Degree project

Using machine learning to reveal doxorubicin induced cardiotoxicity pathways in cardiomyocytes

Master Degree Project in Informatics with a specialization in Data Science

Second Cycle 15 credits

Spring term 2024

Student: John Baxter

Supervisor: Dario Melguizo Sanchis and Göran Falkman

Examiner: Alexander Karlsson

Abstract

Doxorubicin is an effective anti-cancer drug, but its use is limited by dose-dependent cardiotoxicity, leading to long-term heart damage. This thesis explores the application of machine learning and deep learning models to analyse single-cell RNA sequencing data from cardiomyocytes exposed to doxorubicin. Models found that Drug_D2 samples were easily classified, achieving an AUC of ~0.9, while distinguishing between Control_D2/14 and Drug_D14 samples proved more challenging, with AUCs ranging from ~0.7 to 0.8. Classic ML models, such as Logistic Regression, Random Forest, and XGBoost, showed higher consistency in identifying important features linked to known cardiotoxic pathways compared to artificial neural networks. However, ANNs may uncover previously unexplored features significant for understanding doxorubicin-induced cardiotoxicity. Future work should explore DL models' types, integrate multi-omics data, and explore unprocessed data to enhance feature extraction, aiming to improve prediction models for patient-specific cardiotoxicity risks and guide new therapeutic strategies.

Table of Contents

Abbrevations 1
1 Introduction 2
1.1 Problem definition 3
2 Background 4
2.1 scRNA Sequencing 4
2.2 Machine Learning Methods 4
2.3 AI and Deep Learning 6
2.4 Data Splitting 7
2.5 Hyperparameter optimisation 8
2.6 Classification Metrics 8
2.7 Explainability 9
2.8 Using deep learning to analyse scRNA data 11
2.8.1 TSNE and UMAP 11
2.9 Doxorubicin-induced Cardiotoxicity 12
3 Method 13
4 Implementation 16
4.1 Dataset 16
4.2 Preparation of scRNA data 16
4.3 Data Shuffling and Feature Selection 17
4.4 Model Building 19
4.4.1 Model 1 – Logistic Regression 20
4.4.2 Model 2 – Random Forest Classifier 21
4.4.3 Model 3 - Gradient Boosting Classifier (XGBoost) 21
4.4.4 Model 4 – Artificial Neural Networks – Sequential 21
4.4.5 Model 5 – Artificial Neural Networks – Multilayer Perceptron Classifier 22
4.4.6 Model 6 – Convolutional Neural Networks and Image Classification 22
4.5 Explainable AI - Important Model Feature Discovery 22
4.6 Biological Understanding of features 23
5 Results 24
5.1 Preparation of scRNA data and clustering 24
5.2 Model building and comparison 25
5.2.1 Model 1 - Logistic Regression 25
5.2.2 Model 2 – Random Forest Classifier 26
5.2.3 Model 3 - Gradient Boosting Classifier (XGBoost) 27

5.2.4 Model 4 – Artificial Neural Networks: Sequential 28
5.2.5 Model 5 – Artificial Neural Networks: Multilayer Perceptron 29
5.2.6 Model 6 – Convolutional Neural Networks and Image Classification 30
5.3 Classification Summary 31
5.4 Biological Feature Discovery 31
5.5 Biological Implications 33
5.5.1 Biological Term Enrichment 34
6 Discussion 36
6.1 Previous Research 36
6.2 Methods, implementation and results 37
6.2.1 Pre-processing 37
6.2.2 Model building 37
6.2.3 CrossFold validation 38
6.2.4 Model results 38
6.2.5 Explainability: Global and local importance 39
6.2.6 Biological results discussions 40
6.2.7 Local important difference with CNN image classification 41
6.3 Ethical and societal aspects 42
7 Conclusion 42
7.1 Future work 43
References 44

Abbrevations

AI: Artificial Intelligence

ANN: Artificial Neural Network

AUC: Area Under the Curve

CNN: Convolutional Neural Network

Control_D2/D14: Cardiomyocytes which were not exposed to doxorubicin, RNA extracted two or 14 days after exposure

CVD: Cardiovascular Disease

Drug_D2/D14: Cardiomyocytes which were exposed to doxorubicin, RNA extracted two or 14 days after exposure

iPSC: induced Pluripotent Stem Cells

LIME: Local Interpretable Model-agnostic Explanations

MI: Myocardial infarction

ML: Machine learning

MLP: Multi-layer Perceptron

PCA: Principal Component Analysis

ROC: Receiver Operating Characteristic curve

scRNA: Single Cell RNA

SHAP: SHapley Additive exPlanations

t-SNE: t-distributed stochastic neighbour embedding

UMAP: Uniform Manifold Approximation and Projection

XGBoost: eXtreme Gradient Boosting

1 Introduction

"Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from noisy, structured and unstructured data, and apply knowledge from data across a broad range of application domains" (Cao, 2018). In this thesis Data Science processes and models are applied to process noisy, high dimensional biological sequencing data, extracting important biological insights.

Data science is a multidisciplinary field that combines statistics, computer science, and domain knowledge to extract meaningful insights from data. Within data science, machine learning (ML) techniques are essential for identifying patterns and making predictions from large datasets, often beyond the capacity of human analysis. Machine learning involves algorithms learning from data to recognize trends and predict future occurrences, providing powerful tools for data-driven decision-making (Géron, 2023).

This thesis explores the application of ML models to distinguish between heart cells exposed to the anti-cancer drug doxorubicin and those that have not been exposed. Doxorubicin is widely used for its efficacy in treating various cancers, but its use is limited due to its dose-dependent, cumulative cardiotoxic effects, which predominantly impact the heart (Rawat et al., 2021; Renu et al., 2018; von Hoff et al., 1979). By applying machine learning to gene expression data, this study aims to identify the specific genetic features that differentiate between exposed and unexposed cells, contributing to our understanding of the drug's cardiotoxicity.

The data science problem addressed in this thesis is a classification task, aiming to classify samples based on their exposure to doxorubicin and to identify the genetic features that are most predictive of exposure status. The dataset comprises four conditions: Drug_D2, Drug_D14, Control_D2, and Control_D14. 'Drug' or 'Control' indicates whether the samples were exposed to doxorubicin, while 'D2' and 'D14' denote the time points two days and fourteen days post-exposure, respectively.

As with most data science projects, data preparation is a crucial step. The R package 'Seurat' (Hao et al., 2023a) was employed to clean the dataset, removing outliers, normalizing and scaling the data, and assessing how the samples cluster. After preparation, various supervised learning models were trained to classify the samples. Feature selection and extraction were performed using both intrinsic model metrics and external methods like permutation importance to determine which genetic features were most relevant.

In addition to traditional machine learning models, deep learning techniques such as artificial neural networks (ANNs) and multilayer perceptron (MLPs) were explored to evaluate their performance and feature selection capabilities. These deep learning models, which consist of multiple layers that can learn complex representations of data, often perform well on high-dimensional data but are usually less interpretable than simpler models. To address this, the study employed explainability tools such as permutation importance, SHAP (SHapley Additive exPlanations) values and LIME (Local Interpretable Model-agnostic Explanations) to interpret model predictions and better understand the underlying biological mechanisms.

The results indicated that Drug_D2 samples were easily distinguishable from the others, reflecting their distinct expression profiles shortly after exposure. In contrast, Drug_D14, Control_D2, and Control_D14 showed more overlap, which aligns with the expectation that cells might recover from doxorubicin exposure over time, as suggested by Holmgren et al.,

(2018). Feature extraction across different models revealed several key genetic markers related to cardiovascular processes, highlighting potential pathways for therapeutic intervention to mitigate cardiotoxicity.

Notably, the simplified models using only the top 25 features, even with limited hyperparameter tuning, performed comparably to more complex models, highlighting the importance of feature selection and model interpretability. By integrating explainability methods, this thesis demonstrates how understanding the internal workings of ML models can provide valuable insights into biological data, ultimately contributing to more transparent and trustworthy applications of machine learning in biology and healthcare.

1.1 Problem definition

Doxorubicin, an anti-cancer drug from the anthracycline group, is effective against various cancer forms but is limited by its adverse side effects, particularly its well-documented long-term cardiotoxicity. This cardiotoxicity, characterized by cumulative, dose-dependent, and progressive heart damage, can result in congestive heart failure (Rawat et al., 2021; Renu et al., 2018). Although the biological pathways contributing to doxorubicin-induced cardiotoxicity are suspected, more research is needed to identify and understand the cellular changes leading to this toxic effect.

In this thesis, I aim to address the gaps in current knowledge regarding doxorubicin-induced cardiotoxicity by focusing on the research question: What are the key genetic and transcriptomic features driving cardiomyocyte response to doxorubicin exposure, and how can machine learning models, including deep learning, be use to reveal novel insights into the biological pathways involved?

To achieve this, analysis of single-cell RNA (scRNA) sequencing data from cardiomyocytes exposed to doxorubicin, with sequencing performed at two and 14 days after exposure. This data provides a transcriptomic profile of how heart cells react to the drug and describing the cellular processes involved. However, scRNA sequencing data is highly complex, with significant dimensionality and variation, making it challenging to analyse by simple human analysis. To be able to analyse such complex data, a combination of classic machine learning models and neural networks (NN) will be used to classify and analyse the data. A key objective is to assess how well these models capture trends in noisy scRNA data and determine which approach is more effective in identifying important features. By comparing the results from different models, a consensus on the key genes and pathways contributing to doxorubicin-induced cardiotoxicity can be found.

Additionally, ML explanations will rank the features learned by these models to determine which biological pathways are most influential in distinguishing between cells exposed to doxorubicin and control cells. This analysis will provide a better understanding of the underlying mechanisms of toxicity and identify the most critical pathways, potentially leading to new insights into the prevention and treatment of doxorubicin-induced heart damage.

To summarise, this thesis will contribute to the research field by attempting to:

- Identify key pathways and genetic features involved in doxorubicin-induced cardiotoxicity.
- Combine and compare multiple ML and deep learning models to achieve consensus on these features.

- Evaluate the effectiveness of classic ML and NN approaches in capturing trends in noisy scRNA sequencing data.
- Rank the importance of identified pathways to enhance the understanding of doxorubicin's toxic effects on cardiomyocytes.

2 Background

2.1 scRNA Sequencing

Humans are made up for approximately 30 trillion cells (Hatton et al., 2023). Traditional cellular classification relies upon histology, microscopy or cell surface markers. These can identify cells, however; misclassification is frequent and requires extensive expertise. Furthermore, through simple identification it is not possible to see the heterogenicity reported within cell populations (Munsky et al., 2012). Single cell RNA sequencing (scRNA) is a technique wherein individual cell types are captured and their transcriptomic profile is sequenced. These cells were captured using ICELL8, whereby cells types are confirmed by human specialist visualisation and then sequenced.

scRNA sequencing provides unparalleled resolution to analyse cellular processes, being able to capture specific cell types. scRNA transcriptomic profiles can then be analysed, detailing cellular processes, roles and trajectory. scRNA is an improvement over bulk RNA sequencing which contains a population of cell types and therefore a high degree of biological heterogenicity making analysis and comparison difficult. scRNA has shown great potential when discovering new cellular types, tracking developmental trajectories, identifying gene regulatory systems and disease understanding and diagnosis (Grün et al., 2015; Petropoulos et al., 2016; Villani et al., 2017).

In this thesis supervised machine learning models will be used to classify cellular experimental types based on scRNA data. Supervised machine learning, as the name suggests is learning from data in which the outcome is known or pre-determined i.e. supervised data. By learning inputted data patterns with labelled outputs, future cases where the labelling is unknown can be predicted (Géron, 2023). The most commonly used supervised learning methods include: Logistic Regression, K-Nearest-Neighbours, Naïve Bayes and Decision Trees. However, as computational power increases deep learning neural networks have become common place in machine learning techniques.

2.2 Machine Learning Methods

Logistic Regression is one of the simplest and most common classification methods. It is simple to implement and easy to interpret (Sperandei, 2014). Logistic regression provides probabilities for each class using multinomial logistic regression extensions. In logistic regression the model predicts the probability that the dependent variable Y equals 1 (the class) as a function of the independent variables X. This can be seen in the equation.

$$P(Y = 1|X) = \sigma(\beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta k * nk + \alpha)$$

In this equation $\beta 0$ is the intercept, $\beta 1, \beta 2, \ldots, \beta k$ are the coefficients of the prediction features X. The advantage of this equation is that the coefficients values of each feature can be extracted helping to understand the most important features influencing model classifications.

K-Nearest Neighbours (KNN) is a classifier that relies on the location of samples in relation to its proximity of other labelled samples using a majority vote. Put simply, if a new data point has four close neighbours labelled as class X, and one as class Y, then the prediction of the new class is X. It is one of the most popular and simplest to understand. However, it does suffer from assumption that similar points must be the same, which isn't always true. KNN suffers from scaling issues wherein the larger the dataset the more inefficient and inaccurate the model becomes (G. Guo et al., 2003).

Naïve Bayes is a group of classification methods that applies Bayes' theorem with a 'naïve' assumption of conditional independence between every pair of features (Webb & Bayes, 2017). Although simple, they have worked well in document classification and spam filtering. Naïve Bayes is a rapid learning method, requiring low data volumes. This supervised learning technique is also quite good at dealing with high dimensionality as each distribution is treated independently. Naïve Bayes are powerful but not well suited to this thesis' problem as unlike discriminative classifiers, Naïve Bayes does not learn important discriminative features used to classify (Webb & Bayes, 2017). This means it is not possible to discover the important biological pathways surrounding doxorubicin cardiotoxicity.

Decision Tree algorithms create a hierarchical tree structure consisting of roots, branches and leaf nodes. These structures depend on a dividing strategy in which samples from a root node are split based on feature parameters to leaf nodes. This structure can contain many branches and layers grouping samples into their correct labelled categories. A great advantage of decision trees is that it is a 'white box' model wherein discriminatory features can be easily explained. These models work well with relatively small data sizes but as the data grows in complexity, so does the tree. These large trees can result in overfitting whereby data samples are fragmented into small leaf nodes. Pruning less important nodes and employing random forest ensemble techniques can improve or maintain accuracy (Géron, 2023; Kotsiantis, 2013).

Random Forest Classifier is a development of decision trees. Random forest uses many decision trees, thus creating a forest of classifiers. Random forest is an ensemble method where the final result is made up of many other classifiers coming to a consensus. As an ensemble method, the tree classifiers are based on random sub datasets to improve overfitting and accuracy. This method has better success than simple decision trees with highly dimensional and noisy data making it a better candidate for scRNA data (Breiman, 2001; Pouyan & Kostka, 2018).

Extreme Gradient Boosting has become the leading machine learning algorithm in recent publications. It has won a host of Kaggle data competitions and named Technology of the year in 2019 (Heller, 2019). Extreme gradient boosting is a scalable, distributable gradient boosted decision tree. It is similar to random forest in such that it uses multiple decision trees. Where random forest bootstrapping samples sub-sets of data to create trees. Gradient boosting then combines weak models to create a stronger model. Gradient boosting iteratively trains an ensemble of weak trees using the earlier decision tree error values to update future trees, thus minimising errors (Hastie et al., 2001; Nvidia.).

Despite these useful and common place machine learning techniques, various studies have concluded that traditional machine learning techniques do not capture the complexity of scRNA data. However, a promising avenue for scRNA analysis is the field of deep machine learning and neural network classifiers (Brendel et al., 2022; Flores et al., 2022). Single cell technologies allow the generation of hundreds to thousands of datapoints, meaning that this type of data is perfect for deep learning methodologies which can capture complex data

patterns (Oskolkov, 2019).

2.3 AI and Deep Learning

Artificial intelligence (AI) is the process of making computers think or act in a manner that we would otherwise expect a human to do. Machine learning is a subset of AI which works on creating programmes or algorithms which learn data patterns and improve from experience, without every step being programmed explicitly (Mosavi et al., 2019; Schweidtmann & Mitsos, 2019). Deep learning is a subset of machine learning which uses artificial neural networks to process and analyse information (Géron, 2023). Deep learning is especially used in visual or speech recognition but also within genomics and medicine due to the highly dimensional nature of their data (Angermueller et al., 2016; Jin et al., 2021; Sapoval et al., 2022). Neural networks are layers of computation nodes that consist of input layers, hidden layers and output layers, which result in the classification. In a neural network, each node connects to other nodes and has its own associated weight and threshold value. Should the values input into the node reach or exceed a threshold value then the node is activated and information is passed to the next layer (Géron, 2023). A neural network with more than two hidden layers is considered as deep, hence the name "deep learning". As seen in Figure 1, deep neural networks can have a complex hidden layer structure this aids in capturing complex patterns, but makes the classification decision structure opaque (Géron, 2023). The learning that occurs in an ANN is based upon risk minimisation of the difference between the predicted and labelled cases in the training data (Vapnik, 1995). This ultimately updates weights between nodes being depending on the error value with the aim of minimising the difference between the predicted and true class states. ANNs can trace their original design back to biological neural networks in which a stimulus is passed along a neurone and if above a threshold can pass the message onwards to connected neuronal layers (Suzuki, 2011).

ANN as a concept has been around for almost 50 years and is now experiencing something of a renaissance thanks in part to increased computational power, the decreasing cost of storage, large increases in the volume of data, and development of more advanced learning algorithms

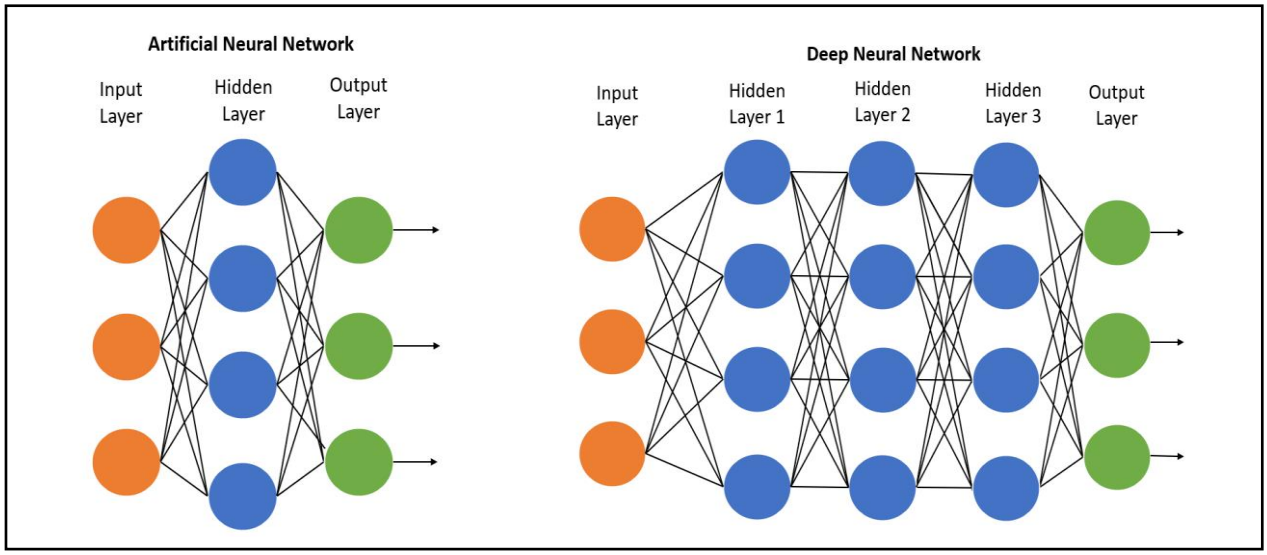

Figure 1. Artificial neural networks (left) consist of three-layer types. An input layer, where feature values are input, a hidden layer in which weight is applied to each input and an output whereby the input material is classified. In deep neural networks (right) consists input, at least two hidden layers and then an output layer.

(Abdolrasol et al., 2021; Medium, 2023; Rumelhart et al., 1986). ANNs have also gained popularity due to their ability to capture complex data pattens as well as scalability through parallelization, handling large data volumes and testing hyperparameters (Zappone et al., 2019). Enhancing this popularity is the democratisation and ease of use through application user interfaces (APIs) in multiple programming languages such as Python, C++, R, Spark and Matlab (Abdolrasol et al., 2021). These out-of-the-box ANNs make developing deep learning algorithms accessible to all without requiring detailed programming knowledge.

Of artificial neural networks there are various types. Feedforward networks are those in which input data flows forward until the output later is reached. An example of this is Kera's Sequential model (Moolayil, 2019). The sequential model allows you to build layer by layer, with each layer has one input and output tensor. It is straightforward to use and layers have been added sequentially to add complexity. Recurrent neural networks are those that 'remember' previous layers helping decide output layers. These typically work with time series or sequential data. (Moolayil, 2019). Multilayer perceptron's is type of ANN that consists of multiple layer of neurons with each layer fully connected to the last one. As an ANN, it has an input, hidden, and output layer. The algorithm trains using backpropagation, whereby errors in the classification are calculated via optimization 'cost' algorithms and then fed back into updating the weights to improve classifications. These cost functions can take various forms and can also be a hyperparameter. A common cost function for classification tasks is the Categorical Cross Entropy loss. It penalises incorrect classifications and can handle multiple classes efficiently (Géron, 2023). The aim of these cost functions is to minimise the cost function by adjusting model parameters, thereby improving accuracy. The final model mentioned here is the Convolution Neural Networks. These types of models have become one of the most common in recent artificial intelligence solutions (Moolayil, 2019). These networks apply filters to input information to condense data, it then pools those data points and uses these to make decisions. CNNs have great applications for video and image-based data but have also shown promise in other fields such as scRNA data analysis (Vasighizaker et al., 2021).

2.4 Data Splitting

In model building, it is important to prevent overfitting and reduce data leakage which can influence the model performance. Overfitting is the process where the model works well but fits too close to the training data. When faced with a new data point the model struggles to accurately process them as it is not robust to differences. Data leakage is when information from the test data set is unintentionally shown to the model creation steps and leads to incorrect model performance scores (Müller & Guido, 2017).

To prevent these problems several strategies can be used. The most common is to split the data into three groups; training, validation and testing (Müller & Guido, 2017). By splitting the data, the models can be built and properly evaluated without repeatedly using the exact same data points twice or more. Training is the largest group, typically 60-80%, and is used for training the model to recognise patterns in the data. Validation data, typically 30-10%, is used to test how good the model is at classifying new data points. Testing data, typically 30-10%, should be completely independent and used once validation data results show satisfactory results in order to test an unseen data set as an example of how future new data much be classified.

Overfitting can be prevented through cross fold validation, shorter training times and introducing more data to the model training (Müller & Guido, 2017). This keeps the model more general so that it fits new data better, rather than becoming overly fitted on training data.

It is therefore important to consider how data is split. As mentioned, most commonly it is split into training data; used to train the model, validation data; to help confirm the good model parameters and testing data; used to evaluate the model's predictions. Cross validation is an important method in the use of this split data (Nti et al., 2021). In this way the data is divided into k folds. The model is trained on k-1 folds and then validated on the remaining fold. This is repeated k times (number of folds) with each fold being used once as the validation set. The validation performance is an averaged metric of the k trails. Cross validation reduces the risk of overfitting by repeatedly using data sets but with new diversity each time. This ensures a more robust model in different data groupings for future use (Géron, 2023).

2.5 Hyperparameter optimisation

Hyperparameters in model building refers to the options available in how machine learning models conduct their learning processes. These options vary between model types, these parameters are not learned from the data but are set before the learning begins. There are two overall types of hyperparameters. Model specific, those which define the model architecture such as the activation functions. The others are training related hyperparameters such as the learning rate, epochs, and dropout rate (Géron, 2023). Hyperparameters have a defining impact on the performance of models and how they learn. It is of vital importance that the right hyperparameters of the learning job are found. RandomSearchCV and GridSearchCV are used in this thesis for model hyperparameter testing.

Random and Grid search work with the same goal in mind, to test hyperparameters to find the best combination of settings to generate the most successful model. How they do this is different. RandomSearch with a user defined range randomly sample a fixed number of parameters as a type of shot-in-the-dark parameter test, within the range. This random search aims to reduce the computational expenses at the starting point when the range of hyperparameters is unknown and the parameter space is large. This creates a rough framework for starting a GridSearch. GridSearch is a more specific, defined search in which stated parameters are exhaustively searched, evaluating all possible combinations of the defined. Both search methods use cross validation (CV). By using RandomSearchCV and GridSearchCV, the starting parameters can be narrowed down, followed with a more exhaustive search of the optimal region. Whilst at the same time cross fold validation counters overfitting.

2.6 Classification Metrics

The success of classification models is based on the confusion matrix. The confusion matrix is a table that describes the results of the model's classification of test data of which the true values are known (Grandini et al., 2020). These values show the proportion of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). From these values, several other values can be created. Accuracy is the ratio of correct prediction against all other prediction outcomes.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is the ratio of correct predictions against total positive predictions, false positives included.

$$Precision = \frac{TP}{TP + FP}$$

Recall is also known as sensitivity and is the ration of correct classifications against the true classifications.

$$Recall = \frac{TP}{TP + FN}$$

F1 score is a combination of the precision and recall metric.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

These metrics are the baseline in which classification models will be compared and contrasted. An easily interpretable classification metric is known as the Receiver Operating Characteristic curve (ROC) and Area Under Curve (AUC) (J. Huang & Ling, 2005). This produces a graph which shows the performance of the model using the Recall True Positive and False Positive Rate.

$$False\ Positive\ Rate = \frac{FP}{FP + TN}$$

As seen in Figure 2, these two metrics are plotted and the total area under that curved line measures the performance of the model. The greater the area, the better the model is at predicting true data point instances.

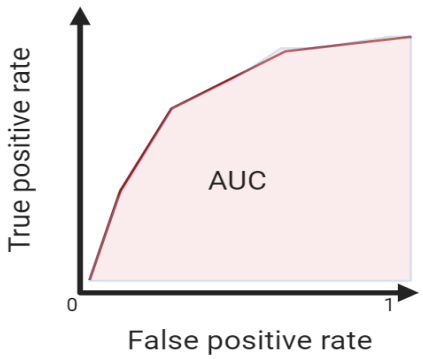

Figure 2. ROC example showing AUC of true cases

2.7 Explainability

Once a classifier has appropriate classification metrics it can be a good idea to open-the-lid of the machine learning and understand the important features learnt for decision making. This is call AI explainability. AI explainability is of vital importance, especially as AI is increasingly integrated into society. This can be a big problem within AI research and application as for example if an AI makes a healthcare decision, what is that based on? Can it be critically evaluated? How can the user know if it might be correct or a mistake? Without the information about how this decision is made one would be very reluctant to trust and follow through with the AI's decision (Srinivasu et al., 2022).

Machine learning models and AI techniques are often called 'Black-boxes'. This is because although an answer or output is achieved, it is not explained how the model came to that conclusion (Rai, 2020). Learning the important features used to make decisions in ML models in called opening the lid. There are two types of explainability: global importance and local importance. Global importance investigates the importance of each feature over the entire dataset. Answering the question "What features are most important for the model in general?" (Botari, Hvilshj, et al., 2020). Local importance however, focuses more on individual predictions. For example: "What features were important for classification of sample number 1?". Local importance extractors tend to rely on complimentary packages such as LIME (Ribeiro et al., 2016a) and SHAP (Lundberg et al., 2017).

To overcome this several techniques have been developed to understand how or what has the ML model made a conclusion based on what information. Some of these techniques are in-built within models based on equation inputs such as with Logistic Regression or data splitting with Random Forest. This type of explainability is known as 'White-box' or 'glass box' as they are easily extracted (Rai, 2020). However, 'black box' models such as ANN and MLP which contain weights can be harder to extract and require additional packages such as SHAP, LIME and permutation importance (Altmann et al., 2010; Lundberg et al., 2017; Ribeiro et al., 2016a).

LIME works by selecting a single data point, generates synthetic data perturbing the data values from the original, observes the changes these tweaks have on the data, ranking the effect of these changes per feature. Those changes closer to the original output instance are determined as higher importance to the model effectiveness (Ribeiro et al., 2016b).

Permutation importance works by shuffling the values for every feature across all the test samples and measuring the model performance after shuffling. Shuffling prevents the association between the selected feature and target variable but all other features remain the same. Shuffling the feature makes it less informative in the model as it loses its structure. The will most likely degrade the model's performance as the feature is now useless. if the models' performance was greatly degraded, then that feature was important, if no change is observed then that feature had no impact on the model in the first place. This degradation is recorded and the process is repeated for all features. Features are then ranked, showing their degradation effect on the model if shuffled, denoting their importance (Altmann et al., 2010; Kaneko, 2022).

SHAP, on the other hand, operates based on Shapley values from cooperative game theory, attributing the prediction of the model to individual features by evaluating their contribution across all possible combinations of features (Lundberg et al., 2017). For each feature, SHAP calculates its marginal contribution by comparing how the model's output changes when the feature is included versus when it is excluded. These contributions are then summed to recreate the final prediction, with higher SHAP values indicating greater feature importance (Lundberg et al., 2017).

2.8 Using deep learning to analyse scRNA data

Using deep learning to solve some of scRNA data processing problems have become a hot topic amongst bioinformatians. Deep learning performs better than traditional machine learning approaches as they are better able to capture complexity found in high dimensional data (Flores et al., 2022). This is because of the inherent nature of neural networks (NNs) which contain thousands to millions of trainable input parameters, especially useful in biology which involves thousands to millions of samples or genes (Brendel et al., 2022).

Deep learning has been used to discover cell clusters (Tian et al., 2019), cell type identification (Jia et al., 2023), gene annotation (Yuan et al., 2022) as well as pre-processing tasks such as batch correction (Li et al., 2020), experimental error detection (Bernstein et al., 2020; Mcginnis et al., 2019) and autoencoding (Wang & Gu, 2018). Flores et al., 2022, reviewed 25 deep learning methods concluding that deep learnings ability to scale, transferability and flexibly will cement deep learning as viable tools to analyse scRNA data.

scRNA data suffers from biological and technical factors such as batch effects, genetic dropout, high background noise, sparse gene coverage and high variance (Lähnemann et al., 2020). Fortunately, a large number of novel statistical approaches have been developed to handle these challenges. Specialist pipelines such as Seurat for R (Hao et al., 2023a) and ScanPy (Wolf et al., 2018) for Python include a range of specific commands to streamline data cleaning and processing. These commands include scaling, normalisation, dimensionality reduction, clustering and command specific for processing scRNA data. These packages include not only preparation commands but also visualisation, differential expression and annotation. In fact, there is an explosion of analyse and processing methods that makes standardisation difficult (Luecken & Theis, 2019).

scRNA sequencing data analysis involves three main steps. Firstly, the data produced is a gene by cell matrix. In this count matrix, raw gene expression counts are aligned with each cell. These counts are quality controlled using a mix of covariates and normalised via typical standardisation approaches such as log transform (Luecken & Theis, 2019). The second step involves annotation, cell cycle removal, doublet removal, denoising, and batch effect removal and finally dimensionality reduction (Brendel et al., 2022; Luecken & Theis, 2019).

2.8.1 TSNE and UMAP

scRNA dimensionality reduction has multiple options such as PCA, t-SNE, and UMAP. t-distributed Stochastic Neighbour Embedding (t-SNE) attempts to visualise high dimensional data by reducing it to a few dimensions. It aims to minimise the Kullback-Leibler divergence by converting data points similarity to low dimension space (van der Maaten & Hinton, 2008). It is useful in retaining the data's local structure keeping similar data points together. It is not effective however, at preserving global structural topology. Uniform Manifold Approximation and Projection (UMAP) achieves a similar dimensionality reduction for plotting clusters but uses manifold learning techniques and topological data analysis. It has been reported to be more effective at retaining local and global data structures (Yang Yang et al., 2021). UMAP is commonly used for downstream analysis as it captures the original data structure better than t-SNE (Yang Yang et al., 2021). For this reason, UMAP will be used for image creation with CNN models and used where possible to make conclusions on the data.

Finally, downstream data analysis can occur with cell clustering, cell type annotation, trajectory analysis to answer biological questions. As previously mentioned, deep learning can be used both in the pre-processing and biological understanding of scRNA data. Some pre-processing models will be used but not discussed in depth in this thesis. This thesis's focus is on the biological data analysis and feature extraction results.

2.9 Doxorubicin-induced Cardiotoxicity

Doxorubicin is a common and effective antineoplastic agent for various cancer types including paediatric cancer, leukaemia and breast cancer (Rawat et al., 2021). It suffers, however, from adverse side effects of interest here is its induced cardiotoxicity. Various cardio-toxicity pathways have been suggested. Cellular apoptosis mediated loss of cardiomyocytes and oxidative stress (Nebigil & Désaubry, 2018), altered levels of Sarocoplasmic Reticulum Ca^{2+} (Zhao et al., 2020), mitochondrial biogenesis inhibition (Henriksen, 2018), and reactive oxygen species (ROS) are suggested to play a role in Doxorubicin-induced cardiotoxicity. Other dysregulation and dysfunction mechanisms have also been suggested as important pathways leading to cardiotoxicity (Renu et al., 2018).

Doxorubicin-induced cardiotoxicity can appear years after treatment, and is known as long-term cardiotoxicity (Steinherz et al., 1991; von Hoff et al., 1979). In this study sets of induced cardiomyocytes pluripotent stem cells (iPSCs) were exposed to Doxorubicin and then underwent scRNA sequencing two and 14 days after exposure. scRNA sequencing has therefore captured the cellular transcription profiles responding to exposure at different time points. By building a classifier to classify unexposed healthy controls and different post exposure sample along their time scales, it should be possible to extract the biological features discriminating between cell types and discover more about the cellular mechanisms effected by Doxorubicin.

3 Method

In this section is described the overall method of the thesis. In data science there are various frameworks that describe the working process of a data science life cycle. One of the well-known and used in this thesis is Cross Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM is described comprising of four levels of abstraction. In order from general to specific they are phases, generic tasks, specialized tasks and process instances (Wirth & Hipp, 2000). According to this model of data science iterative cycles of processing and use of data are applied. As showing in Figure 3, first it is important to start with business or problem understanding, this is interchangeable with data understanding which informs the previous and influences the latter. Next is data preparation which also includes a back-and-forth with modelling. Once models are built, they are evaluated, which can influence business/problem understanding and again feed into the data understanding, preparation and updating of models. Once models are sufficient, they can be deployed.

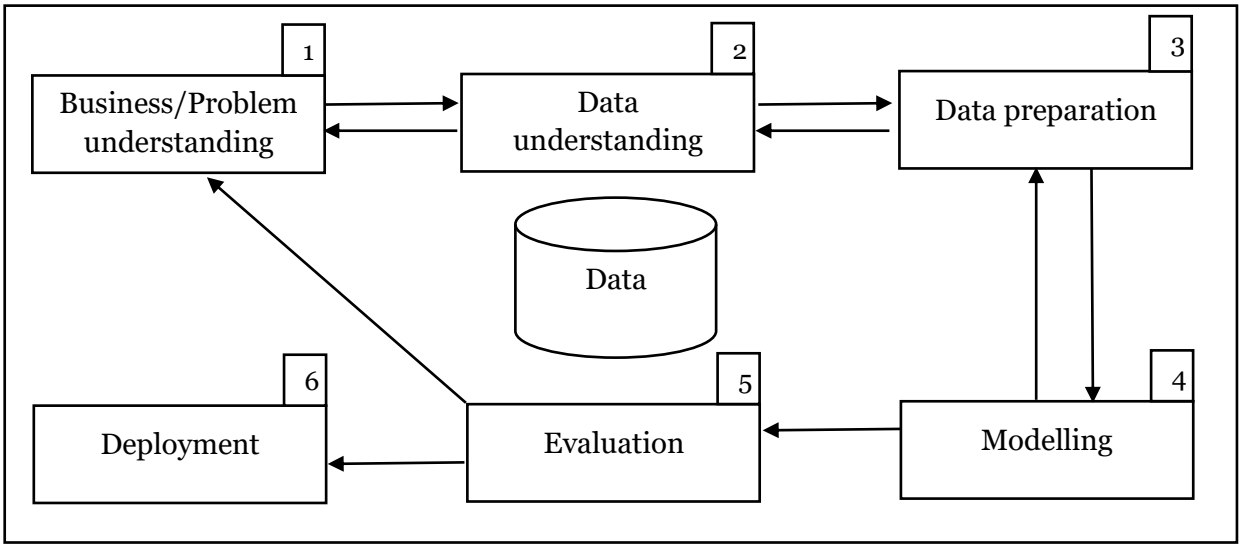

Figure 3. CRISP-DM data science methodology. Each step consists of various sub-steps. Steps 1-5 feed backwards and into each other, creating an iterative system where how steps are completed update based on prior and sequential results. Step 6 is the deployment of a model once the model cannot be improved. Figure adapted from Wirth & Hipp, 2000.

Step 1, business and problem understanding were first built on interviews with the data creator Dr. Dario Melguizo Sanchis. Projects under Dr. Sanchis include AI data analysis and creation of deep learning models with various iPSC types. Through Dr. Sanchis guidance understanding of the experiment set up, type of exposure and previous research was defined. Further understanding of the effects of doxorubicin on hearts as well as machine learning with cardiovascular and scRNA sequencing data was done using Google Scholar, Scopus and IEEE.

Step 2, understanding data was done through exploration in the R package Seurat (Hao et al., 2023b). This is a well-known standard tool for the exploration and processing of scRNA data (Luecken & Theis, 2019). This package and language was chosen for ease of use and Seurat appears the most robust and comprehensive to process scRNA data compared to ScanPy, the Python based scRNA data processing package. Following recommendations by Dr. Sanchis and the paper by Luecken and Theis the cells were filtered and data processed in step 3 and then data was used in step 4, modelling.

Modeling was done in Google Colab in a Python environment; this is due to the simpler API functionality for model building and flexibility with the Python language. Python APIs include

many ease-of-use functions for cross-validation methods, neural network building, and other machine learning functions, simplifying model building. Modeling was diverse and included many classic machine learning classification models such as KNN, SVM, and Random Forest to test general robustness. Based on several iterations of modeling, some models did not make it to further hyperparameter tuning, and the number of models tested was reduced. The final models tested included Random Forest, XGBoost, Logistic Regression, Artificial Neural Networks, Multilayer Perceptron, and Convolutional Neural Networks.

Random Forest models are ideal for scRNA data due to their ability to handle high-dimensional datasets and its robustness to noise, which is common in biological data (X. Chen & Ishwaran, 2012). Additionally, its feature importance ranking helps identify important gene features. XGBoost is known for its efficiency and high predictive performance, XGBoost is particularly suited for structured data like scRNA. It can handle missing data and is effective at capturing complex, non-linear relationships between features (T. Chen & Guestrin, 2016). It is commonly used with scRNA data (Abu-Doleh & Al Fahoum, 2024; Martinez et al., 2024) . Logistic Regression models provides simplicity and interpretability whilst being able to capture surprisingly complex patterns in scRNA such as distinguishing between cell types or disease states (J. Guo & Gao, 2022).

ANNs such as Kera's sequential NN are models capable of capturing non-linear relationships in the data, making them suitable for complex biological datasets like scRNA. They can adapt to the high dimensionality of gene expression data, potentially revealing intricate cellular patterns (Suzuki, 2011). Multilayer Perceptrons are a type of ANN. MLP is a robust model that can learn patterns in scRNA data (Feng et al., 2023; T. Zhang et al., 2024). It is able to model non-linear interactions across genes features and is effective for tasks such as classification and feature extraction due to simple weight important extractions. MLPs are particular powerful as they feature back propagation which allows for efficient learning to occur (Géron, 2023).

Although CNNs are traditionally used for image analysis, they were included to explore how transforming scRNA data into spatial or image structured formats can take advantage of CNN's feature extraction capabilities. This approach shows promise for not only cellular classification but also extracting biologically meaningful patterns from gene expression data (Flores et al., 2022; Sharma et al., 2019).

During modeling data splitting and minimizing data leakage between training and testing data was carefully considered. During the data preparation stage, the data was randomized and divided into six parts. Five parts would become the training and validation data, the remaining the test data. Important features in pre-processing were identified using the training/validation data, these features were then applied to the test data separately. This minimized peaking at, and potentially extracting different important features from the test data, ensuring more independence from selected model features. Training and validation data were saved as a CSV followed by a separate test CSV data file. For model hyperparameter tuning, the training and validation data was split 80:20 used for model creation and evaluation respectively. Once the model parameters were set then the combined training and validation data was used to build or fit the model. Adding the validation data to the fitting should reduce overfitting by introducing more data. Finally, once build the unseen test data was used for model prediction and evaluation.

Step 5 evaluation was performed after using various model hyperparameter tuning and used classification metrics for every class of the four. Of these metrics average weighted accuracy

and the individual F1-scores were considered the most important. This is because it is important to see on average, how the model performs on all classes but as a multiclass problem, individual classes scores are important to understand the biological problem. When top evaluation metric models could not be improved further, feature extraction took place. Feature extraction is the process of identifying the data features having the strongest impact on the model performance.

Global features were extracted using the native feature extractions present in ML models. These methods were selected as they are simple and use the inherent structure of the model's algorithm. Logistic regression used coefficient values, Random Forest and XGBoost the feature importance parameter highlighting features which best split the samples. Despite their simple to implement nature, there have been reported feature biases when analysing scRNA data (Nicodemus, 2011) .

MLP used learnt input weights, and ANN used permutation importance. Permutation Importance is useful as it can be used with classification and regression models. However, there are some drawbacks, it is computationally expensive and some have reported that the calculation is unstable when the sample number is low. Furthermore, strongly correlated features can create increased false positives due to unimportant but highly correlated features as well as reduced importance of jointly correlational features compared to independent features (Kaneko, 2022; Musolf et al., 2022). To attempt to overcome computation issues, a sampling method was used. Whilst the sample was relatively large (n = 50, features = 500) it should be noted that this may affect the performance of Permutation Importance in ranking features.

CNN used another package called Class Activation Maps (CAM) feature section using activated image pixel metrics which correspond to gene features. This method of pixel selection is highly sensitive to activation thresholds so results should be taken with a degree of skepticism.

Local feature importance was tested with SHAP, LIME and permutation importance (Altmann et al., 2010; Lundberg et al., 2017; Ribeiro et al., 2016a). Local importance was tested as a proof of concept as local importance will be vital in future applications of AI where understanding the individual outcome of a model case is required for trust and understanding (Rai, 2020; Srinivasu et al., 2022). These packages were used as they are relatively commonly applied to explainable AI and are model-agnostic, meaning they can be applied to machine learning models regardless of model structure. A mix of explainability methods were used in order to test the consistency of results and overcome drawback associated with a single model. Permutation methods suffer from over emphasis on correlated features (Hooker et al., 2021). Some have reported that LIME can be incoherent (Främling et al., 2021). This could be due to multiple factors such as randomness of sampling, variation with sampling proximity and variation in explained model credibility for different data points (Y. Zhang et al., 2019). Lastly, SHAP values tend to focus on influence, rather than importance (Främling et al., 2021).

Based on global features, the biological interpretation was possible as well as being able to create simpler models with fewer parameters with the top 25 global features being used to re-train models. These top 25 features were then explored for potential biological function.

Step 6, deployment was not performed in this thesis.

4 Implementation

4.1 Dataset

Takara Bio's ICELL8 Single-Cell System was used to obtain single cells for RNA-sequencing in a process as described by Holmgren et al., 2018. This system enables high-throughput processing of hundreds of single cells into individual nanowells. This system can obtain high quality scRNA data by separating multi-cellular tissue samples into single cells with minimal damage and cellular stress (Takara Bio). It is of vital importance that each nanowell contained one cell. The number of genes contained in each well sample sequenced was analysed. Abnormally high frequencies of genes were understood to be nanowells containing multiple cells also known as doublets (Bloom, 2018). These doublets were excluded from the dataset and subsequent downstream analysis.

This data contained the raw RNA counts from each cell sequenced in the form of a CSV file. This file contained 1771 biological cells rows and 58735 gene columns. Along with these two text files with meta data about the cell sequenced was combined and added to the CSV file. Data in the meta file included descriptions for each cell the most important being 'Sample_Type'. Sample types included annotations for whether the cell was from Control or Drug samples and time after exposure. These are named: ControlDay 2, Control Day 14, Drug Day 2 and Drug Day 14.

Ethical approval for the use of hES-CM stem cell line used for cardiomyocyte generation was not required since the cells used in the study were originated using a commercial cell line, Cellartis Pure hES-CM, purchased from Takara Bio Europe AB Swedish Filial (Gothenburg, Sweden).

4.2 Preparation of scRNA data

The data was uploaded and processed in R studio ("Ocean Storm" 2023.12.1). Technical control samples were removed and a Seurat Object (Hao et al., 2023b) was created using the CSV as count data and the text files as the meta data object. Min cells = 50 was used to include features found in at least 50 cells. This removes some noise and unlikely important gene feature as they are found in very few cells.

Using the biomaRt package (Durinck et al., 2009), ensembl gene IDs were converted to gene names. Ensembl gene IDs which did not have a corresponding gene name were left with their ensembl gene codes. Percentage of mitochondrial genes were calculated for quality control using the "^MT-" pattern, finding all mitochondrial genes which start with MT.

The Seurat object was filtered, retaining cells with number of features = 200 – 10,000 and cells with counts 500 to 3,000,000. Percentage of mitochondrial genes accepted between 5 - 95%. The high threshold was accepted as cardiomyocytes likely contain high percentage mitochondria content compared to other cells. Doublets were identified and removed using DoubletFinder (McGinnis et al., 2019) package which detects suspected doublets based on artificial nearest neighbours. Results in a filtered Seurat Object containing remaining samples.

The filtered gene counts were normalised and scaled using the default Seurat commands and parameters. The gene count dimensions were reduced using Seurat PCA commands and an elbow plot was used to determine the appropriate number of principle components.

To identify separation between the expression profile of the different sample types Seurat clustering was performed with the 'FindNeighbors' and 'FindClusters' functions using six PCA dimensions and resolutions 0.1, 0.3, 0.5, 0.7, 1. Dimplots were used to identify the most appropriate resolution for setting cluster identity using the 'Idents' function. UMAP and TSNE modelling was performed using Seurat 'RUN' command using six PCA dimensions.

4.3 Data Shuffling and Feature Selection

In order to create testing and training data, cell indices were shuffled and divided into six folds. Training and testing Seurat objects were created from the original Seurat object based on folds indices, folds 1:5 for training and 6 for testing. Venn diagrams were used to ensure that no indices were found in both sets.

Feature selection, selected genes of interest from the training data based on the 'FindVariableFeatures' Seurat command. Default selection parameter 'vst' and nfeatures was 3000. This selected 3000 genes of notable variance in the training data. These selected features were captured in another variable and were used to select the same 3000 features in the testing data. This minimises data leakage between the training and testing data sets. The scaled featured gene counts were then extracted from the training and test data and were written to new CSV files for model creation and testing. These CSV files contained sample type status and scaled gene counts for every cell in their respective training/test dataset. A summary of this pre-processing pipeline can be seen in Figure 4.

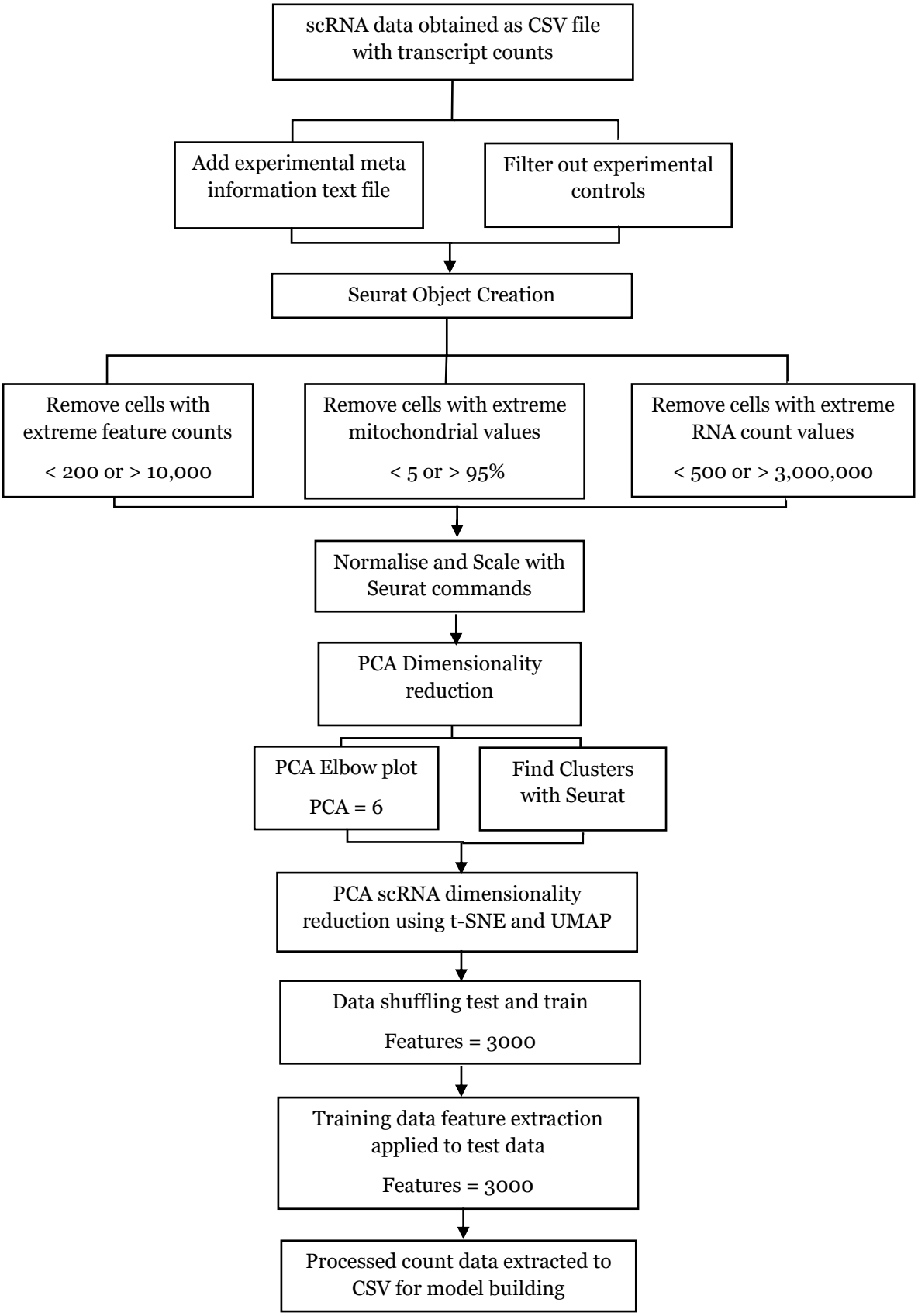

Figure 4. scRNA Preparation pipeline overview. From obtaining csv file with scRNA transcript counts to CSV of processed counts. 3000 features are scaled, normalised, modelled with dimensionality reduction and extracted. Preparation performed in R studio with Seurat package

4.4 Model Building

Models built in this thesis are categorised as supervised learning classification models. Various models were tested as described below. The output of these models is a multi-category classification of: Control_D14, Control_D2, Drug_D14, and Drug_D2. 'Control' being healthy unexposed cells and 'Drug' denoting those exposed to doxorubicin. D2 and D14 is two and fourteen-days after 'exposure'.

After preparation that data, CSV files were read in Google Colab with Python. Sample type was transformed using LabelEncoder for model classification. Data was split into train (n = 1123), validation (n = 281) and test (n =281) using random state. Model hyperparameter testing used sklearn.model_selection RandomSearchCV followed by GridSearch for exhaustive search of the optimal region found by random search. Hyperparameter search used cross fold validation set to three. Validation data was used to evaluate the random and grid hyperparameter search. Model evaluation was from the sklearn.metrics package classification_report and ROC. The classification report includes the metrics of precision, recall, F1-score and their number of support cases. When fitting the final model parameters, combined training and validation data was used (n = 1404). Final evaluation of the models used the prediction from the test set data.

Models' creation first used the 3000 features captured from Seurat data pre-processing. After model creation, important features were extracted using global importance metrics. In a goal to reduce model complexity and potentially improve model performance, new models were created using only the top 25 features from each model. Old and new model creation following hyperparameter searches stated above were used in an attempt to test the robustness of the larger model. Simple 25 feature models with no hyperparameter testing and use of default settings were tested if a simpler model resulted in better or worse outcomes. A selection of these 25 features extracted were summarised and biologically interpreted. A general flow for created models is shown in Figure 5.

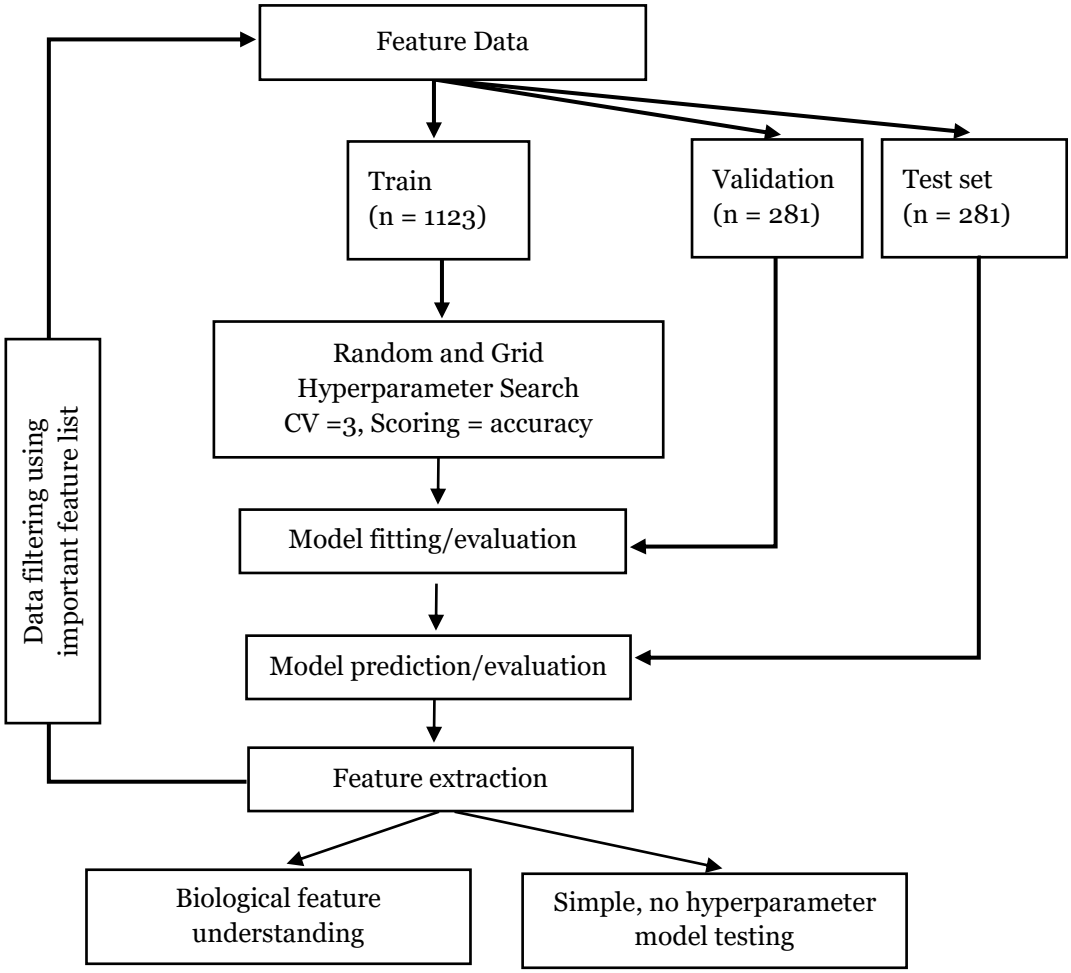

Figure 5. Model building pipeline. 3000 feature scRNA cell counts are split into train, validation and tests sets. Training data was used to random7grid search. Cross validation = 3 and scoring = Accuracy. Model predictions on the test data. Important features from the model were extracted. Features were used to filter features and then tested in the previous model and to re-model. Important features are explored on a biological level. Model creation was performed in Google Colab with Python.

4.4.1 Model 1 – Logistic Regression

Logistic regression classifier was imported from sklearn.ensemble and the optimal parameters were tested using model design options described in Table 1.

Table 1. Tested model parameters for the Logistic regression

Hyperparameter	**Testing range**
C	Uniform(loc=0.01, scale = 100)
Penalty	'L1', 'L2'
Solver	'Lbfgs', 'saga'

4.4.2 Model 2 – Random Forest Classifier

Random Forest classifier was imported from sklearn.ensemble and the optimal parameters were tested using model design options described in Table 2.

Table 2. Tested model parameters for the Random Forest Classifier

Hyperparameter	**Testing range**
n_estimators	randint 50, 200
Max_depth	randint 3, 10
min_samples_split	randint 2, 20
min_samples_leaf	randint 1, 20
max_features	'sqrt', 'log', 'None'

4.4.3 Model 3 - Gradient Boosting Classifier (XGBoost)

Gradient Boosting classification was imported from sklearn.ensemble and the optimal parameters were tested using model design options described in Table 3.

Table 3. Tested model parameters for the XGBoost Classifier

Hyperparameter	**Testing range**
N_estimators	50, 100
Learning rate	0.0.1, 0.1
Max_depth	3, 5

4.4.4 Model 4 – Artificial Neural Networks – Sequential

ANN classification was imported from tensorfow.keras. The sequential model was used with early stopping from keras.callbacks. Optimal parameters were tested using model design options described in Table 4. Early stopping was set to patience 10, including multiple layers 1-3 were included.

Table 4. Tested model parameters for the ANN Classifier Sequential

Hyperparameter	**Testing range**
N_estimators	50, 100
Learning rate	0.001, 0.01, 0.02
Neurons	5, 15, 30, 60
Optimizer	'SGD', 'Adam', 'relu'
Dropout_rate	0.0, 0.2, 0.4

4.4.5 Model 5 – Artificial Neural Networks – Multilayer Perceptron Classifier

ANN Multilayer Perceptron (MP) classification was imported from sklearn.neural_network. The model was used Random search. Optimal parameters were tested using model design options described in Table 5

Table 5. Tested model parameters for the ANN MLP Classifier

Hyperparameter	**Testing range**
Hidden_layer_sizes	50, 100, 150, 200
Activation	'relu', 'tanh', 'logistic'
Alpha	Uniform(0.0001, 0.01)
Learning_rate	'constant', 'adaptive'
Learning_rate_init	Uniform 0.001, 0.1

4.4.6 Model 6 – Convolutional Neural Networks and Image Classification

Images for CNN classification were created using pyDeepInsight package (Sharma et al., 2019). Using UMAP reducer due to reports that UMAP reduction results in superior reduction for scRNA data (Yang Yang et al., 2021). Pixel size set to 224, images for each cell were created based on gene feature values. The CNN RseNet50 was used to classify scRNA images. Batch size and epochs were tested for sizes 10, 20, 30, 40, 50. A reduced model was tested by using 45 features combined from the top 25 features from other models. This is because the global feature ranking was not possible with the CNN.

4.5 Explainable AI - Important Model Feature Discovery

Discovery of features used to classify cells was done using a variety of methods encompassing local and global importance feature extraction techniques.

Logistic regression global coefficients were extracted to sort and list the important features. Random forest feature importance was extracted with Gini importance. The higher the score the higher to importance in splitting the classes. XGBoost used the get_booster() function using weight as the feature importance score. MLP and ANN used permutation_importance (Altmann et al., 2010) from sklearn. ANN permutation importance used a reduced dataset (n = 50) and only the top 500 features extracted from preparation 'vst' method. This was due to limited computation processing requiring a reduced dimensionality. CNN classification extraction was based on activation pixels representing gene features. Using the 'mean' function with the CAMFeatureSelector which extracts the important features learnt with the trained PyTorch model. Local importance of individual class was explored using LIME (Ribeiro et al., 2016a). This revealed the feature weights effecting individual classifications.

4.6 Biological Understanding of features

Common features found from the model were collected using various model methods as mentioned above. Common important model features were searched via g:Profiler (Kolberg et al., 2023) and GO-term search. g:Profiler and GO-term are functional enrichment tools which allows gene name searches across multiple biological databases. These techniques interpret gene codes to group and understand their function and structure.

g:Profiler and GO-term biological process enrichment was performed using top 25 important feature lists extracted from each model. GO-term used Fisher's Exact and Bonferroni correct for multiple testing. Some proteins participate in multiple pathways, so the tests are not completely independent of each other and the Bonferroni correction is conservative. For ontology terms, the simple Bonferroni correction becomes extremely conservative reducing and simplifying the GO-terms summary.

Common gene features used by the machine learning models are search via scientific databases with the key term 'cardiovascular', 'cardio', 'cardiac', 'cardiotoxicity' or 'doxorubicin' to reveal if these genes are associated with cardiovascular function.

5 Results

5.1 Preparation of scRNA data and clustering

Seurat object created with a minimum number of cells = 50 and filtering resulted in 1685 cells to be used. PCA elbow plots identified the number of PCA components to be six. Figure 6 showing PCA 1:2 shows Drug_D2 groupings distinct from other samples. Figure 7 shows the UMAP and tSNE dimensional reduction by sample type identified Drug_D2 cells as distinctive to other sample types. Control and Drug_D14 were heterogeneously mixed, however Drug_D14 is clustered closer to day 14 neighbours, whereas controls are scattered throughout other groups.

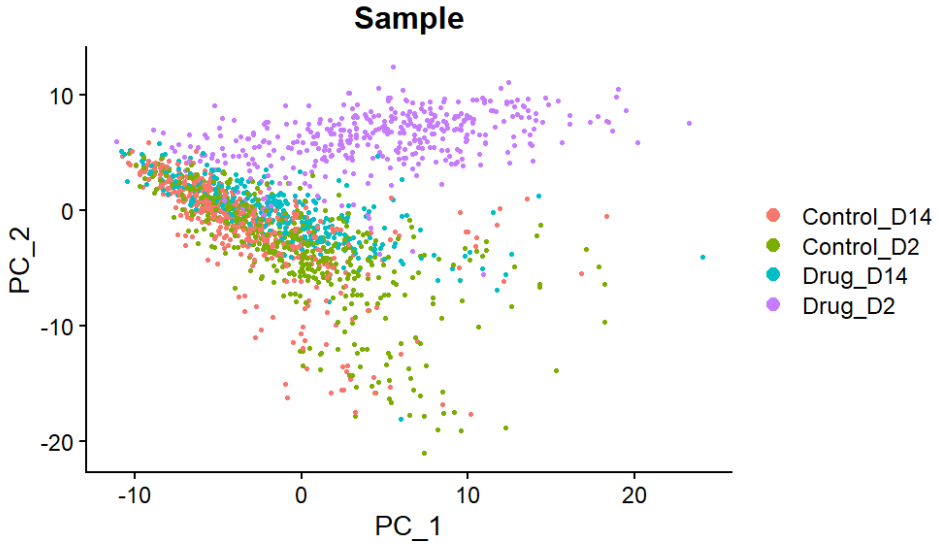

Figure 6. PCA 1 and 2 showing sample type groupings. Drug_D2 shows clear clustering with variation explained by PC_1. Other samples show mixed variation explained by PC_1 and PC_2.

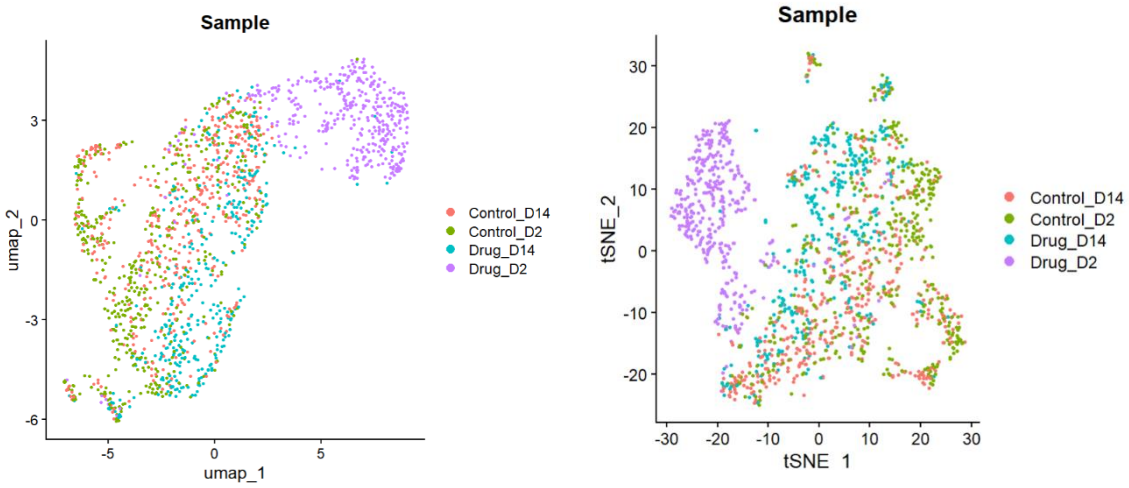

Figure 7. UMAP and tSNE groupings show clear clustering of Drug_D2 and more heterogenous spread of other samples with Drug_14 have a reduced spread compared to control sample types, Control_D2/D14.

5.2 Model building and comparison

In total six machine learning models were tested thoroughly. The best parameters are noted for each model followed with the classicisation report. The model's classification followed a similar pattern. Drug_D2 obtained high classification accuracy (~0.9) whilst Drug_D14, Control_D2/D14 obtained low scores (~0.6). Re-running filtered features and retaining the models did not improve the models substantially unless noted in the individual model description. Each model created is described below with the hyperparameters and ROC/AUC results.

5.2.1 Model 1 - Logistic Regression

Logistic regression model hyperparameter tuning found C = 1.2605, penalty = 'L2' and solver = 'saga' to be the optimal parameters. These resulted in the best scores as shown in Table 6 with a mean accuracy = 0.7. Both 3000 and 25 feature models behaved similar (Figure 8).

Table 6. Classification report for 3000 feature Random Forest Classifier.

Sample_Type	Precision	Recall	F1-Score	Support
Control Day 14	0.67	0.62	0.64	74
Control Day 2	0.56	0.70	0.60	69
Drug Day 14	0.71	0.68	0.70	56
Drug Day 2	0.93	0.87	0.90	82

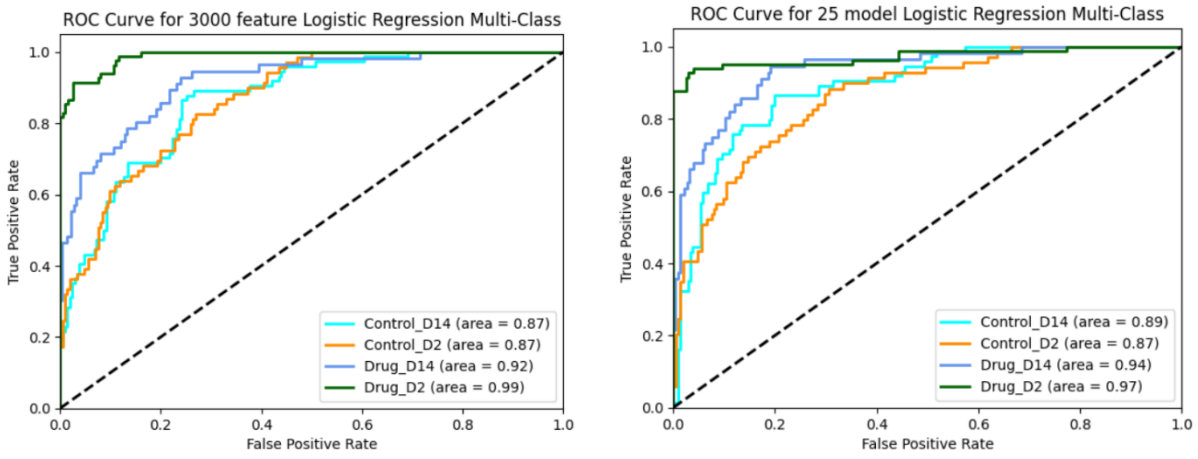

Figure 8. ROC of the Logistic Regression classifier with 3000 feature model (left) and 25 feature model (right). ROC is calculated using One vs Rest.

5.2.2 Model 2 – Random Forest Classifier

Random Forest classifier hyperparameter tuning found n_estimators = 168, max_depth = 8, min_samples_split = 11, min_samples_leaf = 9, max_features = None. These resulted in the best scores as shown in Table 7 with a mean accuracy = 0.7. Both 3000 and 25 feature models behaved similar (Figure 9).

Table 7. Classification report for 3000 feature Random Forest Classifier.

Sample_Type	Precision	Recall	F1-Score	Support
Control Day 14	0.55	0.57	0.56	74
Control Day 2	0.53	0.70	0.60	69
Drug Day 14	0.78	0.64	0.71	56
Drug Day 2	0.99	0.83	0.90	82

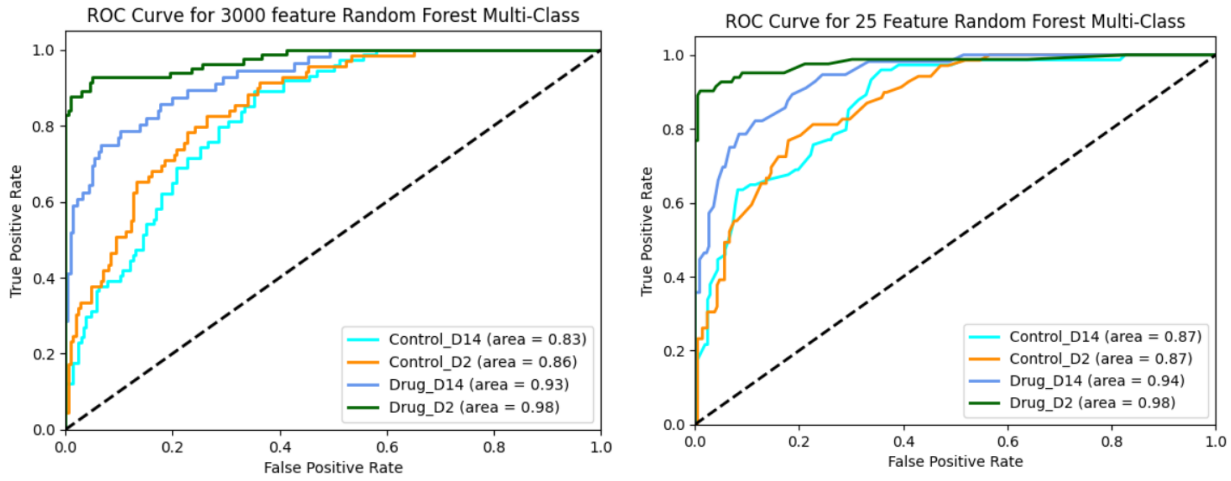

Figure 9. ROC of the Random Forest classifier with 3000 feature model (left) and 25 feature model (right). ROC is calculated using One vs Rest.

5.2.3 Model 3 - Gradient Boosting Classifier (XGBoost)

Extreme Gradient Boosting hyperparameter tuning found the best parameters to be learning rate = 0.5, max_depth = 5, and n_estimators = 121. Table 8 shows the classification report with a mean accuracy of 0.71. Both 3000 and 25 feature models behaved similar (Figure 10).

Table 8. Classification report for 3000 feature XGBoost.

Sample_Type	Precision	Recall	F1-Score	Support
Control Day 14	0.67	0.63	0.65	89
Control Day 2	0.59	0.72	0.65	80
Drug Day 14	0.83	0.68	0.75	63
Drug Day 2	0.89	0.86	0.88	59

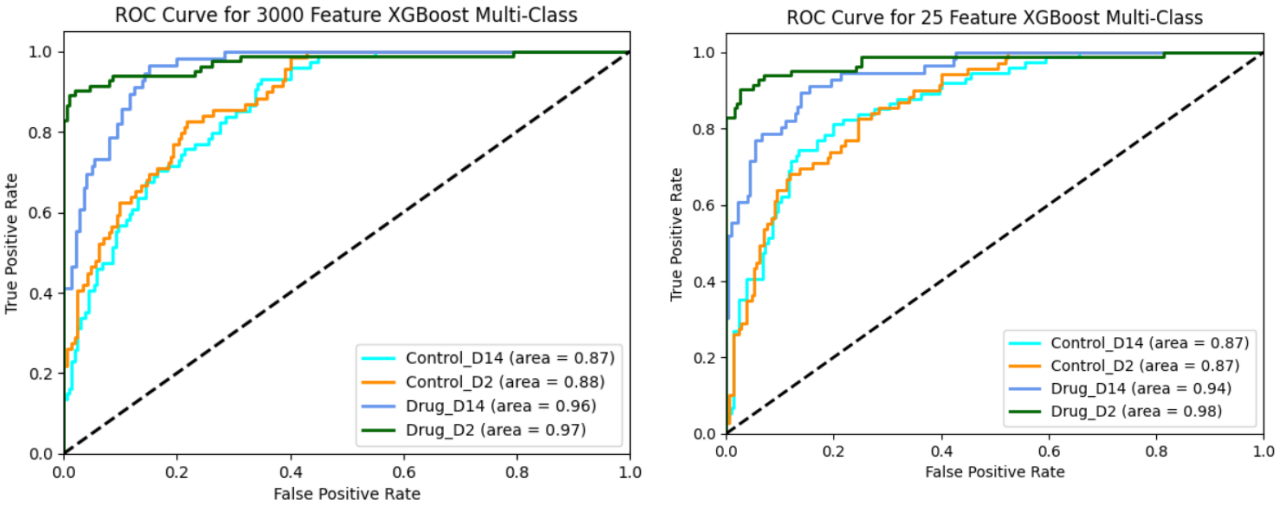

Figure 10. ROC of the XGBoost classifier with 3000 feature model (left) and 25 feature model (right). ROC is calculated using One vs Rest.

5.2.4 Model 4 – Artificial Neural Networks: Sequential

ANN hyperparameter tuning found the best parameters to be optimizer = 'Adam', neurons = 80, dropout_rate = 0.0, batch_size = 40, activation = 'uniform'. Table 9 shows the classification report with a mean accuracy of 0.63. Of note, RandomSearch and Grid Search showed conflicting best results. Rerunning the search parameters resulted in wide varying results. Early stopping found the epochs to be ~30 and additional layers had a negligible impact. 3000 feature model performed better than the 25 feature model (Figure 11).

Table 9. Classification report for ANN.

Sample_Type	Precision	Recall	F1-Score	Support
Control Day 14	0.80	0.43	0.56	74
Control Day 2	0.59	0.70	0.64	69
Drug Day 14	0.60	0.84	0.70	56
Drug Day 2	0.94	0.93	0.93	82

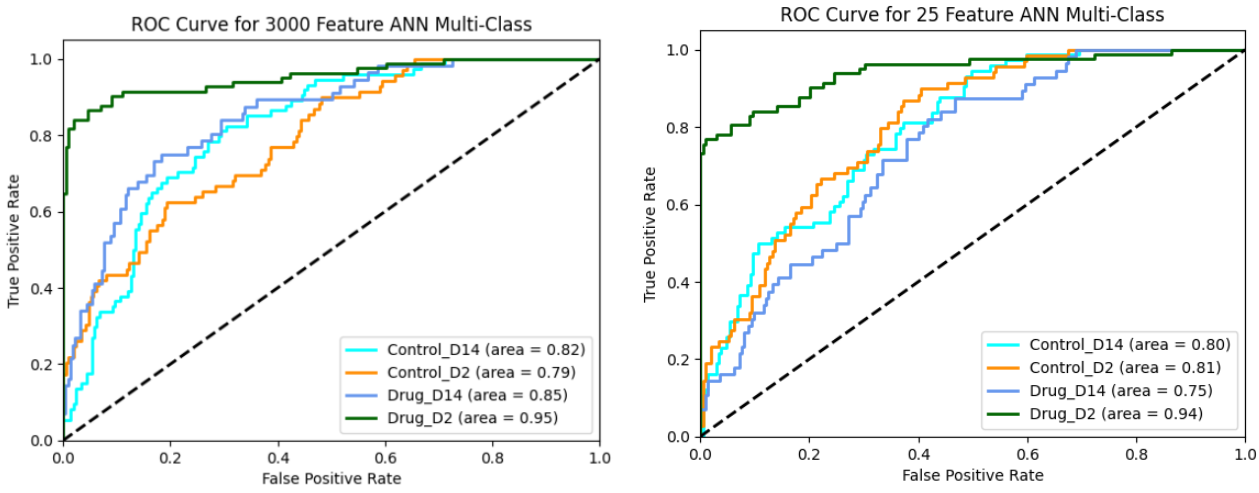

Figure 11. ROC of the ANN classifier with 3000 feature model (left) and 25 feature model (right). ROC is calculated using One vs Rest.

5.2.5 Model 5 – Artificial Neural Networks: Multilayer Perceptron

ANN Multilayer Perceptron hyperparameter tuning found the best parameters to be activation = 'logistic', alpha = 0.002, hidden_layer_sizes = 150, learning_rate_init = 0.04, max_iter = 1000, learning_rate = 'adaptive'. Table 10 shows the classification report with a mean accuracy of 0.68. Both 3000 and 25 feature models behaved similar (Figure 12).

Table 10. Classification report for MLP.

Sample_Type	Precision	Recall	F1-Score	Support
Control Day 14	0.61	0.58	0.60	74
Control Day 2	0.50	0.64	0.56	69
Drug Day 14	0.71	0.64	0.67	56
Drug Day 2	0.93	0.82	0.87	82

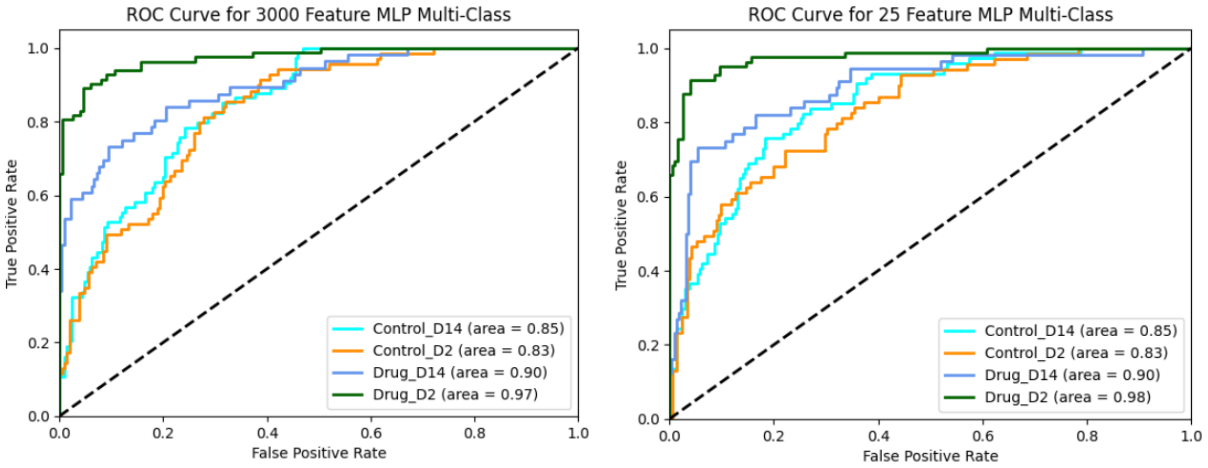

Figure 12. ROC of the MLP classifier with 3000 feature model (left) and 25 feature model (right). ROC is calculated using One vs Rest.

5.2.6 Model 6 – Convolutional Neural Networks and Image Classification

The CNN pyDeepInsight model generates images from scRNA data. This image (Figure 13), is then used to classify the cell sample type based on this image. Important pixel regions can be discovered highlighting the features used to discriminate the samples, Figure 13. Following default package set up with batch size = 20 and epoch = 20. Additional epochs or changing of the batch size had no real impact on the over accuracy of the classification. Table 11 shows the classification report with a mean accuracy of 0.69. Reducing the model for 45 features, the CNN model resulted in a slight drop in mean accuracy (0.65) and AUC (Figure 14).

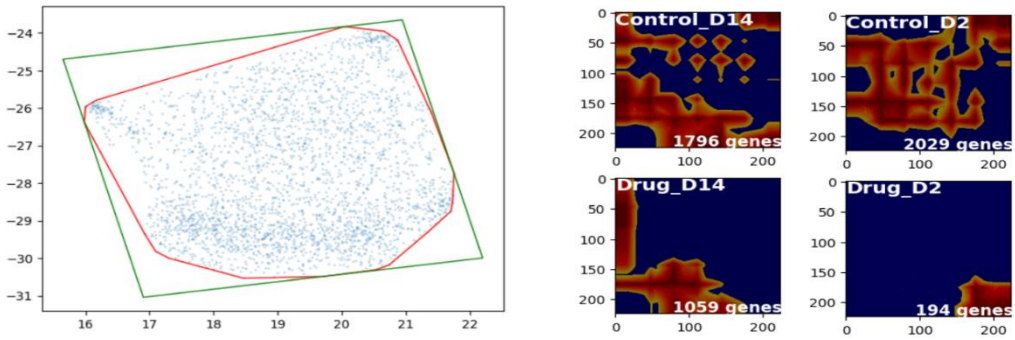

Figure 13. Sample of the images generated from the pyDeepInsight package for scRNA data (left). Pixel size set to 224. The activated pixel regions show overlap between sample types (right), with Drug_D2 being the most distinctive.

Table 11. Classification report for 3000 feature CNN model.

Sample_Type	Precision	Recall	F1-Score	Support
Control Day 14	0.45	0.69	0.54	48
Control Day 2	0.78	0.50	0.61	108
Drug Day 14	0.55	0.84	0.67	37
Drug Day 2	0.94	0.88	0.91	88

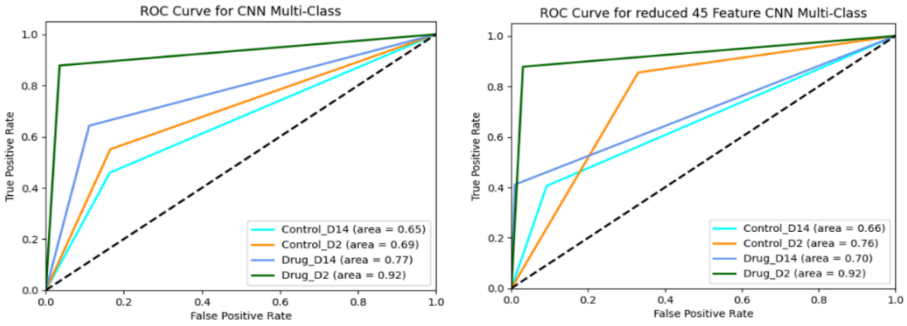

Figure 14. ROC of the MLP classifier with 3000 feature model (left) and 45 feature model (right). ROC is calculated using One vs Rest.

5.3 Classification Summary

A common theme from the models were the poor discrimination between Controls Day 2 and 14 which is to be expected as they should be functionally the same. Drug_D14 also presents poor discrimination. Drug_D2 however was clearly distinguished from other classes in most instances.

5.4 Biological Feature Discovery

The top 25 globally important features from all models were extracted and compared. Figure 15 shows there was considerable overlap between these models. Classical ML models have considerable overlap whilst NN models (ANN and MLP) diverged the most in important features. The ANN showed the greatest difference in feature importance. The 25 top features were checked for similarity and overlap shown in Table 12. A selection of these will be explored in greater detail.

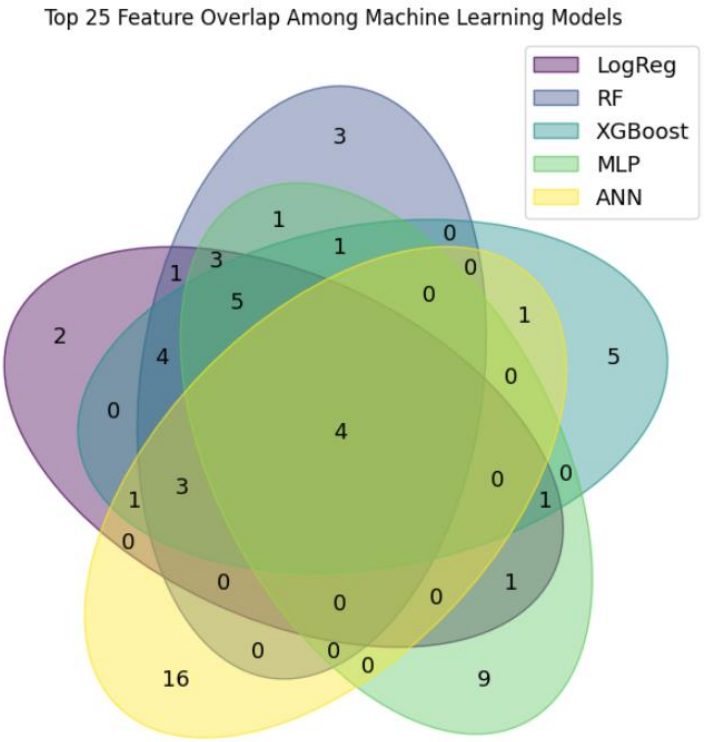

Figure 15. Top 25 genes from each model was extracted and overlaps reveal common patterns in important global features. Four are common between all models. NN models (ANN and MLP) had the least in common with each other. ANN model diverged the most in selected important features.

Table 12. Overlapping global features and top model features to discriminate between sample type classes. CNN is not included here as features are not ranked, only recorded.

Feature	Common Features	Logistic Regression	Random Forest	XGB	ANN	MLP
1	TMSB4X	LDHA	ACTC1	CDKN1A	H2AC8	TMAB4X
2	BANCR	CDKN1A	CDKN1A	ACTC1	GDF15	LDHA
3	NPPA	MYL7	LDHA	MYL7	A2M	SSR3
4	COX7A1	ACTA1	ACTA1	ACTA1	ZSCAN31	BANCR
5		MT.ND2	MYH7	LDHA	CAM1D	MYL7
6		B2M	RPS27L	TMSB4X	GABRA4	RPS27L
7		ACTC1	MYL7	B2M	ANO10	C1orf105
8		MYL2	TMSB4X	MT.ND2	RASL12	RFC1
9		TMSB4X	RGS5	ENO1	GOLGA8B	ENO1

Local feature importance

LIME was used to test individual case importance and permutation importance was also tested. As expected from the global classification report, individual cases of class Control_D14, Control_D2 and Drug_D14 showed small features influence on their classification. Local features were very clear for the Drug_D2 class which had much higher confidence classification scores.

Using the threshold 0.65, the CNN model found many genes overlaps between the different classes. In Table 13, overlap features can be seen with some interesting overlap between the Drug_D2 and Drug_D14 classes.

Table 13. Gene feature overlaps between classes with the CNN model using pyDeepInsight.

	Control_D14	Control_D2	Drug_D14	Drug_D2
Control_D14		1399	1021	94
Control_D2	1399		722	73
Drug_D14	1021	722		10
Drug_D2	94	73	10	

5.5 Biological Implications

There are various genes that consistently appeared as important features. The list is quite exhaustive but five common gene will be explored below. LDHA encodes lactate dehydrogenase A. This enzyme is a catalysis for the conversion of pyruvate to lactate with NADH in anaerobic glycolysis (NCBI, 2024). CDKN1A encodes for the vital cyclin-dependent kinase inhibitor protein. This protein, through binding, inhibits cyclin-cyclin—dependent-kinase complexes. Its function acts as a regulator for cell cycle progression. It is tightly controlled by the tumour suppressor p53. CDKN1A is involved in the response to DNA damage (NCBI, 2024a). ACTA1 belongs to the actin muscle group of proteins. It is a group highly conserved throughout nature and are involved in cell motility, structure and integrity. Diseases related to this gene result in muscle fibre defects and muscle weakness (NCBI, 2019). MYL7 is known as myosin light chain 7. As the name suggests, this gene codes for part of the protein myosin. Myosin's counterpart is actin, together they make up the structure of muscle fibres. This protein is also known for of calcium ion binding and almost all eukaryotic cells contain myosin. ACTC1 is another gene encoding for actin. ACTC1 is known as Actin Alpha Cardiac Muscle 1. It is commonly associated with cardiomyopathy and other heart muscle diseases.

The CNN model noted 10 genes overlapping between Drug_D14 and D2 aiding in classification. These important feature genes could contribute to the long-term cardiotoxicity seen in doxorubicin treatment. These genes potential is discussed below.

APOBEC3F – is a duplicated gene and members of this family encode for proteins thought to the involved in enzymes editing RNA which have roles in cell cycle control. Diseases associated with APOBEC3F are linked to immune and cancer related diseases. This gene is involved in a range of DNA and RNA modifications including demethylation, DNA damage responses and gene expression (Cervantes-Gracia et al., 2021).

NMRAL1 - encodes of a NADPH sensor protein which senses NADPH levels and migrated to inhibit phosphorylation of p65, a multi role protein. It therefore acts as a transcription regulator, controlling cellular processes. Reduced expression of NMRAL1 increases nitric oxide production and reduced cell viability (Qi et al., 2021).

TST – limited information can be found on this gene. TST proteins are found in the mitochondria and catalyses a range of reactions (NCBI, 2012).

TRAPPC8 – limited information can be found on this specific gene. TRAPP proteins are known to play a role trafficking the endoplasmic reticulum to Golgi apparatus organelle (Lamb et al., 2016). Dysfunction in TRAPP proteins have been associated with cardiovascular disorders through poor trafficking outcomes (Papaioannou et al., 2023).

TTC28.AS1 – an antisense RNA. They do not encode proteins but control expression and modification of other genes through interactions with sense RNA. It is most likely involved in the cell cycle (Human Protein Atlas, 2022). This antisense RNA has associations with proteins regulating cardiovascular cell function and which has implications in cardiovascular disease (Bu et al., 2022).

PAXBP1 – this gene encodes for the protein predicted to enable DNA binding activity. It is predicted to be involved in positive regulation of myoblast and regulation of skeletal muscle. It was found upregulated in calcified aortic heart valves (Latif et al., 2015).

EXOSC1 – this gene is involved in the regulation of RNA. Other members of the EXOSC family

have been implicated in complete heart blockage. EXOSC5 variants have been implicated in this heart disorder, suggesting that RNA dysregulation may play a role in cardiotoxicity (Latif et al., 2015). EXOSC1 was found upregulated two days after MI (Y. Wang et al., 2018).

MMP11 – are a part of a wider family called MMP related to the breakdown extracellular matrixes in normal processes such as embryonic development and tissue remodelling (NCBI, 2008). MMPs are abundant in cardiomyocytes and have been implicated in doxorubicin related cardiotoxicity in multiple studies as an early event. Oxidative and nitrosative stress results in MMPs activation and has been shown to contribute to dilated cardiomyopathy and congestive heart failure (Bai et al., 2004). (Chan et al., 2021) have suggested that MMP inhibitors could be a therapeutic agent to reduce the effects of doxorubicin induced cardiotoxicity. By inhibiting MMPs intracellular and extracellular matrix remodelling can be prevented, possibly leading to a reduction in the harmful effects of doxorubicin.

CSNK1E – this gene is involved in control of cellular processes and interacts with a circadian rhythm gene (NCBI, 2014). Methylation of these circadian rhythm genes have been implicated in increased cardiometabolic risks (Ahmadi et al., 2022).

SLC5A3 – this gene is a potassium channel regulator. It is predicted to work with the regulation of reactive oxygen species and therefore oxidative stress (NCBI, 2022). Information about how this gene is involved in CVD is unknown but this gene is found near genes loci with association to CVD (Beaney et al., 2017).

5.5.1 Biological Term Enrichment

The top 25 features extracted from each model were enriched to find common biological processes there are involved with (Table 14). As expected, the majority of terms are related to cellular stress, but also the morphology of cells and the function of cardiac muscle fibres.

Table 14. Number of significant biological processes features from top 25 important features used in the ML models. Enriched using GO term biological processes and Bonferroni correction. Counts may not be exactly 25 as features may be multi-role or role unknown.

Significant biological process	**Logistic Regression**	**Random Forest**	**XGBoost**	**ANN**	**MLP**	**CNN***
Muscle filament sliding	11	11	8			
Cardiac muscle hypertrophy in response to stress	4	4	4			
Regulation of the force of heart contraction	4	8	4		4	
Ventricular cardiac muscle tissue development	14		14		6	
Myofibril assembly	8					
Cardiac muscle contraction	9		7	2		
Regulation of striated muscle contraction	6		5			
Regulation of heart rate	4	4	4			
ATP metabolic process	7	7				
Cardiac muscle tissue morphogenesis		13				
Heart contraction		8				
Sarcomere organisation			10			
Cardiac tissue development					7	
Striated muscle contraction				3		
Unmapped IDs	3	2	3	1	3	

* CNN has ten local importance gene overlapping Drug_D14 and Drug_D2. GO and g: Profiler found no hits.

6 Discussion

Overall, the models showed a similar pattern. Accuracy scores were ~0.7. Individual classification scores revealed a class divide in discrimination power. Control_D14/D2 and Drug_D14 were poorly classified (~0.6) whilst Drug_D2 was easily classified (~0.9). This suggests that the expression profile of the control samples and Drug_D14 were highly similar, leading to poor discrimination. Drug_D2 having experienced doxorubicin most recently, showed a clear expression profile different in response to the drug and was easily classifiable Figure 6. This is not unexpected as it has been reported that after exposure cells return back to normal expression after two weeks (McSweeney et al., 2019). Furthermore, UMAP and t-SNE analysis showed clear cluster discrimination between Drug_D2 and the other samples which were highly heterogenous (Figure 7).

It is not clear from this work, the potential long-term causes of cardiotoxicity after doxorubicin treatment. Whilst the models' feature importance revealed global discriminatory features, the poor classification of Drug_D14 means that these features were only useful in discriminating between Drug_D2 and the other classes. Local feature analysis confirms this, where LIME cases show poor feature scores when classifying Control_D14/D2 and Drug_2 classes.

6.1 Previous Research

Papers exploring the used of machine learning models to understand or predict biological functions are wide ranging. From predicting cardiotoxicity with patient data to exploring iPSC cardiomyocytes as in this thesis.

One paper used a combination of patient data and biological data containing features of interest from doxorubicin treatment to create a cardiotoxicity risk prediction random forest model with a higher accuracy than just biological or patient data alone (Chaix et al., 2020). Chaix et al, used 31 genes as biological features. Interestingly, of these features none of the top 25 features discovered in this paper were included (Figure 15). Their use of random forest with sequencing data demonstrates its robustness with sequencing data types. In this study they found that the genes which were linked to protect against cardiotoxicity were useful for prediction metrics (Chaix et al., 2020).

In the paper by Alimadadi et al., (2020), bulk RNA sequencing data (mixed cell type samples) from human left heart ventricle tissues from patients suffering with cardiomyopathy were collected. Cardiomyopathy is the disorder effecting the heart, where the heart loses its ability to pump blood. Many of the biological pathways identified in Table 14 contribute to such a disorder. They trained support vector machines, decision trees, neural networks with PCA, elastic net and random forest. This machine learning methods resulted in accuracy scores of 80 - 93%. 50 highly important genes were then extracted and used to retain the models resulting in an improved accuracy for random forest and neural networks with PCA. There is some small overlap seen between features in (Alimadadi et al., 2020) and this work. Accuracy improvement was not seen in the models rebuilding in this thesis. It could be due to the highly heterogenous nature of cell classes Drug_D14 and Control_D14/D2. The data used in Alimadadi et al., was RNA bulk sequencing of different disease types which may have better separation between the classes.

6.2 Methods, implementation and results

6.2.1 Pre-processing

Preparation scRNA followed rough guidelines from (Luecken & Theis, 2019) and Dr. Dario Melguizo Sanchis. Filtering of cell outliers was done following these recommendation and contextual biological information. For example, typically high mitochondrial gene percentage is an indicator of cell death and should not be included in downstream analysis. However, cardiomyocytes as heart cells have high energy requirements and therefore have high mitochondrial counts. It was justifiable that high percentage mitochondrial gene cells were included.

Influencing the model creation was the use of the extract feature option using the default 'vst' function and 3000 features. There aren't clear guidelines in how many features to extract within scRNA literature, based on global importance learnt a much lower number would have been adequate. Furthermore, testing other extract important feature settings might have revealed different or the same genes found in the model feature extractions. The data randomised before splitting into training and test sets was investigated for balance. The data is slightly unbalanced with fewer Drug_D2/D14 samples. However, they were not sufficiently unbalanced to cause concern.

6.2.2 Model building

Performing random search followed with grid search was computationally expensive and time consuming. Although it likely found the optimal hyperparameter regions for classification, it was not exhaustive, many default parameters were used. Hyperparameter tuning could be shortened by first running categorical hyperparameters and running random search/grid search with the numerical options. This would reduce the number of combinations and speed up model building.

Many types of models were built and tested. A 3000-feature model followed by two types of reduced model. A local reduced model based on the individual model's important feature scale and a global reduction model where all the models was re-built based on the top 25 features by the model's local importance ranking. The global reduction models were required to provide a consistent benchmark across models as their local important ranking can be interpreted in different ways and are not consistent between models. In all cases, using old and updated model parameters, their performance remained similar.

This likely means that the models were resilient to noisy features not contributing to the model. However, it also reveals that hyperparameter searching as not efficient. Hyperparameter searching was extensive and time consuming but these models only show merger improvements over the default models' settings.

3000 features were used from the pre-processing stage, however as seen in the reduced models, this number could have been greatly reduced with whilst achieving the same classification scores. There are no strict guidelines with how many features to capture in the pre-processing stage, in retrospect, it would have been beneficial to test a very limited number of features first followed by expanding the number to see if it had an impact on the scores. This would reduce computational requirements and encourage faster hyperparameter searching. Furthermore, a reduction in features could have been tested from the start using

sklearn.feature_selection options such as recursive feature elimination with cross-validation to select features (Misra & Singh Yadav, 2020). These could have been used to reduce dimensionality on datasets, potentially improving model performance.

If the models were to be deployed as is in the final stage of the CRISP-DM data science model then the reduced models would be the most useful if we were to consider speed and simplicity for future classifications.

6.2.3 CrossFold validation

Crossfold validation was implemented with hyperparameter search. Crossfold aims to prevent overfitting in the model by splitting the data into sets, training the model with the majority of sets and then validating the performance with the remaining set. Generally speaking, the higher the K fold values (the more folds) the better. However, more K is computationally expensive (Wong & Yeh, 2020). Nested fold validation was attempted but this became slow and inefficient using Google Colab. If more resources were available then the process could be distributed across multiple platforms and results collated.

Nested cross validation could have been an improvement during the hyperparameter investigation stage as it incorporates an additional robustness through data diversity. However, similar to findings by Wainer & Cawley (2021), cross validation is often enough for hyperparameter and model performance. The reduced models with no hyperparameter testing had similar performance to 3000 feature hyperparameter tested models. Therefore, nested validation was not a requirement to improved performance and could be omitted in interest of time and computational restrictions.

6.2.4 Model results

Models created showed similar classification scores throughout. Drug_D2 was easily classified with F1-score ~0.9. This high classification success is not replicated with the other sample types; controls and Drug_D14. However, this is to be expected. UMAP and t-SNE modelling (Figure 7) show that Control_D14/D2 and Drug_14 are highly heterogenous with no clear expression clustering, whereas Drug_D2 shows clear separation from the other classes. In the way, the models are great at discriminating between Drug_D2 but the other classes are almost 50/50 chance as been classed as each other.

Random forest is able to working with a range of data types and tasks. In this task gene features and expression were used to split samples and make classifications. Random forest possesses advantages by using multiple decision tress to make a ‘forest’ mitigating the weaknesses associated with decision trees. Random forest helps with overfitting and it is simple to extract important features via investigating splits in the decision trees. The model was simple to implement and classification scores were on par with other models. Random forest is known to be able to process unscaled and unnormalized data (Lei et al., 2020). It would have been interesting to apply random forest to the unprocessed data to see if similar conclusions were reached. If so, a random forest would be a good candidate for processing scRNA data which requires extensive processing whilst not hindering scores.

As mentioned, logistic regression is one of the simplest machine learning models. Logistic regression assumes a linear relationship and models the log-odds of an event (Sperandei, 2014). Important features were easily extracted with the coefficients. It would seem that the data was linear as logistic regression was able to perform well. The features extracted from the

model were used in the ANN, thus features determined linearly important were captured in a form of feature reduction. Although sensitive to outliers, it is hoped that the large sample size (n = 1401) and removal of some biological cell outliers have mitigated that. The solver saga and lbfgs was tested for multiclass. Saga was determined to be best that might be due to the scale of the data as saga is for lager datasets (Defazio et al., 2014). Saga is designed to converge faster than standard gradient descent by reducing gradient variance, making it more suitable for large datasets.

XGBoost is a popular choice of model. It combines multiple weaker models and combines them. It behaves similar to decision trees but requires less resources. Determining the number of boosting stages is n_estimatators, which over the default 100 setting. This shouldn't be a problem as gradient boosting is robust to overfitting. Most useful, was the simplicity that important features can be extracted (Friedman, 2001). Although XGBoost can require less resources by default, it did take a long time due to the high dimensional data used for training.

scRNA sequencing data is noisy. The hope was that the ANN would be able to deal with this as ANNs are good at capturing complex data patterns. However, the ANNs developed did not offer greater classification than classical ML methods. Incorporating additional layers did not help in classifications. When running random and grid search, it was apparent that the best hyperparameters seemed to vary wildly with every run. This would suggest that the model is struggling to capture trends and is overly complex. The random initialisation might be having strong effects on hyperparameter selection. Regardless, the model's performance did not vary significantly despite the changes.

MLP did rank important features differently to the other models. This is likely due to the learning dynamics of MLP backpropagation and gradient descent which can lead to different top features highlighted due to the stochastic nature. Initial weights of the MLP NN may also contribute to this (Géron, 2023).

The conversion of scRNA data to images is an interesting method to aid in classification of cells. The CNN used to classify images showed a wide range of features separating the classes. The model performed as well as other models. Changes in learning epochs or batch size made no impact on the model performance. Improvements could have been made by including more features, potentially increasing image complexity. Further advances in the image transformation such as in the paper by Jia et al., (2023) use the relationship between features to add image complexity. This additional biological and image complexity could reveal more biological relevant patterns for exploration. Testing of CNN augmentation could have been explored further in this work as it showed improved success in other pyDeepInsight papers (Grafton et al., 2021). However, it is suspected that as found in other ML methods, the difficult classification due to class overlap means that augmentation would not improve the model prediction.

6.2.5 Explainability: Global and local importance

Global feature importance found some similar features from models. Global importance did have some conflicting ranking, however that should be expected as the models learn in different ways. It should be considered that feature importance can be calculated in different ways sometimes within the same model such as XGBoost. The most commonly used metrics were chosen in this thesis work for simplicity. It was clear from Figure 15, that the classic machine learning methods (Logistic Regression, Random Forest and XGBoost) shared highly similar important features, whilst NN models show divergence. This could be due to a few

reasons. Firstly, the weight initialisation and structure of the models might have strong influences on the feature importance. Secondly it could be due to differences in importance extraction. Whilst the majority of models used in-built feature extractions, the MLP and ANN model used permutation importance. That being said permutation importance was tested with some classic ML models can the same features were found as from their native feature extractions.

Local importance was most useful via LIME which was able to show prediction values for important features. As expected from the high heterogenicity of Control_D14/D2 and Drug_D14, the local prediction of these classes showed very low feature influence values. Drug_D2 being the most distinct expression group had clear feature influence values. With time invested it would be possible to make summaries of the influencing factors of these classes to reveal the positive or negative class effects. Local permutation importance was implementing in some models but not all so won't be discussed at length, but showed similar patterns to LIME. A paper by Främling et al., (2021) argues that LIME and SHAP explanations actually estimate influence rather than importance and found that LIME results tended to be unstable and incoherent. In an extensive local importance review by Hooker et al., (2021), they state that when feature exhibit statistical dependence, permute and predict methods can be highly misleading when applied to the model.

An advance on the local importance tools such as permutation importance, LIME, and SHAP is the package meLIME (Botari, Hvilshøj, et al., 2020). These packages create visualisations for the local decisions so that individual cases can be viewed and understood. Unfortunately, time did not allow this method to be explored and applied to all models. Regardless, important model features have clear biological links to heart toxicity and/or heart malfunction. These are explored in the section below.

6.2.6 Biological results discussions

Various papers have reported the negative effects of high LDHA levels after cardiovascular events and suggest the LDHA levels can be used for predicting long term changes in cardiac structure and function after heart attacks (H. Zhang et al., 2024). Specifically, LDHA has been associated with enlarged hearts and reduced left ventricular ejection (H. Zhang et al., 2024) as well as arterial stiffness and 10-year cardiovascular disease (Zhu et al., 2022).

CDKN1A also known as p21 is tightly linked with the cellular cycle control and thus will likely have a big impact of cellular proliferation and overall health. Studies has linked CDKN1A to antioxidant defence systems in response to cellular stress stimuli in cardiomyocytes (W. Chen et al., 2009). CDKN1A expression changes have been linked to another drug induced cardiotoxicity (Liu et al., 2020). Jiang et al., 2021) investigated the used of micro-RNA as a therapeutic agent of myocardial infarction. They found that these agents target the CDKN1A protein influencing the progression of myocardial infarction. All such evidence points to CDKN1A/p21 being an important discriminator involved in doxorubicin and indeed general cardio-toxicity.

A paper by McSweeney et al., (2019) investigated the role of doxorubicin on cardiomyocytes with exposure followed with RNA sequencing. Similar to work in this thesis, extractions had two-time framings: 7 days and 14 days after exposure. They explored gene expression change and followed with gene annotation. 1290 genes were significantly up and down regulated from drug exposure. McSweeney found that on day seven 1290 genes and day 14 showed 315 genes that were dysregulated compared to controls. The most common pathways found in the

downregulation set was those in the cell cycle and cell division including the CDKN1A gene. Similar to genes found in this thesis, this indicated that cell division mechanisms are slowed down by changes in the expression of cell cycle progression genes. McSweeney's paper found that 14 day longer term cardiomyocytes had continued cell cycle gene dysregulation, whilst other earlier term exposure genes were returning to normal expression. This can contribute to the poor classification results of the Drug_D14 cells used in model creation. They have mainly returned to baseline expression matching the controls.

200 different mutations in ACTA1 gene are known to cause skeletal muscle disease (Nowak et al., 2013). Therefore, it is clear that ACTA1 dysregulation can have far reaching consequences on cellular health. Interestingly according to the 2015 paper by (Donkervoort et al., 2015) cardiac involvement is rarely seen in ACTA1 related muscle weakness or degeneration. However, according to the machine learning feature extractions it appears to be one of the most important and frequent features to discriminate between classes. The heart losing its ability to pump blood is called cardiomyopathy. In the journal letter by Reza et al., (2018) found a unique case of cardiomyopathy associated with novel ACTA1 variant. Highlighting the role of ACTA1 dysfunction resulting in cardiac health problems.

MYL7 (Myosin light chain 7) has been shown to the critical of regulation of cardiac function (Grimm et al., 2005) relating to heart contraction and atrial function. Mice embryos injected with MYL7 inactivators resulted in fatal cardiac consequences (C. Huang et al., 2003). Deng et al., (2007) treated mice with doxorubicin and found that chronic treatment resulted in mortality, cardiac fibrosis and cardiac insufficiency. A range of gene expression changes were discovered including that of MYL7, ACTA1 and MYL4.

ACTC1, actin alpha cardiac muscle 1, has be clearly linked to heart disease. In studies reduced expression has been suggested to induce cardiomyocyte apoptosis, cell death (H.-K. Jiang et al., 2010). An additional study investigating doxorubicin and cardiomyocytes also found changes via downregulation of myosin and actin related genes (Holmgren et al., 2018).

In summary, it appears that some of the global important features learnt by ML algorithm show that doxorubicin interferes with genes involved in cardiac stability, contraction, cellular cycles, arterial stiffness and apoptosis. There were many more genes that are not here explored in depth but further study could reveal biomarkers and cardiotoxicity mechanisms.

6.2.7 Local important difference with CNN image classification

CNN image classification found ten genes which overlapped between Drug_D14 and Drug_D2. Through literature searching, many of these genes have been shown to have a link to cardiovascular disorders and dysfunction. However, in many cases it is not clear exactly how. These genes may also play a role in the cardiotoxicity of Drug_D2 but were not mapped in such a way that the CNN would include them in Drug_D2 classification. It has been recorded that gene expression returns to normal after doxorubicin expression (Holmgren et al., 2018). These gene features can provide a starting point to discover longer term changes in gene expression after doxorubicin exposure. Conversely, they could be a product the scRNA data to image transformation and further testing would be required to check if these genes appear consistently and warrant further biological testing.

6.3 Ethical and societal aspects

Work performed here contributes to knowledge surrounding cardiotoxicity with doxorubicin treatment. Such information could be useful for future researchers in identifying pathways associated with doxorubicin treatment. This work contributes to UN development goal #3; Good health and well-being (UN, 2015)

Human embryonic stem cell-derived cardiomyocytes, Cellartis Pure hES-CM, were obtained from Takara Bio Europe AB Swedish Filial (Gothenburg, Sweden). These cells are licensed for use of scientific research and are derived from inner cell mass of blastocyst-stage embryos. These cell lines are commonly used for model cell cultures. However, many studies report the divergence of cell lines to form individual cell types which may not represent *in vivo* conditions or those that could be replicated (Hu et al., 2010; Yoshihara et al., 2016). This could lead to conclusions not representative of 'real-world' experiences. This is an on-going problem within cell lines research and not solvable here. It is also important to note that iPSC cardiomyocytes only replicate 2D or single cell biological behaviours can findings from research should not the extended to discoveries of the whole heart functions and structure as an organ (Nijak et al., 2021).

7 Conclusion

This thesis demonstrates the effectiveness of machine learning models in identifying crucial features from noisy, high-dimensional scRNA sequencing data. The findings show that classic machine learning models, such as Logistic Regression, Random Forest, and XGBoost, displayed higher consistency in identifying important features compared to the artificial neural networks. Convolutional neural networks (CNNs) also emerged as a novel and promising method for classifying scRNA data, offering an alternative approach that may warrant further exploration.

The study revealed that Drug_D2 samples were easily classified, achieving an AUC of ~0.9, indicating a clear distinction between these samples and controls. In contrast, distinguishing between Control_D2/14 and Drug_D14 samples proved challenging, with AUCs ranging from ~0.7 to 0.8. This difficulty suggests that Drug_D14 expression profiles are very similar to controls, complicating the understanding of long-term effects of doxorubicin.

The key features extracted by these models have strong associations with known biological pathways involved in doxorubicin-induced cardiotoxicity. However, ANNs identified fewer significant hits in biological process databases like GO Terms and g:Profiler. This suggests that ANNs may be identifying less well-established pathways and hold potential for uncovering previously unexplored biological features important to doxorubicin-induced cardiotoxicity. These findings highlight the potential necessity of using a consensus approach, combining insights from multiple models to capture a comprehensive set of important features. Relying solely on one model type, such as ANNs, could miss critical features linked to known pathways, whereas integrating results from both classic ML models and deep learning approaches might provide a fuller picture of the biological processes at play.

Additionally, the results indicate that simple models with few features, even without extensive hyperparameter tuning, performed comparably to more complex models with hyperparameter optimization. This suggests that simplicity may offer advantages in model interpretability and computational efficiency without sacrificing performance.

Overall, this work contributes to a deeper understanding of the complex nature of doxorubicin-induced cardiotoxicity. It highlights the value of combining machine learning models to identify and rank key genetic and transcriptomic features, ultimately enhancing knowledge of both established and potentially novel biological mechanisms driving cardiomyocyte response to doxorubicin exposure.

7.1 Future work

Future work could focus on refining deep learning models, particularly ANNs, to better identify biologically relevant features in scRNA data. This could include exploring advanced architectures, such as graph neural networks (J. Wang et al., 2021) . To avoid potential biases introduced by pre-processing, experimenting with unprocessed data (e.g., non-scaled and non-normalized) with ANNs might be an avenue of research preventing bias. Additionally, research could investigate the expression of cardiotoxicity features across varying levels of doxorubicin exposure and different patient conditions to predict the likelihood of cardiotoxicity. Integrating multi-omics data, such as proteomics or metabolomics, could further enhance understanding of doxorubicin-induced cardiotoxicity. Deep learning approaches appear to have a useful approach in combining these different data types (Athaya et al., 2023; Miao et al., 2021).

Identified features could be combined with other patient data to create improved models for predicting patient response to doxorubicin and the likelihood of cardiotoxicity. Developing consensus-based approaches that combine outputs from multiple models would also ensure a more comprehensive capture of key features. Finally, experimental validation of these features is crucial for confirming their role in cardiotoxicity and guiding new therapeutic strategies.

References

Abdolrasol, M. G. M., Suhail Hussain, S. M., Ustun, T. S., Sarker, M. R., Hannan, M. A., Mohamed, R., Ali, J. A., Mekhilef, S., & Milad, A. (2021). Artificial Neural Networks Based Optimization Techniques: A Review. *Electronics 2021, Vol. 10, Page 2689*, *10*(21), 2689. https://doi.org/10.3390/ELECTRONICS10212689

Abu-Doleh, A., & Al Fahoum, A. (2024). XgCPred: Cell type classification using XGBoost-CNN integration and exploiting gene expression imaging in single-cell RNAseq data. *Computers in Biology and Medicine*, *181*, 109066. https://doi.org/10.1016/J.COMPBIOMED.2024.109066

Ahmadi, S. A., Tranmer, J. E., Ritonja, J. A., Flaten, L., Topouza, D. G., Duan, Q. L., Durocher, F., Aronson, K. J., & Bhatti, P. (2022). DNA methylation of circadian genes and markers of cardiometabolic risk in female hospital workers: An exploratory study. *Chronobiology International*, *39*(5), 735–746. https://doi.org/10.1080/07420528.2022.2032729

Alimadadi, A., Manandhar, I., Aryal, S., Munroe, P. B., Joe, B., & Cheng, X. (2020). Machine learning-based classification and diagnosis of clinical cardiomyopathies. *Physiological Genomics*, *52*(9), 391–400. https://doi.org/10.1152/physiolgenomics.00063.2020

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, *12*(7), 878. https://doi.org/10.15252/MSB.20156651/ASSET/4A945DE9-C07A-47FB-B2F1-430ADD4361AC/ASSETS/GRAPHIC/MSB156651-FIG-0005-M.PNG

Athaya, T., Ripan, R. C., Li, X., & Hu, H. (2023). Multimodal deep learning approaches for single-cell multi-omics data integration. *Briefings in Bioinformatics*, *24*(5). https://doi.org/10.1093/bib/bbad313

Bai, P., Mabley, J., Liaudet, L., Virág, L., Szabó, C., & Pacher, P. (2004). Matrix metalloproteinase activation is an early event in doxorubicin-induced cardiotoxicity. *Oncology Reports*. https://doi.org/10.3892/or.11.2.505

Beaney, K. E., Smith, A. J. P., Folkersen, L., Palmen, J., Wannamethee, S. G., Jefferis, B. J., Whincup, P., Gaunt, T. R., Casas, J. P., Ben-Shlomo, Y., Price, J. F., Kumari, M., Wong, A., Ong, K., Hardy, R., Kuh, D., Wareham, N., Kivimaki, M., Eriksson, P., … Consortium, U. (2017). Functional Analysis of the Coronary Heart Disease Risk Locus on Chromosome 21q22. *Disease Markers*, *2017*, 1–10. https://doi.org/10.1155/2017/1096916

Bernstein, N. J., Fong, N. L., Lam, I., Roy, M. A., Hendrickson, D. G., & Kelley, D. R. (2020). Solo: Doublet Identification in Single-Cell RNA-Seq via Semi-Supervised Deep Learning. *Cell Systems*, *11*, 95-101.e5. https://doi.org/10.1016/j.cels.2020.05.010

Bloom, J. D. (2018). Estimating the frequency of multiplets in single-cell RNA sequencing from cell-mixing experiments. *PeerJ*, *2018*(9), e5578. https://doi.org/10.7717/PEERJ.5578/SUPP-4

Botari, T., Hvilshj, F., Izbicki, R., & De Carvalho, A. C. P. L. F. (2020). *MeLIME: Meaningful Local Explanation for Machine Learning Models*. https://github.com/tiagobotari/melime.

Botari, T., Hvilshøj, F., Izbicki, R., & de Carvalho, A. C. P. L. F. (2020). *MeLIME: Meaningful Local Explanation for Machine Learning Models*. https://arxiv.org/abs/2009.05818v1

Breiman, L. (2001). *Random Forests* (Vol. 45). https://doi.org/10.1023/A:1010933404324

Brendel, M., Su, C., Bai, Z., Zhang, H., Elemento, O., & Wang, F. (2022). Application of Deep Learning on Single-Cell RNA Sequencing Data Analysis: A Review. *Genomics, Proteomics & Bioinformatics*, *20*(5), 814–835. https://doi.org/10.1016/J.GPB.2022.11.011

Bu, S., Nguyen, H. C., Michels, D. C. R., Rasheed, B., Nikfarjam, S., Singh, R., Wang, L., Patel, D. A., Singh, S., Qadura, M., & Singh, K. K. (2022). Transcriptomics of angiotensin II-induced long noncoding and coding RNAs in endothelial cells. *Journal of Hypertension*, *40*(7), 1303–1313. https://doi.org/10.1097/HJH.0000000000003140

Cao, L. (2018). Data Science. *ACM Computing Surveys*, *50*(3), 1–42. https://doi.org/10.1145/3076253

Cervantes-Gracia, K., Gramalla-Schmitz, A., Weischedel, J., & Chahwan, R. (2021). APOBECs orchestrate genomic and epigenomic editing across health and disease. *Trends in Genetics*, *37*(11), 1028–1043. https://doi.org/10.1016/j.tig.2021.07.003

Chaix, M. A., Parmar, N., Kinnear, C., Lafreniere-Roula, M., Akinrinade, O., Yao, R., Miron, A., Lam, E., Meng, G., Christie, A., Manickaraj, A. K., Marjerrison, S., Dillenburg, R., Bassal, M., Lougheed, J., Zelcer, S., Rosenberg, H., Hodgson, D., Sender, L., … Mital, S. (2020). Machine Learning Identifies Clinical and Genetic Factors Associated With Anthracycline Cardiotoxicity in Pediatric Cancer Survivors. *Cardio Oncology*, *2*(5), 690–706. https://doi.org/10.1016/J.JACCAO.2020.11.004

Chan, B. Y. H., Roczkowsky, A., Cho, W. J., Poirier, M., Sergi, C., Keschrumrus, V., Churko, J. M., Granzier, H., & Schulz, R. (2021). MMP inhibitors attenuate doxorubicin cardiotoxicity by preventing intracellular and extracellular matrix remodelling. *Cardiovascular Research*, *117*(1), 188–200. https://doi.org/10.1093/cvr/cvaa017

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785/SUPPL_FILE/KDD2016_CHEN_BOOSTING_SYSTEM_01-ACM.MP4

Chen, W., Sun, Z., Wang, X.-J., Jiang, T., Huang, Z., Fang, D., & Zhang, D. D. (2009). Direct Interaction between Nrf2 and p21Cip1/WAF1 Upregulates the Nrf2-Mediated Antioxidant Response. *Molecular Cell*, *34*(6), 663–673. https://doi.org/10.1016/j.molcel.2009.04.029

Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, *99*(6), 323–329. https://doi.org/10.1016/J.YGENO.2012.04.003

Defazio, A., Bach, F., & Lacoste-Julien, S. (2014). SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. *Advances in Neural Information Processing Systems*, *2*(January), 1646–1654.

https://arxiv.org/abs/1407.0202v3

Deng, S., Kulle, B., Hosseini, M., Schlüter, G., Hasenfuss, G., Wojnowski, L., & Schmidt, A. (2007). Dystrophin-deficiency increases the susceptibility to doxorubicin-induced cardiotoxicity. *European Journal of Heart Failure*, *9*(10), 986–994. https://doi.org/10.1016/j.ejheart.2007.07.016

Donkervoort, S., Yang, M., Leach, M., Medne, L., Yum, S., Hotchkiss, L., Rutkowski, A., Chatfield, K., Auerbach, S., Miyamoto, S., & Bönnemann, C. (2015). Cardiomyopathy in patients with ACTA1-myopathy. *Neuromuscular Disorders*, *25*, S287. https://doi.org/10.1016/j.nmd.2015.06.361

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols 2009 4:8*, *4*(8), 1184–1191. https://doi.org/10.1038/nprot.2009.97

Feng, X., Xiu, Y. H., Long, H. X., Wang, Z. T., Bilal, A., & Yang, L. M. (2023). Advancing single-cell RNA-seq data analysis through the fusion of multi-layer perceptron and graph neural network. *Briefings in Bioinformatics*, *25*(1), 1–13. https://doi.org/10.1093/BIB/BBAD481

Flores, M., Liu, Z., Zhang, T., Hasib, M. M., Chiu, Y. C., Ye, Z., Paniagua, K., Jo, S., Zhang, J., Gao, S. J., Jin, Y. F., Chen, Y., & Huang, Y. (2022). Deep learning tackles single-cell analysis—a survey of deep learning for scRNA-seq analysis. *Briefings in Bioinformatics*, *23*(1). https://doi.org/10.1093/BIB/BBAB531

Främling, K., Westberg, M., Jullum, M., Madhikermi, M., & Malhi, A. (2021). Comparison of Contextual Importance and Utility with LIME and Shapley Values. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12688 LNAI*, 39–54. https://doi.org/10.1007/978-3-030-82017-6_3/FIGURES/8

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189–1232. http://www.jstor.org/stable/2699986

Géron, A. (2023). *Hands On Machine Learning with Scikit Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly.

Grafton, F., Ho, J., Ranjbarvaziri, S., Farshidfar, F., Budan, A., Steltzer, S., Maddah, M., Loewke, K. E., Green, K., Patel, S., Hoey, T., & Mandegar, M. A. (2021). Deep learning detects cardiotoxicity in a high-content screen with induced pluripotent stem cell-derived cardiomyocytes. *ELife*, *10*. https://doi.org/10.7554/ELIFE.68714

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: an Overview*. https://arxiv.org/abs/2008.05756v1

Grimm, M., HAAS, P., WILLIPINSKISTAPELFELDT, B., ZIMMERMANN, W., RAU, T., PANTEL, K., WEYAND, M., & ESCHENHAGEN, T. (2005). Key role of myosin light chain (MLC) kinase-mediated MLC2a phosphorylation in the ?-adrenergic positive inotropic effect in human atrium. *Cardiovascular Research*, *65*(1), 211–220. https://doi.org/10.1016/j.cardiores.2004.09.019

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., & Van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature 2015 525:7568*, *525*(7568), 251–255.

https://doi.org/10.1038/nature14966

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN Model-Based Approach in Classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2888*, 986–996. https://doi.org/10.1007/978-3-540-39964-3_62

Guo, J., & Gao, J. (2022). Comparison of different machine learning algorithms on Cell Classification with scRNA-seq after Principal Component Analysis. *2022 7th International Conference on Intelligent Computing and Signal Processing, ICSP 2022*, 1476–1479. https://doi.org/10.1109/ICSP54964.2022.9778439

Hao, Y., Stuart, T., Kowalski, M. H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., & Satija, R. (2023a). Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology 2023 42:2*, *42*(2), 293–304. https://doi.org/10.1038/s41587-023-01767-y

Hao, Y., Stuart, T., Kowalski, M. H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., & Satija, R. (2023b). Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology 2023 42:2*, *42*(2), 293–304. https://doi.org/10.1038/s41587-023-01767-y

Hastie, T., Friedman, J., & Tibshirani, R. (2001). *The Elements of Statistical Learning*. https://doi.org/10.1007/978-0-387-21606-5

Hatton, I. A., Galbraith, E. D., Merleau, N. S. C., Miettinen, T. P., Smith, B. M. D., & Shander, J. A. (2023). The human cell count and size distribution. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(39), e2303077120. https://doi.org/10.1073/PNAS.2303077120/SUPPL_FILE/PNAS.2303077120.SD02.XLSX

Heller, M. (2019). *InfoWorld's 2019 Technology of the Year Award winners | InfoWorld*. InfoWorld. https://www.infoworld.com/article/3673770/infoworld-s-2019-technology-of-the-year-award-winners.html?page=2

Henriksen, P. A. (2018). Anthracycline cardiotoxicity: an update on mechanisms, monitoring and prevention. *Heart*, *104*(12), 971–977. https://doi.org/10.1136/HEARTJNL-2017-312103

Holmgren, G., Sartipy, P., Andersson, C. X., Lindahl, A., & Synnergren, J. (2018). Expression Profiling of Human Pluripotent Stem Cell-Derived Cardiomyocytes Exposed to Doxorubicin—Integration and Visualization of Multi-Omics Data. *Toxicological Sciences*, *163*(1), 182–195. https://doi.org/10.1093/toxsci/kfy012

Hooker, G., Mentch, L., & Zhou, S. (2021). Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, *31*(6), 1–16. https://doi.org/10.1007/S11222-021-10057-Z/FIGURES/9

Hu, B. Y., Weick, J. P., Yu, J., Ma, L. X., Zhang, X. Q., Thomson, J. A., & Zhang, S. C. (2010). Neural differentiation of human induced pluripotent stem cells follows developmental principles but with variable potency. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(9), 4335–4340. https://doi.org/10.1073/PNAS.0910012107/SUPPL_FILE/PNAS.200910012SI.PDF

Huang, C., Sheikh, F., Hollander, M., Cai, C., Becker, D., Chu, P.-H., Evans, S., & Chen, J. (2003). Embryonic atrial function is essential for mouse embryogenesis, cardiac morphogenesis and angiogenesis. *Development*, *130*(24), 6111–6119. https://doi.org/10.1242/dev.00831

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *17*(3), 299–310. https://doi.org/10.1109/TKDE.2005.50

Human Protein Atlas. (2022). *TTC28 protein expression summary - The Human Protein Atlas*. https://www.proteinatlas.org/ENSG00000100154-TTC28

Jia, S., Lysenko, A., Boroevich, K. A., Sharma, A., & Tsunoda, T. (2023). scDeepInsight: a supervised cell-type identification method for scRNA-seq data with deep learning. *Briefings in Bioinformatics*, *24*(5), 1–12. https://doi.org/10.1093/BIB/BBAD266

Jiang, H.-K., Qiu, G.-R., Li-Ling, J., Xin, N., & Sun, K.-L. (2010). Reduced ACTC1 Expression Might Play a Role in the Onset of Congenital Heart Disease by Inducing Cardiomyocyte Apoptosis. *Circulation Journal*, *74*(11), 2410–2418. https://doi.org/10.1253/circj.CJ-10-0234

Jiang, W., Song, Q., Lu, Z., Wang, S., Liu, T., Wang, X., & Wang, M. (2021). Myocardial Infarction-Associated Extracellular Vesicle-Delivered miR-208b Affects the Growth of Human Umbilical Vein Endothelial Cells via Regulating CDKN1A. *BioMed Research International*, *2021*. https://doi.org/10.1155/2021/9965639

Jin, S., Zeng, X., Xia, F., Huang, W., & Liu, X. (2021). Application of deep learning methods in biological networks. *Briefings in Bioinformatics*, *22*(2), 1902–1917. https://doi.org/10.1093/BIB/BBAA043

Kaneko, H. (2022). Cross-validated permutation feature importance considering correlation between features. *Analytical Science Advances*, *3*(9–10), 278–287. https://doi.org/10.1002/ANSA.202200018

Kolberg, L., Raudvere, U., Kuzmin, I., Adler, P., Vilo, J., & Peterson, H. (2023). g:Profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkad347

Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, *39*(4), 261–283. https://doi.org/10.1007/s10462-011-9272-4

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S. O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B. de, Cappuccio, A., … Schönhuth, A. (2020). Eleven grand challenges in single-cell data science. *Genome Biology*, *21*(1). https://doi.org/10.1186/S13059-020-1926-6

Lamb, C. A., Nühlen, S., Judith, D., Frith, D., Snijders, A. P., Behrends, C., & Tooze, S. A. (2016). TBC1D14 regulates autophagy via the TRAPP complex and ATG9 traffic. *The EMBO Journal*, *35*(3), 281–301. https://doi.org/10.15252/EMBJ.201592695

Latif, N., Sarathchandra, P., Chester, A. H., & Yacoub, M. H. (2015). Expression of smooth muscle cell markers and co-activators in calcified aortic valves. *European Heart Journal*, *36*(21), 1335–1345. https://doi.org/10.1093/EURHEARTJ/EHT547

Lei, Y., Harms, J., Wang, T., -, al, Nadler, E. O., Mao, Y.-Y., & Zhu, T. (2020). Analysis on the

Applicability of the Random Forest. *Journal of Physics: Conference Series*, *1607*(1), 012123. https://doi.org/10.1088/1742-6596/1607/1/012123

Li, X., Wang, K., Lyu, Y., Pan, H., Zhang, J., Stambolian, D., Susztak, K., Reilly, M. P., Hu, G., & Li, M. (2020). Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nature Communications 2020 11:1*, *11*(1), 1–14. https://doi.org/10.1038/s41467-020-15851-3

Liu, R., Li, D., Sun, F., Rampoldi, A., Maxwell, J. T., Wu, R., Fischbach, P., Castellino, S. M., Du, Y., Fu, H., Mandawat, A., & Xu, C. (2020). Melphalan induces cardiotoxicity through oxidative stress in cardiomyocytes derived from human induced pluripotent stem cells. *Stem Cell Research and Therapy*, *11*. https://doi.org/10.1186/s13287-020-01984-1

Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, *15*(6), 8746. https://doi.org/10.15252/MSB.20188746/SUPPL_FILE/MSB188746-SUP-0003-DATASETEV1.ZIP

Lundberg, S. M., Allen, P. G., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*. https://github.com/slundberg/shap

Martinez, S., Garduno, G. ;, Toloue, A. ;, Ostadgavahi, A. ;, Hewins, B. ;, Dutt, M. ;, Kumar, A. ;, Martin-Loeches, I. ;, Kelvin, D. J., Sganzerla Martinez, G., Garduno, A., Ostadgavahi, A. T., Hewins, B., Dutt, M., Kumar, A., Martin-Loeches, I., & Kelvin, D. J. (2024). Identification of Marker Genes in Infectious Diseases from ScRNA-seq Data Using Interpretable Machine Learning. *International Journal of Molecular Sciences 2024, Vol. 25, Page 5920*, *25*(11), 5920. https://doi.org/10.3390/IJMS25115920

Mcginnis, C. S., Murrow, L. M., & Correspondence, Z. J. G. (2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Systems*, *8*, 329-337.e4. https://doi.org/10.1016/j.cels.2019.03.003

McGinnis, C. S., Murrow, L. M., & Gartner, Z. J. (2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Systems*, *8*(4), 329-337.e4. https://doi.org/10.1016/J.CELS.2019.03.003

McSweeney, K. M., Bozza, W. P., Alterovitz, W.-L., & Zhang, B. (2019). Transcriptomic profiling reveals p53 as a key regulator of doxorubicin-induced cardiotoxicity. *Cell Death Discovery*, *5*(1), 102. https://doi.org/10.1038/s41420-019-0182-6

Medium. (2023). *Artificial Neural Networks in Machine Learning* . https://tirendazacademy.medium.com/artificial-neural-networks-in-machine-learning-fa653d74b1a1

Miao, Z., Humphreys, B. D., McMahon, A. P., & Kim, J. (2021). Multi-omics integration in the age of million single-cell data. *Nature Reviews Nephrology*, *17*(11), 710–724. https://doi.org/10.1038/s41581-021-00463-x

Misra, P., & Singh Yadav, A. (2020). Improving the Classification Accuracy using Recursive Feature Elimination with Cross-Validation. *International Journal on Emerging Technologies*, *11*(3), 659–665. www.researchtrend.net

Moolayil, J. (2019). *Learn Keras for Deep Neural Networks A Fast-Track Approach to*

Modern Deep Learning with Python-Jojo Moolayil. https://doi.org/10.1007/978-1-4842-4240-7

Mosavi, A., Salimi, M., Ardabili, S. F., Rabczuk, T., Shamshirband, S., & Varkonyi-Koczy, A. R. (2019). State of the Art of Machine Learning Models in Energy Systems, a Systematic Review. *Energies 2019, Vol. 12, Page 1301*, *12*(7), 1301. https://doi.org/10.3390/EN12071301

Müller, A. C. ., & Guido, Sarah. (2017). *Introduction to machine learning with Python : a guide for data scientists*. 378.

Munsky, B., Neuert, G., & Van Oudenaarden, A. (2012). Using Gene Expression Noise to Understand Gene Regulation. *Science*, *336*(6078), 183–187. https://doi.org/10.1126/SCIENCE.1216379

Musolf, A. M., Holzinger, E. R., Malley, J. D., & Bailey-Wilson, J. E. (2022). What makes a good prediction? Feature importance and beginning to open the black box of machine learning in genetics. *Human Genetics*, *141*(9), 1515–1528. https://doi.org/10.1007/S00439-021-02402-Z/FIGURES/5

NCBI. (2008). *MMP11 matrix metallopeptidase 11 [Homo sapiens (human)] - Gene - NCBI*. https://www.ncbi.nlm.nih.gov/gene/4320#summary

NCBI. (2012). *TST thiosulfate sulfurtransferase [Homo sapiens (human)] - Gene - NCBI*. https://www.ncbi.nlm.nih.gov/gene/7263#summary

NCBI. (2014). *CSNK1E casein kinase 1 epsilon [Homo sapiens (human)] - Gene - NCBI*. https://www.ncbi.nlm.nih.gov/gene/1454#summary

NCBI. (2015). *CDKN1A cyclin dependent kinase inhibitor 1A [Homo sapiens (human)] - Gene - NCBI*. https://www.ncbi.nlm.nih.gov/gene/1026#summary

NCBI. (2019). *ACTA1 actin alpha 1, skeletal muscle [Homo sapiens (human)] - Gene - NCBI*. https://www.ncbi.nlm.nih.gov/gene/58#summary

NCBI. (2022). *SLC5A3 solute carrier family 5 member 3 [Homo sapiens (human)] - Gene - NCBI*. https://www.ncbi.nlm.nih.gov/gene/6526#summary

NCBI. (2023). *LDHA lactate dehydrogenase A [Homo sapiens (human)] - Gene - NCBI*. NCBI. https://www.ncbi.nlm.nih.gov/gene/3939#summary

Nebigil, C. G., & Désaubry, L. (2018). Updates in anthracycline-mediated cardiotoxicity. *Frontiers in Pharmacology*, *9*(NOV), 415242. https://doi.org/10.3389/FPHAR.2018.01262/BIBTEX

Nicodemus, K. K. (2011). Letter to the Editor: On the stability and ranking of predictors from random forest variable importance measures. *Briefings in Bioinformatics*, *12*(4), 369–373. https://doi.org/10.1093/BIB/BBR016

Nijak, A., Saenen, J., Labro, A. J., Schepers, D., Loeys, B. L., & Alaerts, M. (2021). iPSC-Cardiomyocyte Models of Brugada Syndrome—Achievements, Challenges and Future Perspectives. *International Journal of Molecular Sciences*, *22*(6), 2825. https://doi.org/10.3390/ijms22062825

Nowak, K. J., Ravenscroft, G., & Laing, N. G. (2013). Skeletal muscle α-actin diseases (actinopathies): pathology and mechanisms. *Acta Neuropathologica*, *125*(1), 19–32. https://doi.org/10.1007/s00401-012-1019-z

Nti, I. K., Nyarko-Boateng, O., Aning, J., & Justice, A. (2021). Performance of Machine Learning Algorithms with Different K Values in K-fold Cross-Validation. *Article in International Journal of Information Technology and Computer Science*, *6*, 61–71. https://doi.org/10.5815/ijitcs.2021.06.05

Nvidia. (n.d.). *XGBoost – What Is It and Why Does It Matter?* Retrieved May 13, 2024, from https://www.nvidia.com/en-us/glossary/xgboost/

Oskolkov, N. (2019). *Deep Learning for Data Integration*. Towards Data Science. https://towardsdatascience.com/deep-learning-for-data-integration-46d51601f781

Papaioannou, P., Wallace, M. J., Malhotra, N., Mohler, P. J., & El Refaey, M. (2023). Biochemical Structure and Function of TRAPP Complexes in the Cardiac System. *Basic to Translational Science*, *8*(12), 1599–1612. https://doi.org/10.1016/J.JACBTS.2023.03.011

Petropoulos, S., Rd, D. E., Linnarsson, S., Sandberg, R., Lanner, F., Edsgä, D., Rn Reinius, B., Deng, Q., Panula, S. P., Codeluppi, S., & Reyes, A. P. (2016). Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos Resource Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell*, *165*, 1012–1026. https://doi.org/10.1016/j.cell.2016.03.023

Pouyan, M. B., & Kostka, D. (2018). Random forest based similarity learning for single cell RNA sequencing data. *Bioinformatics*, *34*(13), i79–i88. https://doi.org/10.1093/bioinformatics/bty260

Qi, Y., Wang, X., Li, W., Chen, D., Meng, H., & An, S. (2021). Pseudogenes in Cardiovascular Disease. *Frontiers in Molecular Biosciences*, *7*. https://doi.org/10.3389/fmolb.2020.622540

Rai, A. (2020). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Rawat, P. S., Jaiswal, A., Khurana, A., Bhatti, J. S., & Navik, U. (2021). Doxorubicin-induced cardiotoxicity: An update on the molecular mechanism and novel therapeutic strategies for effective management. *Biomedicine & Pharmacotherapy*, *139*, 111708. https://doi.org/10.1016/J.BIOPHA.2021.111708

Renu, K., V.G., A., Tirupathi, T. P., & Arunachalam, S. (2018). Molecular mechanism of doxorubicin-induced cardiomyopathy – An update. *European Journal of Pharmacology*, *818*, 241–253. https://doi.org/10.1016/J.EJPHAR.2017.10.043

Reza, N., Garg, A., Merrill, S. L., Chowns, J. L., Rao, S., & Owens, A. T. (2018). *ACTA1* Novel Likely Pathogenic Variant in a Family With Dilated Cardiomyopathy. *Circulation: Genomic and Precision Medicine*, *11*(10). https://doi.org/10.1161/CIRCGEN.118.002243

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). "Why Should I Trust You?" *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Proceedings of the Demonstrations Session, 97–101. https://doi.org/10.18653/v1/n16-3020

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Sapoval, N., Aghazadeh, A., Nute, M. G., Antunes, D. A., Balaji, A., Baraniuk, R., Barberan, C. J., Dannenfelser, R., Dun, C., Edrisi, M., Elworth, R. A. L., Kille, B., Kyrillidis, A., Nakhleh, L., Wolfe, C. R., Yan, Z., Yao, V., & Treangen, T. J. (2022). Current progress and open challenges for applying deep learning across the biosciences. *Nature Communications 2022 13:1*, *13*(1), 1–12. https://doi.org/10.1038/s41467-022-29268-7

Schweidtmann, A. M., & Mitsos, A. (2019). Deterministic Global Optimization with Artificial Neural Networks Embedded. *Journal of Optimization Theory and Applications*, *180*(3), 925–948. https://doi.org/10.1007/S10957-018-1396-0/TABLES/8

Sharma, A., Vans, E., Shigemizu, D., Boroevich, K. A., & Tsunoda, T. (2019). DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture. *Scientific Reports 2019 9:1*, *9*(1), 1–7. https://doi.org/10.1038/s41598-019-47765-6

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, *24*(1), 12–20. https://doi.org/10.11613/BM.2014.003

Srinivasu, P. N., Sandhya, N., Jhaveri, R. H., & Raut, R. (2022). From Blackbox to Explainable AI in Healthcare: Existing Tools and Case Studies. *Mobile Information Systems*, *2022*, 1–20. https://doi.org/10.1155/2022/8167821

Steinherz, L. J., Steinherz, P. G., Tan, C. T. C., Heller, G., & Murphy, M. L. (1991). Cardiac Toxicity 4 to 20 Years After Completing Anthracycline Therapy. *JAMA*, *266*(12), 1672–1677. https://doi.org/10.1001/JAMA.1991.03470120074036

Suzuki, K. (2011). *Artificial Neural Networks-Methodological Advances and Biomedical Applications*. www.intechopen.com

Takara Bio. (n.d.). *ICELL8 cx Single-Cell System*.

Tian, T., Wan, J., Song, Q., & Wei, Z. (2019). Clustering single-cell RNA-seq data with a model-based deep learning approach. *Nature Machine Intelligence 2019 1:4*, *1*(4), 191–198. https://doi.org/10.1038/s42256-019-0037-0

UN. (2015). *THE 17 GOALS | Sustainable Development*. https://sdgs.un.org/goals

van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*.

Vapnik, V. Naumovich. (1995). *The nature of statistical learning theory*. 188.

Vasighizaker, A., Zhou, L., & Rueda, L. (2021). *Cell Type Identification via Convolutional Neural Networks and Self-Organizing Maps on Single-Cell RNA-Seq Data*. https://doi.org/10.1145/3459930.3471171

Villani, A. C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., Jardine, L., Dixon, D., Stephenson, E., Nilsson, E., Grundberg, I., McDonald, D., Filby, A., Li, W., De Jager, P. L., … Hacohen, N. (2017). Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, *356*(6335).

https://doi.org/10.1126/SCIENCE.AAH4573/SUPPL_FILE/AAH4573_VILLANI_SM.PDF

von Hoff, D. D., Layard, M. W., Basa, P., Davis, H. L., Von Hoff, A. L., Rozencweig, M., & Muggia, F. M. (1979). Risk factors for doxorubicin-induced congestive heart failure. *Annals of Internal Medicine*, *91*(5), 710–717. https://doi.org/10.7326/0003-4819-91-5-710

Wainer, J., & Cawley, G. (2021). Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Systems with Applications*, *182*, 115222. https://doi.org/10.1016/j.eswa.2021.115222

Wang, D., & Gu, J. (2018). VASC: Dimension Reduction and Visualization of Single-Cell RNA-Seq Data by Deep Variational Autoencoder. *Genomics, Proteomics & Bioinformatics*, *16*(5), 320–331. https://doi.org/10.1016/J.GPB.2018.08.003

Wang, J., Ma, A., Chang, Y., Gong, J., Jiang, Y., Qi, R., Wang, C., Fu, H., Ma, Q., & Xu, D. (2021). scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. *Nature Communications*, *12*(1), 1882. https://doi.org/10.1038/s41467-021-22197-x

Wang, Y., Huang, Y., Zhang, M., Zhang, X., Tang, X., & Kang, Y. (2018). Bioinformatic Analysis of the Possible Regulative Network of miR-30a/e in Cardiomyocytes 2 Days Post Myocardial Infarction. *Acta Cardiologica Sinica*, *34*(2), 175–188. https://doi.org/10.6515/ACS.201803_34(2).20170926A

Webb, G. I., & Bayes, N. (2017). Naïve Bayes. *Encyclopedia of Machine Learning and Data Mining*. https://doi.org/10.1007/978-1-4899-7502-7_581-1

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, *1*, 29–39.

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, *19*(1). https://doi.org/10.1186/S13059-017-1382-0

Wong, T.-T., & Yeh, P.-Y. (2020). Reliable Accuracy Estimates from *k* -Fold Cross Validation. *IEEE Transactions on Knowledge and Data Engineering*, *32*(8), 1586–1594. https://doi.org/10.1109/TKDE.2019.2912815

Yang Yang, A., Sun, H., Zhang, Y., Yang, Y., Wu, D., Yu Correspondence dwu, D., Yang, Y., Zhang, T., Gong, J., Wei, Y., Duan, Y.-G., Shu, M., & Yu, D. (2021). Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. *CellReports*, *36*, 109442. https://doi.org/10.1016/j.celrep.2021.109442

Yoshihara, M., Hayashizaki, Y., & Murakawa, Y. (2016). Genomic Instability of iPSCs: Challenges Towards Their Clinical Applications. *Stem Cell Reviews and Reports 2016 13:1*, *13*(1), 7–16. https://doi.org/10.1007/S12015-016-9680-6

Yuan, M., Chen, L., & Deng, M. (2022). scMRA: a robust deep learning method to annotate scRNA-seq data with multiple reference datasets. *Bioinformatics*, *38*(3), 738–745. https://doi.org/10.1093/BIOINFORMATICS/BTAB700

Zappone, A., Di Renzo, M., Debbah, M., Lam, T. T., & Qian, X. (2019). Model-Aided Wireless Artificial Intelligence: Embedding Expert Knowledge in Deep Neural Networks for

Wireless System Optimization. *IEEE Vehicular Technology Magazine*, *14*(3), 60–69. https://doi.org/10.1109/MVT.2019.2921627

Zhang, H., Kang, K., Chen, S., Su, Q., Zhang, W., Zeng, L., Lin, X., Peng, F., Lin, J., & Chai, D. (2024). High serum lactate dehydrogenase as a predictor of cardiac insufficiency at follow-up in elderly patients with acute myocardial infarction. *Archives of Gerontology and Geriatrics*, *117*, 105253. https://doi.org/10.1016/J.ARCHGER.2023.105253

Zhang, T., Wu, Z., Li, L., Ren, J., Zhang, Z., & Wang, G. (2024). CPPLS-MLP: a method for constructing cell–cell communication networks and identifying related highly variable genes based on single-cell sequencing and spatial transcriptomics data. *Briefings in Bioinformatics*, *25*(3), 198. https://doi.org/10.1093/BIB/BBAE198

Zhang, Y., Song, K., Sun, Y., Tan, S., & Udell, M. (2019). "Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations. *Arxiv-Machine-Learning*.

Zhao, N., Li, Q., Sui, H., & Zhang, H. (2020). Role of Oxidation-Dependent CaMKII Activation in the Genesis of Abnormal Action Potentials in Atrial Cardiomyocytes: A Simulation Study. *BioMed Research International*, *2020*. https://doi.org/10.1155/2020/1597012

Zhu, W., Ma, Y., Guo, W., Lu, J., Li, X., Wu, J., Qin, P., Zhu, C., & Zhang, Q. (2022). Serum Level of Lactate Dehydrogenase is Associated with Cardiovascular Disease Risk as Determined by the Framingham Risk Score and Arterial Stiffness in a Health-Examined Population in China. *International Journal of General Medicine*, *15*, 11. https://doi.org/10.2147/IJGM.S337517