The Application of XAI with Integrated Gradients on Schizophrenia Detection from EEG Signals

[redacted]

Start Date: 2nd September 2024
1st Proposal submission

Dataset: https://repod.icm.edu.pl/dataset.xhtml?persistentId=doi:10.18150/repod.0107441

1 Background, Domain and Scope

Schizophrenia is a mental health disorder that is characterised by delusions, hallucinations and impaired cognitive ability(Patel, Cherian, Gohil, & Atkinson, 2014). Explainable AI(XAI) will be applied to Schizonet(Grover, Chharia, Upadhyay, & Longo, 2023) which aims to detect schizophrenia in EEG signals. A large reason why detecting schizophrenia using AI is important is due to the multitude of benefits from shortening the duration of untreated psychosis such as avoiding social consequences, improving recovery and reducing harm to the schizophrenic person or those around them. (Marshall et al., 2014)

The use of XAI, in this case Integrated gradients(IG),which is a technique that explains the relationship between the predictions in terms of features and allows better insights into how a model reaches a prediction(Schwegler, Müller, & Reiterer, 2023)., is very necessary in the medical sector as clinicians need to trust black box models such as Schizonet.(Tjoa & Guan, 2021).Schizonet achieved 99.84% accuracy in predicting schizophrenia, a very high accuracy and thus a motivating factor in applying XAI to the model.

2 Problem Description

Models such as the dense neural network used in Schizonet are black box models. Black box models are machine learning models *" which are very hard to explain and to be understood by experts in practical applications "* (Loyola-González, 2019).The issue with this in a medical and diagnostic sense such as diagnosing schizophrenia is that these models may not be trusted by psychiatrists especially if they cannot be understood. EEG data also is very subject to noise(Hassani & Karami, 2015) Deep learning interpretability such as integrated gradients however can help with identifying this noise along with other important features that could inform psychiatrists and experts as to how the schizophrenia in this case is being "diagnosed" (Cui, Yuan, Wang, Li, & Jiang, 2023).

In this project, the assumptions made are that the EEG data has been correctly labelled and taken from schizophrenic/non-schizophrenic people in the same manner and that explainable AI Interpretability techniques such as Layer wise relevance propagation(LRP) have shown promising results in interpretability for MRI results with convolutional neural networks thus showing this should have applications to EEG data with dense neural networks and similar XAI techniques using gradient-based saliency maps such as Integrated gradients. The limitations of this research are the lack of a large dataset to test XAI on in a neural network context as only 14 people's EEG data is available in the original dataset. The delimitations of this research are that it will only focus on the XAI technique integrated gradients and will only consider schizophrenia as portrayed in EEG data. A completely agnostic XAI method will not be considered for example.

3 Literature Review

3.1 Approaches to solve the problem

There are numerous state of the art approaches to interpretability, these include feature importance based approaches (Rashed-Al-Mahfuz et al., 2021)and methods such as Layer-wise relevance propagation(LRP) and Integrated gradients which allow for individual pixels and time series regions to be examined for interpretability which can be more useful in terms of understanding predictions for both clinicians and researchers in EEG and MRI data.

Approaches in the evaluation of XAI in these cases generally are split into human centred evaluations and more quantitative methods using data perturbation for example.

Schoonderwoerd et al. argues for the use of human centred design approaches to XAI and for the fact that explanations need to be specifically derived from clinical support in the case of clinical decision support systems (Schoonderwoerd, Jorritsma, Neerincx, & van den Bosch, 2021).Muddamsetty et al. looks at an approach to evaluate how expert evaluations align with the XAI evaluations using expert annotations and input (Muddamsetty, Jahromi, & Moeslund, 2021).

Although these are valid methods and allow for the furthering of explainable AI techniques in the sensitive domain of medicine, in this project, they are not as viable as quantitative methods. The quantitative methods of evaluating the usefulness of XAI are described below.

Three papers that use and inform on strict quantitative statistical tests that are promising are (Mayor

Torres, Medina-DeVilliers, Clarkson, Lerner, & Riccardi, 2023), (Sánchez-Hernández, Torres-Ramos, Román-Godínez, & Salido-Ruiz, 2024) and (Hooker, Erhan, Kindermans, & Kim, 2019), although some require a large amount of domain knowledge. Many such as (Jin, Li, Fatehi, & Hamarneh, 2023) rely on measures such as professional feedback through surveys and conclude that their novel quantitative method is no match for experts. Other methods such as (Partamian et al., 2021) do not look at statistical tests but rather model performance.

3.2 Gaps in Research

The literature of EEG data with neural networks leans towards convolutional neural networks, perhaps due to their proficiency with spatial and temporal data (Han, Li, & Zhu, 2019). These have proven to be successful with EEG data and XAI techniques such as in the paper (Böhle, Eitel, Weygandt, & Ritter, 2019) .

One thing that could cause issues with this project is that papers showing the usefulness of dense neural networks regarding EEG data is quite sparse. There are some examples with dense neural networks, most notably in this case, Schizonet (Grover et al., 2023). There are also some examples with dense convolutional networks such as (Jana, Bhattacharyya, & Das, 2019).

These examples signify that the use of dense neural networks could be viable and XAI formed around these could be used to add more interpretability and improve the credibility of these. XAI methods throughout the literature focus mainly on post-hoc methods and novel methods such as (Hussain et al., 2023) and (Mayor Torres et al., 2023). These are seen to provide good results.

There are examples of Interpretability techniques also - most notably LRP (Böhle et al., 2019) and deepLift (Apicella, Isgrò, Pollastro, & Prevete, 2024) - that show promising signs and allow for further exploration. Although, integrated gradients seem less widely used, the fact that LRP, a technique that would also highlight important pixels/regions has had success, allows for the use of integrated gradients as a technique to be viable. This study in particular (Cui et al., 2023), justifies integrated gradients along with deepLift and LRP as XAI methods.

Another gap is that the quantitative methods for evaluating XAI are not standard. As stated in the approaches section, there are numerous ways that XAI can be evaluated, such as using data perturbation. Usually, this involves the re-training of neural networks which is computationally expensive although some studies such as have looked into the use of data perturbation without re-training (Rong, Leemann, Borisov, Kasneci, & Kasneci, 2022). There is simply no unified way of evaluating XAI – quantitatively or quantitatively, although clinical guidelines such as in (Jin et al., 2023) show there is a growing interest in more standard XAI used for health.

4 Research Question

"To what extent can the interpretability, measured by the correlation of average overlap of connectivity indices and accuracy in saliency maps for schizophrenia detection using dense neural networks with Electroencephalogram(EEG) data be enhanced by the implementation of Integrated Gradients? "

5 Hypothesis

H_0 :The correlation between overlap in gradient based saliency maps with integrated gradients and accuracy on connectivity measures using integrated gradients is not strongly negative, and therefore
$\rho \geq -0.8$

H_a :If the connectivity measures Phase Locking Value (PLV), Partial Directed Coherence (PDC), Directed Transfer Function (DTF), Phase Lag Index (PLI), Synchronisation Likelihood (SL), and Pearson Correlation (COR) are trained 100 times accounting for schizophrenic and non-schizophrenic patients, then the correlation of accuracy vs. average overlap of each connectivity measure will be very negative
$\rho < -0.8$

In this scenario, each model instance of the connectivity measures – PLV, PDC etc will be trained 100 times for schizophrenic and non-schizophrenic with integrated gradients. Each time they are trained a gradient based saliency map with integrated gradient outlines showing feature importance will be given. The average pixel overlap from each map will be calculated and the accuracy of each test will be recorded

also. These will be recorded as a scatter plot and the Pearson correlation will be used to define the exact correlation. The alternative hypothesis in this case assumes that the correlation will be less than -0.8 as indicated as strongly negative by the national library of medicine (*Finding and Using Health Statistics — nlm.nih.gov*, n.d.)

6 Research objectives and experimental activities

The study will start with the Schizonet model (Grover et al., 2023). The different connectivity measure models are already trained. There is architecture for each. The project will be split into three main parts for each connectivity measure. First the explainable AI technique will be used on the pre-trained architecture on the test data for schizophrenic/non-schizophrenic, then the explainable images generated for both will be tested for their amount of overlap using Pearson Correlation/Structural Similarity Index Measure(SSIM) and the accuracy of the model ran will be noted. Then, the average overlap and accuracy will be plotted.

The first step will be to load in the pre-trained architecture such as PLV net from the Schizonet code and the data. A Keras implementation of the Integrated gradients algorithm as introduced in the paper(Sundararajan, Taly, & Yan, 2017)will be implemented from the Keras documentation(Nain, 2020) to be used on the EEG data. The packages iNNvestigate introduced in (Alber et al., 2019) and keras-vis were considered at first also, however these were decided against due to iNNvestigate not inherently supporting softmax layers which are a large part of the neural networks in use in the original Schizonet project.This was found out during some preliminary testing. Keras-vis could be a good option also, however it doesn't have built-in integrated gradients support. The closest is smooth grad which may still be used to verify some integrated gradients results but won't be the main way of explaining the models.

Then the explainable images of each will need to be checked for the amount of overlap. This will be done with Pearson correlation and SSIM.Pearson correlation for image comparison captures linear relationships between images and SSIM which also is used for image similarity, may be a better indicator of human perception of image similarity due to how it correlates with observers cognitions.(Maruyama, 2023) Pearson correlation could be considered better than SSIM per the study. (Starovoitov, E.E., & K.T., 2020). However as the images will be inspected visually, the human perception aspect of SSIM may be useful and has been used in XAI saliency map similarity.(Sixt, Granz, & Landgraf, 2019) . Pearson correlation will be done using the numPy package and SSIM with the skimage package from scikit-image.The accuracy and the scores from these two will be used.

Finally, after 100 runs of the above, the points will be plotted on a scatter plot. This will firstly give a visual on whether it is correlated, negatively or otherwise although a more thorough test will have to be completed as will be described in the next section.

The dataset used contains raw EEG data however in this case, the connectivity measures will only be used and are available as NumPy files with them being matrices that are representative of schizophrenia or not, from the Schizonet study. The architecture built from these will be used.

6.1 Evaluation of designed solution

The evaluation of this project will be done on the average overlap and accuracies given on the scatter plot. The Pearson correlation will be implemented using the scipy.stats module and then the correlation coefficient given will be compared against the -0.8 as referenced in the National Library of Medicine (*Finding and Using Health Statistics — nlm.nih.gov*, n.d.). This means then that if the coefficient is greater than or equal to -0.8, the correlation is not strongly negative enough to reject the null hypothesis. Otherwise, the null hypothesis can be rejected.

The logic behind this is that the less overlap between schizophrenia and non-schizophrenia will result in more accuracy as they should be different enough to ensure they can be differentiated by Schizonet with 99.84% accuracy. If it is strongly negatively correlated in this way, the null hypothesis can be rejected.

The saliency maps will also be visually inspected to see whether there are particular aspects of the signals that are being focused on. This can help define if it is noise that is causing the high prediction for example.

References

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., ... Kindermans, P.-J. (2019). innvestigate neural networks! *Journal of Machine Learning Research*, *20*(93), 1–8. Retrieved from http://jmlr.org/papers/v20/18-540.html

Apicella, A., Isgrò, F., Pollastro, A., & Prevete, R. (2024). *Toward the application of xai methods in eeg-based systems.*

Böhle, M., Eitel, F., Weygandt, M., & Ritter, K. (2019). Layer-wise relevance propagation for explaining deep neural network decisions in mri-based alzheimer's disease classification. *Frontiers in Aging Neuroscience*, *11*. Retrieved from https://www.frontiersin.org/articles/10.3389/fnagi.2019.00194 doi: 10.3389/fnagi.2019.00194

Cui, J., Yuan, L., Wang, Z., Li, R., & Jiang, T. (2023, aug). Towards best practice of interpreting deep learning models for eeg-based brain computer interfaces. *Frontiers in Computational Neuroscience*, *17*. Retrieved from http://dx.doi.org/10.3389/fncom.2023.1232925 doi: 10.3389/fncom.2023.1232925

Finding and using health statistics — nlm.nih.gov. (n.d.). https://www.nlm.nih.gov/oet/ed/stats02-300.html#:~:text=For%20example%2C%20if%20r%3D%20%2D,than%20%2D0.7%20are%20considered%20strong. (Accessed: 2024-05-12)

Grover, N., Chharia, A., Upadhyay, R., & Longo, L. (2023). Schizo-net: A novel schizophrenia diagnosis framework using late fusion multimodal deep learning on electroencephalogram-based brain connectivity indices. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *31*, 464-473. doi: 10.1109/TNSRE.2023.3237375

Han, H., Li, Y., & Zhu, X. (2019). Convolutional neural network learning for generic data classification. *Information Sciences*, *477*, 448-465. Retrieved from https://www.sciencedirect.com/science/article/pii/S0020025518308703 doi: https://doi.org/10.1016/j.ins.2018.10.053

Hassani, M., & Karami, M. R. (2015, July). Noise estimation in electroencephalogram signal by using volterra series coefficients. *J. Med. Signals Sens.*, *5*(3), 192–200.

Hooker, S., Erhan, D., Kindermans, P.-J., & Kim, B. (2019). *A benchmark for interpretability methods in deep neural networks.*

Hussain, I., Jany, R., Boyer, R., Azad, A., Alyami, S. A., Park, S. J., ... Hossain, M. A. (2023). An explainable eeg-based human activity recognition model using machine-learning approach and lime. *Sensors*, *23*(17). Retrieved from https://www.mdpi.com/1424-8220/23/17/7452 doi: 10.3390/s23177452

Jana, R., Bhattacharyya, S., & Das, S. (2019). Epileptic seizure prediction from eeg signals using densenet. In *2019 ieee symposium series on computational intelligence (ssci)* (p. 604-609). doi: 10.1109/SSCI44817.2019.9003059

Jin, W., Li, X., Fatehi, M., & Hamarneh, G. (2023). Guidelines and evaluation of clinical explainable ai in medical image analysis. *Medical Image Analysis*, *84*, 102684. Retrieved from https://www.sciencedirect.com/science/article/pii/S1361841522003127 doi: https://doi.org/10.1016/j.media.2022.102684

Loyola-González, O. (2019, 10). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, *7*, 154096-154113. doi: 10.1109/ACCESS.2019.2949286

Marshall, M., Husain, N., Bork, N., Chaudhry, I. B., Lester, H., Everard, L., ... Birchwood, M. (2014). Impact of early intervention services on duration of untreated psychosis: Data from the national eden prospective cohort study. *Schizophrenia Research*, *159*(1), 1-6. Retrieved from https://www.sciencedirect.com/science/article/pii/S0920996414003624 doi: https://doi.org/10.1016/j.schres.2014.07.005

Maruyama, S. (2023). Properties of the ssim metric in medical image assessment: correspondence between measurements and the spatial frequency spectrum. *Physical and Engineering Sciences in Medicine*, *46*(3), 1131–1141. Retrieved from https://doi.org/10.1007/s13246-023-01280-1 doi: 10.1007/s13246-023-01280-1

Mayor Torres, J. M., Medina-DeVilliers, S., Clarkson, T., Lerner, M. D., & Riccardi, G. (2023). Evaluation of interpretability for deep learning algorithms in eeg emotion recognition: A case study in autism. *Artificial Intelligence in Medicine*, *143*, 102545. Retrieved from https://www.sciencedirect.com/science/article/pii/S0933365723000593 doi: https://doi.org/10.1016/j.artmed.2023.102545

Muddamsetty, S. M., Jahromi, M. N. S., & Moes-

lund, T. B. (2021). Expert level evaluations for explainable ai (xai) methods in the medical domain. In A. Del Bimbo et al. (Eds.), *Pattern recognition. icpr international workshops and challenges* (pp. 35–46). Cham: Springer International Publishing.

Nain, A. K. (2020). *Keras documentation: Model interpretability with Integrated Gradients — keras.io.* `https://keras.io/examples/vision/integrated_gradients/`. ([Accessed 12-08-2024])

Partamian, H., Khnaisser, F., Mansour, M., Mahmoud, R., Hajj, H., & Karameh, F. (2021, 09). A deep model for eeg seizure detection with explainable ai using connectivity features. In (Vol. 8). doi: 10.5121/ijbes.2021.8401

Patel, K. R., Cherian, J., Gohil, K., & Atkinson, D. (2014, September). Schizophrenia: Overview and treatment options. *P T*, *39*(9), 638–645.

Rashed-Al-Mahfuz, M., Moni, M. A., Uddin, S., Alyami, S. A., Summers, M. A., & Eapen, V. (2021). A deep convolutional neural network method to detect seizures and characteristic frequencies using epileptic electroencephalogram (eeg) data. *IEEE Journal of Translational Engineering in Health and Medicine*, *9*, 1-12. doi: 10.1109/JTEHM.2021.3050925

Rong, Y., Leemann, T., Borisov, V., Kasneci, G., & Kasneci, E. (2022). *A consistent and efficient evaluation strategy for attribution methods.*

Schoonderwoerd, T. A., Jorritsma, W., Neerincx, M. A., & van den Bosch, K. (2021). Human-centered xai: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human-Computer Studies*, *154*, 102684. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1071581921001026` doi: https://doi.org/10.1016/j.ijhcs.2021.102684

Schwegler, M., Müller, C., & Reiterer, A. (2023). Integrated gradients for feature assessment in point cloud-based data sets. *Algorithms*, *16*(7). Retrieved from `https://www.mdpi.com/1999-4893/16/7/316` doi: 10.3390/a16070316

Sixt, L., Granz, M., & Landgraf, T. (2019, December). *When explanations lie: Why many modified bp attributions fail.* (preprint)

Starovoitov, V., E.E., E., & K.T., I. (2020, 01). Comparative analysis of the ssim index and the pearson coefficient as a criterion for image similarity. *Eurasian Journal of Mathematical and Computer Applications*, *8*, 76-90. doi: 10.32523/2306-6172-2020-8-1-76-90

Sundararajan, M., Taly, A., & Yan, Q. (2017). *Axiomatic attribution for deep networks.* Retrieved from `https://arxiv.org/abs/1703.01365`

Sánchez-Hernández, S. E., Torres-Ramos, S., Román-Godínez, I., & Salido-Ruiz, R. A. (2024). Evaluation of the relation between ictal eeg features and xai explanations. *Brain Sciences*, *14*(4). Retrieved from `https://www.mdpi.com/2076-3425/14/4/306` doi: 10.3390/brainsci14040306

Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(11), 4793-4813. doi: 10.1109/TNNLS.2020.3027314

7 Activities

- **Week 1: Model Verification and Code Adjustment**
  - Ensure the dense neural network model is working as well as in the previous experiment(Grover et al., 2023).
  - Modify the codebase to facilitate the integration of XAI methods, enhancing adaptability and ease of future modifications if necessary.
- **Week 2-6: Implementation of XAI Methods**
  - Implement interpretability tools using keras on the connectivity models.
  - Consult existing literature and utilise implementations to guide and validate the implementation process.
- **Week 6-8: Evaluation Using Pearson and SSIM**
  - Use Pearson/SSIM to calculate the average overlaps and note the accuracy also.
  - Plot these results to be analysed.
- **Week 8-9: Results Interpretation**
  - Interpret the results using the Pearson coefficient to understand any correlations.
  - Visually inspect saliency maps to identify any indications of noise or specific regions critical for predictions.
- **Week 9-13: Dissertation Writing**

– Compile research findings and analyse the impact of implemented XAI methods on model interpretability and write the dissertation.

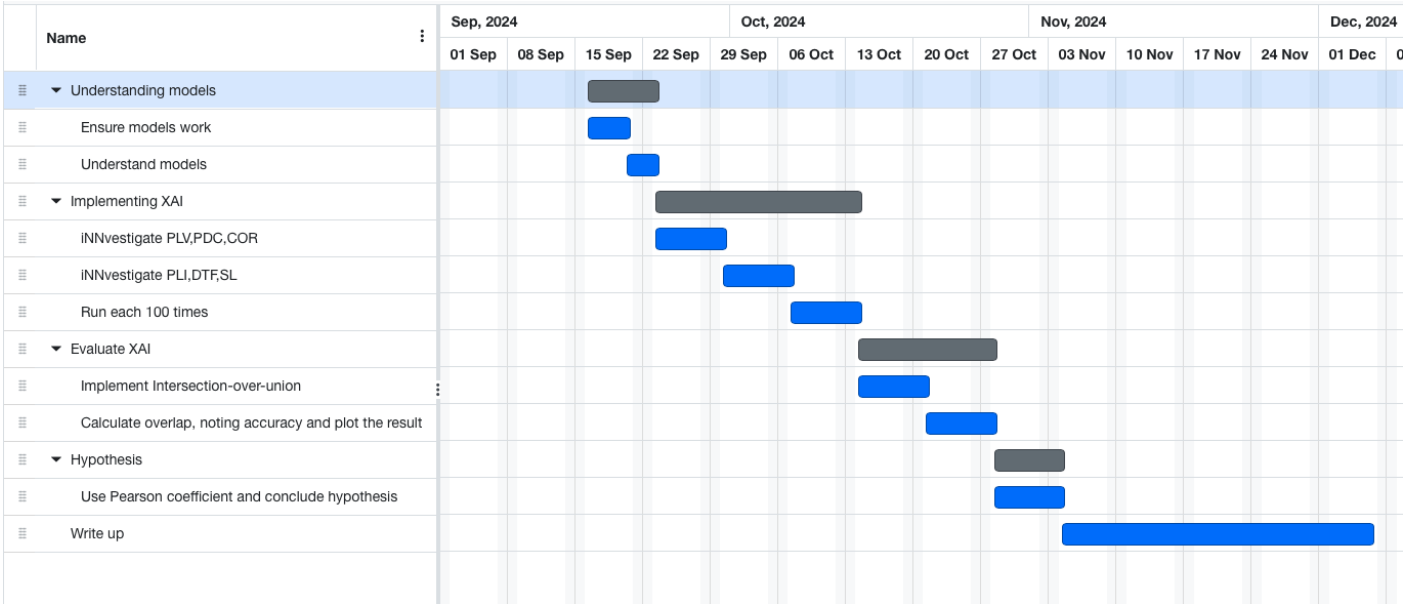

Figure 1: *Gantt Chart for Activities*