Developing an Interpretable Schizophrenia Deep Learning Classifier on fMRI and sMRI using a Patient-Centered DeepSHAP

Jacob Reiter
Department of Computer Science
Stanford University
CS 335 - FAccT Deep Learning
jreiter@stanford.edu

Abstract

This paper aims to explore the applications of interpretable methods to the task of schizophrenia classification. We use the Mind Research Network's data which consists of functional network connectivity values (FNC) and source-based morphometry (SBM) loadings. These features are derived from fMRI and sMRI, respectively. We design a deep neural network for the classification task and achieve high AUC for the literature. Then, we explore various iterations of DeepSHAP, a form of SHapley Additive exPlanations or estimations of Shapley values, as our interpretable method. We find that Standard DeepSHAP is a successful tool for schizophrenia interpretability. However, we note that a Label-Consolidated DeepSHAP more accurately portrays collective feature importance for FNC values. Lastly, we conduct initial needfinding for a Patient-Centered DeepSHAP concept; from a human-centered perspective, we explore a patient experience with our DeepSHAP results.

1 Introduction

For most modern day psychiatric disorders, physicians are left to subjective measures in order to diagnose patients. With no blood test for mental illness, psychiatrists look to the Diagnostic and Statistical Manual of Mental Disorders (DSM) in order to make a diagnosis. Yet, even the requirements laid out by the DSM have been critiqued as arbitrary in nature and continue in a reliance on subjective reporting from the patient.

Achieving a correct diagnosis is critical, as it indicates the optimal treatment for a patient. This problem is even more apparent for those suffering from psychosis, delusions, and hallucinations. Schizophrenia, schizoaffective disorder, delusional disorder, schizotypal disorder, and schizophreniform disorder all fall within the family of schizophrenia spectrum disorders. However, psychosis might often be an indication of bipolar disorder with psychotic features; the same is seen with major depressive disorder. These minute, but critical distinctions, often leave psychiatrists guessing for a correct diagnosis. The dangers of prescribing a patient mood stabilizers when anti-psychotics are actually necessary are great: the side effects of anti-psychotics can be debilitating and mood stabilizers would not provide the preventative measures for an impending fully developed schizophrenia. The converse is also true, however, almost half of patients diagnosed within the family of schizophrenia spectrum disorders do not actually have any [1]. Thus, it is critical to develop new methods of accurately diagnosing schizophrenia in patients.

Even more so, for such a technology to exist, it is critical that physicians can make use of the predictions given. Current research in the literature for schizophrenia classification is rooted in

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

black-box models that fail to provide transparency for physicians. Physicians should be able to see from where decisions are derived, and choose how to use that information in combination with their own examination and training. We also will explore how psychiatrists view this topic, and consider how these same questions arise for patients.

It is in this way that I argue developing a transparent, interpretable, and explainable model for diagnosing schizophrenia is of importance. This paper will employ a deep learning neural network for the schizophrenia classification task. DeepSHAP will be explored as the interpretable method. Modifications to DeepSHAP will be considered in order to meet physician and patient needs through a human-computer interaction lens.

2 Data

This research makes use of the Mind Research Network's publicly available data through the IEEE International Workshop on Machine Learning for Signal Processing. The data was released to the public in 2014, and centers on both functional magnetic resonance imaging (fMRI) and structural magnetic resonance imaging (sMRI).

The data concerns 144 patients, 69 received diagnoses of schizophrenia. The data has been pre-processed into functional network connectivity (FNC) maps and features, as well as source-based morphometry (SBM) loadings and maps. FNC values and SBM loadings derive from fMRI and sMRI, respectively.

FNC values are a type of functional modality feature, and indicate correlation between brain maps over time. SBM loadings are built upon voxel-based morphometry (VBM), and use independent component analysis to identify groupings. SBM can be understood as a method of measuring gray matter in the brain.

Below is an example of the Mind Research Networks' fMRI cross-sectional brain map.

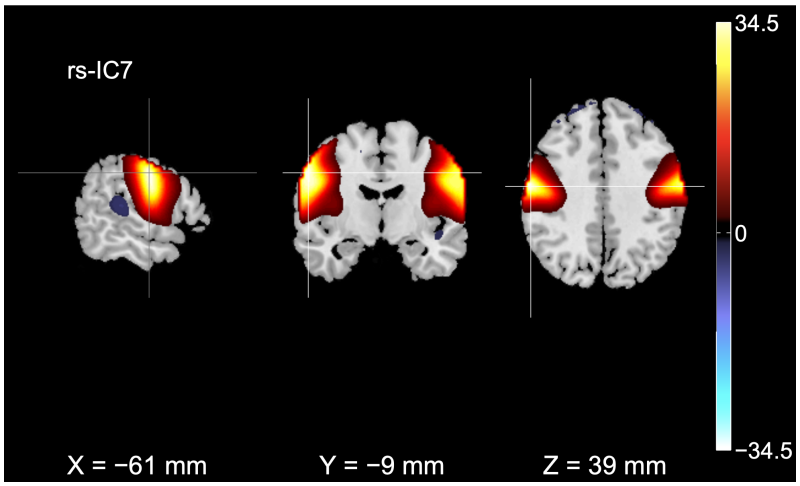

Figure 1: Cross-sectional fMRI brain map.

The 32 SBM loading features are derived from the 32 cross-sectional sMRI maps. The 378 FNC values, which represent brain map correlation, correspond to each pairwise combination of the 28 cross-sectional fMRI maps.

Research now posits that "FNC in schizophrenia patients shows less hemispheric asymmetry compared to that of the healthy controls" [2]. In addition, studies have also shown that SBM can indicate presence of schizophrenia [3], as gray matter presence can largely be understood as brain operating power in correlated regions of the brain.

This is also a High Dimensional Low Sample Size (HDLSS) dataset, in which there are minimal subjects, but a large number of features. This informs design decisions for the deep learning model presented.

3 Related Work

The data this paper considers has been tackled in some other notable works. Two Finland researchers at Aalto University have produced the best results thus far using Gaussian process classification [4], where one considers observations as coming from Bernoulli distributions. Additionally, a team of doctoral researchers in India published in the Journal of Applied Science and Computations explain that they used a Random Forest classifier and stayed away from deep learning approaches [5].

Other approaches include that from Karolis Koncevicius of Lithuania's Vilnius University in which he used distance weighted discrimination to achieve a high AUC score [6].

Almost every paper cited avoids the usage of a deep learning model and explores some other method. Of all the published studies explaining approaches to solving this schizophrenia classification problem, only one makes notable mention of using neural networks or deep learning. Kamini Jodha's research [7] shows his exploration of only a maximum of 3 hidden layers in a schizophrenia classification task, however he does not dive deeper in optimization. Other research not using the MRI data discussed thus far has shown that deep learning models for the diagnosis of schizophrenia can be very promising. For example, work published in the Schizophrenia Research Journal explores the application of deep learning to EEG features and data [8].

No research thus far in the literature has aimed to make these schizophrenia classifiers interpretable, transparent, or explainable. None consider the specific needs of patient and physician.

4 Approach

4.1 Pre-Processing and Environment

A significant consideration for our research is the problem of a High Dimensional Low Sample Size dataset. Several approaches are considered in this paper, firstly we explore a simple training, validation, and test set split. Alternatively, we explore a K-fold cross validation method, in which more of the training and validation data can be used.

Data across FNC and SBM loadings are consolidated by subject for training, and errors and missing values are cleaned in pre-processing. Data manipulation is done in NumPy and pandas.

This research is conducted with use of Google Cloud Platform's Compute Engine, using a VM instance with one NVIDIA Tesla K80 GPU with a virtual environment.

4.2 Baselines

4.2.1 Interpretability Baseline

For interpretability, the true baseline that we use is the finalized deep neural network prior to any interpretable methods being appended or added. These set a clear expectation for the current baseline interpretability that exists for deep learning models.

To continue, we as well develop baselines for the deep neural network.

4.2.2 DNN Baseline

While the focus of this paper is to explore and develop an interpretable schizophrenia classifier, it is critical to ensure that the classification itself is accurate and successful. Thus, a significant portion of this paper is devoted to developing a deep neural network on which we will apply an interpretable method. The primary metric for development is accuracy, but additional metrics are considered. For medical predictions, minimizing false negatives and false positives is of utmost importance.

For the deep neural network, we chose to develop with Keras and TensorFlow. We create a simple model with only two dense layers. First, we split the data on a 0.8 to 0.2 training to validation split. Data is sliced into TensorFlow batches of size 16.

This baseline model has one dense layer of 128 units, followed by a prediction dense layer with sigmoid activation. The starting learning rate is a simple and constant 0.3, binary cross entropy loss,

and accuracy as a metric. There are 48,641 trainable parameters. We achieved a validation accuracy of 0.7059 after 15 epochs.

When running this model on the actual test data that is hidden by the data owners, all that can be seen is the area under the ROC curve score, for which this very simple model gets an AUC of 0.62053.

4.3 DNN Final Implementation

Further network architecture research was conducted on the deep neural network.

Firstly, we adjust several components of the data pre-processing, including shifting the labels from 0 and 1 to a centered -0.5 and 0.5. Additionally, we coalesce the FNC and SBM features.

Secondly, we now use 4 dense layers. Beginning with 128 units, followed by two 256 unit layers, and a final prediction dense layer with sigmoid activation.

Thirdly, we continue with our usage of an Adam optimizer, but lower our learning rate to 0.001.

4.4 DeepSHAP

4.4.1 Standard DeepSHAP

In making the model transparent, we apply the model-agnostic method of DeepSHAP to the deep neural network

DeepSHAP is a form of SHapley Additive exPlanations (SHAP)– estimations of Shapley values. Shapley values are rooted in coalition game theory, in which the Shapley value of each feature is the average of its contributions to every possible combination of features– or "coalitions" [9].

When we deal specifically with HDLSS datasets, exploring every possible feature combination is too computationally costly, and thus DeepSHAP is a powerful tool to estimate Shapley values in deep learning models. We are able to keep the computational power and accurate results of deep learning models, while keeping the interpretative power of Shapley values.

For SHAP estimations, we can represent Shapley values as additive feature attribution methods. For SHAP, we represent the "coalition vector" z' through the explanation model g as:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

Here, we take the sum from feature $j = 1$ to the maximum coalition size, M, of the feature attribution for j, ϕ_j, multiplied by the coalition vector of j, z'_j.

The "coalition vector" $z' \in \{0, 1\}^M$. Each feature represented in the coalition vector can either be "present" or "absent" in the combination of features, represented as either 1 or 0 [9].

4.4.2 Label-Consolidated DeepSHAP

As explained in the data section, this paper makes use of a HDLSS dataset. This informs not only the design of the neural network, but also of the interpretability experience. Even in the usage of standard DeepSHAP, a user may very well be overwhelmed by the possibility of over 400 features. Additionally, the feature space of FNC is exponentially greater than that of SBM given that the very nature of FNC is to represent the pairwise combination of the 28 cross-sectional fMRI maps.

Due to these conditions, we reverse the labels of the FNC features to correspond again to the FNC maps. For example, we would convert the feature label of "FNC_322" to a label of "FNC_12->29", representing the feature that shows the correlation between cross-sectional fMRI maps 12 and 29.

In addition, we then explore consolidating the FNC labels for each of the fMRI maps, as well as the corresponding Shapley values. For example, all features that involve cross-sectional fMRI map 12, will be consolidated into a feature of "FNCMerged_12".

For simplicity, we choose to take the mean of the feature values for the merged value, however, we take the sum of Shapley values to preserve the combined influence. We considered taking the sum of

the absolute values of the Shapley values, yet decided that the influence depicted in the force plots is more adequately represented as the sum of signed influences.

4.4.3 Patient-Centered DeepSHAP

Any discussion of interpretability is void without considering the perceptions of those who will have to interpret and make use of the technology in the real world. We began initial needfinding, and conducted a user interview with a Stanford University psychiatrist. The goal of this user interview was to perform some amount of needfinding.

There are several key findings that informed a third iteration of our interpretable model, this one centering more on melding the deep learning model combined with DeepSHAP with user experience and human-computer interaction.

We found verification in the need for a correct diagnosis, and subsequently the correct treatment and medication. However, we were surprised to discover that interpretability for the patient seemed almost more important than interpretability for the physician. Often, initial psychotic episodes and symptoms may be indicative of a severe psychosis and schizophrenia to come. Prescribing anti-psychotics early can drastically improve the prognosis for patients, slowing and even perhaps preventing full-blown psychosis and schizophrenia. However, we find that a key struggle was in convincing patients to take a medication with severe side-effects for a disease that may or may not develop. Providing patients with transparent information will help inform their own choices to engage with treatment.

There are several key methods found in our user interview for improving transparency for patients that would mimic key techniques already employed by psychiatrists. Firstly, we focus on providing additional context through comparison against other patient predictions (while ensuring privacy). Secondly, we develop routine methods for explanations, as if a psychiatrist was explaining the very prediction model. Lastly, we provide methods of further exploration and detail to inform decision making.

5 Results

5.1 DNN Results

In order to make an interpretable model, we wanted to ensure that we had a successful model to build upon. On our final 4 dense layered network, we were able to achieve a validation accuracy of 0.8235. On the test data, our model receives an AUC of 0.81282.

5.2 DeepSHAP Results

5.2.1 Standard DeepSHAP Results

We manage to produce various interpretable results through the Standard DeepSHAP packages. Firstly, as seen in Figure 2, we produce a graph of stacked SHAP force plots. On the x-axis are individual data instances, or individual patients. The y-axis shows the model's prediction output. Each slice of blue depicts an individual feature's corresponding Shapley value. Blue marks features that decreased the model's prediction, while red features increased the model's prediction. Where the two meet is the final prediction for that user. In Figure 2, the users are clustered by similarity.

We can sort by model output, as is shown in Figure 3. In this example, we show that in interactive versions of this tool, hovering on a specific user depicts the feature names and their corresponding values which contributed to the model's prediction. In bold, we see we are evaluating the index 15 which has predicted value of 0.9549.

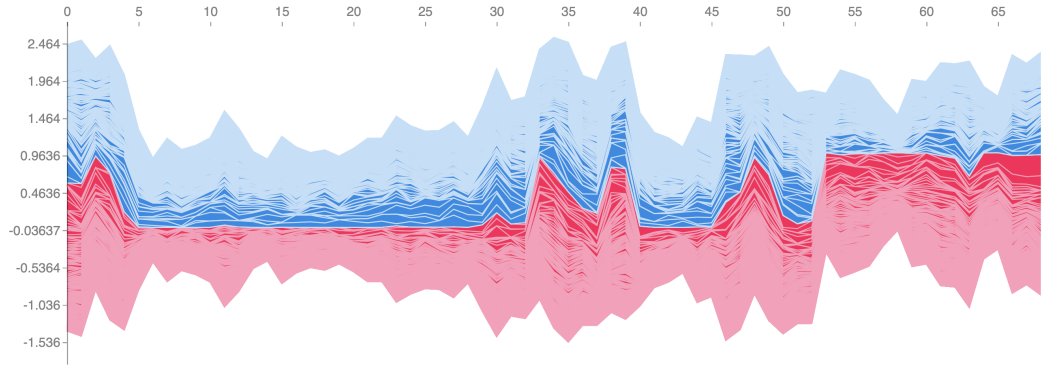

Figure 2: Stacked Standard DeepSHAP force plots clustered by similarity.

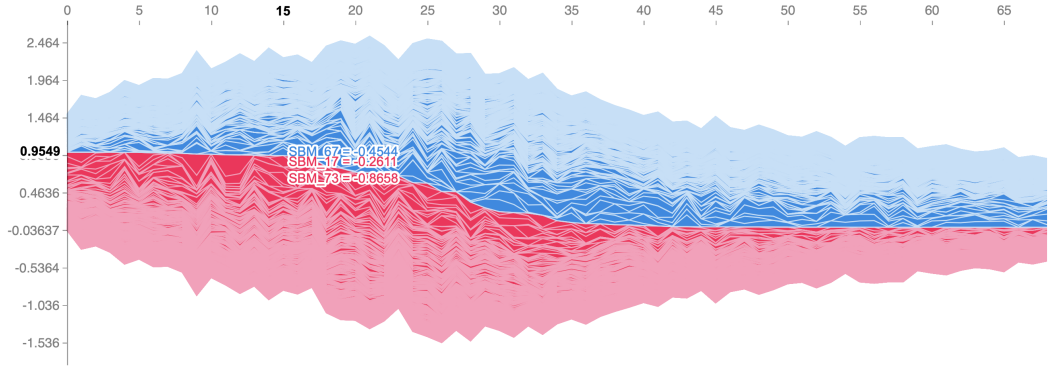

Figure 3: Stacked Standard DeepSHAP force plots by descending output value.

We also produce force plots for individual instances and users. In Figure 4, we see one such example. The predicted output is 0.93, as compared to the base value of 0.4636. Like in the stacked graphs, we see the different red and blue feature forces that inform this prediction. These examples are greatly dominated by SBM features, with only one derived from FNC.

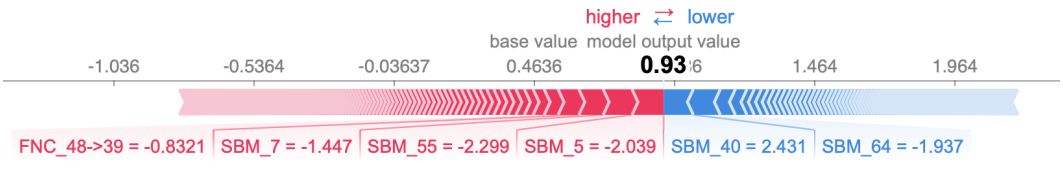

Figure 4: An example of an individual Standard DeepSHAP force plot.

In addition, we manage to find an even closer analysis of the individual features. The scatter plot in Figure 5 helps show correlation between feature value and SHAP value.

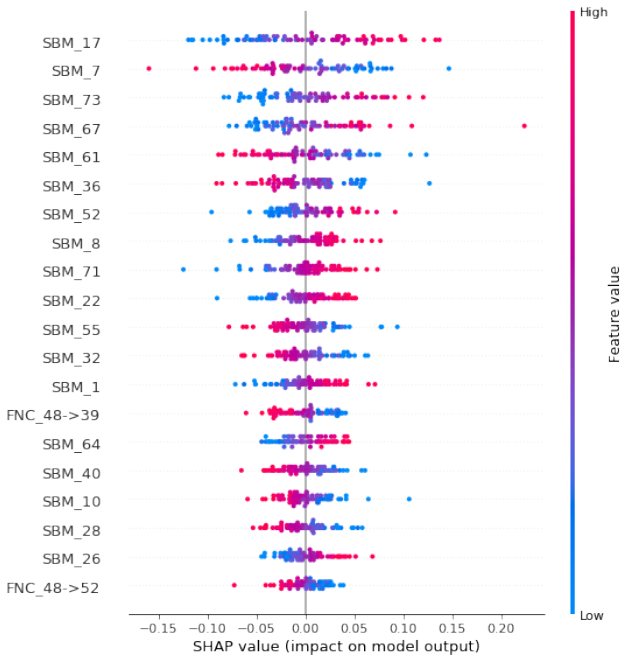

Figure 5: Scatter plot showing correlation between feature value and Standard DeepSHAP values.

5.2.2 Label-Consolidated DeepSHAP Results

The most critical distinction discovered through our Label-Consolidated DeepSHAP is the effect on the highest ranking features by mean(|SHAP value|). This can be understood as depicting what features held the most impact on the model's output. In Figure 6, we can see the difference in bar graph representation between the Standard and Label-Consolidated DeepSHAP models.

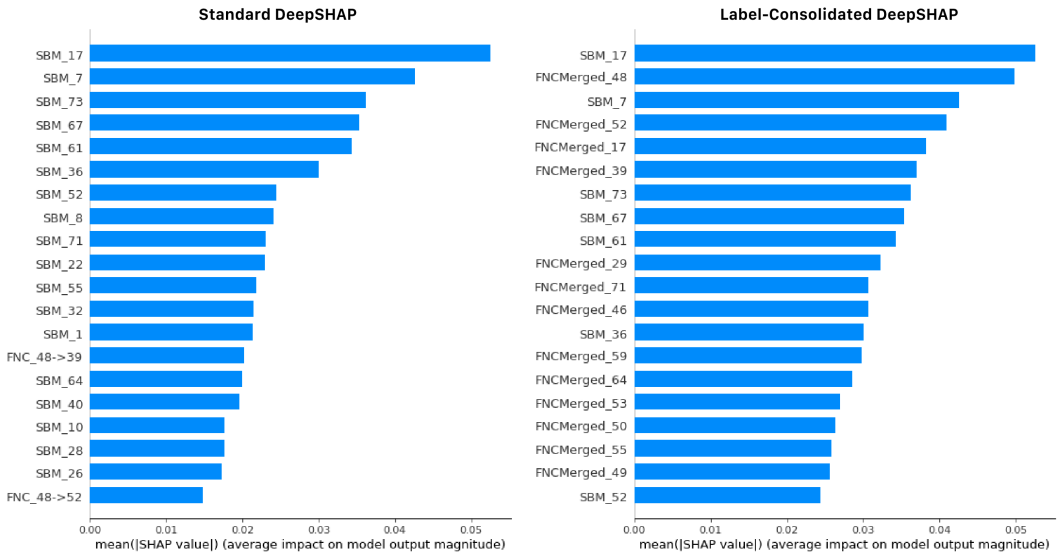

Figure 6: Bar graph comparison between Standard and Label-Consolidated DeepSHAP feature importance, ranked by mean(|SHAP value|).

We can see that consolidating FNC features by their individual map reveals that the collective impact from FNC values is actually significant to the model. We can see this same impact in the force plots for individual instances and users. In Figure 7, we provide one such example. The predicted output is

0.35, below the base value of 0.4636. FNCMerged features completely dominate both the positive and negative forces that make up the final prediction.

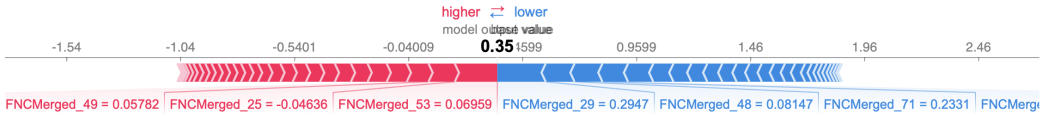

Figure 7: An example of an individual Label-Consolidated DeepSHAP force plot.

5.2.3 Patient-Centered DeepSHAP Results

The results of this section are largely exploratory and preliminary. As will be discussed, these concepts must be tested and used by actual patients, not solely psychiatrists as was done in this paper.

In line with our needfinding, we first hoped to mimic the technique already employed by psychiatrists of providing context to the patient about the realistic status of their condition. We find that making use of the stacked DeepSHAP force plots in descending output value is very interpretable to patients. This concept can be seen in Figure 8. The stacked force plot helps visually show the patient how their prediction compares to others. We avoid classic violations of Jakob Nielsen's Usability Heuristics by converting language of the Standard DeepSHAP into more digestible language. We give the user control and freedom to explore through interactive buttons for deeper exploration.

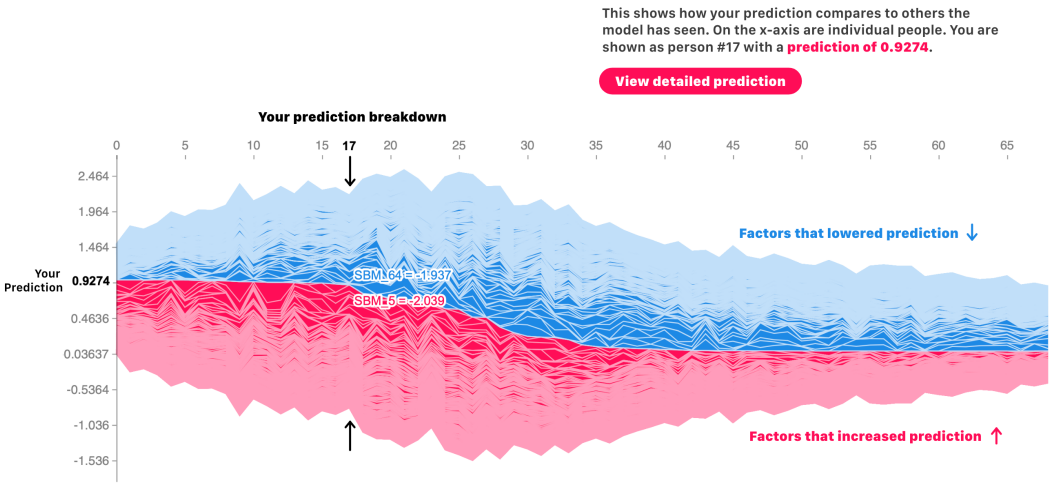

Figure 8: Patient-Centered DeepSHAP concept with stacked force plot.

Through user exploration of the "view detailed prediction" button, we make use of the Label-Consolidated DeepSHAP force plot. This concept can be seen in Figure 9. We believe the Label-Consolidated view adds a level of abstraction that is helpful in reducing complexity. Again, we aim to reduce possibility of errors through helpful guidance, as well as provide clarity through language and rhetoric.

While we hope to reduce complexity, there remains an importance to provide the patient with the ability to get as detailed into the prediction explanation as they choose. Thus, users can look at an individual feature and learn more about its derivation. In the example shown in Figure 9, the user chooses to look at the specific feature SBM_17. A pop-up provides added explanation and also shows the specific cross-sectional map that corresponds with SBM_17. We found in our needfinding that there was a specific desire to see brain maps as a source of scientific ethos. Concepts marked in red are interactable and can provide further explanation for the user.

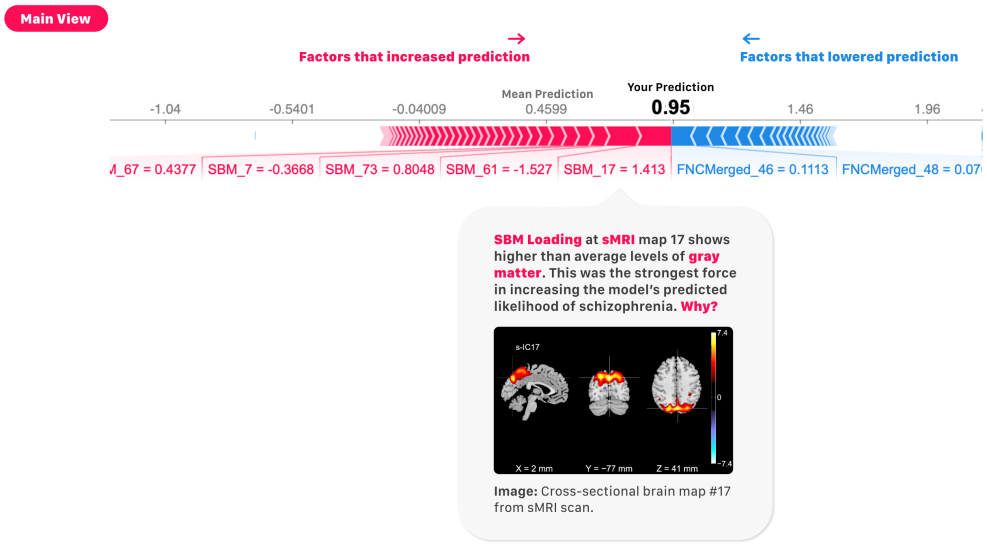

Figure 9: Patient-Centered DeepSHAP concept with individual force plot, with detail pop-up on specific feature SBM_17.

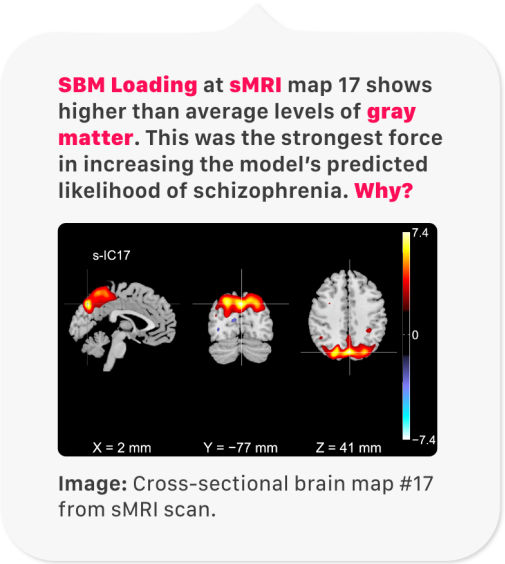

Figure 10: Close view of detail pop-up on specific feature SBM_17.

6 Discussion

To begin, we believe it is important to discuss the distinctive results across Standard and Label-Consolidated DeepSHAP. It is clear that Standard DeepSHAP shows SBM values as dominating the output prediction. At first glance, the interpretation of the Standard DeepSHAP would be that SBM loadings are more important than FNC to the classification of schizophrenia. However, our Label-Consolidated DeepSHAP shows this is incorrect. The very fact that there are exponentially more FNC values than there are SBM loadings dilutes the importance of each individual FNC value. Consolidating the FNC values shows that they are indeed critical to the classification of schizophrenia. This conclusion lines up with previous work demonstrating the connection between FNC and schizophrenia.

However, the Label-Consolidated DeepSHAP is not error free. The individual merged FNC values do not overcount, however the many merged FNC values together do. This is due to the fact that an FNC value corresponding to the correlation between maps 12 and 17 is incorporated into the merged FNC values of both 12 and 17. This is an important benefit to using Standard DeepSHAP.

We also find several unique discoveries worth discussing from the stacked force plots. We notice that the model particularly likes making binary choices, this is because it is of course trained on binary inputs. However, one can see that the times where the model does not make a binary choice, and instead falls in the middle, the width of the stacked graph is greatest. What this intuitively tells us, is that there is comparatively a greater number of both positive and negative factors to other individual instances. Because of this, the model can essentially not make up its mind, and chooses to fall somewhere in the middle.

We also note from our individual force plots that our base value is highly skewed. This is due to the data pre-processing methods employed that ensured a more even split of schizophrenia and neurotypical patients. In real world implementations, the significantly lower true base value will only emphasize the value of high predictions.

The scatter plot also unveils unique discoveries in the research of schizophrenia classification. Again in line with other research predicting a connection between gray matter and schizophrenia, we see similar evidence in Figure 5. What we find uniquely novel is that for some cross-sectional brain maps, low amounts of gray matter had high SHAP values, or increased the model's prediction. However, for other cross-sectional brain maps, high amounts of gray matter also had high SHAP values and increased the model's prediction. The literature highly suggests that low amounts of gray matter is associated with schizophrenia, however our research depicts that it perhaps is a combination of both low and high amounts of gray matter at different cross-sectional maps.

Lastly, we do see Patient-Centered DeepSHAP as an impactful and exciting concept for the future of human-centered artificial intelligence. Our needfinding revealed a unique importance in providing patients with an explainable experience to inform their decision to engage with treatment and pursue anti-psychotic medication. This topic should be the focus of further work, as only psychiatrists and physicians were consulted in needfinding.

7 Conclusion

This paper demonstrates not only the growing need for more robust technologies in the world of psychiatry, but that in these developments there remains a commitment to ensuring fairness, accountability, and transparency. We find that DeepSHAP serves as a powerful tool in making a deep neural network classifier for schizophrenia interpretable. However, it also becomes clear that additional modifications enhance the transparency of DeepSHAP. Consolidating and modifying the SBM loadings and FNC values derived from fMRI and sMRI substantially influence the results of DeepSHAP. Additionally, we see the importance of a human-centered approach to this interpretable model. Needfinding with psychiatrists and physicians revealed the real world needs of interpretability for patients and illuminate the need for further user research.

References

[1] Coulter, Chelsey, Baker, Krista K., Margolis, Russell L. Specialized Consultation for Suspected Recent-onset Schizophrenia Diagnostic Clarity and the Distorting Impact of Anxiety and Reported Auditory Hallucinations. Journal of Psychiatric Practice, 2019 DOI: 10.1097/PRA.0000000000000363

[2] Agcaoglu O, Miller R, Damaraju E, et al. Decreased hemispheric connectivity and decreased intra- and inter-hemisphere asymmetry of resting state functional network connectivity in schizophrenia. Brain Imaging and Behavior. 2018 Jun;12(3):615-630. DOI: 10.1007/s11682-017-9718-7.

[3] Xu, L., Groth, K. M., Pearlson, G., Schretlen, D. J., Calhoun, V. D. (2009). Source-based morphometry: the use of independent component analysis to identify gray matter differences with application to schizophrenia. Human brain mapping, 30(3), 711–724. https://doi.org/10.1002/hbm.20540

[4] Solin, A., Sarkka, S. (n.d.). Mlsp 2014 Schizophrenia Classification Challenge: Winning Model Documentation . Aalto University.

[5] Bombade, B., Hanwate, A., Jadhav, S. (n.d.). Diagnose Schizophrenia through Ginni-Index along With Random Forest Classification on Multi-Modal Brain Magnetic Resonance Imaging.

[6] Karolis Koncevicius, K. K. (n.d.). MLSP 2014 Schizophrenia Classification Challenge: 3rd Position. Retrieved from https://github.com/KKPMW/Kaggle-MLSP-Schizo-3rd

[7] Jodha, K. (n.d.). Retrieved from http://homepages.cae.wisc.edu/ ece539/project/f17/Jodha$_r pt.pdf$

[8] Shim, M., Hwang, H.-J., Kim, D.-W., Lee, S.-H., Im, C.-H. (2016). Machine-learning-based diagnosis of schizophrenia using combined sensor-level and source-level EEG features. Schizophrenia Research, 176(2-3), 314–319. doi: 10.1016/j.schres.2016.05.007

[9] Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.