Initial Plan: HIV Neuroimaging Data Analysis and Supervised Machine Learning

Supervisor: Matthias Treder

Co-supervisor: Georgina Spies

Module Number: CM3202

Module Title: One Semester Individual Project

Module Credits: 40

Robert Taylor | C1630630 | 27/01/2020

Contents

Ethics

In this project I will be using neuroimaging data of subjects with HIV provided by Stellenbosch University, Tygerberg. This neuroimaging data has been anonymised but still contains participant biometric data. Through my supervisor, I have been permitted by associates at Stellenbosch University to work with this data. For this data, I have been given this ethical consideration statement:

"The proposed study will form part of a study headed by Professor Soraya Seedat at Stellenbosch University, Tygerberg. The study has already received ethical approval by the IRB/ Committee for Human Research at the University of Stellenbosch (Ethics reference no: N07/07/153). This study will assess the brain and behavioural effects of childhood abuse and HIV among South African women. The study will utilise neuropsychiatric, neuropsychological, neuroimaging, and molecular genetic assessments."

As the data I am working with contains sensitive information, I have taken the required precautions to ensure the safety of the data. This includes forwarding the ethical consideration from Stellenbosch University to COMSC Ethics before my project begins. No other areas in this project require ethical approval.

Project Description

HIV is one of the deadliest infections in the world, with almost 40 million people in the world being affected. In 2014, an estimated 5.51 million South Africans were infected with human immunodeficiency virus (HIV), and for women aged 15–49, an estimated 18.5 % of the population was HIV-positive (Statistics_South_Africa, 2014). Although detecting HIV using blood tests is often accurate, reliable and cost-effective, understanding the full effect of HIV on the brain is not easy to determine. HIV penetrates the blood-brain barrier early in the course of infection and infects nerve cells, resulting in neuroinflammation and neuronal death. HIV-infected individuals commonly have cognitive deficits called HIV-associated neurocognitive disorders (HANDs). These deficits have been found to occur in psychomotor skills, verbal and visual memory, working memory, information processing speed, and executive functioning. HANDs among a sample of HIV-positive women has previously been investigated. The results demonstrated that the sample of HIV-infected women scored poorly on measures of speed of information processing and executive functioning. (Osowiecki, et al., 2000)

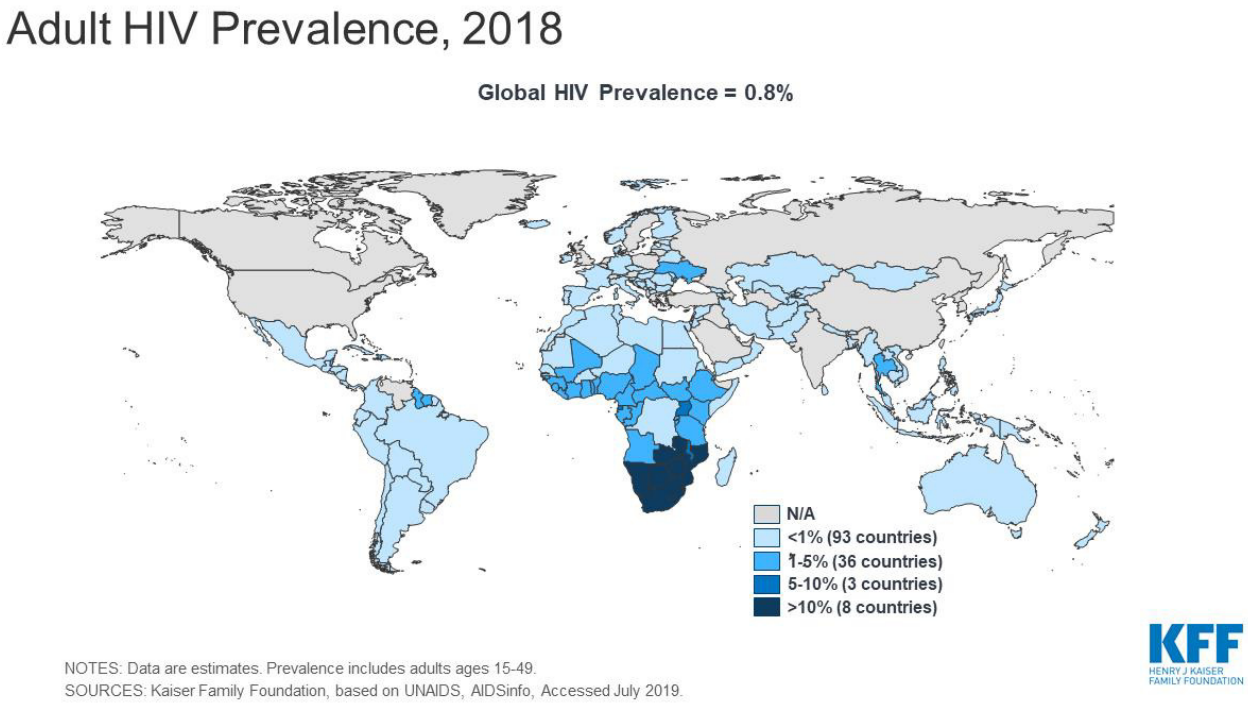

Figure 1 HIV World-map (Kaiser, 2019)

In this project, I will be analysing neuroimaging data acquired using a 3T Magnetom MRI scanner. This dataset has been acquired from Stellenbosch University, Tygerberg, who are looking for results with a strong inference between neuroimages and HIV. MRI neuroimages are the most reliable form of brain imaging as it allows us to see the size and quantity of grey-matter and white-matter within the different areas of the brain. With a more advanced HIV disease stage and greater immunocompromise there is often more neuroimaging abnormalities and neurocognitive impairments. Even with effective therapy, individuals who are HIV-infected continue to demonstrate ongoing aberrations in white and grey matter. An increase in brain white matter and subcortical grey matter abnormalities are also linked to immunodeficiency recovery among infected individuals. (Spies, 2015) This project will offer thoughtful insight towards understanding the brain tissue and cognitive variables for HIV positive women in South Africa.

The project will have a heavy focus on analysing the HIV dataset provided by Stellenbosch University, using professional analytical techniques to produce tangible results. I will assess the HIV dataset and select the most appropriate method for analysis of variance in the neuroimages. I intend to utilize the use of classification and statistical analysis to provide insight about the HIV status and neuroimaging abnormalities of the participants. The current dataset is not very comprehensive, so results from the data analysis will be transformed into detailed visualisations which will have an increase in interpretability for myself and the collaborating university.

Using machine learning techniques, I will be developing a predictive neural network model to detect the shared characteristics of HIV positive women in the neuroimages. By performing a systematic comparison of different classical machine learning models, I will be able to identify the best classification performance and apply that model. The predictive model should be able to interpret the tabular MRI data and then suggest if the individual is HIV positive and, if possible, which areas of the brain indicate the presence of HIV. I can then compare the results of my predictive model with the conclusions that had previously been discovered by Stellenbosch University. Highlighting any discoveries that the supervised machine learning model has found. Any new results found by my project can then be utilized by Stellenbosch University in their next publication on the subject.

Project Aims and Objectives

This project aims to use data analysis and machine learning techniques to identify HIV in the brain of South African women.

Aims

- Further my technical knowledge of data science regarding data analysis and predictive machine learning.
- Analyse and visualise the neuroimaging dataset to identify key areas of the brain that signify HIV infection.
- Create classical machine learning models and evaluate the model's peak classification performance to determine the best model to utilize.
- Develop the predictive model to closely predict if a participant has HIV based on neuroimaging data.
- Identify any tangible results that could contribute to the Stellenbosch University research project.
- Reflect, evaluate and document the one semester individual project.

Objectives

1. **Plan Project**:
    i. Identify the problem and research more about HIV and neuroimaging.
    ii. Learn about analytical, machine learning and deep learning practices.
    iii. Obtain the dataset and check ethical considerations.
    iv. Set objectives & milestones with approval from supervisor.
2. **Explore/Initialise Data**:
    i. Look through the sample dataset of participants checking the demographics, size and clinical characteristics.
    ii. Analyse the descriptive data provided in the dataset (percentages, averages, frequency, standard deviation).
    iii. Sanitize the dataset to only be relevant for the verdicts needed in my project.
3. **Model Data**:
    i. Perform data analysis using classification or clustering and statistical analysis techniques (T-test, ANOVA, REML) to find correlations, associations and variance between the neuroimaging data.
    ii. Create data visualisations from the results of the data analysis using applications such as matplotlib & seaborn.
4. **Implement Predictive Learning**:
    i. Design a machine learning model to handle the neuroimaging data.
    ii. Create the predictive model and apply a train/test split from the HIV dataset.
5. **Develop Convolutional Neural Network**:
    i. Use the optimal model by evaluating the model's peak classification performance.
    ii. Build up the machine learning model using cross validation to minimise error.
    iii. Create a Neural Network to make accurate predictions from the HIV neuroimaging data.
6. **Document**:
    i. Gather results from the project and model any significant discoveries.
    ii. Evaluate findings, comparing them to the study of HIV neuroimaging at Stellenbosch University.
    iii. Write up final report.

Work Plan

This project will take place over the course of the Spring semester of Cardiff University. This gives me 12 weeks plus Easter recess to complete the project and deliverables. Currently, I have scheduled weekly meetings with my supervisor at 14:00 every Tuesday. During this time, I will have the chance to discuss my current developments, raise concerns and ask for advice with the supervisor and other students. I will also be having 2 core review meetings during the project. One in week 5 and a second review meeting in week 11.

Deliverables

The crucial deliverables for my project are:

- Initial Plan
- Data visualisations and analytics from HIV dataset
- Classical Machine Learning models trained on the neuroimaging data
- Fully functional Neural Network model trained on the neuroimaging data
- Results identified from the analysis and evaluation of the predictive learning model's performance
- Final Report

Milestones/Timescale

Gantt Chart highlighting the objectives, milestones & deliverables:

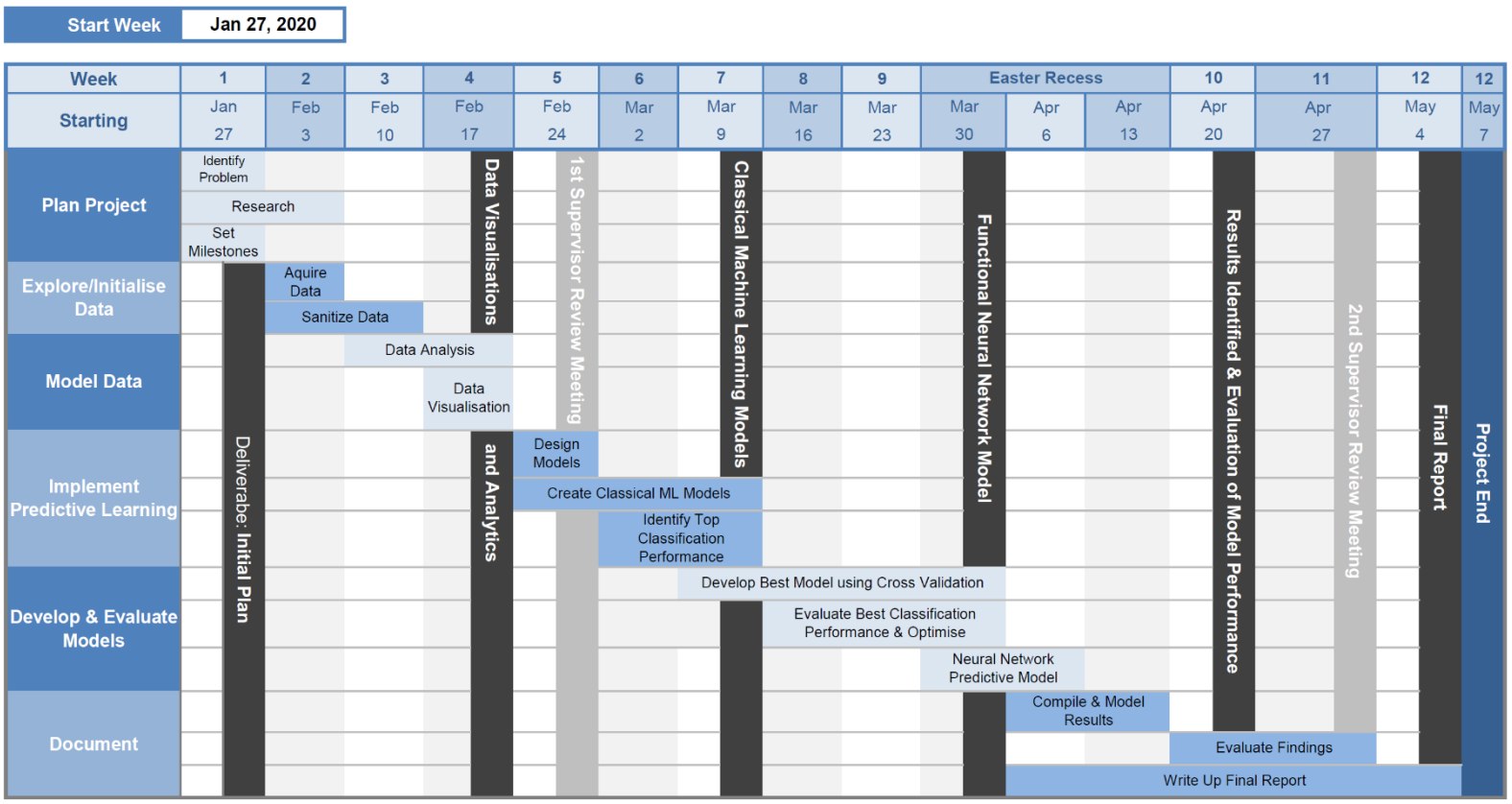

Figure 2 Project Timescale Gantt Chart

Descript weekly tasks/project timeline:

- **Week 1** (27/01/2020):
  - Rename project and description to be a more precise problem.
  - Research topics: HIV, Neuroimaging
  - Create the Initial Plan
  - **Deliverable**: Complete the Initial Plan and get supervisor approval
- **Week 2** (03/02/2020):
  - Research technical information: Data Analysis, Machine Learning, Deep Learning
  - Acquire the research data
  - Look through the sample dataset of participants checking the demographics, size and clinical characteristics.
  - Sanitize the dataset to only be relevant for the verdicts needed in my project.
- **Week 3** (10/02/2020):
  - Use data management techniques to appropriately store, order and access the HIV data.
  - Perform data analysis using classical machine learning techniques: supervised learning (classification, regression) and if possible, evaluate unsupervised learning techniques (clustering, association).
  - Use statistical data analysis techniques on the HIV dataset including T-test, Decision Tree, K-nearest neighbour.
- **Week 4** (17/02/2020):
  - Perform analytical tests such as ANOVA and REML to discover relations in the neuroimaging data.
  - Create data visualisations from the results of the data analysis using applications such as matplotlib & seaborn.
  - **Deliverable**: Data visualisations and analytics from HIV dataset
- **Week 5** (24/02/2020):
  - Document my findings thus-far.
  - Design the supervised machine learning models.
  - Organise dataset to use in pre-processing.
  - Format the Final report and add the supporting documents to the relevant sections.
  - **1st Review meeting**
- **Week 6** (02/03/2020) & **Week 7** (09/03/2020):
  - Create classical machine learning models: Discriminant Analysis (LDA), Support Vector Machines (SVM), Random Forest.
  - Apply a train and test split from the HIV dataset for the supervised machine learning models.
  - Identify the top classification performance for each model created.
  - Build up the best model using cross validation to minimise error.
  - **Deliverable**: Classical Machine Learning models trained on the neuroimaging data
- **Week 8** (16/03/2020) & **Week 9** (23/03/2020):
  - Assess accuracy of the predictive models in comparison with the other models.
  - Evaluate effectiveness of the machine learning model, looking for the best classification performance and the performance of the other models.
  - Explore alternative methods for identifying features. E.g. Feature importance for random forests & activation patterns for linear classifiers.
- **Easter Recess** (30/03/2020):
  - Optimise program by testing/tuning the model to be more efficient than the classical machine learning models.
  - **Deliverable**: Functional Neural Network model trained on the neuroimaging data
  - Use deep learning and neural network models to make accurate predictions from the HIV neuroimage data.
  - Gather results from the project and produce visualizations on any significant discoveries.

- Start the first sections of the final report: Introduction, Background.

❖ **Week 10** (20/04/2020):
- Continue writing up the final report: Approach, Implementation, Results and Evaluation.
- Evaluate the performance of my chosen model with other models, identifying the top classification performance.
- Evaluate the findings, comparing them to the current study of HIV neuroimaging at Stellenbosch University.
- **Deliverable**: Results identified from the analysis and evaluation of the predictive learning model's performance

❖ **Week 11** (27/04/2020):
- Finish the last sections of the final report: Future Work, Conclusion, supporting docs.
- **2nd review meeting.**

❖ **Week 12** (04/05/2020):
- Review the final report.
- Hand-in the final report.
- **Deliverable**: Final Report

Project Risks & Matrix

Risk Number	Risk	Likelihood (1-3)	Impact (1-3)	Mitigation Strategy
1	Injury or Illness, having to take leave to recover	Possible (2)	Major (3)	Focus on critical deliverables (project crashing) and file for extenuating circumstances if injury/illness requires significant recovery.
2	Bias, leading to inaccurate models	Unlikely (1)	Moderate (2)	Ensure data is valid and relevant to the project using data management processes. Have my supervisor review my produced data models.
3	Data, having minimal data or bad data for project	Possible (2)	Moderate (2)	Check the data provided and adjust the project to fit with the given dataset.
4	Unethical data present, protected participant data	Unlikely (1)	Major (3)	Forward ethical considerations to the COMSC Ethics department for evaluation.
5	Loss of project, documentation & code	Possible (2)	Major (3)	Back-up all files after every session to both an external hard drive and google drive.
6	Inconsistent assumptions or judgments, leading to unsuitable data modelling	Unlikely (1)	Minor (1)	Keep assumptions consistent with already existing practices and ask supervisor or collaborators for assistance and assessment of work.
7	Erroneous interpretation of the output and models	Unlikely (1)	Minor (1)	Log/monitor the outputs regularly and test for errored interpretations.

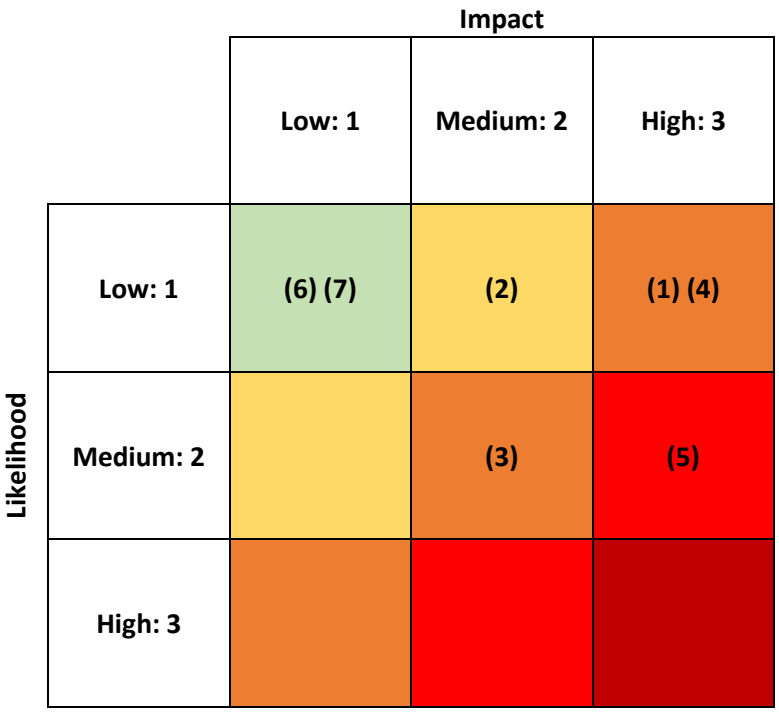

Figure 3 Risk Matrix

References

Kaiser, F. F., 2019. *The Global HIV/AIDS Epidemic.* [Online]
Available at: https://www.kff.org/global-health-policy/fact-sheet/the-global-hivaids-epidemic/
[Accessed 29 01 2020].

Kuhn, T. et al., 2018. *An Augmented Aging Process in Brain White Matter in HIV,* Los Angeles: Hum Brain Mapp.

Lundervold, A. S., 2019. *An overview of deep learning in medical imaging focusing on MRI.* [Online]
Available at: https://www.sciencedirect.com/science/article/pii/S0939388918301181
[Accessed 29 01 2020].

Osowiecki, D. et al., 2000. *Neurocognitive and psychological contributions to quality of life in HIV-1 infected women,* s.l.: s.n.

Spies, G., 2015. *Effects of HIV and childhood trauma on brain morphometry,* s.l.: Springer.

Statistics_South_Africa, 2014. *Mid-year population estimates.* [Online]
Available at: http://www.statssa.gov.za/publications/P0302/P03022014.pdf