**Deciphering On- and Off-Tissue Sampling of Oxysterols in µLESA-MSI Analysis: A Machine Learning Approach**

**Link to github repository –**

**Introduction**

The human brain contains 25% of the total body cholesterol by mass (1). Here, cholesterol serves a whole host of functions including forming an essential component of myelin sheaths (2), modulating both pre- and post-synaptic properties (3), as well as a substrate of metabolism into bioactive molecules, including oxysterols. Oxysterols are oxidized products of cholesterol and its precursors (4). As cholesterol metabolism is dysfunctional in many neurodegenerative conditions (5), analysis of cholesterol and oxysterols has become of particular interest in the study of health and disease. A major method for oxysterol analysis is mass spectrometry (MS). An advanced technique of MS is microscale Liquid Extraction Surface Analysis – MS Imaging (µLESA-MSI), which enables analysis of the spatial distribution of oxysterols, in a section of tissue. µLESA-MSI is normally coupled to Liquid Chromatography enable easy quantitation of oxysterols.

One of the major drawbacks of µLESA-MSI is that the sample preparation causes slight delocalization of metabolites to the glass slide immediately surrounding the glass slide that the samples are mounting on. This means it is not always easy to determine whether the data collected from sampling in µLESA-MSI is from an On- or Off-Tissue position, without comparing a post-analysis image of the sample to a pre-analysis image. In this report I develop a code utilizing peak areas the for abundant oxysterols in sections of mouse brain (acquired via µLESA-MSI), which uses a machine learning approach to classify samples into either On- or Off-Tissue.

**Methods**

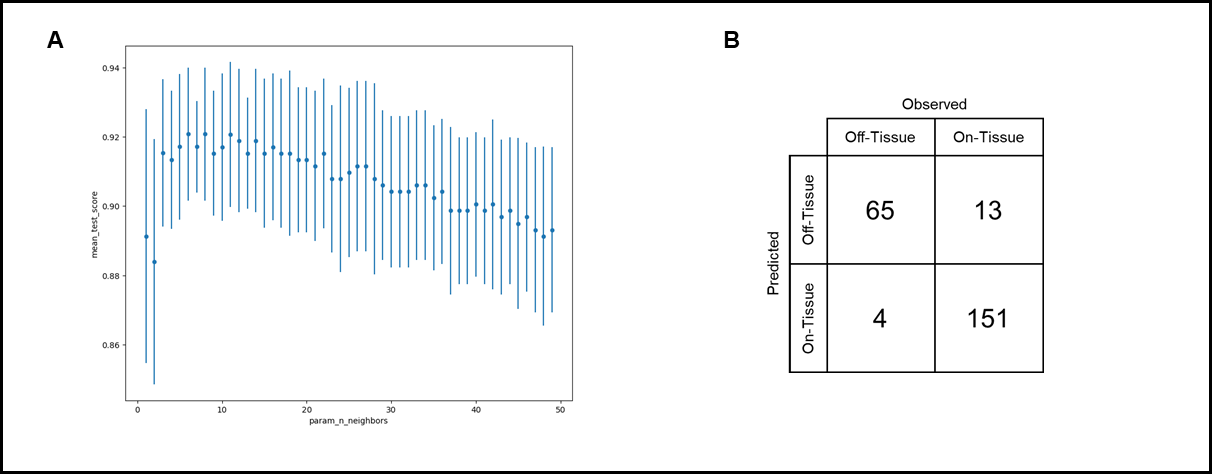
Dataset

The dataset used for this code is comprised of the area of the chromatographic peaks for 3 oxysterols which are abundant in the brain; cholesterol ([Chol]), desmosterol ([Des] the main precursor for cholesterol biosynthesis in the adult brain) and 24S-hydroxycholesterol ([24S-HC] the major product of cerebral cholesterol metabolism). Also included in the dataset were the peak areas for deuterated internal standards of the 3 oxysterols (Chol-D7 for Chol, Des-D6 for Des and 24R/S-HC-D7 for 24S-HC), which are deposited on the sample prior to MS analysis, as well as whether the analysed sample was from an On- or Off-Tissue position. 766 different samples were included in the dataset, 497 of which were sampled from On-Tissue positions and 279 Off-Tissue samples.

Code

After installation of required packages (pandas and sklearn functions), data cleaning was performed to replace NaN values (denoting no peak found in range) with 0, as NaN effectively means there is 0 peak area. This cleaned data was then outputted as a new cleaned dataset. This cleaned dataset was then split into train (70%) and test (30%) to mitigate limit overfiiting.

The selected model to be applied was a K Nearest Neighbours (KNN) algorithm. As the data has a wide range of values, a scaler was applied to enable feature variables to exert equal effect on the predictive model and allow for better classification into groups. Testing of n\_neighbours was conducted in a range of 1-50, to optimize model performance, with n\_neighbours = 8 yielding the best mean test score (see fig 1A). The KNN was then constructed using the mentioned parameters.



**Figure 1:** (**A**) The cross validation results showing mean test score output for different n\_neighbours values. (**B**) - A confusion matrix summarising the KNN model evaluation comparing the model predicted groupings compared to the ‘true’ observed groupings

**Results & Discussion**

Fig 1B summarises the results of model evaluation. Overall, the model performance was good, with an accuracy of 0.93 (F1 score, n = 233). However, the model was better at correctly predicting On-Tissue samples (F1 score = 0.95, n = 155) than Off-Tissue samples (F1 score = 0.88). The model recall of On-Tissue samples was also greater than Off-Tissue samples (0.97 and 0.83 respectively). However the model precision is high in both Off- and On-Tissue samples (0.94 and 0.92 respectively).

There are several potential reasons why the model was stronger at predicting On-tissue samples than Off-Tissue samples. Firstly, there was a greater number of On-Tissue samples in the original dataset, which is hard to avoid because the primary aim of µLESA-MSI is to analyse tissue sections, and therefore more On-Tissue samples will be present. A second reason, is there was a spread of peak areas observed Off-Tissue. This could also just be due to the technique, as in the immediate vicinity of the tissue (but still off-tissue) strong peaks, and therefore large peak areas, will be observed due to this slight delocalization of oxysterols introduced during sample preparation, whereas further from the tissue edge the peak areas will be very low, sometimes 0.

In conclusion, the constructed KNN model performed well, with an overall model accuracy of 0.93. However, the model was much better at correctly predicting On-Tissue spots than Off-tissue spots, which may be inherently due to the flaws of the technique for analysing samples in µLESA-MSI.

**References**

1 - Björkhem I, Meaney S. Brain cholesterol: long secret life behind a barrier. Arteriosclerosis, thrombosis, and vascular biology. 2004 May 1;24(5):806-15.

2 - Hussain G, Wang J, Rasul A, Anwar H, Imran A, Qasim M, Zafar S, Kamran SK, Razzaq A, Aziz N, Ahmad W. Role of cholesterol and sphingolipids in brain development and neurological diseases. Lipids in health and disease. 2019 Dec;18:1-2.

3 - Korinek M, Gonzalez-Gonzalez IM, Smejkalova T, Hajdukovic D, Skrenkova K, Krusek J, Horak M, Vyklicky L. Cholesterol modulates presynaptic and postsynaptic properties of excitatory synaptic transmission. Scientific Reports. 2020 Jul 28;10(1):12651.

4 - Griffiths WJ, Wang Y. Oxysterol research: a brief review. Biochemical Society Transactions. 2019 Apr 30;47(2):517-26.

5 - Martín MG, Pfrieger F, Dotti CG. Cholesterol in brain disease: sometimes determinant and frequently implicated. EMBO reports. 2014 Oct;15(10):1036-52.