MSc Data Science projects - Maths

Project 1: Automated spectral classifier for Wiserep

Supervisors: Prof Paolo Mazzali (ARI) and Dr Ivo Siekmann (CSM)

Project background:

Modern astronomical surveys detect hundreds of new transients every night, and the number is going to increase with the advent of Vera C. Rubin Observatory / Large Synoptic Survey Telescope (LSST) next year. For each of the transients, spectra are recorded which give the radiation emitted from a supernova over a range of wavelengths. The flux for specific wavelengths gives information about the chemical elements ejected in the explosion. Different classes have been defined based on the presence or absence of flux at specific wavelengths. However, many events do not fall clearly in any of the known classes.

In this project the student will use machine-learning algorithms to define an optical spectroscopy classification scheme to be used in Wiserep, the IAU repository for transient data.

Aims/objectives:

- Develop a method for automatic classification of optical spectroscopy data
- Gain additional insight into the differences between classes of spectra
- Optional: Integration of the method in Wiserep

Dataset(s):

Supernovae spectra are publicly available: https://wiserep.weizmann.ac.il/

Spectra provide flux measured over a range of wavelengths. Each spectrum on Wiserep is classified as a type such as Ia, II.

For the analysis, the wavelengths are discretised in a few hundred bins. This yields a multivariate data set of continuous variables (fluxes) and a categorical variable for the classification of each spectrum.

Methodological approach:

The performance of different classification algorithms such as

- logistic regression
- multi-layer perceptron (MLP)
- Support Vector Machines (SVM)
- Random Forest

on a training data set is compared. The resulting models are analysed to gain insight into the differences of the classes.

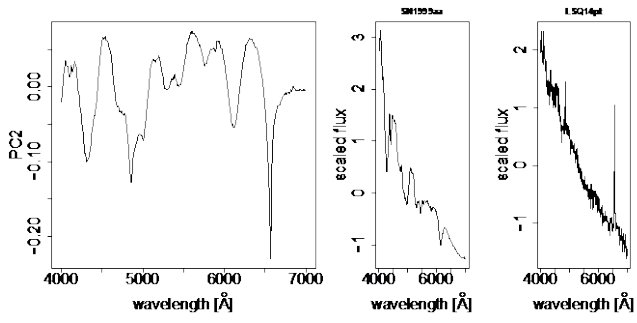

Figure project 1: In a preliminary study, principal component analysis (PCA) was applied to a dataset consisting of type Ia and IIP spectra. The PCA results indicate that the two types can be distinguished based on the presence or absence of a spike at 6562.8 Å which indicates the presence or absence of a Hα line. The project will continue this study using a supervised learning approach.

Resources required:

No specialised software or computing facilities required.

Project 2: Guessing genes - Detecting changepoints along a DNA string

Supervisor: Dr Ivo Siekmann (CSM)

Project background:

It has long been known that the four bases adenine (A), thymine (T), guanine (G) and cytosine (C) are not distributed uniformly over DNA sequences. For example, the GC content, the relative frequency of the bases G and C, can be highly variable in different parts of a DNA strand. At the same time GC content is, for example, often elevated in coding regions of the genome and is also associated with other questions from genetics and evolutionary biology.

In a previous study a Bayesian changepoint analysis approach has been successfully applied for finding segments of approximately constant GC content in viral genomes. In this project this work will be extended by accounting for frequencies of all four nucleotides A, C, G and T.

Aims/objectives:

- Apply an extension of the Markov chain Monte Carlo (MCMC) method by Siekmann et al. (2014) to reference genomes.
- Infer segments based on frequencies of nucleotides A, C, G and T accounting for uncertainties.
- Relate inferred segments to biological genomic structures such as genes.
- Compare with results obtained previously based on GC content.

Dataset:

Genomic data are freely available from the database of the National Center for Biotechnology Information (NCBI): https://www.ncbi.nlm.nih.gov/genome

Methodological approach:

An extension of the MCMC method by Siekmann et al. (2014) will be applied. Code will be available. The project requires knowledge of Bayesian statistics and MCMC.

Resources required:

No specialised software or computing facilities required.

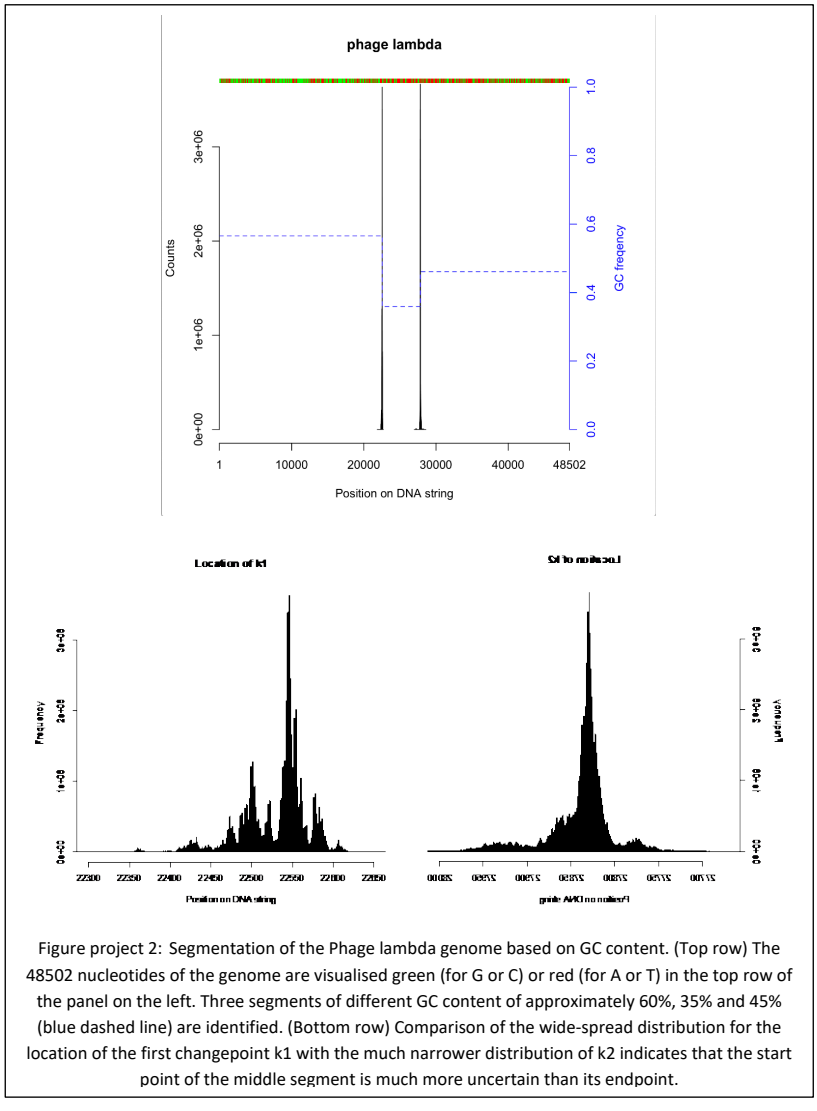

Figure project 2: Segmentation of the Phage lambda genome based on GC content. (Top row) The 48502 nucleotides of the genome are visualised green (for G or C) or red (for A or T) in the top row of the panel on the left. Three segments of different GC content of approximately 60%, 35% and 45% (blue dashed line) are identified. (Bottom row) Comparison of the wide-spread distribution for the location of the first changepoint k1 with the much narrower distribution of k2 indicates that the start point of the middle segment is much more uncertain than its endpoint.

Project 3: Wait a minute! - Stochastic models with delay

Supervisor: Dr Ivo Siekmann (CSM)

Project background:

Continuous-time Markov models can be used to model physical, chemical and biological systems that stochastically change between different states. They can, for example, be used for modelling chemical reactions, biomolecules like ion channels or even represent the hunting strategies of predators in ecosystems. The parameters of Markov models usually depend on the current state of the environment, but many systems respond to their environment with a delay. The goal of this project is to develop a model that accounts for these delays and test its behaviour via stochastic simulations. The model will be applied for investigating the delayed response of an ion channel to varying concentrations of calcium.

Aims/objectives:

- Investigate the properties of a Markov model with time delay such as first-passage times and dwell-time distributions.
- Investigate the delayed response of an ion channel to varying calcium concentrations by fitting the Markov model with time delay to the data by Mak et al. (2007).

Dataset(s):

Distributions of first-passage times from Mak et al. (2007)

Methodological approach:

- Theory of Markov models
- Simulation of Markov models using the Gillespie algorithm.

Resources required:

No specialised software or computing facilities required.

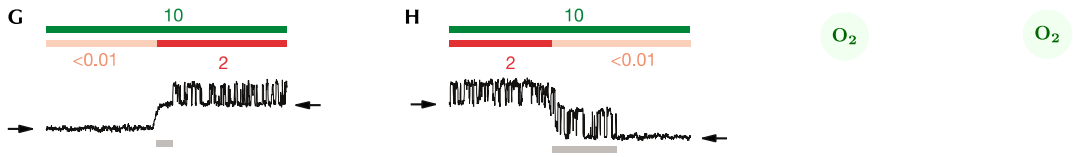

Figure project 3: (G and H) Delayed response of an ion channel. Rather than instantaneously adjusting to a change of the calcium concentration, data from Mak et al. (2007) show that the response of the IP_3 receptor to a rapid increase from 0.01μM to 2μM as well as a rapid decrease from 2μM to 0.01μM is delayed, indicated by grey bars. (Middle and Right) Rather than fitting a model with time-dependent rates $q_{12}(t)$ and $q_{21}(t)$ the aim of the project is to fit a model with delayed rates $q_{12}(t)$ and $q_{21}(t)$.

Project 4: Mechanistic data-driven modelling with differential equations (ODE) - a Bayesian approach

Supervisor: Dr Ivo Siekmann (CSM)

Project background:

Many real systems in physics, chemistry and biology can be described by systems of ordinary differential equations (ODE). These models are designed based on assumptions on underlying physical, chemical or biological processes. They help us understand how the behaviour of a system emerges from underlying processes. From models of infectious diseases we can see, for example, how the spread of an infection arises from contacts between individuals. When ODE models are fitted to time-series data we obtain estimates for parameters such as growth rates of populations, transmission rates of infections or rates of chemical reactions. Usually, parameter estimates are found by optimisation approaches but it becomes increasingly clear that simply obtaining the "best fit" is not enough - individual parameters might give similarly good fits over a wide range or there might be multiple alternative parameter sets consistent with the data. It is therefore crucial to gain insight into parameter uncertainty. Bayesian statistics provides a framework that allows us to address this challenge because it enables us to calculate probability distributions for the parameters of a model. There are several potential applications including infectious diseases, pharmacology and chemical reaction networks.

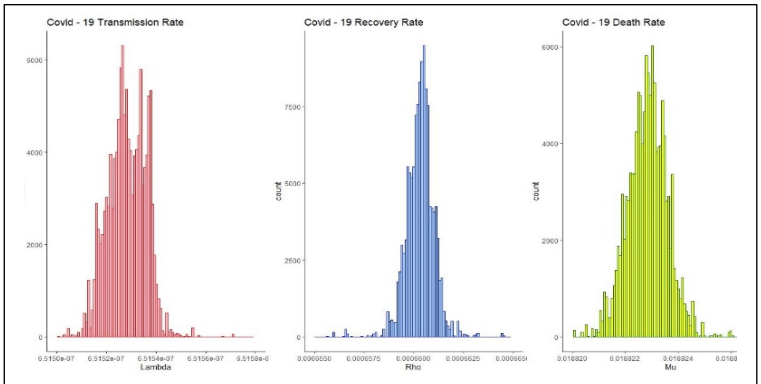

Figure project 4: Parameter estimates for the transmission rate λ, the recovery rate ρ and the death rate μ after fitting the SIR model to COVID-19 data, from Meredith (MSc Thesis, 2020).

Aims/objectives:

- Obtain estimates with uncertainties for the parameters of mechanistic ODE models
- Study of non-identifiability similar to the approach from Siekmann et al. (2012)

Dataset(s):

Depending on the application considered in the project, different data sets can be used:

- COVID-19: https://github.com/CSSEGISandData/COVID-19.git
- chemical reaction networks: https://www.synapse.org/#!Wiki:syn1720047/ENTITY/56061

Methodological approach:

The project requires knowledge of Bayesian statistics and MCMC:

- Parameter inference based on Metropolis-Hastings sampling from a likelihood derived from a suitable error distribution for the time series data
- Study of non-identifiability similar to the approach from Siekmann et al. (2012)

Resources required:

No specialised software or computing facilities required.

Project 5: What makes a pretty face? - Recognising the parts of a face using non-negative matrix factorisation

Supervisor: Dr Ivo Siekmann (CSM)

Collaborators: Dr Ivan Olier (CSM) and Dr Sandra Ortega (CSM)

Project background:

Non-negative matrix factorisation (NMF) is a data mining method that allows for unsupervised separation of source signals, also known as blind source separation. It has been applied to the extraction of features from image data, the identification of topics in text corpora or the deconvolution of data generated by mixing a number of sources. The method is based on factorising a non-negative data matrix X in two non-negative factors W and H i.e. $X = W H$.

We will investigate a Bayesian approach to NMF and develop a Markov Chain Monte Carlo (MCMC) algorithm for sampling from the posterior distribution. The solution of finding a factorisation is not unique. A Bayesian framework will enable us to explore alternative solutions apparent in the posterior distribution as well as constraining possible factorisations by choosing a suitable prior distribution.

Aims/objectives:

- Develop a Markov Chain Monte Carlo (MCMC) method for non-negative matrix factorisation (NMF) based on Moussaoui et al. (2013)
- Investigate the non-identifiability of NMF
- Extend the MCMC method for NMF to convex-NMF.

Dataset(s):

This project is mostly about methods development. In the first instance, test data constructed by linear mixing of images will be sufficient.

Methodological approach:

- Markov chain Monte Carlo (MCMC) approach based on Moussaoui (2013)
- Study of non-identifiability similar to the approach from Siekmann et al. (2012)

Resources required:

No specialised software or computing facilities required.

Project 6: Detection of person's identity from blood vessels of the retina.

Supervisor: Dr Gabriela Czanner (CSM)

Collaborators: Dr Bryan Williams (Lancaster University)

Project background:

There is increasing interest in the identification of persons via the presentation of their blood vessels, which is useful in cases of violent crime, security systems, and identification for medical purposes. While blood vessels show strong potential for identification, their variation and robustness as a biometric is not well understood. Retinal fundus cameras offer an excellent opportunity to obtain a clear view of the retinal vasculature and allow it to be investigated in detail.

Aims/objectives:

The aim is to investigate a dataset of retinal images to determine measures for identifying where two images come from the same individual and to explore their accuracy and variation.

Dataset(s):

We have two datasets available for this project. From Dataset 1, we have 124 images from 39 participants. From Dataset 2, we have 153 images from 59 participants. In all cases, there are at least two images per participant and in many cases, there are more than 2. Some preprocessing may be required, including semi-supervised segmentation of the vasculature, to obtain the required vessel patterns from the images.

Methodological approach:

This project involves semi-automated image and shape analysis as well as investigation of graph comparison methods.

Resources required:

Software R and possibly Matlab for pre-processing.

Project 7: Statistical modelling methodology toward detection and progression of Keratoconus.

Supervisor: Dr Gabriela Czanner (CSM)

Collaborators: Dr Vito Romano (Consultant Ophthalmic Surgeon at The Royal Liverpool University Hospital Associate Professor of Ophthalmology at University of Liverpool) and Professor Steven Kaye (Professor of Ophthalmology and Consultant Ophthalmologist, Faculty of Medicine, Institute of Life Course and Medical Sciences, University of Liverpool)

Project background:

Keratoconus is a significant cause of visual loss in young adults. It is a progressive, usually bilateral corneal disease, accounting for more than 25% of corneal transplants undertaken in Europe and the United States (Al-Yousuf et al., Br J Ophthalmol. 2004;88(8):998-1001.) It is characterized by progressive ectasia and thinning of the cornea leading to a reduction in vision. Emerging new treatment strategies and imaging technology have improved the management of keratoconus. Corneal collagen cross-linking (CXL) is perhaps the most recent promising innovation in the management of keratoconus. This procedure can delay or halt further progression of keratoconus and reduce the need for corneal transplantation. Early detection of progressive disease in keratoconus is an important consideration when offering treatment with CXL. The introduction of corneal topography and tomography has improved the ability to diagnose keratoconus by increasing the ability to identify ectatic change at an earlier stage than has been previously possible (Belin et al., Eye Contact Lens. 2014;40(6):326-330). Despite advances in the field of imaging technology, however, measurement imprecision remains an important issue in identifying and discriminating change as a marker of disease progression. In particular, it can be difficult to discern true change from measurement imprecision and treatment is not without risk and cost. For example, the development of microbial keratitis is one of the most serious complications following CXL (Galvis et al.,Clin Ophthalmol. 2017;11:657-668).

Aims/objectives:

(1) To develop and characterise N-dimensional multivariate distributions of corneal tomographic indices; and to validate retrospectively and prospectively to patients with keratoconus and healthy subjects. (2) To develop refractive data transformation probability maps and to apply these to characterise refractive data of patients with and without keratoconus. (3) To develop a mathematical modelling approach for detection of keratoconus progression. This will be inherently interpretable approach.

Dataset(s):

I need to get the details from the collaborators.

Methodological approach:

Multivariate normal distribution, Blant-Altman analysis, Shape modelling, Monte-Carlo simulation, Statistical Inference and Classification, Empirical Bayes.

Resources required:

Software R.

Project 8: The effect of COVID-19 on A&E attendance in the North West with a focus on violence and self-harm.

Supervisor: Dr Ian Jarman (CSM)

Collaborators: Jennifer Germain, PHI project manager

Project background:

The Trauma and Injury Intelligence Group (TIIG) was established by the Public Health Institute (formerly the Centre for Public Health) in 2001 to develop an injury surveillance system to enable systematic data collection and sharing across the North West of England. TIIG collects and reports on reliable injury and violence information from emergency departments (EDs) in Merseyside, Cheshire, Cumbria, Lancashire and Greater Manchester, in addition to warehousing data collected by the North West Ambulance Service (NWAS).

Aims/objectives:

Working with Public Health Institute at LJMU, this project will investigate A&E admissions in the North West with a focus on the changes in hospital admissions after the COVID-19 lockdown in March 2020. The profile of admissions will be investigated month by month, pre- and post-lockdown, with a specific focus on violence and self-harm. As well as the time series investigation an exploratory analysis of A&E admissions will also be undertaken with supplementary demographic information to identify different profiles (clusters) of Patients at higher risk of admission to A&E due to violence or self-harm.

Dataset(s):

The data TIIG collects includes (for EDs) data on patient demographics including age, sex, ethnicity and LSOA of residence, information concerning the visit to ED such as arrival mode, referral mode and outcome, and data on the type of injury sustained. For assault presentations, TIIG work with EDs to encourage them to collect the Information Sharing to Tackle Violence minimum dataset which comprises attendance date and time, assault date and time, location of incident and details of weapon used. For NWAS data we collect demographic data (age, sex), date and time of callout, location of callout, type of injury sustained and whether the patient was conveyed to hospital or not. All data is record level non-patient identifiable and received on a monthly basis.

Methodological approach:

Appropriate time series investigation techniques given the research question. Appropriate pattern recognition algorithms (for instance, clustering) to describe sub groups of attendees to A&E.

Resources required:

Expectation is to use standard software available within the LJMU (Matlab, R, SPSS etc).

Project 9: Optimization of methods used to read and extract information from digitized electrocardiograms (image recognition).

Supervisor: Dr Elon Correa (CSM)

Project background:

Due to an ageing population, the UK is facing a sharp rise in avoidable illnesses. Undetected atrial fibrillation (AF) is a major health concern with life-threatening complications, and it is the most common arrhythmia in the elderly population. The availability of sophisticated cardiac implanted electronic devices (CIEDs), either dual chamber pacemakers or biventricular pacemakers (for cardiac resynchronisation therapy) facilitates the detection of atrial high-rate episodes (AHRE) or subclinical AF (SCAF). Automated screening from ECGs coupled with clinical and lab data analysis is a complex task often convoluted by low-quality digitized images. Optimization of such screening tools will enable more intense ECG monitoring and benefit patients.

Aims/objectives:

This project will focus on the optimization of methods used to read and extract information from ECGs. After an ECG is digitized, the data will be filtered using image processing, denoising and signal recovery methods. An improved, higher quality, ECG will then be reconstructed.

Dataset(s):

We will initially study an existing dataset of CIEDs from Liverpool Heart & Chest Hospital (LHCH) using 12 lead ECG data, to derive a machine learning (ML)/AI model for predicting prevalent AHRE/SCAF (derivation cohort).

Methodological approach:

ECG reconstruction will use deep learning models and wavelet transforms. Decision curve analysis and calibration will evaluate the diagnostic ability of the ML classifiers and account for the clinical utility of models.

Resources required:

For the data analysis, it is expected that you will need to use a specialised software package of your choice such as R, MATLAB, SPSS, SAS, Minitab, Python, etc.

Project 10: A comparison of different techniques for handling missing values: a simulation study.

Supervisor: Dr Elon Correa (CSM)

Project background:

Many real-world data, such as health care questionnaires, survey of customers shopping habits, etc., will contain some missing information. In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.

Aims/objectives:

First, this project will use full datasets, with no missing values, to randomly generate copies of these datasets containing different proportions of missing values. Second, many different imputation techniques (the process of replacing missing data with substituted values) will be used to compare the accuracy of identical statistical models applied to the original datasets, with no missing values, and their respective imputed data versions. The results of these comparisons will give us real insights into the pros and cons of each imputation technique applied.

Dataset(s):

It is expected that for this project you will produce your own simulated datasets so that the occurrence of missing values can be easily controlled for further analysis.

Methodological approach:

Several different traditional missing values methods will be compared such as listwise or case deletion, pairwise deletion, mean substitution, regression imputation and maximum likelihood. You will also be encouraged to develop other novel and suitable methods for dealing with missing values.

Resources required:

For the data analysis, it is expected that you will need to use a specialised software package of your choice such as R, MATLAB, SPSS, SAS, Minitab, Python, etc.

Project 11: An investigation of General Practitioner drug prescribing patterns in England from 2014 – 2020 linked to Mental Health conditions and the impact of Covid19.

Supervisor: **Dr Ian Jarman** (CSM)

Collaborators: Dr Lucy Astle, Dr Tim Smith both GP's NHS East Lancs

Project background:

There is interest in the NHS regarding drug prescription and the pattern of prescribing in General Practices over time especially since the first lockdown in March 2020 due to Covid19. Of particular interest is the change in drugs prescribed linked to Mental Health conditions

The English Prescribing dataset contains detailed information on prescriptions issued in England that have been dispensed in England, Wales, Scotland, Guernsey, Alderney, Jersey, and the Isle of Man.

The dataset combines elements of the Detailed Prescribing Information (DPI) data previously released by NHS Business Services Authority (NHSBSA) via the Information Services Portal, and the Practice Level Prescribing in England (PLP) data released by NHS Digital via their website. It is intended to replace both of those sources.

These datasets have been brought together to provide end users with a single comprehensive, consistent and accessible source of prescribing information.

Aims/objectives:

The project will investigate the time series patterns of monthly drug prescriptions in General Practices in East Lancashire and how they differ from North West and England as a whole. The main focus will be on drugs prescribed linked to mental health conditions with a specific interest in the impact Covid19 has had on prescribing patterns.

In addition, data from GP Practice profiles consisting of general health and deprivation indicators will be used to derive clusters driven by distinct data patterns for all GPs in England. These will be mapped to the prescribing data with an interest if there is any relationship between GP practice cluster membership and prescribing patterns.

Dataset(s):

All data is publicly available

English Prescribing Dataset: Welcome - Open Data Portal BETA (nhsbsa.net)

GP Practice Profiles: National General Practice Profiles - PHE

There will be a need to map different years of data together and some data integrity checking but this will not form a significant aspect of the project.

Methodological approach:

Time series analysis of prescription patterns.

Clustering Methodology as appropriate to the form of the GP Practice data.

Resources required:

Standard software packages can be used that are available free to LJMU students (Matlab, SPSS, R, etc).

Project 12: Extending the Orthogonal Search Rule Extraction algorithm to make use of different forms of data and classification architectures.

Supervisor: Dr Ian Jarman (CSM)

Collaborators: Dr Terence Etchells inventor of OSRE

Project background:

Trained Artificial Neural Networks (ANN) are justifiably criticised for their lack of transparency. Yet, the black-box nature of neural network inferences is often cited as a critical limitation for their validation especially in safety critical applications, for example in a medical environment, assigning patients to different risk groups which have radically different treatment regimes. The extraction of understandable rules from an inference model enables the user of the model to open the black box, and readily interpret the decisions of the model.

A principled rule extraction methodology is Orthogonal Search-based Rule Extraction (OSRE). OSRE extracts conjunctive rules using categorical ordinal/nominal data from smooth decision surfaces derived from a Multi-Layer Perceptron which is a popular Artificial Neural Network configuration.

Aims/objectives:

This project looks to update and extend the OSRE methodology when the explanatory variables are continuously valued and also update the architecture to enable the use of different classification algorithms, be they derived from traditional statistical models which are linear-in-the-parameters, such as logistic regression, or with generic non-linear approximations to decision surfaces, as is the case for the wide range of ANN architectures.

The practical value of the rule extraction paradigm will be demonstrated with direct visualisation of decision surfaces in well-separated data subspaces, enabling inferences about individual data points to be contextualised within the grouped decisions characterised by Boolean rules.

Dataset(s):

This is a more theoretical project with additional development of the algorithm, any publicly available classification datasets will be appropriate.

Methodological approach:

Theoretical understanding of OSRE enabling extensions and updates to the current version.

Resources required:

Matlab.

Project 13: Using machine learning to model acute care conditions in intensive care units.

Supervisor: Dr Ivan Olier (CSM)

Project background:

Patients admitted to intensive care units (UCI) are acutely unwell. Since they require round-the-clock monitoring, things in ICU happen at a much faster pace than other hospital areas and other healthcare settings. After a few hours, there have been collected a considerably amount of patient-related data, mainly vitals in the form of times series. This offers exciting challenges to data mining and machine learning: ICU data is typically untidy, highly heterogenous, and time-dependent. Many clinical questions can be addressed using machine learning models.

This project seeks to identify ICU phenotypes as defined as the most acute care conditions (e.g., acute myocardial infarction, pneumonia, septicaemia, etc).

Aims/objectives:

This project aims to develop ML models of ICU phenotypes.

O1: To build a multi-class ML model for phenotyping.

O2: To produce a phenotype visualisation map based on the developed classifier.

O3: To find possible links between factors (vitals, clinical history, etc) and phenotypes.

Dataset(s):

We will use the MIMIC-III database (https://mimic.physionet.org/). The database is already installed in our servers. We have run several projects using the database already. The student will receive an extract of the data which requires missing value imputations and data cleaning. The data comprises tabular and time series.

Methodological approach:

The project requires the use of a range of ML methods to ensure the highest possible model is implemented. Handling time series will require the use of state-of-the-art ML methods such as deep convolutional neural networks and/or recurrent neural networks.

Also, coding skills, especially in Python, are expected.

Resources required:

Python and Keras.

Project 14: Machine learning modelling of ECG for detection of atrial fibrillation.

Supervisor: Dr Ivan Olier (CSM)

Project background:

Atrial fibrillation (AF) is the most common form of irregular heart rhythm (arrhythmia) worldwide. It is commonly diagnosed using electrocardiograms (ECGs). ECGs are biomedical waveforms that have become attractive to the machine learning (ML) community as they comprise very noisy multivariate time series. There was a specific challenge to address the problem of detecting AF in the ECG automatically: the Physionet/CinC challenge 2017.

A common methodological question is whether the ML model should take in hand-crafted (manually extracted) ECG features or let the model to automatically extract them to make predictions. The former is yields models easier to explain as features are typically derived from well-known morphological aspects of the ECG. The latter is typically implemented using deep learning and could potentially deliver higher performing models. However, there are the questions on whether the automatically extracted features are dataset-dependant, whether they relate to the hand-crafted ones, and/or whether they could identify ECG features that have remained undetectable to the human eyes.

Aims/objectives:

This project aims to develop ML models that detect AF in ECG waveforms.

O1: To build a four-class ML for detecting AF, other arrhythmias, normal sinus rhythm and noise based on hand-crafted features of the ECG.

O2: To build a similar classifier as O1 but using deep learning for feature extraction.

O3: To perform an exhaustive comparative model evaluation to provide insights into which approach could be more beneficial in clinical settings.

Dataset(s):

We will use the Physionet/CinC challenge 2017 database and the MIMIC-Waveform database. Both databases are already installed in our servers. Data cleaning and pre-processing are required in both databases.

Methodological approach:

The project requires the use of a range of ML methods to ensure the highest possible model is implemented. Handling time series will require the use of state-of-the-art ML methods such as deep convolutional neural networks and/or recurrent neural networks.

Also, coding skills, especially in Python, are expected.

Resources required:

Python and Keras.

Project 15: Recognising myocardial infarction from ECG images using machine learning.

Supervisor: Dr Ivan Olier (CSM)

Project background:

Many electrocardiograms (ECGs) are recorded as scanned copies. Identifying cardiac conditions form this kind of ECG images requires a different methodological approach from waveform analysis. In this sense, ECG-diagnosis becomes an image analysis problem rather than a multivariate time series one. Although many cardiac events can be identified in the ECG, this project will focus on detecting myocardial infarction (MI).

Aims/objectives:

This project aims to develop an ML to identify MI from ECG images.

O1: To build a binary classifier ML for MI detection in ECG images.

O2: To identify image features that could be more relevant to MI identification.

O3: To perform model evaluation and position its performance with respect to physicians.

Dataset(s):

We will use the PTB Diagnostic ECG Database (https://www.physionet.org/content/ptbdb/1.0.0/). The database is already available from our servers.

Methodological approach:

The project requires the use of a range of ML methods to ensure the highest possible model is implemented. Handling time series will require the use of state-of-the-art ML methods such as deep convolutional neural networks and/or recurrent neural networks.

Also, coding skills, especially in Python, are expected.

Resources required:

Python and Keras.