

Configuration Manual

MSc Research Project
Msc Data Analytics

Aleena Paul
Student ID: 23285583

School of Computing
National College of Ireland

Supervisor: Sheresh Zahoor

National College of Ireland
MSc Project Submission Sheet



School of Computing

Aleena paul
Student Name:
23285583
Student ID:
Msc Data Analytics **Year:** 2025
Programme:
Research Practicum
Module:
Sheresh Zahoor
Lecturer:
11-12-2025
Submission Due Date:
Early recognition of sepsis based on explainable time-series deep
Project Title: learning and machine learning models on real ICU data
.....

Word Count:903..... **Page Count:**4.....

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Aleena Paul
Signature:
10-12-2025
Date:

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Aleena Paul
Student ID: 23285583

1 Introduction

This configuration manual gives a fairly good overview of the arrangement needed to operate the early sepsis prediction system generated in this research. It describes the technical context, software requirements, data structure, directory hierarchy and procedures to recreate the entire analysis pipeline. This is meant to be an uncomplicated but comprehensive reference hence that the researchers and evaluators can run the notebook and replicate results without confusion.

2 System Requirements

The project was created and tested on a windows11 platform with an Intel i3 processor with 8GB of RAM, but any modern computer with at least 16GB of RAM is likely to be compatible. The storage space is suggested to be 10 GB or more that will store the dataset, visual output, and temporary cache files. An optional CUDA-enabled graphics card is useful to speed up deep-learning models and in particular when training recurrent or convolutional neural networks.

2.1 Software Requirements

Python 3.12 or the later version and the usage through Jupyter notebook or Visual Studio Code are required to run the project. The following libraries and packages will have to be installed:

- Numpy 1.26.4
- Pandas 2.2.2
- Tqdm 4.64.1
- Matplotlib 3.9.2
- scikit-learn 1.6.1
- XGBoost 3.1.1
- LightGBM 4.6.0
- PyTorch 2.6.0

3 Dataset

The dataset of the present project is the PhysioNet Sepsis Challenge 2019¹, which is located in the training_setA and training_setB folder. Thousands of patient-specific.psv files of

¹ <https://physionet.org/content/challenge-2019/1.0.0/#files-panel>

hourly ICU time-series measurements of patients can be found in these folders. Both files contain vital signs, lab values, and demographic indicators, which allow forecasting the development of sepsis. To make the process faster, a cache folder is created. The first loading of the .psv files of both training setA and training setB is merged and stored as a single stored Parquet/CSV file in the cache directory. The next time the notebook runs, this cache file is used hereby reading thousands of .psv files in a short period of time, thus this makes the data processing much quicker in other runs. The Dataset.xlsx file is also there to summarise variable definitions and metadata. All preprocessing like treatment of missingness, generating of fixed-length windows are carried out in the main notebook.

4 Directory Structure

```
archive/
|
|— cache/
|— training_setA/
|— training_setB/
|— Dataset.xlsx
|— Sepsis_Prediction.ipynb
```

5 Code Execution

- **Step 1: Open the Notebook file**
Open Sepsis_Prediction.ipynb in the jupyter Notebook or VS code and ensure that all the dataset folders such as trainingsetA, training setB, cache, Dataset.xlsx are all in the same file.
- **Step 2: Run Imports and Setup**
The first cell is to be performed in order to load the necessary libraries. Install any missing requirements as there are any.
- **Step 3: Load and Prepare the Dataset**
Execute the cells which reads the psv files, construct labels and execute the basic EDA tests.
- **Step 4: Preprocess Data**
Do the preprocessing step of dealing with median imputation, scaling, six hour window generation, and split of datasets in training, validation and test sets of both ML and DL workflow.
- **Step 5: Training and Testing Machine-Learning Models.**
Run the cells of the ML pipeline and train such models as logistic regression, random forest, GBM, lightGBM and XGBoost. Print the evaluation matrix scores with the notebook.
- **Step 6: Training and Testing Deep-Learning Models**
Run the DL pipeline cells to train such models as LSTM, GRU, BILTM, CNN1D. Print the evaluation matrix scores with the notebook.
- **Step 7: Run SHAP Explainability**
Run the SHAP explainability cell to identify the top most important features to XGBoost to create SHAP summary plots.
- **Step 8: Calibration and Operating-Point Analysis.**
Calculate the brier scores and obtain the calibration curves, sensitivity-specificity thresholds and ROC curves by running the calibration