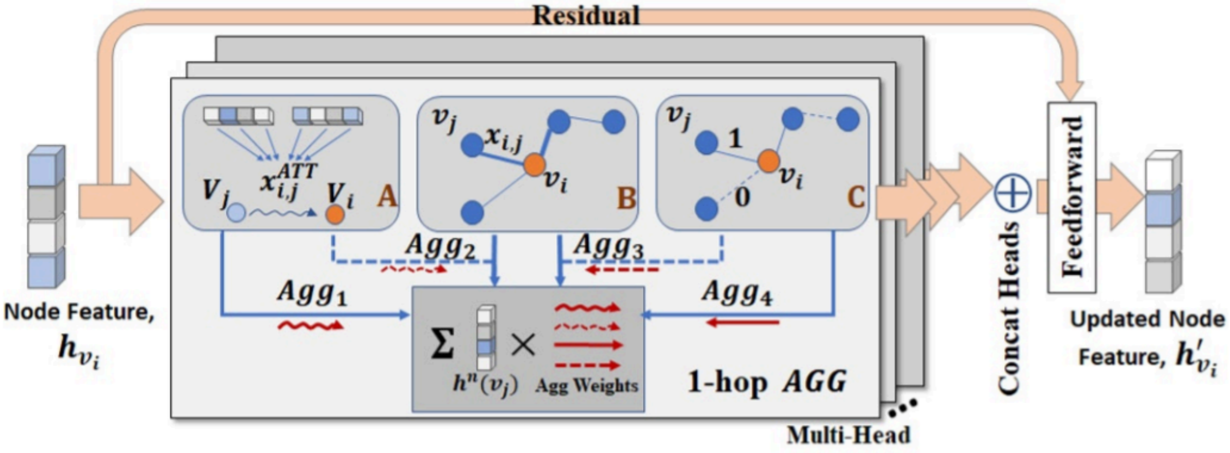


Figure 2. *The multi-graph convolutional kernel (MGCK) used to fuse MRI and fMRI data* (Zhang et al., 2020)

One such study, conducted by Zhang et al., leveraged the availability of both fMRI and MRI data for many subjects in the PPMI dataset, allowing them to perform graph classification using a multimodal model. Their contributions were using deep graph learning to model brain functions evolving from its structural basis and proposing an end-to-end automatic brain network representation framework based on the intrinsic graph topology. It also modeled the cross-modality relationship through a deep graph encoding-decoding process based on a proposed MGCK and drew graph saliency maps subject to the supervised tasks (Zhang et al., 2020). My usage of graphs to represent the network structure of the brain was inspired by this paper.

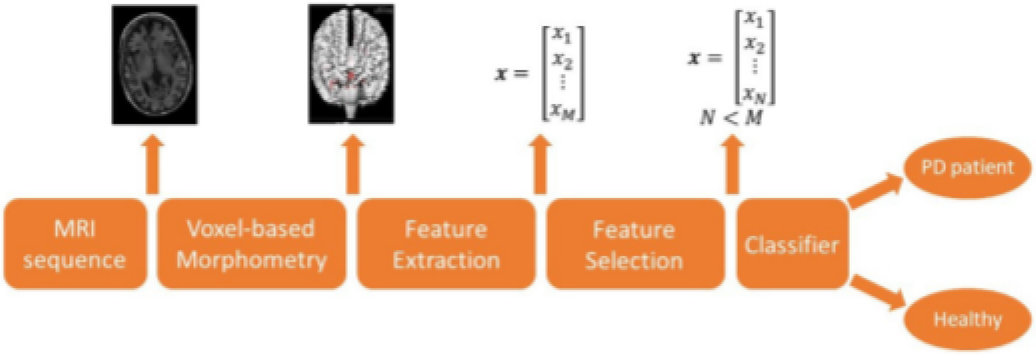

$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix}$
$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$
$N < M$
MRI sequence
Voxel-based Morphometry
Feature Extraction
Feature Selection
Classifier
PD patient
Healthy

Figure 3. *Voxel-based morphometry Parkinson's Disease Detection* (Solana-Lavalle & Rosas-Romero, 2021)

Another PD diagnosis study was conducted by Solana-Lavalle et al. and utilized voxel-based morphometry (VBM) with which they identified regions of interest from MRI scans in order to classify images as shown above. They trained an ensemble of models to maximize performance. Their main contributions were having conducted separate gender-based studies, using multiple classifiers to measure PD detection performance, achievement of high detection accuracy while using a small number of features, and the extraction of ROIs in brain locations different from those usually used for PD detection (Solana-Lavalle & Rosas-Romero, 2021).

However, the current state of automated Parkinson's diagnosis as shown above is not without its shortcomings; some include potentially high computational complexity, low interpretability, and replicability (the ease with which the data can be replicated). One issue was that the papers above did not provide implementations for their models. This affected both the replicability of the model's interpretability (such as the graph saliency maps) and that of the performance results. Additionally, the multimodal representation and ensemble models (Albardan, 2020) may have been unnecessarily complex solutions to the binary classification problem: could comparable performance have been attained with a simpler model? Addressing this question as it pertains to medical contexts, along with the other identified previous limitations, is the primary goal of this research.

2 - Background

Prior to introducing my model, it is helpful to define some preliminary terms, including:

Binary Classification. Predicting one of two classes and multi-class classification involves predicting one of more than two classes.

Hyperparameters. Parameters adjusted from outside the model whose values are used to control the learning process. The rate at which a model 'learns' by adjusting its weights is an example of a hyperparameter.

Features. An individual measurable property or characteristic of a phenomenon.

Labels. Something that describes the correct output (e.g., a picture of a dog might be labeled "dog").

MRI. Magnetic resonance imaging (MRI) is a medical imaging technique used in radiology to form pictures of the anatomy and the physiological processes of the body.

Functional MRI. Also known as fMRI, functional magnetic resonance imaging measures brain activity by detecting changes associated with blood flow; when an area of the brain is in use, blood flow to that region also increases.

K-Fold cross-validation. A procedure superior to creating single training and testing sets for a dataset in estimating the predictive performance of the model on new data. It creates k different training sets each with its own unique test set by identifying k different ways to split the original data.

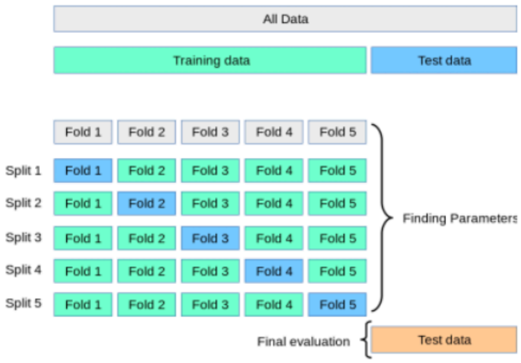


Figure 4. *A visualization of k-fold cross-validation* ("3.1. Cross-Validation: Evaluating Estimator Performance — Scikit-Learn 0.24.2 Documentation," n.d.)

Convolutional Neural Network. Convolutional networks are a specialized type of neural networks that use convolution in place of general matrix multiplication in at least one of their layers. An example of convolution is averaging the values from an NxN pixels size 'filter' as it passes over an image, compressing the information into a smaller grid.

CONN Toolbox. CONN is an open-source MATLAB-based cross-platform software for the computation, display, and analysis of resting state functional MRI data (rsfMRI). rsfMRI is used interchangeably with fMRI in this paper as during all sessions in which fMRI data was obtained patients were at a resting state.

ROI. A region of interest. In this context it refers to regions in the brain relating to connectivity. There are 164 ROIs in the standard atlas for the CON toolbox (add those in networks.info and those in atlas.info)

Performance Metrics. Different metrics are required to fully capture the effectiveness of the model. To account for this accuracy, sensitivity, specificity, and precision, metrics often used in the literature (Solana-Lavalle & Rosas-Romero, 2021), were all used.

Mean Pooling Layer. It compresses input by dividing it into pooling regions and calculating the mean values of each one.

3 - Technical Approach

As was mentioned before, the reasoning behind the technical approach outlined below was based on answering the following questions: is multimodal machine learning (through the advantages and disadvantages that differentiate it from unimodal learning) the better design choice when considering medical diagnosis problems? In order to do so, it is necessary to identify the characteristics of both multimodal machine learning and the needs for optimal automation in medical diagnosis.

The multimodal machine learning models generally have (1) relatively higher predictive performance (D’Mello & Kory, 2012b), (2) high data requirements, computing requirements, and potential for overfitting due to large numbers of weights (Gao et al., 2020), and (3) low interpretability (D’Mello & Kory, 2012a) and replicability due to their complexity (Haibe-Kains et al., 2020). It is relatively difficult to replicate results from multimodal models due to the increased complexity inherent to fusing multiple modalities in learning. The characteristics of automated medical diagnosis (including medical data) that overlap with the differences between multimodal and unimodal learning include relatively small and noisy (Algan & Bakbak, 2021) datasets and an emphasis on inference efficiency, for time-sensitive (Joshi et al., 2021) and mobile (Olveres et al., 2021) diagnoses. Such datasets are problematic for any kind of learning due to the potential for overfitting (Reed & Marksii, 1999); this issue is exacerbated by the need for multiple modalities of data and overfitting inducing noise (Gao et al., 2020). Also important in medical contexts are replicability (for use of models on different diseases) and interpretability (Joshi et al., 2021); as stated before, trustworthy diagnoses are necessary in life-threatening situations. It is better to have a model whose reasoning can be understood and trusted, even if it is less accurate (Kim, et al., 2016).

Per the lists and statements above, multimodal machine learning, despite initially seeming to outperform unimodal, is not necessarily more suitable. Consequently, unimodal learning is a characteristic identified as generally better-suited to medical diagnosis. As such, the model was designed with explainability and efficiency in mind.

3.1 Model

To test the model, we select the diagnosis of Parkinson's Disease, the second most common neurodegenerative disease (Gwinn, 2013), as the main task. Since brains are networks (D'Mello & Kory, 2012a) machine learning with graphs was conducted; specifically, patient fMRI brain scans were represented by graphs. I took a deep learning approach by using a graph convolutional network for the binary classification task of Parkinson's diagnosis.

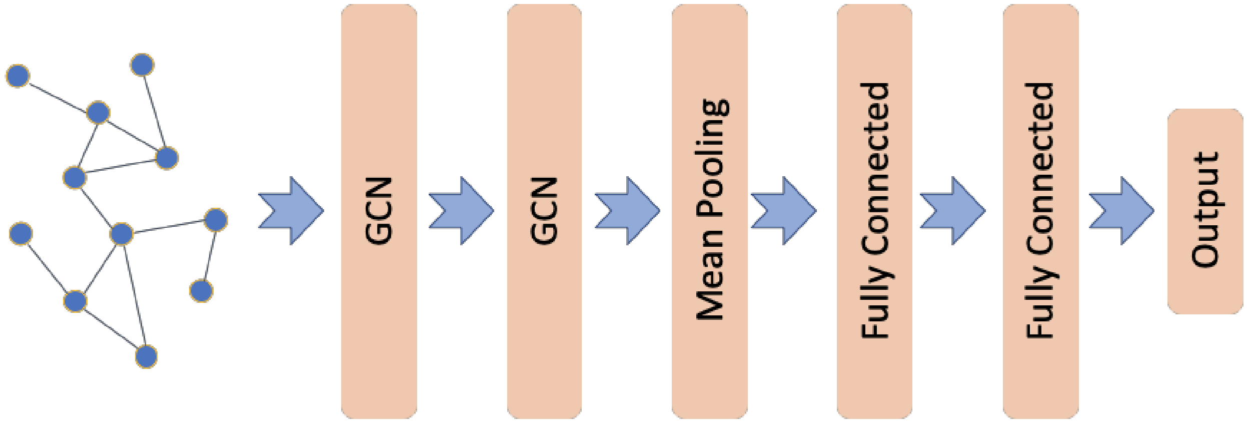


Figure 5. *Proposed model architecture* (CSIRO's Data61, 2018)

The suggested model used for medical diagnosis was a unimodal graph convolutional network with mean pooling trained with cross-validation. Because it is unimodal and uses cross-validation, it is more data efficient in that small-dataset-induced overfitting can be mitigated. As mentioned before, it also is easier to replicate and interpret due to not requiring the fusion of data modalities, such as via a multi-stage graph convolutional kernel (Zhang et al., 2020), allowing it to be more easily used in many contexts. Finally, it will aim to provide better inference time, which is generally valuable in time-sensitive medical contexts, by keeping the model simple and data to one modality.

More specifically, it consists of graph convolutional, mean pooling, and fully connected layers. The final fully connected layer has a 2-dimensional output feature representing the final binary classification (CSIRO’s Data61, 2018). The fully connected layers’ role was to learn features and classify data. The first GCN layer aggregated features from the graph, and the second aggregated from the first GCN layer’s output. Next, after receiving it from a GCN layer, the mean pooling layer compressed the dimensionality of the input even further for the first fully connected layer’s input.

Graph convolutional network, or GCN, layers were used (Aghdam & Heravi, 2017) since it is often impractical to always use fully connected networks due to the high computational cost. Another measure towards reducing complexity was utilizing ReLU as the activation function for the GCN layers and first fully connected layer; its simplicity allowed for the introduction of non-linearity with relatively little loss of training efficiency.

3.2 Training

K-fold cross-validation was used to prevent poor estimates of the efficacy of the model, a common problem caused by the effects of the arbitrary differences in testing-training splits on predictive performance and exacerbated by relatively small datasets like PPMIs due to greater variability between the potential splits. This allowed it to detect overfitting, an issue especially when learning from particularly noisy data (e.g., medical data) since there is a greater proportion of noise for the model to ‘memorize’ relative to the signal contained in the dataset. The number of folds (a hyperparameter) was tuned for the best performance on average, avoiding the shortfall of overfitting (Crowley, 2019). Since diagnosis is a binary classification problem in this case, the loss function was binary cross entropy.

Adam was used as the optimizer. It is fairly robust to the value of the hyperparameters, reducing the significance of tuning for performance and making it easier to use (Kingma & Ba, 2014).

3.3 Evaluation

Providing the classification output, the final layer was a sigmoid activation, as it is the binary version of the softmax function for multi-class classification (Nwankpa et al., 2018). Four established metrics were used to evaluate the model's predictive performance: accuracy, specificity, recall or sensitivity, and precision (Solana-Lavalle & Rosas-Romero, 2021). PageRank was also included; it allowed for interpretability by looking for the regions of interest most active in determining diagnoses.

3.4 New Methods Proposed

In order to improve interpretability by determining which areas of the brain contributed the most to the classifications, the graphs within each group (as classified by the model) were recorded. The corresponding nodes' values (based on PageRank) for each graph were added for all of the graphs of the control and PD groups. In order to account for potential disparities in the numbers of graphs classified as each group (and therefore different total PageRanks in both graphs), each of the node weights were divided by the number of graphs it accumulated its weights from. Finally, the respective node weights from the collective PD and control graphs were subtracted and the magnitude was taken of each node. Afterwards, the top n nodes (as specified by the user) were used to determine which brain regions (via the nodes' labels) contributed the most.

3.5 Datasets

To conduct this research, fMRI images were obtained from the Parkinson's Progression Markers Initiative sponsored by the Michael J. Fox Foundation. One result of this initiative is one of the most robust libraries of clinical and imaging data and biosamples for Parkinson's disease research (*Real Talk from Patients*, n.d.). The biomarkers identified by the foundation could be used in therapeutic studies, which was the ultimate goal for the initiative. The scans were generated via Tesla scanner, with whole sessions lasting 8.5 minutes (Manza et al., 2015). The fMRIs were acquired as a 3D sequence and stored in the NIfTI file format, with 212 being captured over the course of the session. The data of 260 PD patients and healthy controls were used.

3.6 Implementation

The model was created using Python and more specifically utilized Keras (for implementations of the layers, loss function, optimizer, and performance metrics), Pandas (for dataframes), NetworkX (for PageRank), and Stellargraph, while the data processing pipeline was in MATLAB.

First, the fMRI data was converted from 3D NIfTI to 4D by combining all 212 scan files in each patient visit (also called a session) into a single file containing both the 3D functional connectivity data and the times at which they occurred, comparable to a 3D video. CONN's batch functionality was used to automate the creation of graphs for each session.

After being formatted as such, CONN was used to preprocess, denoise, and analyze the sessions, after which it outputted graphs converted from the fMRI data. Smoothing (removal of noise) was avoided during preprocessing in order to preserve as much information as possible for the conversion. The graph's nodes consisted of brain regions and the edges represented brain connections. The fMRI functional graphs initially had edge weights representing the levels of interaction between all pairs of nodes, but to accommodate learning by a simpler model that did not account for weighted edges, the connections were changed to weights of either 0 or 1 based on whether they were above or below a threshold (|z-score| >= 1), essentially creating an unweighted graph.

The outputted graphs and the labels for the nodes (which ROIs they were) were converted via Python from the .mat files outputted by CONN into Pandas dataframes for use in Python along with NetworkX, Keras, and Stellargraph.

A Stellargraph data generator was used throughout the implementation to improve the scalability of the model by loading only a few graphs at a time, a necessity for use of the model on diagnosis problems for different conditions with much larger datasets.

4 - Experiments

The PPMI dataset was used to evaluate the efficacy of the model, data pipeline, and training methods, as well as that of the PageRank interpretability. In order to obtain the best results possible, various hyperparameters were adjusted.

4.1 Adjusting Hyperparameters

The most relevant hyperparameters adjusted were the batch size, number of folds, and learning rate. The hyperparameter values used were as follows: two different batch sizes (10 and 20) chosen due to the advantages of smaller batches (Masters & Luschi, 2018), 2 different numbers of folds (4, 8), and 3 different learning rates (0.001, 0.005, and 0.010). The different numbers of folds were chosen with regards to the small data size. The number of epochs was kept at 500 since the change in the loss per epoch always reached 0 relatively quickly, while the numbers of neurons weren’t changed due to changes causing greater underfitting and overfitting. This resulted in 12 different possible combinations of hyperparameters (2 times 2 times 3) to be tested. Below are tables for the results of the various combinations.

4-Fold Cross-validation

Batch Size	Learning Rate		
	0.001	0.005	0.01
10	Accuracy: 81.0% Precision: 78.1% Recall over: 75.0% Specificity: 87.1%	Accuracy: 92.1% Precision: 98.4% Recall: 85.8% Specificity: 98.3%	Accuracy: 66.9% Precision: 68.8% Recall: 71.2% Specificity: 62.5%
20	Accuracy: 85.2% Precision: 88.1% Recall: 89.2% Specificity: 81.2%	Accuracy: 87.5% Precision: 90.9% Recall: 90.4% Specificity: 84.6%	Accuracy: 74.6% Precision: 81.2% Recall: 61.7% Specificity: 87.5%

Figure 7. *Mean performance metrics over all folds.*

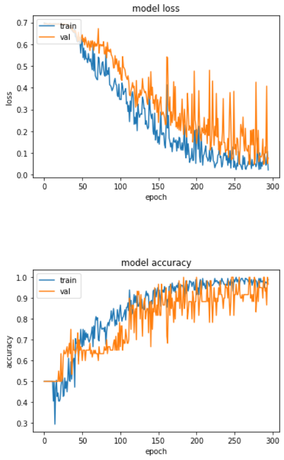
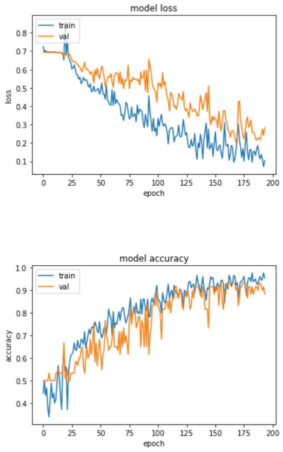
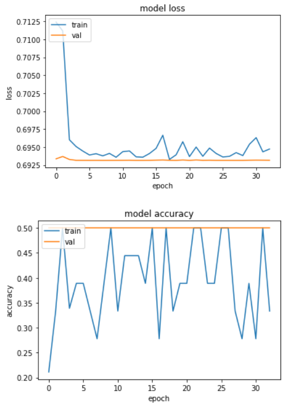
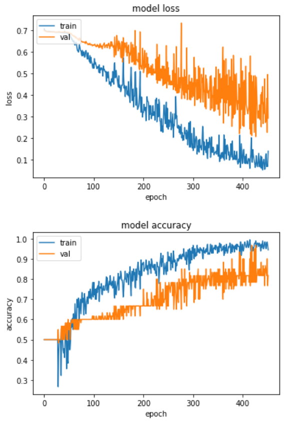
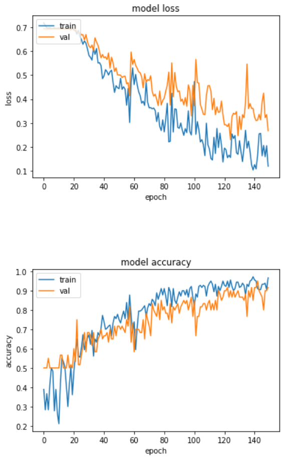
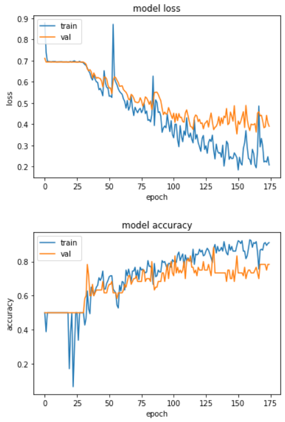
Batch Size	Learning Rate		
	0.001	0.005	0.01
10			
20			

Figure 8. *(Top) Average epoch vs. loss. (Bottom) Average epoch vs. accuracy.*

8-Fold Cross-validation

Batch Size	Learning Rate		
	0.001	0.005	0.01
10	Accuracy: 78.7% Precision: 75.0% Recall: 81.7% Specificity: 75.8%	Accuracy: 87.7% Precision: 96.0% Recall: 82.5% Specificity: 92.9%	Accuracy: 62.7% Precision: 62.5% Recall: 62.9% Specificity: 62.5%
20	Accuracy: 68.3% Precision: 56.2% Recall: 74.2% Specificity: 62.5%	Accuracy: 81.2% Precision: 81.6% Recall: 72.1% Specificity: 90.4%	Accuracy: 74.2% Precision: 84.3% Recall: 62.9% Specificity: 85.4%

Figure 9. *Mean performance metrics over all folds.*

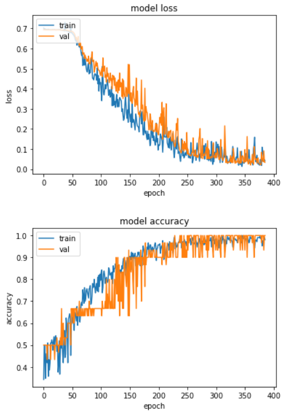
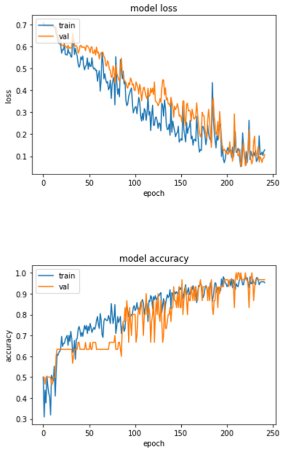
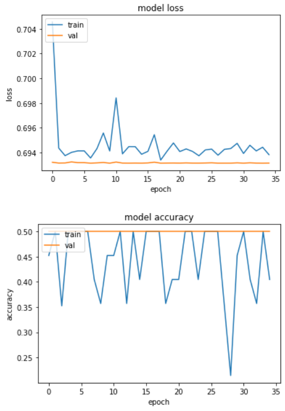
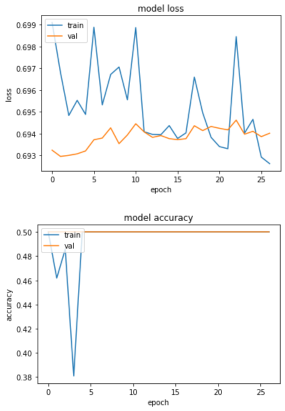
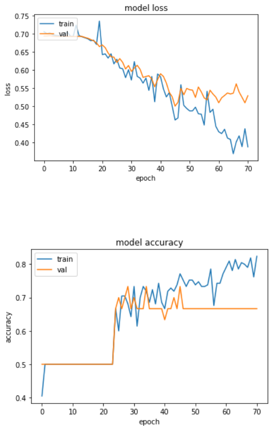
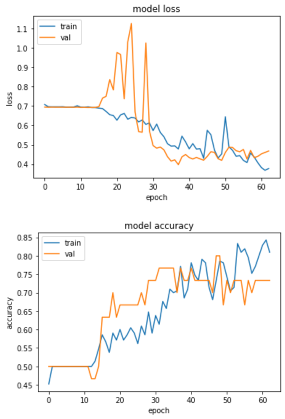
Batch Size	Learning Rate		
	0.001	0.005	0.01
10			
20			

Figure 10. *(Top) Average epoch vs. loss. (Bottom) Average epoch vs. accuracy.*

The hyperparameter combinations that led to poor performance relative to the others through either overfitting or underfitting will be addressed first. Nearly all training performed with a learning rate of 0.01 and/or a batch size of 20 resulted in models that overfit the data. This was indicated by increases in training accuracy or decreases in training loss despite the validation loss and training remaining relatively static, indicating that noise, something useful only for decreasing training loss, was being memorized. As such, the models tended to perform worse than the others on the four performance metrics. Generally 0.01 and 20 hyperparameter combinations failed to converge near a certain accuracy by the end of the training, although some of them with other batch sizes or learning rates.

The combinations with a learning rate of 0.005 and a batch size of 10 resulted in two models on the other end of the performance spectrum. Although smaller batch sizes limit what the model can see at once, the 4-fold cross-validation with 0.005 and 10 hyperparameters achieved the best results. The hyperparameter combinations with the smallest learning rate of 0.001 achieved middling performance, likely resulting from the increased overfitting. Finally, a general conclusion suggested by the results of the experiments is that there appears to be a continuous decrease in performance as any of the hyperparameters move further from the best combination.

4.2 Training Approach

In training the GCN, I used the following values for the hyperparameters: at most 500 epochs, 4 folds, a learning rate of 0.004, and a batch size of 10. The GCN layers each had 64 neurons, the mean pooling layer 32, the first fully connected 16, and the output fully connected layer 1.

4.3 Results and Discussion

Performance Metrics	Percentages
Accuracy	91.2%
Precision	95.5%
Recall	86.7%
Specificity	95.8%

Figure 11. *The performance metrics of the GCN*

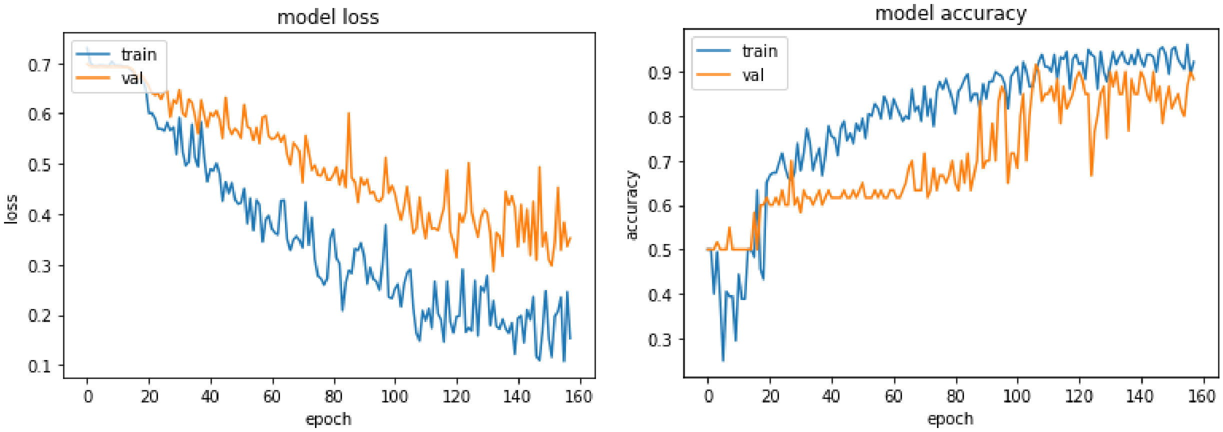


Figure 12. *(Left) average epoch over all folds vs. loss. (Right) Average epoch vs. accuracy.*

Considering the large gap between the test loss and the training loss, the model seems to have overfit the data. However, the epoch vs. accuracy graphs are much more similar, and the metrics indicate that the model performed well enough on the test data to outperform unimodal and multimodal models (Solana-Lavalle & Rosas-Romero, 2021; Zhang et al., 2020).

Finally, in order to assess the PageRank interpretability, brain regions were identified via the aforementioned adding, averaging, and subtracting of brain graphs: the frontal pole, right middle frontal gyrus, left middle frontal gyrus, and inferior frontal gyrus pars triangularis. The frontal gyrus is considered a Parkinson's Disease biomarker (Zhang et al., 2020), indicating that the correctness of the model's diagnosis could be corroborated by professionals.

5 - Conclusion

In this paper I argued for a set of characteristics for medical diagnosis machine learning models and proposed a model exemplifying these characteristics made for Parkinson's diagnosis. Specifically, a relatively efficient and explainable model was used to outperform recently created models for Parkinson's Diagnosis. There is, however, still more work that could be done; for instance, experiments with different diseases, different machine learning models, and real-world applications would help verify the general utility of explainable and simple medical diagnosis models. Additionally, an ablation study could be conducted to help understand the inner workings of the graph machine learning example. I would like future work to continue investigating what was responsible for the model's successes and failures to try to emulate and improve upon it.

6 - References

3.1. Cross-validation: Evaluating estimator performance — scikit-learn 0.24.2 documentation. (n.d.). Https://Arxiv.Org/Pdf/1804.07612.Pdf. Retrieved August 31, 2021, from https://scikit-learn.org/stable/modules/cross_validation.html

Aghdam, H. H., & Heravi, E. J. (2017). *Guide to convolutional neural networks: A practical application to traffic-sign detection and classification*. Springer.

Ahmedt-Aristizabal, D., Armin, M. A., Denman, S., Fookes, C., & Petersson, L. (2021). Graph-Based deep learning for medical diagnosis and analysis: Past, present and future. *Sensors*, *21*(14), 4758. https://doi.org/10.3390/s21144758

Albardan, M. (2020, December 30). Insights on classifier combination. *Towards Data Science*. https://towardsdatascience.com/insights-on-classifier-combination-da56f764fcfa

Algan, G., & Bakbak, B. (2021, February 15). *Deep Learning from Small Amount of Medical Data with Noisy Labels: A Meta-Learning Approach*. Computer Vision and Pattern Recognition; Cornell University. https://arxiv.org/abs/2010.06939

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2017, May 26). *Multimodal machine learning: A survey and taxonomy*. ArXiv.Org. https://arxiv.org/abs/1705.09406

Chen, P.-H. C., Liu, Y., & Peng, L. (2019). How to develop machine learning models for healthcare. *Nature Materials*, *18*(5), 410–414. https://doi.org/10.1038/s41563-019-0345-0

Cleveland, R. J., Eng, S. M., Abrahamson, P. E., Britton, J. A., Teitelbaum, S. L., Neugut, A. I., & Gammon, M. D. (2007). Weight gain prior to diagnosis and survival from breast cancer. *Cancer Epidemiology and Prevention Biomarkers*, *16*(9), 1803–1811. https://doi.org/10.1158/1055-9965.EPI-06-0889

Contributors to Wikimedia projects. (2021, August 27). *Magnetic resonance imaging*. Wikipedia. https://en.wikipedia.org/wiki/Magnetic_resonance_imaging

Crowley, B. G., Mark. (2019). *The theory behind overfitting, cross-validation, regularization, bagging, and boosting: Tutorial*.

CSIRO’s Data61, C. D. (2018). *Supervised graph classification with GCN —*

StellarGraph 1.2.1 documentation. StellarGraph; GitHub. https://stellargraph.readthedocs.io/en/stable/demos/graph-classification/gcn-supervised-graph-classification.html

D'Mello, S., & Kory, J. (2012a). Consistent but modest. *Proceedings of the 14th ACM International Conference on Multimodal Interaction - ICMI '12*. http://dx.doi.org/10.1145/2388676.2388686

D'Mello, S., & Kory, J. (2012b). Consistent but modest. *Proceedings of the 14th ACM International Conference on Multimodal Interaction - ICMI '12*. http://dx.doi.org/10.1145/2388676.2388686

Gao, J., Li, P., Chen, Z., & Zhang, J. (2020). A survey on deep learning for multimodal data fusion. *Neural Computation*, *32*(5), 829–864. https://doi.org/10.1162/neco_a_01273

Gwinn, M. (2013). *Genetics, coffee consumption, and Parkinson's disease*. Genomics & Precision Health; CDC. https://www.cdc.gov/genomics/hugenet/casestudy/parkinson/parkcoffee_view.ht

Haibe-Kains, B., Adam, G. A., Hosny, A., Khodakarami, F., Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., Greene, C. S., Broderick, T., Hoffman, M. M., Leek, J. T., Korthauer, K., Huber, W., Brazma, A., Pineau, J., Tibshirani, R., Hastie, T., … Aerts, H. J. W. L. (2020). Transparency and reproducibility in artificial intelligence. *Nature*, *586*(7829), E14–E16. https://doi.org/10.1038/s41586-020-2766-y

IEEE SPM special issue on explainability in data science: Interpretability, reproducibility, and replicability. (2020, November 12). IEEE Signal Processing

Society. https://signalprocessingsociety.org/blog/ieee-spm-special-issue-explainability-data-science-interpretability-reproducibility-and

Joshi, G., Walambe, R., & Kotecha, K. (2021). A review on explainability in multimodal deep neural nets. *IEEE Access*, *9*, 59800–59821. https://doi.org/10.1109/access.2021.3070212

Kim, B., Khanna, R., & Koyejo, O. (2016). Examples are not enough, learn to criticize! Criticism for Interpretability. *Advances in Neural Information Processing Systems*.

Kingma, D. P., & Ba, J. (2014, December 22). *Adam: A method for stochastic optimization*. ArXiv.Org. https://arxiv.org/abs/1412.6980

Lotankar, S., Prabhavalkar, K. S., & Bhatt, L. K. (2017). Biomarkers for parkinson's disease: Recent advancement. *Neuroscience Bulletin*, *33*(5), 585–597. https://doi.org/10.1007/s12264-017-0183-5

Manza, P., Zhang, S., Li, C.-S. R., & Leung, H.-C. (2015). Resting-state functional connectivity of the striatum in early-stage Parkinson's disease: Cognitive decline and motor symptomatology. *Human Brain Mapping*, *37*(2), 648–662. https://doi.org/10.1002/hbm.23056

Masters , D., & Luschi, C. (2018). *REVISITING SMALL BATCH TRAINING FOR DEEP NEURAL NETWORKS*.

MRI: MedlinePlus medical encyclopedia. (n.d.). Retrieved August 31, 2021, from https://medlineplus.gov/ency/article/003335.htm

Nwankpa, C., Gachagan, , A., Marshall, S., & Ijomah, W. (2018). Activation Functions:

Comparison of trends in Practice and Research for Deep Learning. *Machine Learning*.

Olveres, J., González, G., Torres, F., Moreno-Tagle, J. C., Carbajal-Degante, E., Valencia-Rodríguez, A., Méndez-Sánchez, N., & Escalante-Ramírez, B. (2021). What is new in computer vision and artificial intelligence in medical image analysis applications. *Quantitative Imaging in Medicine and Surgery*, *11*(8), 3830853–3833853. https://doi.org/10.21037/qims-20-1151

Real talk from patients. (n.d.). The Michael J. Fox Foundation for Parkinson's Research | Parkinson's Disease. Retrieved September 1, 2021, from https://www.michaeljfox.org/real-talk-patients

Reed, R., & Marksii, R. J. (1999). *Neural smithing - Supervised learning in feedforward artificial neural networks*.

Sahakian, B. J. (2017, February 9). *Brain scanners allow scientists to "read minds" – could they now enable a "Big Brother" future?* The Conversation. https://theconversation.com/brain-scanners-allow-scientists-to-read-minds-could-they-now-enable-a-big-brother-future-72435

Solana-Lavalle, G., & Rosas-Romero, R. (2021). Classification of PPMI MRI scans with voxel-based morphometry and machine learning to assist in the diagnosis of Parkinson's disease. *Computer Methods and Programs in Biomedicine*, *198*, 105793. https://doi.org/10.1016/j.cmpb.2020.105793

Zhang, W., Zhan, L., Thompson, P., & Wang, Y. (2020). Deep representation learning for multimodal brain networks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* (pp. 613–624). Springer International

Publishing. http://dx.doi.org/10.1007/978-3-030-59728-3_60