Explainable AI applications in Neurologic Disorders classification and prediction models

Interim Report

Georgia Kopalidi

MSc Data Science with Specialisation in Artificial Intelligence

Newcastle University

1. Introduction

In the last years there has been an increase in the development of Artificial Intelligence (AI) models for a wide range of applications. The extensive research and improvement of computational tools have helped the rapid evolvement and expansion of AI. Intelligent models are used to make predictions for friend suggestions in social media, new articles that one might one be interested in, automatically reporting spam incoming emails, among many other examples. However, in most cases there is not a clear explanation why models take certain decisions. This is crucial in AI applications where errors can be critical and can risk someone's life (Bennetot *et al.*, 2021).

The medical field can be benefited greatly by the usage of AI tools and models. Clinicians have to constantly analyse a vast amount of information in order to solve problems in relation to disease understanding and diagnosis, patients' recovery process, etc. AI can assist clinicians in taking faster and more accurate decisions and predict their outcomes (Ramesh *et al.*, 2004). However, there is a significant lack of confidence in AI models, especially when their decisions can affect the outcome of an individual's life when applied in medical scenarios. This is because most models are characterised as "black-boxes". No matter the accuracy of a model, it is important to be able to understand and explain in a great detail how a model learned to classify or predict certain features while being trained, and why it is producing certain outcomes. When building a model, there is information on the classifier's architectures and neural network's structure, how hyperparameters work, which loss functions fit best for the particular problem, but the way a model chooses to output certain results it is not transparent (Xu *et al.*, 2019). Actually, it is known that a model trained again using the same data, can produce different results (Leventi-Peetz and Östreich, 2022). This raises concerns about the reliability of the model.

Explainability includes the reasons the model provides in order to give justification for its functioning (Barredo Arrieta *et al.*, 2020). In literature, explainability tends to get confused with the meaning of interpretability. Interpretable AI focuses on designing new models that are interpretable from the start, whereas explainable AI focuses on giving post hoc explanations on models that are unintelligible to humans (Marcinkevičs and Vogt, 2020).

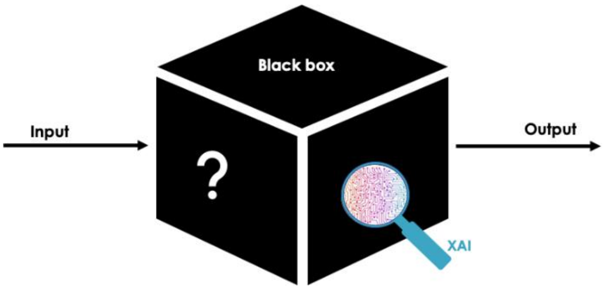


Figure 1: Illustrated example of a black-box analysed by an XAI tool (Griva, 2022)

This research focuses on explainability within the medical field and in particular in Neurological Disorders. Neurological Disorders refer to the illnesses of the central and peripheral nervous system (Lima *et al.*, 2022). The literature regarding these disorders and explainability in AI is still at its novelty and not organised. Neural networks are often chosen to build classification or prediction models, as they are more effective with this type of data. Yet, most techniques applied so far are not transparent on how they reach their outcomes (Fellous *et al.*, 2019).

For epileptic seizure detection, EEG signals are recorded from patients both while having a seizure and while their brain activity is normal. Analysing these signals and their correlation can lead to applications that can predict a seizure before it happens and warn the patient. For this task, Convolutional and Recurrent Neural Networks appear to be used the most as they produce high accuracy results (Craik, He and Contreras-Vidal, 2019). Also, CNNs have been proven to be very accurate in image classification tasks. Hence, MRI data can be used in combination with CNNs to produce highly accurate classification models for Alzheimer's Disease. This type of model architecture includes many layers and parameters which can be altered depending on the seeking results and type of training data. Thus, it is a very complex process that includes a big number of layers, blocks, network weights, constantly updating connections etc (Obaid, Zeebaree and Ahmed, 2020). Explainable AI not only can provide insights on a complex model's training process and why it decided to change specific weights while learning, but it can also help scientists improve the accuracy of a model and mitigate bias (Xu *et al.*, 2019).

Complex models typically achieve better results, but the transparency and explainability becomes significantly lower. There are a couple of different approaches in developing explainable AI methods, such as model-agnostic methods that can be applied in any AI model or model-specific methods that are applied in an individual model and aim to explain its particular learning process. In healthcare, feature importance in data is important for clinicians in order to understand whether a model focused on parts of the data that are actually useful or not.

2. Aim and Objectives

The project aims to study common architectures used for Neurological Disorders classification and prediction models and apply explainable AI (XAI) methods to increase transparency on their learning process and explain their predictions. In fact, XAI tools will be analysed and reviewed based on classification and prediction models that will be built for the purpose of this project. Actually, the most common diagnostic tools for Neurological Disorders are MRI scans, CT scans and Electroencephalogram (EEG). There are hundreds of disorders that can impact the nervous system but considering the time limitation of this project, the focus will be on Alzheimer's, Stroke, Epilepsy and Parkinson's. The reason why these illnesses were chosen, is to include data type variance, as well as AI and ML model variance in order to have sufficient material for explainable AI tools to be applied and analysed. For example, Alzheimer's classification uses MRI images, Stroke Prediction includes binary classification problems depending on certain health factors, Epileptic Seizures are described as EEG signals and Parkinson's Classification includes voice data translated to frequencies.

The objectives of this project are as follows:

- Build four classification models:
    - Alzheimer's Image Classification
    - Stroke Prediction
    - Epileptic Seizure Classification
    - Parkinson's Voice Data Classification
- Use model-agnostic and model-specific XAI tools to explain the predictions of the above models.
- Study the produced explanations and report ways to identify bias, data errors or learning errors, and how it is affecting a model's prediction.
- Review the XAI tools that were used and provide useful insights regarding the efficiency of such tools depending on the scenario. This is important for the purpose of this research, as it can be used to increase reproducibility of such models, by sharing relevant information.

3. Overview of Progress

A detailed literature review has taken place, which was divided into two parts. The first one considered literature about artificial intelligence (AI) and machine learning (ML) techniques in Neuroscience and more specifically, for Epilepsy, Alzheimer's, Stroke and Parkinson's that will be explored in this project. The second part—which can be referred to as the most challenging one—considered literature in regard to Explainable AI, the theory and its applications. In general, the variation of literature in explainable AI is limited. Especially, the progress of explaining models in healthcare is still ongoing. Hence, finding relevant research and experimental material for specific cases within Neurological Disorders applications is rather difficult. Moreover, Vilone and Longo (2020) state that the first literature findings regarding explainable AI appear in the 70s but received more attention as a research topic in the last decade. (Xu *et al.,* 2019)

Up to now, the focus was on finding as much literature as possible in order to identify the most efficient tools for explainability, as well as find parts for improvement. In a short summary, SHAP (SHapley Additive exPlanations) is commonly used for feature importance in a diverse selection of models, along with DeepLIFT (Deep Learning Important FeaTures). Moreover, explainability with attention mechanism is a widely used method in order to show where the model paid more attention in regard to the received inputs. Class activation mapping (CAM) is widely used in healthcare for explainability in models used for image analysis (Yang, Ye and Xia, 2022). Another widely used method in healthcare applications is LIME, which can be used in any model to explain its predictions (Ribeiro, Singh and Guestrin, 2016). In fact, Partamian *et al.* (2021) studied the development of an explainable model that can not only classify EEG signals as part of an epileptic seizure or not, but can provide information about the importance of connectivity measures on the model's outputs. To continue, Soun *et al.* (2021) highlight the importance of early stroke identification and how AI can assist with classification and segmentation of data used for diagnosis. They also mention the limitations of machine learning and deep learning models since they are not easy to understand and state that saliency maps are incorporated to clarify such models. Although many researchers state the need of explainability in order to support models' predictions with high accuracy, XAI applications are still at their novelty. Also, Jahan *et al.* (2022) developed an explainable Alzheimer's prediction model using LIME, which can make a model's outputs comprehensible to the clinician and the patient.

The following have been understood following the research that has taken place:

- Convolutional Neural Networks and Recurrent Neural Networks (such as LSTMs) are widely used for applications in healthcare, especially considering MRI and EEG data.
- The XAI tools that are commonly used with models mentioned above, are LIME, CAM, and SHAP.

This project will continue by finalising the classification models stated above and apply explainability methods using LIME, CAM and Shapley Values to compare their results and efficiency. Furthermore, more tools will be explored for model-specific explanations using visual, rule-based and numerical techniques.

Currently, the model for Alzheimer's Classification is in development using Inception-V3 and TensorFlow. Google Colab is used as this project's IDE and a Pro subscription will be initialised when the experimentation among XAI tools begins, in order to benefit from more RAM size and dedicated GPU allocation. The language that will be used for the development is Python. Notion is heavily used as a tool to keep notes for this project in an organised manner in a Markdown format. Lastly, all documents and code are backed up regularly to OneDrive to avoid any potential loss.

4. Project Plan

The project started on the 25th of April 2022 and finishes on the 23rd of August 2022. A Gantt Chart (Figure 2: Gantt Chart Visualising the Project's Timeline) has been prepared to visualise the project's plans and deliverables in detail. The development will follow an agile approach by regular keeping documentation and making updates when necessary. The literature review that took place until the beginning of June will be organised into a chapter ready to be included in the final thesis document. The code implementation of this project will take place between the 11th of June 2022 and the 9th of August 2022. The last two weeks before the thesis submission and presentation will be used for any last-minute corrections, and final structuring of the written thesis document. Any problems or questions that shall rise through the development of the project will be addressed during the group Supervision Meetings that are regularly taking place each month, or within a private meeting with the project's Supervisor.

5. References

Bennetot, A. *et al.* (2021) 'A Practical Tutorial on Explainable AI Techniques'. Available at: http://arxiv.org/abs/2111.14260.

Barredo Arrieta, A. *et al.* (2020) 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI', *Information Fusion*, 58, pp. 82–115. doi: 10.1016/j.inffus.2019.12.012.

Craik, A., He, Y. and Contreras-Vidal, J. L. (2019) 'Deep learning for electroencephalogram (EEG) classification tasks: A review', *Journal of Neural Engineering*, 16(3). doi: 10.1088/1741-2552/ab0ab5.

Fellous, J. M. *et al.* (2019) 'Explainable Artificial Intelligence for Neuroscience: Behavioral Neurostimulation', *Frontiers in Neuroscience*, 13(December), pp. 1–14. doi: 10.3389/fnins.2019.01346.

Griva, A. (2022) 'To "Black Box" or not to "Black Box"?'. Available at: https://impact.nuigalway.ie/ai-and-creativity/to-black-box-or-not-to-black-box/ Accessed at: 03/06/2022

Jahan, S. *et al.* (2022) 'Explainable AI-based Alzheimer ' s Prediction and Management Using Multimodal Data', (March), pp. 1–16. doi: 10.20944/preprints202203.0214.v1.

Leventi-Peetz, A.-M. and Östreich, T. (2022) 'Deep Learning Reproducibility and Explainable AI (XAI)'. Available at: http://arxiv.org/abs/2202.11452.

Lima, A. A. *et al.* (2022) 'A Comprehensive Survey on the Detection, Classification, and Challenges of Neurological Disorders', *Biology*, 11(3), pp. 1–45. doi: 10.3390/biology11030469.

Marcinkevičs, R. and Vogt, J. E. (2020) 'Interpretability and Explainability: A Machine Learning Zoo Mini-tour', pp. 1–24. Available at: http://arxiv.org/abs/2012.01805.

Obaid, K., Zeebaree, S. and Ahmed, O. (2020) 'Deep Learning Models Based on Image Classification: A Review', *International Journal of Social Science and Business*, 4(October), pp. 75–81. doi: 10.5281/zenodo.4108433.

Partamian, H. *et al.* (2021) 'A Deep Model for EEG Seizure Detection with Explainable AI using Connectivity Features', *International journal of Biomedical Engineering and Science*, 8(4), pp. 1–19. doi: 10.5121/ijbes.2021.8401.

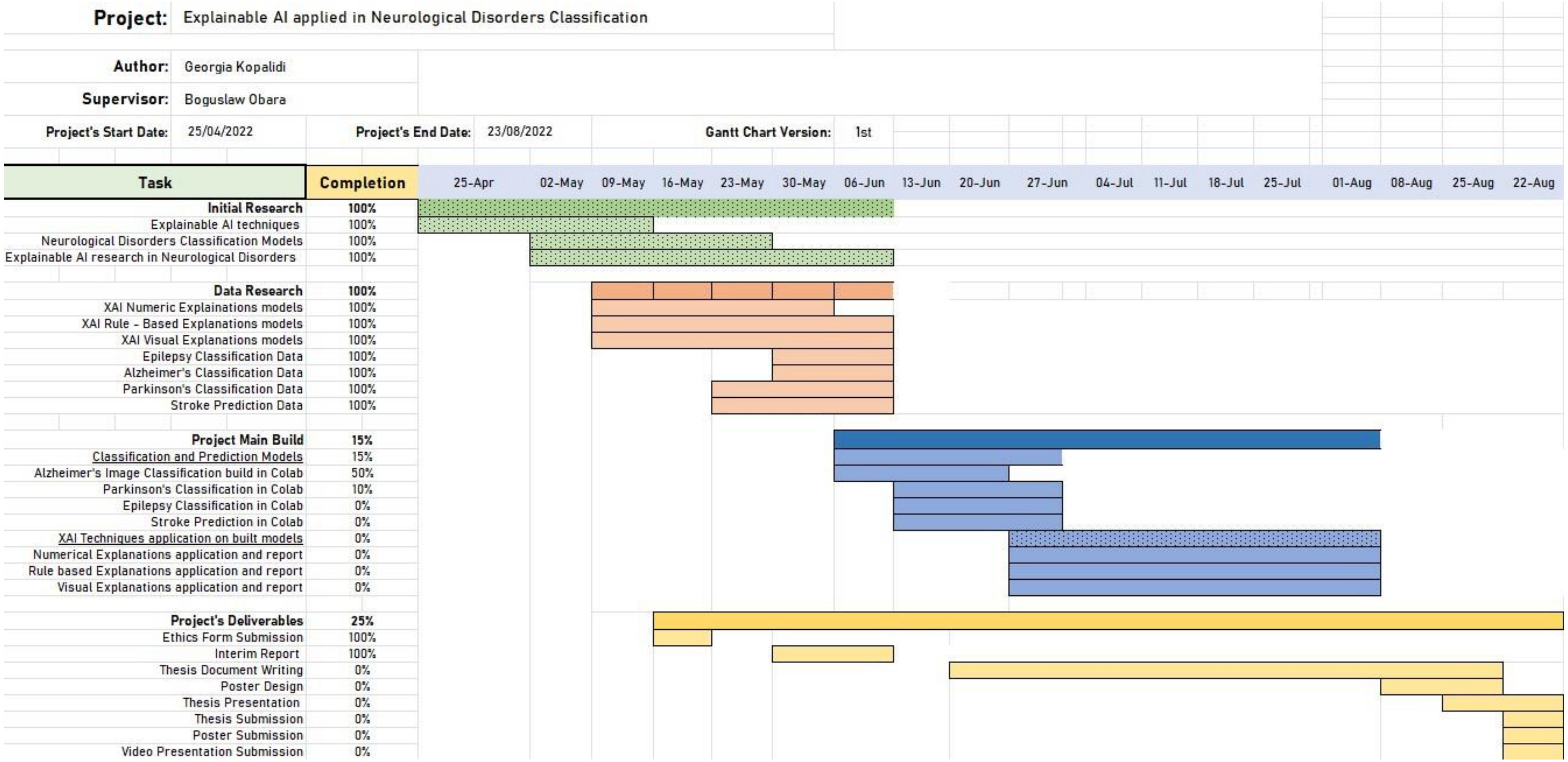
Ramesh, A. N. *et al.* (2004) 'Artificial intelligence in medicine', *Annals of the Royal College of Surgeons of England*, 86(5), pp. 334–338. doi: 10.1308/147870804290.

Soun, J. E. *et al.* (2021) 'Artificial intelligence and acute stroke imaging', *American Journal of Neuroradiology*, 42(1), pp. 2–11. doi: 10.3174/ajnr.A6883.

Vilone, G. and Longo, L. (2020) ‘Explainable Artificial Intelligence: a Systematic Review’, (May). doi: 10.48550/arXiv.2006.00093.

Xu, F. *et al.* (2019) ‘Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges’, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11839 LNAI(September), pp. 563–574. doi: 10.1007/978-3-030-32236-6_51.

Yang, G., Ye, Q. and Xia, J. (2022) ‘Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond’, *Information Fusion*, 77(May 2021), pp. 29–52. doi: 10.1016/j.inffus.2021.07.016.



Project: Explainable AI applied in Neurological Disorders Classification

Author: Georgia Kopalidi

Supervisor: Boguslaw Obara

Project's Start Date: 25/04/2022 **Project's End Date:** 23/08/2022 **Gantt Chart Version:** 1st

Task	Completion
Initial Research	**100%**
Explainable AI techniques	100%
Neurological Disorders Classification Models	100%
Explainable AI research in Neurological Disorders	100%
Data Research	**100%**
XAI Numeric Explainations models	100%
XAI Rule - Based Explanations models	100%
XAI Visual Explanations models	100%
Epilepsy Classification Data	100%
Alzheimer's Classification Data	100%
Parkinson's Classification Data	100%
Stroke Prediction Data	100%
Project Main Build	**15%**
Classification and Prediction Models	15%
Alzheimer's Image Classification build in Colab	50%
Parkinson's Classification in Colab	10%
Epilepsy Classification in Colab	0%
Stroke Prediction in Colab	0%
XAI Techniques application on built models	0%
Numerical Explanations application and report	0%
Rule based Explanations application and report	0%
Visual Explanations application and report	0%
Project's Deliverables	**25%**
Ethics Form Submission	100%
Interim Report	100%
Thesis Document Writing	0%
Poster Design	0%
Thesis Presentation	0%
Thesis Submission	0%
Poster Submission	0%
Video Presentation Submission	0%

Figure 2: Gantt Chart Visualising the Project's Timeline

0. Project title, author, version and date

Project: Explainable AI applications in Neurological Disorders models

Author: Georgia Kopalidi	*Version:* 1st	*Date:* 02/06/2022

1. Description of the data

1.1 Type of study

Researching explainable artificial intelligence techniques and how they can be used to increase the confidence and reliability of classification and prediction models for a variety of Neurological Disorders, such as Alzheimer’s, stroke, Parkinson’s, etc.

1.2 Assessment of existing data

The datasets that will be used in this project, are publicly available and anonymised. Firstly, for epileptic seizure data, this dataset will be used https://www.kaggle.com/datasets/harunshimanto/epileptic-seizure-recognition?datasetId=63117&sortBy=voteCount . It contains EEG values recorded with different time stamps, leading to the development of a time series classification model. Regarding Alzheimer’s classification, the following dataset from Kaggle will be used: https://www.kaggle.com/datasets/tourist55/alzheimers-dataset-4-class-of-images . This dataset will be used to create an image classification model in order to apply XAI techniques for this particular type of classification. Also, the following datasets for Parkinson’s Disease classification https://www.kaggle.com/datasets/vikasukani/parkinsons-disease-data-set and Stroke prediction https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset will be used.

1.2 Types of data

The dataset for Alzheimer’s Classification contains four different classes of MRI images, from non demented patients to moderate demented. The purpose of the creation of this dataset is to be able to understand the stage of Alzheimer’s of an individual’s brain with a high accuracy level. The epileptic seizure recognition dataset includes EEG time series signals. The original dataset is from the UCI Machine Learning Repository, however the dataset that will be used in this project is a simplified and pre-processed version structured in .csv files. To continue, the Parkinson’s dataset contains voice data from 31 individuals, from which 23 were diagnosed with the disease. The dataset allows for binary classification since 0 represents healthy individuals and 1 individuals with the disease. Similarly, the Stroke dataset includes 11 clinical features recorded from approximately 5000 individuals. The correlations between each feature leads to a binary classification problem, with 0 representing individuals who did not have a stroke and 1 those who did.

1.3 Format and scale of the data

- Alzheimer’s: 6400 files, 34.7 MB, split in train and test folders
- Epilepsy: 1 csv file, 11500 rows, 7.62 MB
- Parkinson’s: 1 csv file, 180 columns (time-series format), 7.62 MB
- Stroke: 1 csv file, 5110 observations, 12 attributes, 316.97 KB

2. Data collection / generation

2.1 Methodologies for data collection / generation

No new data will be collected. The datasets mentioned above, will be used to create simple classification / prediction models according to literature’s best proposed AI methods. Then, these models will become the base of the explainable AI techniques that will be applied and analysed, in order to understand which XAI methods are most efficient depending on the type of data used and created model. Any generated data will be hosted in the project’s repository and available upon request.

2.2 Data quality and standards

The data quality of the datasets used to develop the models mentioned above and generate new data will be assessed in accuracy, completeness, consistency, timeliness, validity and uniqueness. Depending on the available time within the project constraints, generated data will be cross-validated with existing research to ensure their quality.

3. Data management, documentation and curation

3.1 Managing, storing and curating data.

The above datasets have been downloaded to my local machine and also uploaded to Google Drive. Any trained models that will be created for this project will be stored in Google Drive and to a local machine.

3.2 Metadata standards and data documentation

Produced trained models architectures details and XAI techniques applications results will be included as Appendices in the final Thesis document. Necessary information in order to reproduce a certain model's accuracy will be included in the GitHub repository of this project. Any features or details necessary to comprehend an XAI algorithm used in this project will be included in the Thesis document Appendices and in the README file of the project’s GitHub repository. Lastly, information regarding training times, benchmarks, XAI processes times etc will be included in the Thesis Appendices, as well as, in the README file of the repository.

4. Data security and confidentiality of potentially disclosive information

4.1 Main risks to data security

All data used for the purpose of this project are publicly available and anonymised. There are not any risks for a potential information disclosure as the data provided can not be linked with the individuals whose measurements were recorded and stored.

5. Data sharing and access

5.1 Suitability for sharing

Yes. The data used are publicly available, hence suitable for sharing. All the techniques and tools of explainable AI that will be used on the trained models, are suitable for sharing in order to increase awareness in the importance of XAI and how it can increase confidence among professionals in the medical field aiming to use AI models for prediction and classification in Neurological Disorders.

5.2 Discovery by potential users of the research data

The project will be stored in a GitHub repository:
https://github.com/georgia-kpl/XAI_NeuroDisorders

5.3 Data preservation strategy and standards

The project will be stored in GitHub without any time restrictions. However, the datasets used for the development of the models or any other public data related to explainable AI tools used in this project, might not be available indefinitely and this depends on their owners.

5.4 Restrictions or delays to sharing, with planned actions to limit such restrictions

Any possible restrictions set in sharing the data and results produced from this project will be stated in the README file in the repository. This project's goal is to contribute in the understanding of explainability techniques and analyse how their applications in neurological disorders models can increase transparency, trust and reliability. This project's researcher / author is not responsible for any malicious use of this information, or misuse of any of the produced models / results.

6. Responsibilities and Resources

Google Colab Pro: £7,89 per month

7. Relevant institutional, departmental or study policies on data sharing and data security

Policy	URL or Reference
Data Management Policy & Procedures	https://www.ncl.ac.uk/media/wwwnclacuk/research/files/ResearchDataManagementPolicy.pdf
Information Security	*https://services.ncl.ac.uk/itservice/policies/InformationSecurityPolicy-v2_1.pdf*
Other	