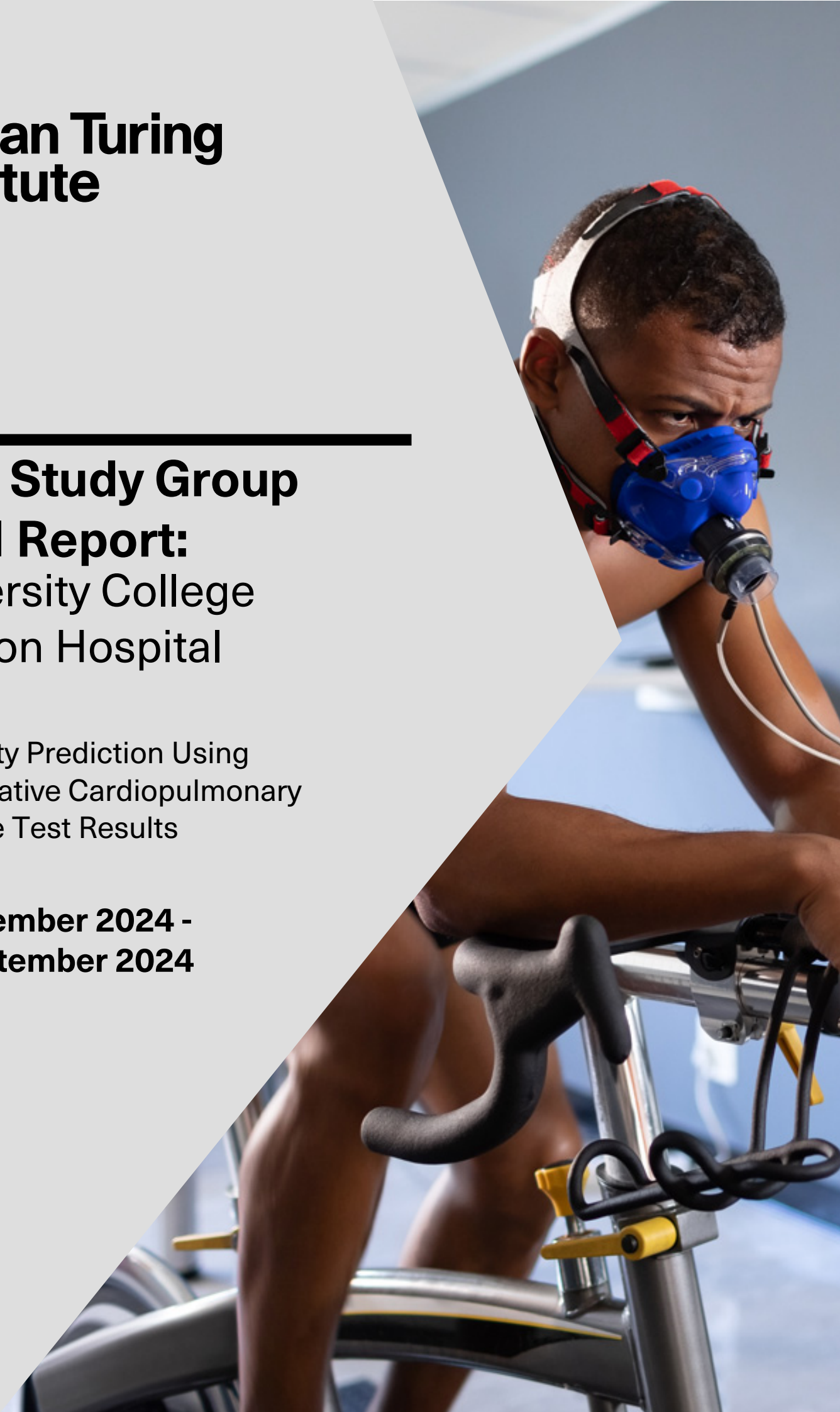
The Alan Turing Institute

Data Study Group Final Report:

University College London Hospital

Morbidity Prediction Using Preoperative Cardiopulmonary Exercise Test Results

9 September 2024 - 13 September 2024



https://doi.org/10.5281/zenodo.15167430

Contents

1 Executive Summary

University College London Hospital (UCLH) has provided a dataset of 148 patients who underwent cystectomy surgery (removal of the bladder). This dataset includes postoperative complication outcomes (morbidities), demographic information, and pre-surgery Cardiopulmonary Exercise Test (CPET) results. CPET is administered to patients scheduled for surgery to determine their exercise capacity, which serves as an indicator of perioperative and postoperative risk. Therefore, CPET can be used to estimate surgical risk. Despite the wealth of physiological data collected during CPET, only a small proportion of this information is currently used when calculating surgical risk.

The purpose of this Data Study Group (DSG) was to apply modern machine learning techniques to develop models predicting postoperative morbidities from CPET data. The DSG objectives included: creating models that are more predictive and interpretable than existing CPET-based risk models; comparing different machine learning algorithms in terms of predictive performance and interpretability; and using these models to derive additional predictive features from CPET data.

The dataset itself consists of high-frequency physiological measurements collected during the CPET sessions for each patient. These measurements include oxygen uptake, carbon dioxide production, heart rate, and blood pressure. Measurements were recorded at one-second intervals over a total test duration of 8–12 minutes per patient. The CPET sessions consisted of stages involving rest, pedalling, and recovery on an exercise bike. CPET session data are linked to demographic information and binary indicators showing whether patients developed cardiovascular (CVS), respiratory, or infection complications after surgery.

It was unclear a priori which machine learning algorithm would perform best on these data, prompting the consideration of several algorithms. Both traditional machine learning (ML) and deep learning (DL) algorithms were applied. For traditional ML algorithms, several handcrafted features were derived from the raw data based on previous literature suggesting their predictive value for postoperative morbidities. These derived features, along with most of the raw features, were input into a ridge classifier, random forest, XGBoost, and Random Convolutional Kernel Transform (ROCKET). Additionally, Granger causality was applied to identify causal relationships within the dataset and to derive p-values for use

as features in machine learning models. These features were evaluated using a k-nearest neighbours (k-NN) classifier.

Several deep learning architectures were also considered. A one-dimensional convolutional neural network (CNN) was applied due to its demonstrated success on time-series data. A recurrent architecture, specifically a Long Short-Term Memory (LSTM) network, was evaluated because of its natural ability to handle sequential data. Finally, a simple multilayer perceptron (MLP) was applied to assess whether predictions could be effectively made using individual time points from the CPET data.

In general, the models performed moderately well on a held-out test set of 16 patients, achieving accuracies ranging between 0.5 and 0.75 for predicting respiratory and infection complications. XGBoost and random forest classifiers generally outperformed the other models in terms of test accuracy and F1 scores. Most models achieved accuracy greater than 0.9 for the CVS complication class; however, this result was primarily due to severe class imbalance, meaning high accuracy could be achieved simply by predicting the majority class.

Due to the absence of benchmark prediction models or typical clinicians' prediction accuracies for this task, it was challenging to determine the practical usefulness of the developed models. Thus, the first research objective could not be conclusively assessed. Additionally, comparing algorithm performance was difficult because test accuracies were calculated differently between some models, preventing the second research objective from being achieved. Lastly, aside from manually engineered features derived from the literature, no additional features were successfully derived from the ML models, meaning the final research objective was also not met.

Interpretability of most models was found to be poor. The deep learning models were inherently difficult to interpret, and most traditional ML models also lacked interpretability. However, the ridge classifier demonstrated inherent interpretability, as it is a linear model, and its learned coefficients provide meaningful insights.

The primary limitations of this study included insufficient data and significant class imbalance, which led to poor generalisation performance. Additionally, the limited duration of the DSG did not allow sufficient time for hyperparameter optimisation or thorough debugging. Nevertheless, this DSG demonstrated that CPET data can be processed into various formats suitable for machine learning models, resulting

in moderate predictive performance.

For future work, the DSG recommends obtaining additional data, allocating more time for hyperparameter optimisation and code quality assurance, standardising the calculation of test accuracy metrics, and employing interpretable machine learning methods such as Shapley Additive Explanations (SHAP) or Local Interpretable Model-agnostic Explanations (LIME).

2 Introduction

2.1 Background

The challenge owner, University College London Hospital (UCLH), is a world-renowned teaching hospital recognized for its cutting-edge medical research and patient care. UCLH is at the forefront of applying advanced technologies, including machine learning and data science, to improve healthcare outcomes.

This challenge is part of a broader initiative to enhance the quality of surgical care, particularly for patients undergoing major operations such as cystectomy. To support this effort, UCLH has made available an anonymized dataset of 148 patients who underwent a preoperative Cardiopulmonary Exercise Test (CPET) followed by cystectomy surgery. Cystectomy, the surgical removal of the bladder, is a major procedure often associated with significant risks.

The goal of this challenge is to leverage data-driven insights to predict post-surgical complications, a crucial step in enhancing patient safety and improving the efficiency of healthcare services.

CPET is a standardized test conducted on an exercise bike to assess a patient's exercise capacity before surgery [7]. The test begins with a 3-minute 'rest stage,' during which the patient remains stationary on the bike. Next, the patient performs 3 minutes of resistance-free pedaling (unloading phase). This is followed by an incremental phase, during which the work rate gradually increases until the patient's tolerance limit is reached, culminating in a recovery phase. The test typically lasts between 8 and 12 minutes, depending on when the patient's tolerance limit is achieved.

Throughout the CPET session, multiple physiological metrics are collected. Various pieces of equipment measure values such as O_2 and CO_2 flow, SpO_2, heart rate, and blood pressure (a full list can be found in [7]). Together, these metrics provide a comprehensive assessment of a patient's exercise capacity.

In this challenge, we aim to develop binary predictive models that forecast post-surgical complications using CPET results. Specifically, we will focus on predicting respiratory, cardiovascular, and infection-related morbidities on days 3 and 5 post-surgery. Given the exploratory nature of Data Study Groups, particular emphasis will be placed on evaluating a wide range of algorithms.

Furthermore, participants are encouraged to consider the explainability and interpretability of their models to foster trust between healthcare professionals and the predictive systems.

This challenge aligns with UCLH's mission to advance personalized medicine by tailoring treatments and interventions to individual patient data, ultimately improving outcomes and optimizing resource allocation within the healthcare system.

In this report, we present a comprehensive overview of the dataset, which comprises anonymized CPET data from 146 patients who underwent cystectomy surgery. The dataset includes high-resolution time-series data capturing detailed cardiopulmonary responses during preoperative testing, as well as binary outcomes indicating the occurrence of post-surgical complications. We also describe the machine learning (ML) methods applied, including preprocessing techniques to address noise and missing data, model selection strategies to optimize performance, and feature engineering approaches designed to enhance predictive accuracy.

Finally, we present the results of our study and discuss the reliability, limitations, and potential impact of the findings.

2.2 DSG Objectives

The challenge questions are:

- Can physiological data from CPET accurately predict post-surgical complications after cystectomy using modern machine learning techniques?
- What machine learning algorithms are most effective for predicting different types of complications (pulmonary, cardiovascular, infections)?
- How can models be designed to balance accuracy with interpretability and uncertainty quantification for medical applications?

3 Data Overview

3.1 Data Description

The dataset comprises time series data from CPET conducted on 148 patients before undergoing cystectomy surgery, along with metadata that contains post-surgical outcomes. In this study we focus on three post-surgical complications, respiratory, cardiovascular, and infection-related morbidities. This dataset was created through routine CPET tests and postoperative assessments conducted at UCLH. Demographic information and duration of hospital stay were also provided for each patient.

The outcomes were labeled in the dataset to indicate whether patients developed complications following cystectomy surgery. The determination of whether a patient developed complications is based on the Post-Operative Morbidity Score (POMS). POMS is a clinical scoring system used to assess whether patients experience specific types of complications, including respiratory, cardiovascular, renal, pain, and infection-related complications. These scores were recorded on days 3 and 5 post-surgery and serve as the primary indicator of the presence (1) or absence (0) of post-surgical complications (Table 1).

The dataset underwent a pseudonymization procedure in order to reduce the risk of personal healthcare data being released into the public domain and to provide more flexibility in terms of data handling. This procedure consisted of removing any demographic data apart from gender and an age bracket. Age bracket specifies an age range that each patient falls into rather than giving an exact age.

3.2 Data Source

The dataset is consolidated and curated at University College London Hospital (UCLH) and maintained in a secure Data Safe Haven (DSH) facility. Given the sensitivity of the data, it was securely transferred to the Turing Research Environment (TRE) for the Data Study Group (DSG). The time series data was collected during CPET assessments, with metadata recorded at different stages of the patient's hospital care at UCLH before and after the cystectomy surgery.

3.3 Data Structure

The dataset is divided into two main components:

CPET time series data: This component of the dataset consists of 17 key physiological variables recorded during the CPET in 1 second intervals, capturing comprehensive measurements of the patient's cardiopulmonary function. These variables include oxygen uptake (VO2), carbon dioxide production (VCO2), heart rate (HR), and ventilation (V'E), among others, offering detailed insights into the patient's respiratory and cardiovascular performance. The data is organised as a time series, covering the distinct phases of the CPET: rest, exercise, and recovery. For each of the 148 patients, these physiological variables were continuously recorded throughout the test, creating a rich dataset that reflects the dynamic physiological changes occurring in response to varying levels of exertion.

The following variables are included in the CPET time series data:

- Timestamp (hh:mm:ss): Time at which the measurement was taken during the CPET.
- Phase: Stage of the CPET test (rest, exercise, recovery).
- Ventilation (V'E): Minute ventilation (L/min).
- PetO2 and PetCO2: End-tidal partial pressures of oxygen and carbon dioxide, respectively.
- Oxygen uptake (VO2) and Carbon dioxide production (VCO2): Measures of oxygen consumption and CO2 production during the test.
- Respiratory exchange ratio (RER): Ratio of CO2 production to oxygen uptake.
- Heart rate (HR): Heart rate in beats per minute.
- Breathing frequency (BF) and Tidal Volume (VT): Number of breaths per minute and the volume of air per breath.
- VO2 per heart beat (V'O2/HR): Oxygen consumption per heartbeat.
- Work rate (WR): Rate of physical work performed during exercise.
- Oxygen Saturation (SpO2): Percentage of oxygen saturation in the blood.

Patient metadata and complication outcomes: This component includes demographic and clinical outcome information for the 148 patients. The metadata includes age bracket, gender, type of surgery, hospital

admission/surgery/discharge dates. The POMS data for respiratory, cardiovascular and infection morbidities on days 3 and 5 post-surgery are presented as binary indicators, where each entry denotes the presence or absence of specific complication. A value of '1' indicates the occurrence of a complication, while '0' signifies no complication. This data serves as the key outcome measure to assess the patients' recovery.

The dataset includes the following metadata and outcome variables:

- Post-Operative Morbidity Score (POMS) data for days 3 and 5, indicating respiratory, cardiovascular, and infection-related complications.
- Age bracket of the patient, for example, 25-40.
- Gender (e.g., Male, Female).
- Type of cystectomy surgery (e.g., robotic, laparoscopy-assisted).
- Dates of hospital admission, surgery, and discharge.

3.4 Data Quality

Overall, the data was typical of real world health data: noisy, missing values and containing demographic and class imbalances. The dataset also had a low number of samples; 148 samples is much smaller than what is typically used in real world machine learning applications.

Appropriateness The dataset was very relevant to the questions posed. It provided very fine grained time series data from the CPET test, which was then used to predict the binary outcomes provided by the POMS data.

Readiness Apart from a few missing values for very particular variables in the CPET time series' (detailed in Section 3.5) the data was mostly complete. It was found that there was 1 patient with CPET data but no POMS outcomes and there were 3 patients with POMS outcomes and no CPET data; the respective data for these patients was removed and not included in the analysis. There was no documentation provided with this data but its content is self explanatory.

Reliability/Bias The dataset had a significant class imbalance issue as illustrated in Table 1 and Figure 1. For all morbidities the number of non-morbidity patients

outnumbers those with a morbidity. However, the difference is most extreme for the cardiovascular morbidities where very few patients have CVS complications. Even though oversampling can be performed to rebalance the data, this initial imbalance can still negatively affect generalisation performance on unseen data.

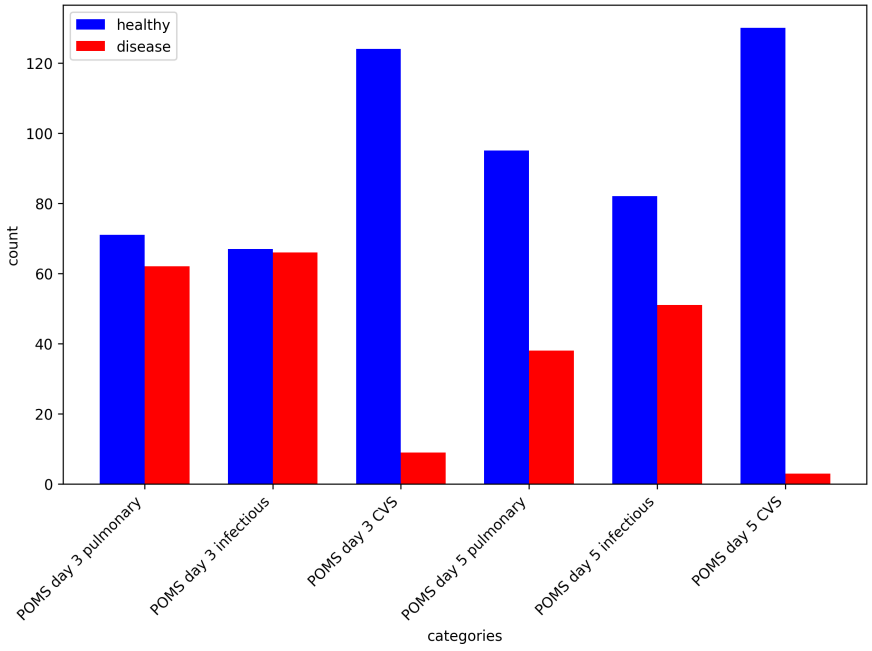


Figure 1: Number of patients with (red) and without (blue) morbidities for day 3 and day 5 post surgery.

It is unknown whether the data represents a diverse set of demographics due to the absence of most of the demographic information. With regards to gender there was a significantly larger number of males represented in the data (99) compared to females (49). The age range split is as follows: $[30, 58] : 23$, $[58, 68] : 32$, $[68, 78] : 56$, $[78, 100] : 37$. The uneven data splits for both gender and age groups will create bias in the final models with a greater generalisation capability expected for the majority classes.

There was found to be a large amount of noise on the CPET data. This noise was likely introduced as the result of inaccurate sensors and therefore could only be remediated by more accurate sensors. Noise reduction filters were applied to the data but this is limited in its effectiveness.

Table 1: Number of patients in the dataset that have and do not have the specified morbidity (complication type).

Complication Type	Number of Patients	
Pul (Day 3)	Healthy	78
	Diseased	70
Inf (Day 3)	Healthy	75
	Diseased	73
CVS (Day 3)	Healthy	138
	Diseased	10
Pul (Day 5)	Healthy	106
	Diseased	42
Inf (Day 5)	Heatlhy	93
	Diseased	55
CVS (Day 5)	Healthy	144
	Diseased	4

Finally, the data was collected between 5 and 7 years ago which might introduce a source of bias. There might be population differences today compared to 5 years ago meaning that the models trained on this old data might not generalise well to patients today. There might also be improvements in medical care compared to 5 years ago meaning that morbidities that would have occurred in certain circumstances then might not occur today, resulting in a larger number of false positives if using the model today.

Sensitivity: Was the data private or confidential? In its original form the data was highly confidential given the fact that it contained personal health information. Subsequent to the pseudonymization procedure detailed in 3.1 the dataset was much less confidential but still sensitive. It was for this reason that it was contained within the Turing Safe Haven.

Sufficiency The dataset has a low number of samples (148) and is therefore not particularly sufficient. This is typical in medical-related settings where willing participants are difficult to come by and sensitive data is hard to collect. Despite this, machine learning techniques, in particular Deep Learning techniques, work better with a sample size larger than 148.

3.5 Data Preprocessing

Effective data preprocessing is an important factor in machine learning, especially when working with real world noisy time series data. Several preprocessing techniques were applied to ensure the dataset was ready for analysis, addressing issues such as normalisation, missing values, and categorical data.

Combining Patient Files The dataset was provided as a single Excel spreadsheet with outcomes and demographic data for *all* patients, and a single Excel sheet for *each* patient containing their CPET results. The outcomes and demographic spreadsheet was split into two, one for training and one for testing. Two single Excel spreadsheets (training and testing) were created by consolidating all of the CPET data - linked with the appropriate demographic and outcomes data via patient ID - into one table.

Standardization All continuous variables, such as heart rate (HR), oxygen uptake (VO2), and minute ventilation (V'E), were standardized to ensure they had a mean of zero and a standard deviation of one (z-score normalisation). Standardization is essential in machine learning because many algorithms, particularly those based on distance metrics or gradient descent, are sensitive to the scale of the input features. By standardizing the data, we ensured that all features contributed equally to model training, avoiding bias from features with larger numerical ranges.

One-Hot Encoding To handle categorical variables such as age group, the phase of the CPET test (rest, exercise, recovery), and type of surgery, a one-hot encoding was applied. This technique converts categorical variables into binary vectors, where each unique category is represented as a separate feature. For example, the type of surgery (robotic, laparoscopic-assisted, or open) was transformed into three binary features, allowing machine learning algorithms to interpret these non-numeric values more efficiently. One-hot encoding is particularly useful in machine learning, as it avoids imposing any ordinal relationships between categories.

Handling Missing Values The dataset contained missing values for certain variables, notably oxygen saturation (SpO2) and other physiological parameters. The SpO2 data contained a significant amount of missing values, marked by a

hyphen ('-'), which were replaced with `NaN`. A total of 59 patients had missing SpO2 data, with 34 of those missing most of their values.

Two approaches were adopted to handle these missing data points.

1. **Padding With Zeros** The first method explored was padding the time series data with zeros to fill in the missing values. Padding is a simple and widely used technique for handling the missing data in time series, especially when the length of the series vary between instances. By filling missing time steps with zeros, we ensured that all time series data has the same length, making it easier for algorithms to process. However, its important to note that this method can introduce distortions, as padding zeros may misrepresent the underlying physiological patterns.
2. **MINIROCKET** Minimally Random Convolutional Kernel Transform (MINIROCKET) [4] is a powerful preprocessing technique designed for time series classification. It works by transforming the time series data into a feature space that can be directly used by models. This has been proven to handle variable length time series effectively and can also deal with missing values by creating feature maps that are invariant to the exact length of the series. This method was favoured because it allowed us to retain as much meaningful information as possible from the time series data, without relying on artificial padding methods.

Smoothing The raw physiological data often exhibited sharp spikes or peaks, which could be attributed to sudden physiological changes or sensor error during the CPET sessions due to machine displacement etc. While these peaks may carry important information, they mostly introduce noise or outliers, potentially confusing models.

To address this, we applied smoothing techniques to reduce the impact of this noise. A moving window average with a window size of 10 was applied to all continuous features. Smoothing is a common technique used in time series analysis to create a more stable and continuous dataset by reducing random fluctuations. By applying smoothing, we were able to capture the underlying trends in the data while minimizing the noise introduced by sudden peaks. This helped the model focus on the more consistent pattern across the dataset rather than being misled by outliers or extreme values.

Figure 2 shows time series data for various physiological variables during a CPET

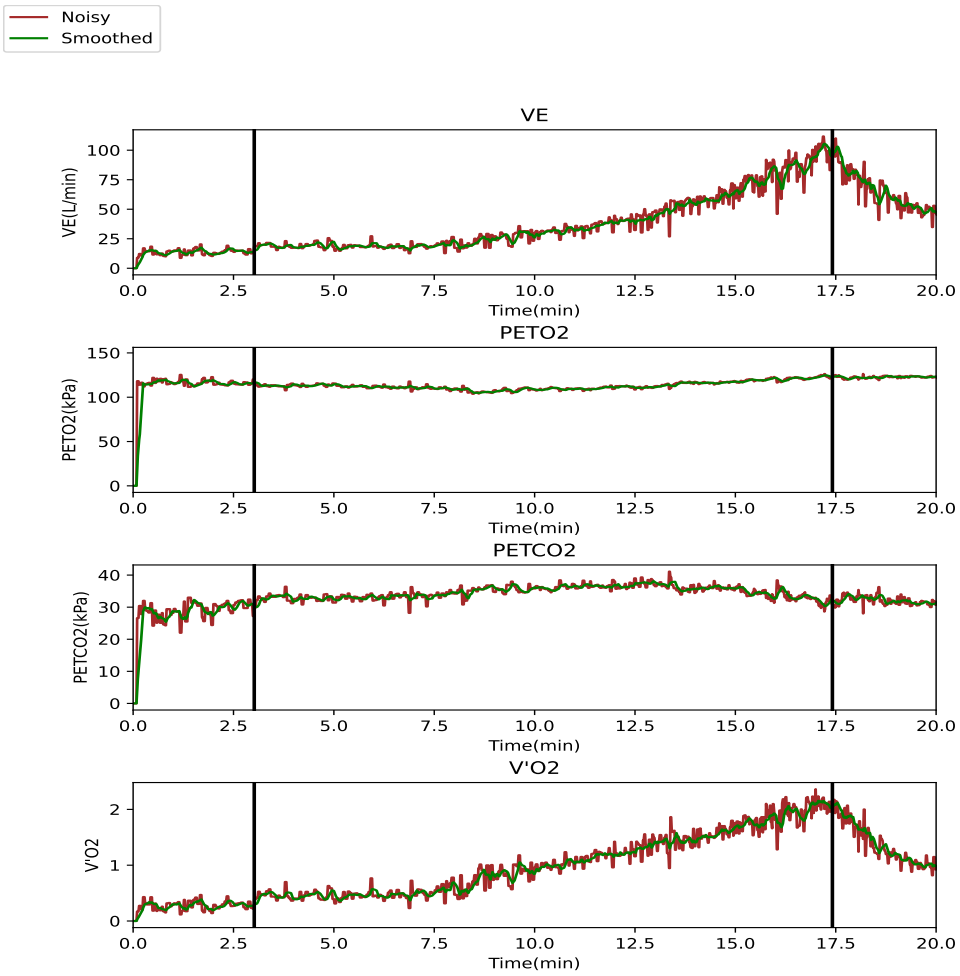


Figure 2: Moving average smoothing applied to four of the CPET variables. The red lines show the raw data and the green lines show the data after smoothing. Vertical lines delineate the rest, exercise, and recovery phases.

test for a single patient, with noisy data (red) and smoothed data (green) for each variable. It can be observed that, for VE, the raw data shows sharp peaks during the exercise phase, especially around the 15-minute mark. The smoothed data effectively removes these abrupt spikes while maintaining the gradual increase and subsequent decrease in ventilation during exercise. This approach ensures that the models built from this data are more likely to generalize well, as they avoid fitting to irrelevant noise or outliers.

Additional Stage In addition to the three stages, rest, exercise and recovery, a distinct sub-stage called "Unload" where the patients pedaled without resistance for three minutes was introduced. This was shown to improve the calculation of scores, such as peak heart rate and time to peak. This additional column (*phase with unl*) was introduced between the 'rest' stage and the 'exercise' phase.

3.6 Feature Engineering

Several features were engineered from the CPET data to enhance the predictive capacity of the machine learning models. Based on recommendations from the reviewed literature [13, 11, 10, 9] the following features were derived from the smoothed time-series variables, separately for each phase:

- Event duration (in seconds).
- Peak values for heart rate (HR) and V'O2.
- Time to peak HR.
- Slopes for HR, V'O2, V'CO2, V'E, V'E/V'CO2: calculated by fitting a regression line on the time series data of each patient.

Additionally, heart rate recovery (HRR) was calculated by subtracting the heart rate measured one minute into the recovery phase from the peak heart rate recorded during exercise. These features contributed to a total of 37 engineered features derived from the time series data - 9 for each of the above features, for each of the 4 phases, plus 1 for HRR.

Despite the richness of these features, none of them, when considered individually, provided linear separation for the outcome variables related to post-surgical complications. For instance, Figure 3 illustrates the distribution of some features across two patient groups: those with and without pulmonary

complications on day 5. Although clear diagnostic separation cannot be observed, certain features, such as peak heart rate during exercise, exhibited variations in their distributions that may offer diagnostic value for machine learning models.

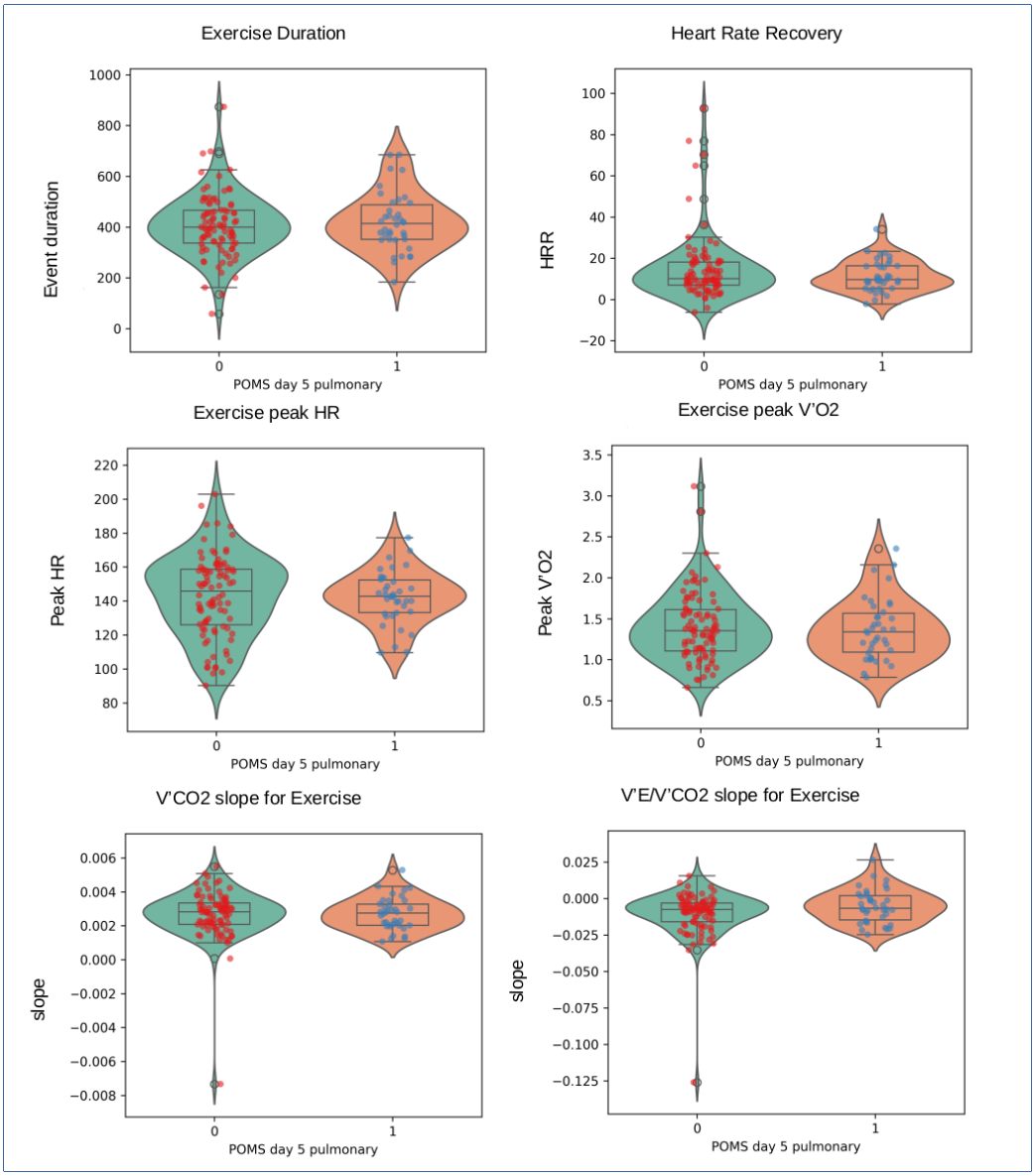


Figure 3: A selection of engineered features plotted separately, focusing on patients with and without pulmonary complications on day 5. Each dot in the plot represents an individual patient, highlighting the distribution of features across the two groups.

The complete set of 37 engineered features was subsequently provided to the

machine learning team as candidate input variables for model development. Despite the presence of outliers in some variables, such as the V'CO2 slope during exercise, no outlier removal was applied at this stage.

3.7 Granger Causality Analysis

Granger Causality (GC) [5] is a technique used to assess whether one time series can predict another, thereby establishing causal relationships between variables. In this study, GC was applied to analyse the CPET time series data, focusing on the relationships between features to enhance model performance and explainability.

We applied GC analysis to the multivariate autoregressive (MVAR) time series derived from the raw data. After normalizing sliding windows for each subject and phase, we calculated GC p-values using model parameters such as window size, overlap, and maxlag. The model order was determined using AIC. Low p-values indicate strong causality from the variable A to B, while high p-values suggest no causal relationship.

Figures 4, 5, and 6 show the GC value evolution across all subjects during rest, exercise, and recovery phases. It can be seen that the casual relationship between some of the features has large variability depending on the phase the subjects are in. Although the computational constraints limited detailed analysis, the preliminary results show that GC provides valuable insights into feature relationships.

The derived p-values can be used to help analyse the relevance of some of the features so we can decide which features are more important and determine whether there are any features that could be removed. They can also be used instead of the raw time series as input features to the predictive models.

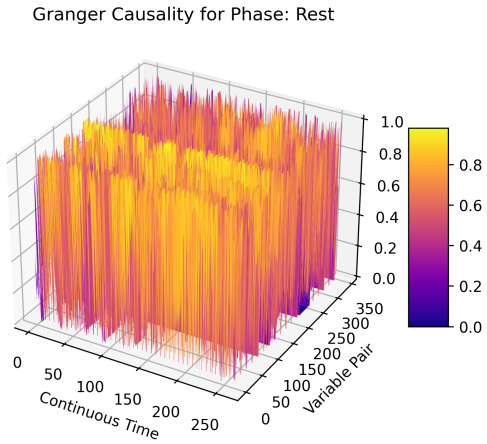


Figure 4: GC values over time for the rest phase. The color scale represents GC values on the Z-axis.

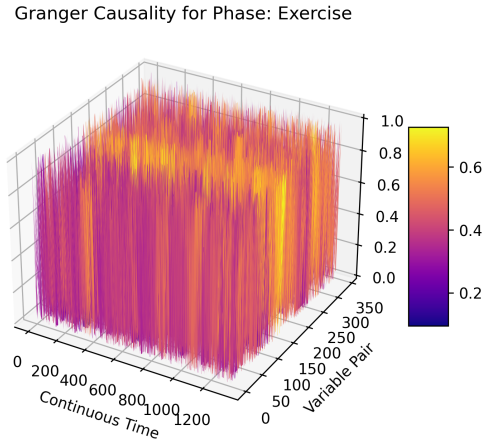


Figure 5: GC values over time for the exercise phase. The color scale represents GC values on the Z-axis.

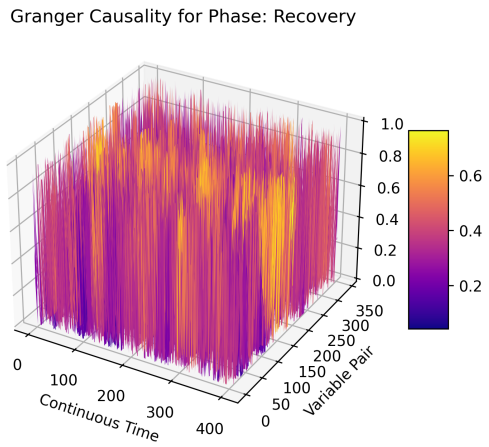


Figure 6: GC values over time for the recovery phase. The color scale represents GC values on the Z-axis.

4 Approach

We used two distinct approaches to build the binary classifiers, traditional machine learning methods and deep learning techniques. Traditional machine learning (ML) models refer to algorithms and techniques that rely on manually derived features and structured datasets for predictive tasks. In this study, key attributes such as peak heart rate and oxygen uptake were extracted from the raw data using domain-specific knowledge, and these feature-engineered variables (as detailed in Section 3.6), along with summary statistics, were employed to classify binary outcomes effectively. These features are expected to enhance the model's ability to focus on the most relevant information, thereby optimizing overall performance. Traditional methods in general, are effective in scenarios where derived features help in improving both model interpretability and explainability.

In contrast, Deep Learning models, (e.g. LSTM networks) were applied directly to the raw time series data from the CPET tests. These models do not typically require manual feature selection, as they are capable of automatically learning hierarchical patterns from the data. LSTMs are particularly effective for capturing long-term dependencies in sequential data, making them well-suited for time series analysis, while 1D CNNs excel at detecting localized temporal patterns. Although DL models perform exceptionally well on large, high-dimensional datasets, they typically require more computational resources and larger datasets for optimal performance.

Evaluation Metrics Models were compared based on accuracy, and for certain models, F1 score, using the same subset of data consisting of 16 patients that were held out for testing. Evaluation on the held out test data ensured that the models were evaluated on unseen data and prevented overfitting. Accuracy and F1 score are evaluation metrics often used to determine the quality of a binary classifier with the following definitions:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$
$$Precision = \frac{TP}{TP + FP}$$
$$Recall = \frac{TP}{TP + FN}$$
$$F_1Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

where TP = true positives, FP = false positives, TN = true negatives, and FN = false negatives.

F1 score makes up for certain failings of accuracy in determining the quality of classifiers with large class imbalances. For example, a classifier that guesses the positive class for all test instances on a dataset where 99% of instances are of the positive class still achieves a 99% accuracy. This accuracy might make the classifier sound impressive but if the negative class indicates the presence of a disease, then all patients with the disease are classified incorrectly, which can have devastating consequences. F1 score is important for incorporating these false positives into the evaluation metric.

4.1 Machine Learning

Binary classifiers were built using 6 different types of model. Each model was selected based on its suitability for the structure of the data and the complexity of the problem.

RidgeClassifierCV RidgeClassifierCV is a linear classifier that uses regularization (ridge regression) to prevent overfitting. It is well-suited for high dimensional data where multicollinearity between features is a concern, as it penalizes large coefficients and ensures the model generalises better. For this study, we used a range of alpha values defined by the logarithmic space between 10^{-3} and 10^{3}. This range allows the model to explore a wide spectrum of regularization strengths. Additionally, the RidgeClassifierCV evaluates each of these alpha values to find the one that strikes the best balance between underfitting and overfitting. This automatic selection of alpha through cross-validation improves the model's performance by the fine-tuning this key

hyperparameter, leading to a more generalized solution that performs well on unseen data.

XGBoost XGBoost is a powerful boosting algorithm that builds an ensemble of weak learners (decision trees) to create a strong predictive model. It applies gradient boosting technique to optimize model performance iteratively by minimizing the errors made by previous trees. XGBoost was ran here with $n\_estimators = 100$, which defines the number of boosting rounds or trees built during the training process. XGBoost is widely known for its robustness and high performance on structured data, particularly when dealing with large datasets and imbalanced classes. The algorithm's use of regularization and advanced techniques such as handling missing values and tree pruning make it particularly effective for binary classification tasks like those in this study.

Random Forest Random Forest is an ensemble learning method that constructs multiple decision trees using different subsets of the data. Each tree is trained on a random subset of features and samples, and the final classification is made by averaging the outputs of all trees. This method reduces the risk of overfitting and improves model robustness, especially in cases where the dataset is noisy or contains outliers. Random Forest is typically well-suited for datasets with a mix of numerical and categorical variables and can handle the noise present in patient data.

ROCKET (RandOm Convolutional KErnel Transform) ROCKET [3] is a method specifically designed for time series classification. It works by applying a large number of random convolutional kernels to the time series data and using the resulting transformations as input features to a linear classifier. ROCKET was chosen for this study due to its ability to efficiently handle high-dimensional time series data, with zero need for feature engineering. Its computationally efficient and scalable nature make it particularly well-suited for handling large volumes of time series data, such as the CPET tests, where capturing temporal patterns is critical. The number of kernels used for these experiments was 500.

Multi-Output Classification with Ensemble Learning Technique (MOCEL) This MOCEL model is an advanced machine learning approach that combines multi-output classification with an ensemble learning technique to address

complex prediction tasks where multiple target variables need to be predicted simultaneously.

The intuition behind the multi-output classification is that, unlike traditional approaches where the classification model predicts single target variable in a single prediction, multi-output classifier predicts multiple output for each instance without using multiple models. This approached reduces computational complexity. The rationality behind the ensemble learning technique is that it improves the accuracy of predictions through a weighted average of the individual model predictions, thus providing an improved overall performance.

In the MOCEL architecture the multi-output classifier is ensembled to enhance prediction power and the robustness of model. Figure 7 shows that in level 0 the base models with default parameters comprises of Random Forest, XGBoost and a Support Vector Classifier. Each of these base models are trained independently. In Level 1, XGBoost is used to ensemble and then used as a meta model. During the training process base models make predictions on training data and the output from them is used as input features for the meta model to learn and output prediction results.

This stacking approach captures the temporal pattern learned by each base model and the meta model will learn to weigh these pattern and weights appropriately. Since time series data can be noisy and subject to fluctuation, stacking makes it robust to handling non-stationarity in time series data. The entire architecture is presented in Figure 7.

k-NN Classifier To verify the use of GC values as features, a KNN classifier ($k = 10$) was trained using the GC p-values, validated using 10-fold cross-validation.

Each of these algorithms was chosen based on its strengths in handling different aspects of the dataset, including time series features, high collinearity, regularization ability, and scalability. By applying a diverse set of algorithms, we aimed to identify the best performing model for predicting post-operative complications using CPET data.

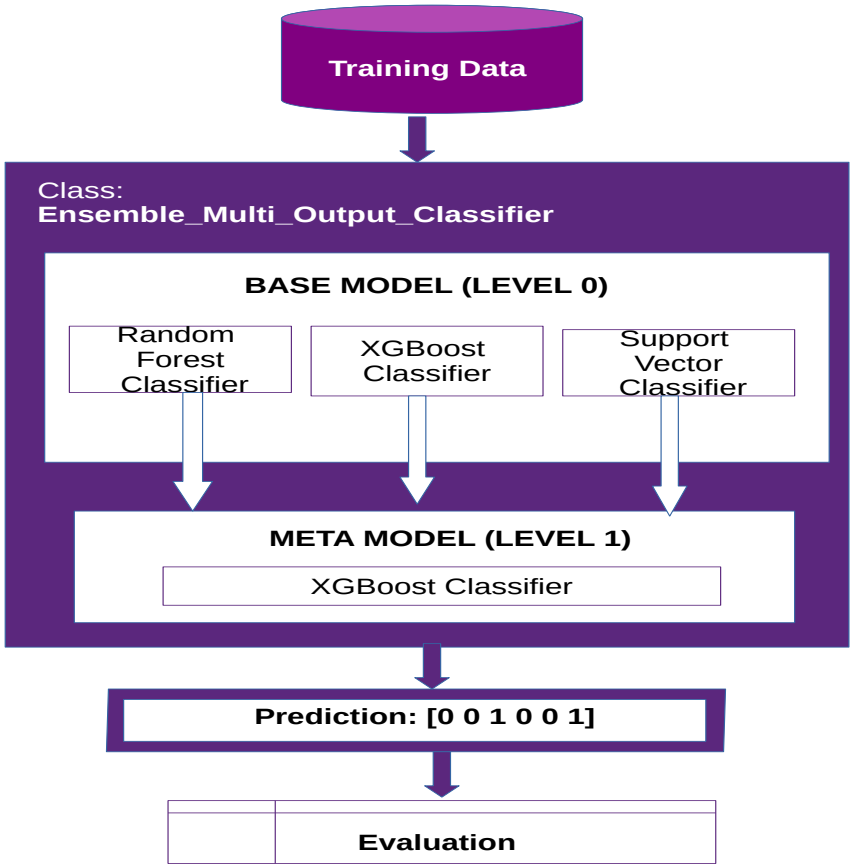


Figure 7: MOCEL architecture

4.2 Deep Learning

Deep learning (DL) has the ability to classify raw data without the need for feature extraction, saving time and pre-processing steps. This subsection explores the specific pre-processing applied to the data with regards to the DL approaches and how the DL models were developed.

4.2.1 Convolutional Neural Network

A 1-dimensional Convolutional Neural Network (CNN) was first explored because it has been shown to perform well on time-series data [4, 15], including on CPET data [14, 1]. A CNN is a type of neural network that applies multiple filters, or kernels, to its input. Each kernel is repeatedly applied to the input in small windows, in this way the number of network weights is much less compared to a multilayered neural network. Previous usage of CNNs demonstrates how different kernels learn to detect different features, such as lines and corners in images. CNNs are particularly suited to processing data that can be considered grid-like, examples include images, time series (1 dimensional grid), and video.

One-hot Encoding Through initial exploration of the data, it was found that there were challenges related to the combination of time-series and categorical data. The categorical data includes features that could be beneficial for the prediction, such as age bracket and gender. Therefore, each of the categorical data was one-hot encoded so that it could be fed into the DL models.

Oversampling After initial exploration of the data, it was discovered that for two of the classification challenges - Day 3 CVS and Day 5 CVS - the class imbalance was extremely high with the vast majority of patients lacking the morbidity (Table 1). The lack of balanced training data makes it challenging to train DL models; therefore, the data was oversampled in the CNN experiments in order to create new synthetic data for the minority class (positive cases of CVS).

The SMOTE package was used for oversampling. SMOTE [2] works by creating new values that are similar to the existing data by not only duplicating existing data but also creating new values that are close to the minority class. The synthetic data is created by randomly selecting one or more K-nearest neighbours for each sample in the minority class. Then a line between the current data point and a selected neighbour is drawn and a new point is randomly selected along that line.

This process is repeated and the new samples are added to the dataset.

Normalisation For the CNN experiments the time-series data was normalised using z-score normalisation. Z-score normalisation transforms the data in a way that helps the model generalise across different features. As CPET data contains different measurements with different scales it ensures each feature has a mean of 0 and a standard deviation of 1. Furthermore, z-score normalisation helps adjust for differences in scale between the train and test dataset which other normalisation techniques may not account for.

Windowing When using convolutional deep learning approaches for time-series data it is common to split the data into windows of samples that includes an overlap between the windows. This enables the more efficient extraction of temporal features from the data because time-series data patterns may not appear in a limited number of time-steps. Furthermore, overlap also enables more training data which is useful for datasets such as those in the medical domain that can often be limited.

Windowing the data should also help the convolutional model to generalise better by exposing the model to different sequences of the data to learn representative features rather than the model trying to fit to the entire data sequence. A range of window and overlap sizes were manually explored ranging from small windows of 20 samples to large windows of 600 samples with overlaps of 20% to 80% explored. It was found that a window size of 200 with a 50% overlap achieved the best accuracy on the validation sets.

Validation A key consideration when evaluating the model performance is the validation approach. When developing health-care models it is vital to test on a per-participant basis as when using a traditional test-train split it is possible to achieve a higher accuracy but this is using data from all participants which does not simulate real-world scenarios. Instead the models have been evaluated per-participant either by excluding a small number of participants from the training data which are then only used for testing or used for k-fold cross-validation. K-fold cross-validation was the chosen approach due to its capability to test over the entire dataset removing any subjectivity of the selected test split. 5 folds were used to evaluate validation performance which balanced the benefits of k-fold cross-validation with increased training time.

Architectures

CNN A CNN was explored using the entire raw CPET dataset consisting of all time-series features. The data was windowed with a window size of 200 and an overlap of 100. A single label consisting of the mean label assigned to each window was used for training.

A number of 1D CNN architectures were explored ranging from 2 to 6 convolutional layers with each followed by a max pooling layer. Additional layers were explored to reduce model overfitting including dropout layers and batch normalisation layers, these were similarly included after each convolutional block. After the convolutional layers both global average pooling and dense layers were explored.

A range of convolutional filter quantities and kernel sizes were explored resulting in an optimal final selection of 32 filters and a kernel size of 3. Similarly, the max pooling size was selected as 3. For the dropout layers a range of percentage values were explored for the number of neurons to selectively ignore with values increasing for the later dropout layers starting with 0.3 to 0.5. A He Normal initialiser was used to initialise the weights of the kernels.

Additional hyperparameters were also explored within the CNN to prevent overfitting, including dilation rate which is the spacing between elements within a filter and kernel regularization including both L1 and L2 regularization at 0.005.

An example CNN model architecture can be seen in Figure 8.

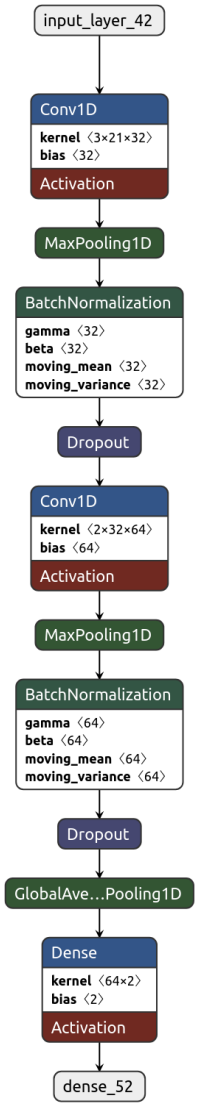


Figure 8: An example convolutional architecture

CNN-LSTM Hybrid Another model architecture explored was a hybrid CNN-LSTM to combine the strengths of both a convolutional architecture in capturing local, short-term features and a recurrent architecture that should better capture long-term temporal dependencies. The aim of this architecture was to improve the generalisability of the model by using a more substantial architecture consisting

of 5 convolutional layers and 3 LSTM layers - exploring both 32 and 64 units. The model architecture can be seen in Figure 9. Similar to the 1D CNN, all of the raw features from the dataset were used as input to the model. The same window size of 200 and overlap of 100 was also used.

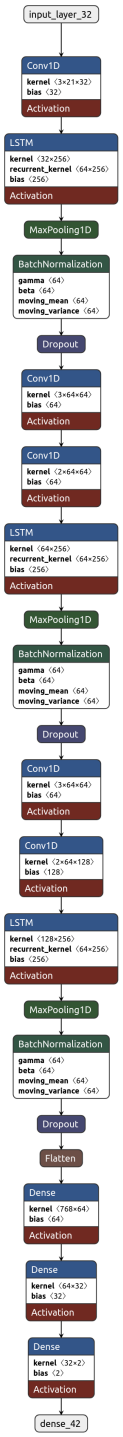


Figure 9: Combined CNN-LSTM network architecture

4.2.2 Long Short-Term Memory

Recurrent Neural Networks (RNN), such as Long Short-Term Memory (LSTM) [6], are commonly used in conjunction with time-series data due to their inherit capability of modelling long term dependencies in sequential data. An RNN is a neural network with a cycle that takes a sequence of data as input. The cycle in the network allows the network to remember information from previous time steps and use this information at the current time step.

Traditional RNNs struggle to learn long-term dependencies due to issues with vanishing and exploding gradients. LSTMs were invented as a way to circumvent the vanishing/exploding gradient issues by including a gating mechanism that has the ability to keep information stored for longer periods of time. Due to their success at learning long-term dependencies in time-series data, LSTMs were utilised for this task.

Preprocessing In order to standardise the number of time steps between patients the data was downsized to the minimum number of time steps of all the patients. The minimum number of time steps over all patient was 499. This downsizing was done by randomly (without replacement) selecting 499 data points from all the time steps of each patient. Then, non-numeric features were removed, such as 't', 'Phase', 'person_id', 'Type_of_surgery', 'Gender' and 'Age'. The training validation split was set at 80:20.

Architecture The architecture chosen consisted of 3 LSTM layers of size 64 each followed by dropout with a rate of 0.2. A dense layer of size 64 with ReLU activation was then appended, followed by another dropout layer with a rate of 0.2, and finally, a dense layer of size 1 with a sigmoid activation function was used as the last layer.

An ADAM optimiser with a learning rate of 1e-3 was used. Binary cross-entropy was used as a loss function.

4.2.3 Multilayer Perceptron

A multilayer perceptron (MLP) is a type of feed forward artificial neural network which consists of multiple layers of neurons. Each neuron within each layer is densely connected to the next layer. This type of architecture was also applied to this task.

Traditional MLP Figure 10 shows the architecture of the traditional MLP model that was used. Data at each time step is treated as 1D vector where each element is considered a feature i.e. measurements in CPET. In this model, 130030 data samples are used (one for each time step of CPET for every patient) for training and validation with a train validation split of 80:20 with regards to different patients. Z-score normalization was applied to the input data before being fed into this model.

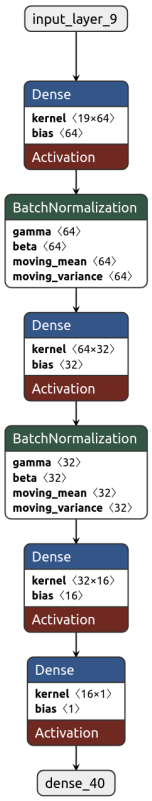


Figure 10: Traditional MLP architecture used during this task

Variational MLP In order to solve the overfitting problem that comes from the data variations between different participants, a variational MLP was built. A variational MLP aims to capture the data variations by reducing noise. Specifically, it maps the input data to a mean and variance that approximates a Gaussian distribution, which can be considered as normalization within the model layers.

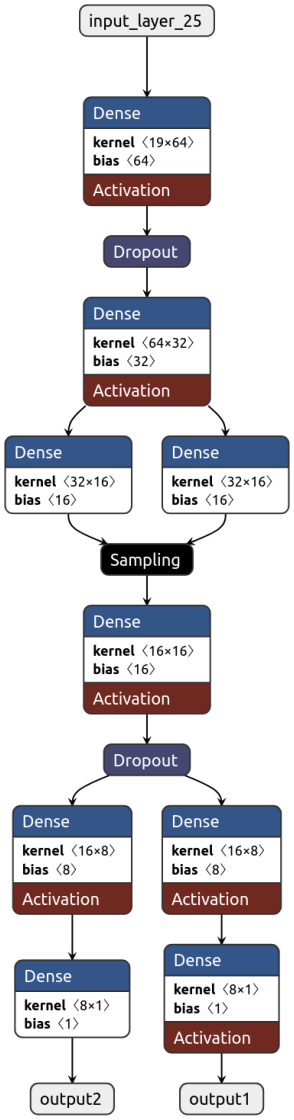


Figure 11: Variational MLP architecture

Dropout layers were included with a rate of 0.2 and so was l2 regularization with a rate of 0.02.

According to the Causal Inference theory, randomization can be used to remove the confounding variables - in this task, they are the participants. Therefore, a

multi-class neural network is built where the first classifier is designed to predict if it is positive/negative for each class and the second classifier is to force the model to be trained on randomized patient IDs. This can be considered as an additional regularization step. Architectural details are shown in Figure 11. Every numerical time series feature (excluding the categorical features) except SpO2 was used to train the model.

5 Results

5.1 Machine Learning

This section evaluates the performance of the applied machine learning models in predicting post-surgical complications. Results are presented in Table 2 for both day 3 and day 5 post-surgical outcomes for respiratory, cardiovascular and infection complications.

5.1.1 Day 3 Results

The first set of results focuses on predicting complications on Day 3 post-surgery. Performance varied across the models, highlighting the challenges posed by imbalanced data and the complexity of the CPET time series.

Ridge Classifier The Ridge Classifier achieved reasonable accuracy (0.60) when predicting pulmonary complications, but a poor F1 score (0.20) indicated difficulties in handling the class imbalance. In the infectious category, the Ridge Classifier performs better, with a validation accuracy of 0.55 and an F1 score of 0.40, suggesting it captures some positive cases. For CVS, the accuracy is very high (0.95), but the F1 score is consistently 0.0, indicating that the model is likely overfitting to the majority class with no correct predictions for the minority class. The model struggled to effectively identify patients with complications, likely favoring the majority class (those without complications).

Rocket Classifier The Rocket Classifier also struggles with the pulmonary classification, showing poor F1 scores despite similar accuracy levels as the Ridge Classifier. For infection cases, the accuracy and F1 score are slightly better on the test set (0.67 accuracy and 0.71 F1 score), indicating that it captures positive cases of infectious better than for pulmonary conditions. However, the

Table 2: Prediction results for Day 3 and Day 5 using various ML classifiers

Models	**Pulmonary**		**Infectious**		**CVS**	
	Accuracy	**F1 Score**	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**
Day 3						
Ridge Classifier						
Validation	0.60	0.20	0.55	0.40	0.95	0.00
Test	0.40	0.40	0.67	0.71	0.93	0.00
Rocket Classifier						
Validation	0.45	0.27	0.55	0.40	0.95	0.00
Test	**0.47**	0.49	0.67	0.71	0.93	0.00
XGBoost Classifier						
Validation	0.45	0.15	0.55	0.40	0.95	0.00
Test	0.40	0.40	**0.80**	**0.80**	0.93	0.00
Random Forest						
Validation	0.55	0.00	0.65	0.53	0.95	0.00
Test	**0.47**	**0.50**	0.60	0.66	0.93	0.00
Day 5						
Ridge Classifier						
Validation	0.60	0.20	0.55	0.40	0.95	0.00
Test	0.53	0.00	0.60	0.50	0.93	0.00
Rocket Classifier						
Validation	0.45	0.27	0.55	0.40	0.95	0.00
Test	**0.67**	0.00	0.60	0.50	0.93	0.00
XGBoost Classifier						
Validation	0.45	0.15	0.55	0.40	0.95	0.00
Test	0.60	0.00	0.73	0.50	0.93	0.00
Random Forest						
Validation	0.55	0.00	0.65	0.53	0.95	0.00
Test	0.60	0.00	**0.80**	**0.67**	0.93	0.00

CVS results again showed a mismatch between accuracy (0.95) and F1 score (0.00). Its ability to handle time series data may have contributed to its improved performance on pulmonary outcomes, although the class imbalance still remained a limiting factor.

XGBoost XGBoost underperformed in predicting pulmonary complications, showing lower accuracy and F1 scores, but achieved an accuracy of 0.8 in the infectious class. CVS results are the same as the previous two models.

Random Forest For pulmonary complications, Random Forest has reasonable validation accuracy (0.55) but has an F1 score of 0. However, it performs slightly better for the infectious class, with an F1 score of 0.53, suggesting some ability to capture positive cases. For CVS, again accuracy is high but the F1 score is 0.00, reflecting poor handling of positive cases.

5.1.2 Day 5 Results

The second set of results focuses on predicting complications on day 5 post-surgery. The models showed some improvement over Day 3, particularly for specific complications.

Ridge Classifier The model maintained similar accuracy for pulmonary and infectious complications, but continued to struggle with low F1 scores. This indicates that while it can correctly classify most patients, it has difficulty identifying the minority class of patients with complications.

Rocket Classifier Rocket Classifier showed improved accuracy in predicting pulmonary complications (0.67), though it still struggled to identify positive cases, as indicated by its modest F1 score. Infectious condition performance remains stable, with some improvement in accuracy (0.60) on the test set and non zero F1 score (0.50). The improvement suggests the model better captures the patterns in the time series, but the class imbalance remained a challenge.

XGBoost It shows slight improvement in accuracy for pulmonary classification on day 5, but the F1 score remained 0, indicating difficulty with positive predictions. For infectious classification, the F1 score remains stable (0.40-0.50) showing consistent but suboptimal performance.

Table 3: Average classification accuracy for each complication category using Granger Causality Features.

Classes	Day 3 Pulmonary	Day 3 Infectious	Day 3 CVS	Day 5 Pulmonary	Day 5 Infectious	Day 5 CVS
Avg. Accuracy	0.536	0.511	0.936	0.667	0.604	0.978

Random Forest This model achieved the best results for predicting infectious complications on Day 5, with the highest accuracy (0.80) and an F1 score of 0.67. Random Forest's ensemble nature likely helped capture complex interactions between features, improving performance for infection outcomes. However, the model's performance on other complications remained poor.

5.1.3 MOCEL Results

The MOCEL model, which employs a multi-output classification approach using ensemble learning, shows promise in improving validation accuracy across multiple outcomes. With more time, this could be demonstrated with metrics such as accuracy and F1 score. It struggled to achieve consistently high performance across all complication types, particularly for more complex outcomes such as cardiovascular issues. This suggests that while MOCEL's ensemble approach is effective for certain outcomes, further optimization may be needed to improve its overall performance.

5.1.4 Granger Causality KNN

As shown in Table 3, the GC-based features perform comparably to the other traditional machine learning algorithms (Table 2) on the test set. The confusion matrices in Figure 12 confirm these findings, suggesting that further refinement is needed to handle unbalanced datasets.

5.2 Deep Learning

5.2.1 Convolutional Neural Network

Given that the models were trained using windowed data, the test set was split into windows of the same dimension that were fed into the model which output a single classification output per time window. The results in Table 4 show the difference in accuracies between each of the folds of the validation data during training and the unseen test data for the class POMS day 3 pulmonary. The results show that

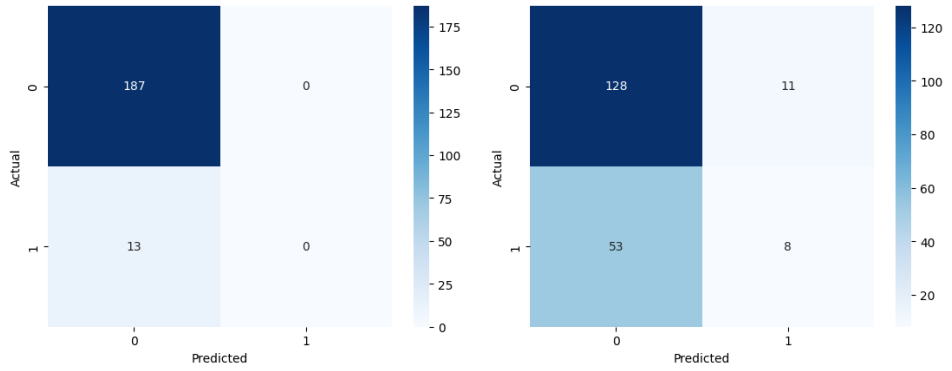


Figure 12: Left: Confusion matrix for Day 3 CVS class, $k-$fold cross-validation accuracy: 0.935. Right: Confusion matrix for Day 5 Pulmonary class, $k-$fold cross-validation accuracy: 0.68.

Table 4: Training accuracy results of the three convolutional models for each k-fold and final test accuracy for day 3 pulmonary.

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Test
CNN 2 layer	58.9%	**60.8%**	**58.5%**	**50.8%**	62.5%	**47.1%**
CNN 3 layer	**61.2%**	53.4%	56.9%	48.9%	55.6%	43.9%
CNN LSTM hybrid	58.1%	46.2%	52.3%	50.4%	**66.4%**	36.8%

there can be a large variation between folds showing the difference in data between participants has a significant impact on model performance. Furthermore, CNN networks with 2 and 3 convolutional layers were explored. The simpler 2 layer convolutional network generally performed best, possibly due to simpler models being less prone to overfitting. Therefore, the 2 layer model was used to evaluate performance on the test dataset across all 6 classes.

The results in Table 5 show the accuracies for both of the convolution CNN models as well as the hybrid CNN-LSTM model. The results show a variation between the classes with day 5 generally achieving higher classification accuracy than day 3. Furthermore, the simpler CNN network consistently outperforms the more complex CNN-LSTM model.

Table 5: Test accuracy results of the 2 layer CNN and the CNN-LSTM hybrid model for all 6 classes

Model	Pul (3)	Pul (5)	Inf (3)	Inf (5)	CVS (3)	CVS (5)
CNN	47.1%	**71.6%**	**60.6%**	**66.5%**	**74.2%**	**93.5%**
CNN-LSTM	47.1%	62.6%	59.4%	45.8%	60.6%	91.2%

5.2.2 LSTM

Table 6 shows the training and validation accuracies for each of the six classes for the pure LSTM model. The LSTM model was not tested on the testing data due to limited time.

Table 6: LSTM training and validation accuracies for each of the six binary outcomes

Target variable	Training accuracy	Validation accuracy
Day 5 CVS	97.51	97.54
Day 5 infectious	60.80	60.50
Day 5 pulmonary	71.07	70.30
Day 3 CVS	93.24	93.11
Day 3 infectious	49.11	49.37
Day 3 pulmonary	54.10	53.65
Average	70.97	70.75

5.2.3 Multilayer Perceptron

Table 7 shows the accuracy when predicting Day 3 pulmonary, Day 3 infectious, Day 5 pulmonary and Day 5 infectious for training, validation and testing for the variational MLP. The accuracy calculated here is per time step (i.e. per row), which means if you have 10 participants and each of them has 1000 time step as the input, then you have 10×1000 outcomes in total.

Figure 13 shows the confusion matrices for both Day 5 Infectious and Day 5 CVS. Both Day 3 CVS and Day 5 CVS show a significantly high accuracy which is over 90%. Figure 13 shows that although Day 5 CVS has a high accuracy it predicts that all data belongs to the positive class. This is likely due to the class imbalance that was *not* addressed using upsampling in this case. We therefore hypothesise

Table 7: Training, validation and testing accuracies for Day 3 pulmonary, Day 3 infectious, Day 5 pulmonary and Day 5 infectious outcomes for the variational MLP

Type	Pul (3)	Inf (3)	Pul (5)	Inf (5)
Train	0.6327	0.5390	0.7787	0.7495
Val	0.5386	0.5291	0.6387	0.7219
Test	0.5446	0.4345	0.7489	0.7395

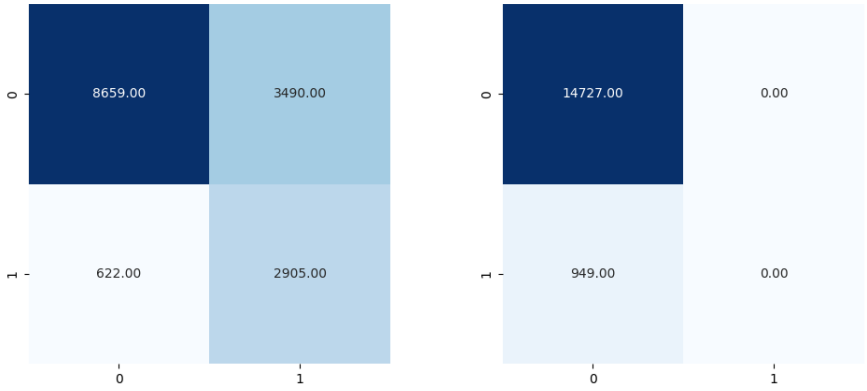


Figure 13: Confusion matrices. Left: Day 5 Infectious with 74% accuracy. Right: Day 5 CVS with 94% accuracy.

that upsampling would have encouraged the model to not predict all cases of Day 5 CVS as positive.

6 Discussion

6.1 Machine Learning

The results highlight the challenge of predicting post-surgical complications from CPET data, especially given the class imbalance in the data. Traditional machine learning models struggled with identifying patients who developed complications, particularly for minority outcomes such as cardiovascular issues. This reflects the limitations of relying on feature-engineered data and standard classification

algorithms for complex and imbalanced medical datasets. Despite this, XGBoost achieved an 80% test accuracy and F1 score on the day 3 infectious complication and the Random Forest acheived a test accuracy of 80% on the day 5 infectious complication.

Models designed specifically for time series data, such as the Rocket Classifier, demonstrated potential by improving accuracy for certain complications, such as pulmonary outcomes. However, the class imbalance continued to impact their overall effectiveness, as indicated by low F1 scores.

The standout performance of Random Forest in predicting infectious complications can be attributed to its ability to handle feature interactions and reduce overfitting through its ensemble structure. MOCEL also showed promise as a multi-output classification model, although further refinement is necessary to handle the more complex patterns in the data.

In conclusion, while these models provide valuable insights into the prediction of post-surgical complications, addressing class imbalance and improving the handling of time series data could further enhance their performance. Future work could focus on balancing techniques and more advanced feature extraction methods to better capture the complexity of the underlying physiological data.

6.2 Deep Learning

The main challenge faced with the development of the deep learning models has been overfitting. The training accuracy has consistently been high in comparison to the low validation accuracy showing that the model is overfitting. This is further demonstrated by the test results, highlighting the difficulty in developing a generalised model that works across new unseen patients.

We aimed to resolve this challenge by exploring techniques to reduce overfitting such as L1 and L2 regularization, dropout layers and kernel initialisers. These parameters have been tuned to help improve overall model performance but further hyperparameter tuning could be completed in the future to improve classification test performance.

Another challenge has been the class imbalance, in particular for the day 3 CVS and day 5 CVS. The training data was oversampled in the CNN experiments to create a 50% data balance between the two classes for both day 3 CVS and day 5 CVS. However, during testing the model accuracy for these classes was high, in

particular for CVS day 5 where both models achieved over 90%. When exploring the confusion matrices for these classifiers, as shown in Figure 14, the vast majority of predictions are for the negative class that contained the vast proportion of the data. Therefore, oversampling the training data had little impact on the model performance potentially due to the synthetic data being similar to the existing limited data which does not generalise to new patients.

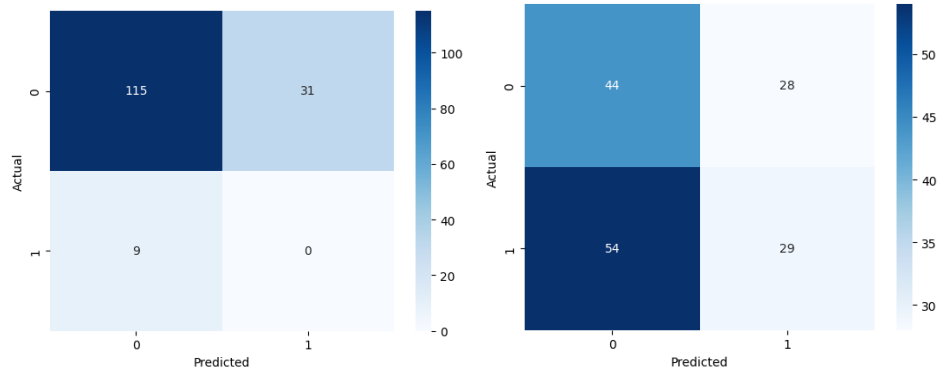


Figure 14: Confusion matrices for CNN model for day 3 CVS (left) and day 5 CVS (right)

Overall, the lack of training data has been the most significant challenge in developing deep learning classification models. The small dataset led to overfitting of models while the imbalanced data resulted in models predicting a single class. However, the classification models developed still show promise, in particular for day 5 pulmonary and day 5 infectious classification.

One interesting observation is the ability of the MLP to achieve testing accuracies greater than 70% despite the fact that the input consists of only one time step of information. This suggests that there is more predictive information in single items of data than was first expected and that ingesting the entire time series might not be necessary.

6.3 Limitations

The primary limitations of the data have been its limited availability and class imbalances.

For all algorithms, there was insufficient time during the Data Study Group for comprehensive hyperparameter optimization. As a result, the chosen hyperparameters may not be optimal, potentially leading to suboptimal test accuracies. Deep learning techniques, in particular, exhibited signs of overfitting, suggesting that they could benefit from further hyperparameter tuning. The time constraints also affected code quality, increasing the likelihood of bugs and suboptimal performance.

A key limitation of the study was the inconsistent calculation of test accuracies across different algorithms. This discrepancy arose due to a misunderstanding among study group members regarding the best method for calculating test accuracy. For example, the MLP model predicted an outcome for each time step of each patient, whereas the CNN model generated predictions for each window. However, test accuracies for all machine learning algorithms were calculated using the same approach, making them directly comparable to each other. This inconsistency in accuracy calculations means that we cannot directly compare the performance of ML models with DL models or even within the DL models themselves. As a result, no definitive conclusions can be drawn about the relative generalization ability of the different algorithms on the test set.

Additionally, we lack a benchmark for evaluating the models produced in this Data Study Group. A useful benchmark could be the accuracy of previous risk models or the typical predictive performance of a doctor on this task. Such benchmarks would provide insight into whether the models developed in this study have practical utility.

Another limitation of the study is the lack of explainability in most of the models. Deep learning models, in particular, function as black boxes, providing no clear indication of how they arrive at their decisions. Some machine learning models, however, offer greater interpretability. For example, the Ridge Classifier is a linear model, meaning that its coefficients provide a quantifiable measure of each variable's contribution to the output.

7 Conclusion

In this DSG, we sought to determine whether post-surgical complications could be accurately predicted using CPET data and modern machine learning techniques.

The results indicate that, in general, the models performed moderately well on the held-out test data, achieving accuracies between 0.5 and 0.75 for the respiratory and infection complication classes. Most models attained an accuracy above 0.9 for the CVS class; however, this was primarily due to class imbalance, meaning that a relatively high accuracy could be achieved simply by predicting the majority class in all cases. Nevertheless, it was challenging to determine whether these accuracies were sufficiently high for practical use, given the lack of benchmarks.

We also aimed to compare a wide range of machine learning algorithms to identify the most effective ones for this task. When evaluating traditional ML algorithms, we found that no single algorithm consistently outperformed the others across all tasks. However, the Rocket Classifier demonstrated superior performance in pulmonary prediction, while XGBoost and Random Forest performed best for infection prediction. For CVS prediction, all traditional ML algorithms performed similarly due to the local minimum of always predicting no morbidity. Comparing the effectiveness of deep learning (DL) models with traditional ML models was challenging since test accuracies were computed slightly differently.

Finally, we sought to balance accuracy and interpretability in these models. We found that most of the algorithms considered offered limited interpretability, particularly the DL models. However, we note that the Ridge Classifier, being a linear model, allows for some interpretability, as its learned coefficients provide valuable insights into the impact of different features on prediction outcomes.

Overall, this DSG successfully demonstrated that predictive models could be developed using CPET data and that these models could achieve an accuracy greater than 0.5, which is the expected accuracy of a random binary classifier. The study also highlighted the need for a larger dataset and brought attention to class imbalances, particularly in CVS complications. Furthermore, it became evident that many modern machine learning algorithms inherently lack explainability, necessitating modifications to enhance transparency and build the trust required for real-world implementation.

7.1 Future Work

Future work will consist of collecting more data and rectifying the class imbalances. It would also be informative to recalculate the accuracies of the DL models in the same manner as the traditional ML models so that a direct comparison can be

performed. Exploring explainable algorithms, such as SHAP [8] and LIME [12], would render the models more useful to physicians who would require a large amount of trust in them for the models to be practically useful. Finally, performing a hyperparameter sweep for all the algorithms would ensure that maximal algorithm performance had been reached.

8 Team members

Vaishnavi Balaji is a COO and Data Scientist in Curenetics, an AI based personalised cancer therapies startup, with a background in data science, machine learning, and artificial intelligence. She contributed to this project by leveraging her expertise in data analytics and predictive modelling, particularly in the domains of healthcare and application development. Her strong foundation in applying machine learning techniques to real-world problems, especially in biomedical research, allowed her to derive actionable insights and develop innovative solutions. Vaishnavi's experience in presenting at international conferences and contributing to research publications further enhanced the project's academic rigor and practical applicability.

Levan Bokeria is a Research Data Scientist at the Alan Turing Institute, specialising at the intersection of health and AI and with a background in cognitive neuroscience research. He contributed to the feature engineering part of the project by researching, extracting and transforming relevant features from the datasets to be used by downstream ML algorithms.

Alicia Falcon Caro is a final-year PhD student at Nottingham Trent University, specialising in the development of advanced signal processing and machine learning techniques applied to physiological time series data, particularly EEG and EMG, for the development of an hyper scanning brain-computer interface (BCI). She contributed to this project by participating as one of the facilitators of the team and through the application of statistical analysis methods, such as Granger Causality, to explore the potential causal relationships between the physiological features, and how these interactions differ across different classes. Additionally, she collaborated with team members on preliminary exploratory analysis and data preprocessing, ensuring the data was optimally prepared for machine learning applications

Dr. Funda Güner is an Assistant Professor at Çankaya University. She holds a

PhD in Industrial Engineering and has completed postdoctoral research at the University of Leicester. Her expertise lies in industrial engineering, with a focus on operations research and mathematical modelling. In this project, she contributed her expertise in modelling and data analysis to support the development of predictive models for morbidity outcomes.

Muhammad Aslam Jarwar is a Senior Lecturer at the School of Computing and Digital Technologies, Sheffield Hallam University. His expertise lies in the Internet of Things (IoT), wearable data processing, and applied artificial intelligence. Jarwar was a participant in the challenge, he contributed to data cleaning, developed the neural network to predict post-surgery risk, and played a key role in writing the report.

Mahreen Kiran is a PhD student at Anglia Ruskin University, Chelmsford, specializing in machine learning with a research focus on type 2 diabetes mellitus (T2DM) prediction using digital twin technology. She contributed to this UCL project, where she was involved in coding, data analysis, model development, and report writing. Her work applies advanced machine learning techniques to enhance predictive accuracy, supporting better clinical decision-making and patient outcomes.

Tochukwu Onyeogulu is a PhD student at Oxford Brookes University, specialising in machine learning and deep learning solutions for big data analysis, action/activity recognition, and multi-omic data analysis. With a background in applied mathematics, he contributed to this project by applying modern machine-learning techniques to build predictive models of morbidity, precisely respiratory complications, cardiovascular complications, and post-surgical infections

Kieran Woodward is a Research Fellow at the University of Nottingham specialising in pervasive computing with a focus on real-world AI applications such as for affect recognition and edge computing, including the development of novel techniques to reduce the size of deep learning models. He contributed to this project by developing the deep learning approaches used to classify the sensitive CPET data, including the integration of multiple models. Furthermore, he also contributed to the data preprocessing necessary for effective machine learning implementation.

Ruoqing Yin is a PhD student at the University College London specialising in time series deep learning analysis. She contributed to this project by feature

engineering, providing useful insights on machine learning model construction.

Amy (Yijie) Zheng was a PhD student at University of Nottingham specialising in computational AI and physics-informed neural networks for fibre optical imaging techniques. She is currently a postdoctoral Research Associate at University of Cambridge. She contributed to this project by facilitating the group to collaborate effectively, providing data-driven solutions using Multilayer Perceptron and writing the draft report.

Vijai Anand is a Research Fellow at the Collaborative Healthcare Innovation using Mathematics, Engineering, and AI (CHIMERA), University College London. Recently, his research focus has been on designing data-driven models to enhance the understanding and prediction of cardiopulmonary interactions in intensive-care patients. He was one of the PIs on this project and prepared the scope and data for the challenge.

James Butterworth is a Research Fellow at the Clinical Operational Research Unit at UCL where he applies modern machine learning algorithms to problems in healthcare. He was one of the PIs on this project and prepared the scope and data for the challenge.

References

[1] Brown, D. E., Sharma, S., Jablonski, J. A., and Weltman, A. Neural network methods for diagnosing patient conditions from cardiopulmonary exercise testing data. *BioData Min. 15*, 1 (Aug. 2022), 16. Publisher: Springer Science and Business Media LLC.

[2] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Int. Res. 16*, 1 (June 2002), 321–357. Place: El Segundo, CA, USA Publisher: AI Access Foundation.

[3] Dempster, A., Petitjean, F., and Webb, G. I. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery 34*, 5 (Sept. 2020), 1454–1495.

[4] Dempster, A., Schmidt, D. F., and Webb, G. I. MiniRocket: A Very Fast (Almost) Deterministic Transform for Time Series Classification. In

Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (New York, 2021), ACM, pp. 248–257.

[5] Granger, C. W. J. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica 37*, 3 (1969), 424–438. Publisher: [Wiley, Econometric Society].

[6] Hochreiter, S., and Schmidhuber, J. Long Short-Term Memory. *Neural Comput. 9*, 8 (Nov. 1997), 1735–1780. Place: Cambridge, MA, USA Publisher: MIT Press.

[7] Levett, D. Z. H., Jack, S., Swart, M., Carlisle, J., Wilson, J., Snowden, C., Riley, M., Danjoux, G., Ward, S. A., Older, P., and Grocott, M. P. W. Perioperative cardiopulmonary exercise testing (CPET): consensus clinical guidelines on indications, organization, conduct, and physiological interpretation. *British Journal of Anaesthesia 120*, 3 (Mar. 2018), 484–500. Publisher: Elsevier.

[8] Lundberg, S. M., and Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems* (2017), I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc.

[9] Metra, M., Faggiano, P., D'Aloia, A., Nodari, S., Gualeni, A., Raccagni, D., and Dei Cas, L. Use of cardiopulmonary exercise testing with hemodynamic monitoring in the prognostic assessment of ambulatory patients with chronic heart failure. *Journal of the American College of Cardiology 33*, 4 (Mar. 1999), 943–950. Place: United States.

[10] Myers, J., Arena, R., Dewey, F., Bensimhon, D., Abella, J., Hsu, L., Chase, P., Guazzi, M., and Peberdy, M. A. A cardiopulmonary exercise testing score for predicting outcomes in patients with heart failure. *American heart journal 156*, 6 (Dec. 2008), 1177–1183. Place: United States.

[11] Myers, J., de Souza, C. R., Borghi-Silva, A., Guazzi, M., Chase, P., Bensimhon, D., Peberdy, M. A., Ashley, E., West, E., Cahalin, L. P., Forman, D., and Arena, R. A neural network approach to predicting outcomes in heart failure using cardiopulmonary exercise testing. *International journal of cardiology 171*, 2 (Feb. 2014), 265–269. Place: Netherlands.

[12] Ribeiro, M. T., Singh, S., and Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2016), KDD '16, Association for Computing Machinery, pp. 1135–1144. event-place: San Francisco, California, USA.

[13] Ross, H. J., Peikari, M., Vishram-Nielsen, J. K. K., Fan, C.-P. S., Hearn, J., Walker, M., Crowdy, E., Alba, A. C., and Manlhiot, C. Predicting heart failure outcomes by integrating breath-by-breath measurements from cardiopulmonary exercise testing and clinical data through a deep learning survival neural network. *European Heart Journal - Digital Health 5*, 3 (Jan. 2024), 324–334. _eprint: https://academic.oup.com/ehjdh/article-pdf/5/3/324/57765185/ztae005.pdf.

[14] Sharma, Y., Coronato, N., and Brown, D. E. Encoding Cardiopulmonary exercise testing time series as images for classification using convolutional neural network. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. 2022* (July 2022), 1611–1614.

[15] Zhao, B., Lu, H., Chen, S., Liu, J., and Wu, D. Convolutional neural networks for time series classification. *Journal of Systems Engineering and Electronics 28*, 1 (2017), 162–169.

The Alan Turing Institute

turing.ac.uk

@turinginst