From Data to Diagnosis in Forensic Pathology

Alexander Svensson

Master's thesis
2026:E65

Lund University

Faculty of Science
Centre for Mathematical Sciences
Computational Science

Centrum Scientiarum Mathematicarum

Abstract

Forensic pathology is a branch of medicine that aims to determine the manner and cause of unnatural, suspicious or unexpected deaths that require a legal investigation. An important part of determining the cause of death is the examination of the body through an autopsy. The purpose of this project is to evaluate whether machine learning algorithms can predict cause of death based on autopsy data. The dataset used contains autopsy variables and cause of death extracted from 23 000 autopsy reports. In order to see if there exists groups within the dataset that correspond with the causes of death, dimension reduction and clustering is performed with UMAP and hierarchical clustering. The classification is performed with four different models tasked with classifying heart failure. The four models are logistic regression, random forest, neural network and TabPFN. The classifiers performed quite similar to each other and all performed AUC scores between 0.75-0.79. Meaning all models performed better than randomly guessing. In conclusion, the project works as a proof of concept for classifying cause of death based on autopsy data. In order to increase the performance of the classifiers, other autopsy variables better capturing differences between the causes of death need to be extracted from the autopsy reports.

Populärvetenskaplig Sammanfattning

Rättsmedicin är en medicinsk gren som utreder icke-naturliga eller misstänkta dödsfall för att fastställa hur och varför en person har dött. Det är dödsfall som kräver en polisutredning och inkluderar mord, självmord, olyckor och obevittnade dödsfall. Kroppen undersöks genom en obduktion och fynden antecknas i en obduktionsrapport. Resultatet från 23 000 obduktionsrapporter har sammanställts till ett stort dataset med variabler som ålder, kön, längd, vikt, organvikter, dimensionsmått på organen, grad av förruttnelse och dödsorsak. Syftet med projektet är att avgöra om det är möjligt att träna en maskininlärningsmodell på obduktionsdata och förutsäga hur en person har dött. Fyra olika klassificeringsmodeller har använts för att klassificera hjärtsvikt som dödsorsak. De fyra modellerna är logistisk regression, random forest, neuralt nätverk och TabPFN. UMAP och hierarkisk klustring används för att undersöka om det existerar grupper i dataset som motsvarar de olika dödsorsakerna. Resultatet från UMAP och hierarkisk klustring visade inga tydliga grupper som motsvarade de olika dödsorsakerna. Klassificeringsmodellerna presterade relativt likt varandra och fick alla ett AUC-värde mellan 0.75-0.79. Det betyder att samtliga modeller är bättre än att slumpmässigt gissa om den avlidne dött av hjärtsvikt eller ej. Modellerna testades på 4400 avlidna, varav 500 avlidna från hjärtsvikt. Modellerna klassificerade korrekt 80-90% av de 3900 som inte avlidit av hjärtsvikt och 40-50% korrekt av de 500 som avlidit av hjärtsvikt. Projektet fungerar som ett proof of concept för att klassificera dödsorsak från obduktionsdata. För att förbättra modellerna hade fler variabler behövt extraheras från obduktionsrapporterna, exempelvis fynd från rättskemiska analyser och resultat av mikroskopisk undersökningar, som i större utsträckning bidrar till differentiering mellan olika dödsorsaker.

Acknowledgements

I would like to express my sincere gratitude to my supervisor and co-supervisors for their invaluable guidance, support, and constructive feedback throughout this thesis.

Supervisor:
Mattias Ohlsson, Department of Earth and Environmental Sciences, Lund University

Co-Supervisors:
Carl Johan Wingren, Department of Forensic Medicine, University of Copenhagen,
Johannes Rødbro Busch, Department of Forensic Medicine, University of Copenhagen,
Ardavan Khoshnood, Department of Clinical Sciences Malmö, Lund University

Contents

1 Introduction **1**
1.1 Purpose and Goal . . . 1

2 Theoretical Background **2**
2.1 k-NN Imputation . . . 2
2.2 Logistic Regression . . . 2
2.3 Random Forest . . . 2
2.4 Neural Network . . . 4
2.5 Hyperparameter Optimization . . . 4
2.6 k-Fold Cross-Validation . . . 5
2.7 TabPFN . . . 5
2.8 UMAP . . . 6
2.9 Hierarchical Clustering and Dendrogram . . . 6

3 Method **8**
3.1 Data Preprocessing . . . 8
3.2 Heart Failure Statistics . . . 10
3.3 Dimension Reduction and Clustering . . . 10
3.3.1 kNN-imputation . . . 10
3.3.2 UMAP . . . 11
3.3.3 Hierarchical clusters . . . 11
3.4 Classification . . . 11
3.4.1 Logistic Regression . . . 12
3.4.2 Random Forest . . . 12
3.4.3 Neural Network . . . 12
3.4.4 TabPFN . . . 13
3.5 Use of Generative AI Tools . . . 13

4 Results **14**
4.1 Heart Failure Statistics . . . 14
4.2 Dimension Reduction and Clustering . . . 15
4.2.1 UMAP . . . 16
4.2.2 Hierarchical Clustering . . . 18
4.3 Classification . . . 21

5 Discussion **26**
5.1 Heart Failure Statistics . . . 26
5.2 Dimension Reduction and Clustering . . . 26
5.3 Classification . . . 27
5.4 Conclusion . . . 27
5.5 Future Work . . . 28

A Appendix **29**

1 Introduction

Forensic pathology is a branch of medicine that aims to determine the manner and cause of unnatural, suspicious or unexpected deaths [1]. These are deaths that require a legal investigation and includes homicides, suicides, accidents and unwitnessed deaths. In order to assess the cause of death, an external and internal examination of the decedent is performed through an autopsy. The autopsy can be supplemented with an ancillary analysis such as toxicology [2] and with the case context. The findings are documented in an autopsy report. Analysing the documents by hand is a complex and time-consuming task, with the risk of variability in interpretations. At the same time, classification is a common task in machine learning and there are many different methods available. The motivation of this project lies in the potential benefits a classification tool could bring to the forensic practice. A tool for classifying cause of death based on the autopsy data would work as a complementary decision making tool for the pathologist. If the classifier and the pathologist agree on the cause of death, the tool would strengthen the case made by the pathologist. If they predict differently, it would be a sign to go back and review the case. This could decrease the variability between forensic pathologists in assessments of cause and manners of death. The tool could also help to find patterns not easily detected through classical analysis methods. In order to create a tool for classifying cause of death, there needs to be a dataset available to train the model on. Historically, the autopsy reports are stored as individual text documents, and not as a large tabular dataset with autopsy variables. This project fills that gap with a tabulated autopsy dataset constructed with autopsy variables and cause of death from 23 000 autopsy reports [3]. The performance of a classification model depends on the quality and relevance of the data on which it is trained. To accurately predict the underlying cause of death, the available autopsy variables must capture characteristics that distinguish between the different outcome categories. If the extracted variables do not contain sufficient information to differentiate between these categories, the classifier will be unable to make reliable predictions. The autopsy dataset used for this project has only extracted a fraction of the available data in the autopsy reports. Some examples include age, sex, height, weight, organ weights, organ dimensions and level of putrefaction. This project will hence work as a proof of concept for classifying cause of death based on the autopsy data, where the dataset could be extended with more autopsy variables in the future.

1.1 Purpose and Goal

The purpose of this project is to evaluate whether machine learning algorithms can classify manner and cause of death and predict cause of death from forensic autopsy data. The goal is to produce a machine learning model that classifies the cause of death. The classification will focus on classifying heart failure as cause of death, based on the autopsy data.

2 Theoretical Background

The following section describes important methods and concepts used for this project. This includes four different classification methods, one dimension reduction method and one clustering method. The classification methods are logistic regression, random forest, neural network and TabPFN. The dimension reduction method is UMAP and the clustering method is hierarchical clustering. Other important concepts explained include kNN-imputation for missing data, k-fold cross validation for validating the models and randomized grid search for parameter optimization.

2.1 k-NN Imputation

Datasets sometimes have missing data. For some machine learning algorithms it is necessary to have a complete dataset for them to function. One solution is to drop all observations with data missing in any of the variables. If there are a lot of observations with missing values, this can lead to a large loss of data. This can particularly be a problem if a large proportion of the missing data belongs to a minority class. Dropping those observations would lead to losing a lot of information for that class and change the class distribution, introducing a bias to the dataset. A different approach is to impute the missing values in order to create a complete dataset. One such method is k Nearest Neighbor imputation [4]. The way it works is that it uses the k number of most similar observations in the dataset to estimate the missing value. The similarity between datapoints is based on the euclidean distance. If a variable is missing in an observation, the imputation is calculated as the weighted average of the k nearest neighbors.

2.2 Logistic Regression

Logistic regression is a method used for classification problems that models the probability that the response variable Y belongs to a specific category [5]. For example, for a binary problem with class 0 and 1, the probability for class 1 given input data X can be written as $Pr(Y = 1|X) = p(X)$. By using the logistic function,

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}, \tag{1}$$

one gets a probability between 0 and 1 for the class belonging. One can then choose a threshold, for example $p(X) > 0.5$, that predicts class 1 and everything less as class 0. By some manipulation of the logistic function one gets

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}. \tag{2}$$

The left hand side of Equation 2 is called the odds. By taking the logarithm of the odds, one gets the log odds,

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p, \tag{3}$$

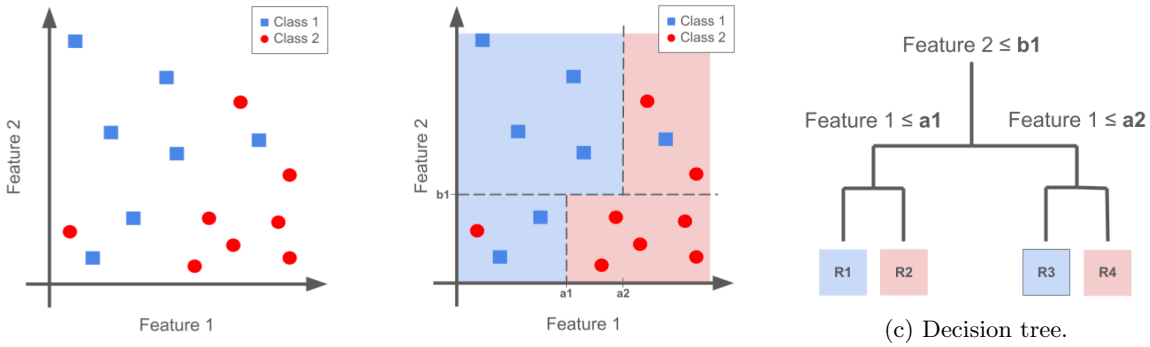
also called logit. The right hand side shows that the logistic regression model has a log odds that is linear in X. Maximum likelihood is then used to estimate the coefficients $\beta_0, \beta_1, ..., \beta_p$, where the estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are chosen to maximize the likelihood function,

$$\ell(\beta_0, \beta_1) = \prod_{i:\, y_i=1} p(X_i) \prod_{i':\, y_{i'}=0} (1 - p(X_{i'}))\,. \tag{4}$$

2.3 Random Forest

Decision trees are so called rule-based models and can be organised in a graph structure of binary trees [6]. It divides the input space into multiple disjoint regions used for the prediction, in this case classification. An example is illustrated in Figure 1 for a decision tree with two numeric input variables and a categorical binary output. Figure 1a displays the feature space with the data collared

with blue squares for class 1 and red circles for class 2. In Figure 1b the feature spaces have been divided into disjoint regions, where a class majority vote decides the class prediction within the region. Figure 1c displays the decision tree, where each split corresponds to the borders of the regions in the feature space. In order to make a prediction for new data, one follows the decision tree from the top and follows the rules at each split until one reaches an end node (leaf), which decides the prediction.



(a) Feature space. (b) Disjoint regions. (c) Decision tree.

Figure 1: Figure (a) displays the featurespace of a dataset with two numeric input variables and a categorical binary output. Figure (b) displays the disjoint regions of the feature space where a majority vote decided the predictive class belonging. Figure (c) displays the decision tree, where each split corresponds to the borders of the feature space in Figure (b). A prediction is made by following the rules in the decision tree from top to bottom.

When building a decision tree, one needs to decide how deep to make the tree. A deeper tree can capture more complex structures, while shallow trees focus on overall structures. One also needs to decide where to make the splits. The main idea is to make the splits such that the regions in the feature space captures the different classes as well as possible. The splits are decided by optimizing the splitting criteria

$$\arg\min_{j,\,s} \left(n_1 Q_1 + n_2 Q_2\right), \tag{5}$$

where n_1 and n_2 are the number of training data points falling to the left respectively right of the splitting node. The cost associated with the split is represented by Q_1 and Q_2. The index of the splitting variable is given by j and the value of the splitting point is given by s. The cost Q_ℓ can be calculated in many different ways. Here the Gini index,

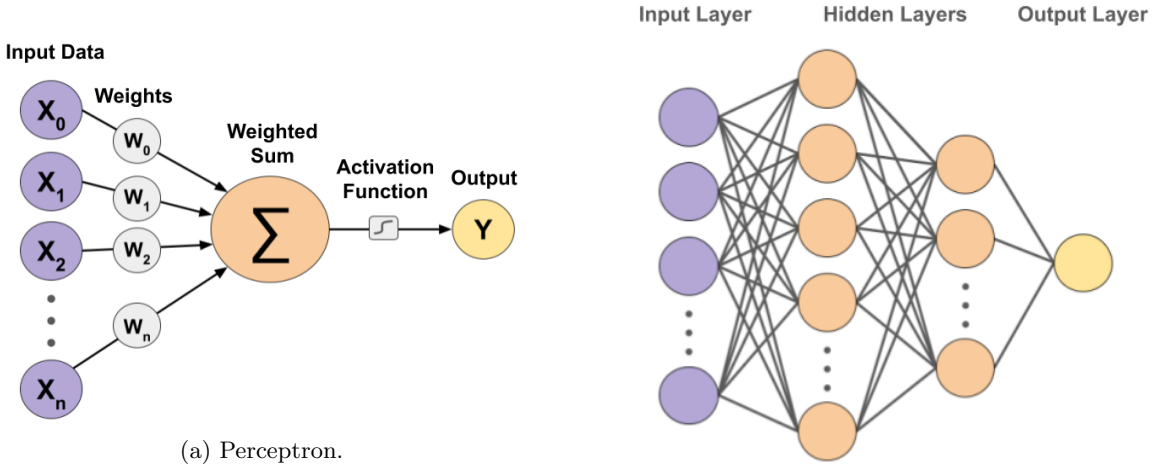
$$Q_\ell = \sum_{m=1}^{M} \hat{\pi}_{\ell m} \left(1 - \hat{\pi}_{\ell m}\right), \tag{6}$$

is used. Where M is the different classes, ℓ is the node and $\hat{\pi}_{\ell m}$ is the proportion of training observations in the ℓth region that belong to the mth class.

In order to reduce the variance in the model, one can use an ensemble method [6]. The main idea is to create several slightly different versions of the model and use the average of the model predictions. This has shown to reduce the variance and without significantly increasing the bias. One such method is bootstrap aggregating, or bagging for short. In bagging, several versions of the training data set are created by drawing data with replacement. Replacement means that after a data point is drawn, it is put back into the pile, such that it can be drawn again. This results in a version of the dataset of the same size as the original, but where some data points appear several times, while others are left out completely. In order to introduce even more randomness to each model one can use the random forest method. It follows the idea of bagging, but with one additional step. For each node split for each tree, it randomly selects a subset of the variables to make the split from. In this way, it perhaps misses the most optimal split at that moment, but it might lead to an even more optimal split further down the tree, that wouldn't have been discovered otherwise. In this way more different trees are explored and the variance is reduced even further.

2.4 Neural Network

The human brain is built up of a network of neurons where electrochemical processes send signals between the neurons [7]. An artificial neural network mimics this principle, with perceptrons representing the neurons. Figure 2a displays an illustration of a perceptron. On the left side is the input layer. Each input is weighted and the weighted sum is calculated. The weighted sum is then plugged into an activation function that decides what signal to send out as the output. By stacking several perceptrons in parallel, a single-layer perceptron is constructed. Combining several such layers in a series results in a multilayered perceptron (MLP), illustrated in Figure 2b. The layers are fully connected, meaning each node is connected to all nodes in the previous and next layer.



(a) Perceptron.

(b) Multi layered perceptron.

Figure 2: Figure (a) illustrates the perceptron. A weighted sum of the input layer is calculated. The weighted sum is then plugged into an activation function, which decides what signal to send out as output. Figure (b) illustrates several perceptrons stacked in parallel and in several layers where each node is connected with all nodes in the previous and next layer. This is called a multi layered perceptron.

A neural network is a parametric model and the parameters are found by minimizing the cost function [6]. In this case the parameters are the weights and biases. The cost function is the average loss over the training data. The loss is a measure of how close the prediction is to the observed data. The first step is to initialize the weights and let the input data propagate forward through the network. The cost and the gradient of the cost function is calculated. Using gradient descent, the parameters are updated by taking a small step in the direction of the negative gradient, which reduces the value of the cost function. During the training one can use batch learning, where the training data is divided up into batches. The network is the trained one batch at a time, where the weights are updated before taking the next batch. The process of training one batch is called one iteration, and when all batches have been used once, an epoch has been completed. Usually one trains for several epochs. By increasing the number of hidden layers and nodes, the flexibility of the network is increased, allowing form capturing more complex relations. However, it also opens up the possibility of overfitting, where it captures the training data, but performs much worse on new data. One way to counter this is through L1 and L2 regularization by adding a penalty term to the cost function to restrict the complexity of the network.

2.5 Hyperparameter Optimization

Some parameters are not learned by the model during training, instead the user needs to make a choice of which parameter values to use [6]. These parameters are called hyperparameters. What is a good choice for a specific hyperparameter depends on the problem and it can be hard to know beforehand what value to choose. A simple approach is to try different values and see which performed the best. The process of testing different parameter values is called grid search. In

case there are several hyperparameters, the number of combinations of hyperparameters increases exponentially [8]. Testing all possible combinations can be computationally inefficient. Instead one can use random search, where a number of random combinations are tested in order to save computation time and still cover many different combinations.

2.6 k-Fold Cross-Validation

In order to estimate the performance of a model one needs to test it. If the model is tested on the same data it was trained on, there is a risk that it overestimates the performance on new unseen data [6]. A solution is to split the data into training and test data. The split is done randomly, in case of any sorting within the dataset. The test score can then be used to report expected performance and to choose between models. If one is not happy with the model's performance on the test data, one would want to change the model. By doing so, one is indirectly optimizing on the test data and loses the ability to estimate the model's performance on unseen data. The solution is to split the training data once again into training and validation data. The validation score can then be used to select hyperparameters and only when it is complete, the test data set is used to estimate the model's performance. During training and evaluation, one would want as much training data and as much validation data as possible. Since the dataset is limited, one can use k-fold cross-validation to use all data during training and to reduce the variance in the validation score. This is done by splitting the training data into k subsets and then use one set at a time as the validation data and the rest as the training data. At the end, the model is trained on all the training data and the estimate of the validation score is the average of the validation scores of the subsets. See Figure 3 for an illustration of the data split and k-fold cross-validation.

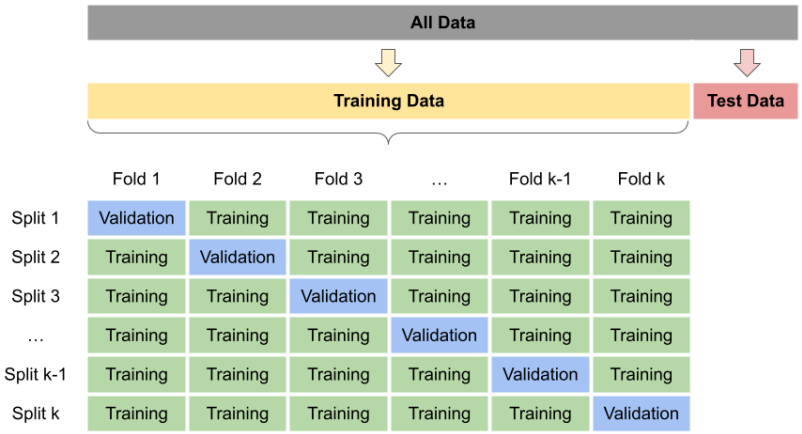


Figure 3: Illustration of the data split and k-fold cross validation process. The full dataset is split into training data (yellow) and test data (red). The training data is then split into k-folds. For each split, one fold is the validation data (blue) and the rest training data (green). Which part is the validation data changes for each split until all folds have been the validation data ones. The model is then trained on all the training data and the estimate of the validation performance is the average of all the validation splits.

2.7 TabPFN

TabPFN stands for Tabular Prior-data Fitted Network and is a generative transformer-based foundation model [9]. As the name suggests, it is a neural network that has already been fitted on prior tabular data. The model has learned across millions of synthetic datasets and has shown to outperform all previous methods on small to medium tabular datasets of up to 10 000 samples and 500 features. TabPFN can handle categorical data, missing values and is robust against outliers and unimportant features. Instead of training on a dataset, it trains across millions of datasets of various prediction tasks. Instead of testing it on individual samples, it is tested on a whole new dataset in a single forward pass.

2.8 UMAP

UMAP stands for Uniform Manifold Approximation and Projection and is a dimension reduction technique that takes high dimensional data and gives a low dimensional representation [10]. This is a manifold learning technique, meaning geometric structures and relationships are preserved. In other words, data that are close (similar) in the high dimensional representation, should be close in the low dimensional representation. UMAP can hence be used to visualize high dimensional data. The construction of the UMAP at a high level, starts with uniformly initializing the data onto a low dimensional space. The goal is then to iteratively reorder the data in the low dimensional space such that it minimizes the cross-entropy between the two topological representations. When implementing the UMAP, there are two important hyperparameters chosen by the user [11]. The first one is $n\_neighbours$, which can be seen as the resolution of the algorithm, where large values shift the focus towards large scale structures and smaller values towards more detailed structures. The second, min_dist, decides how close together datapoints can appear to each other, which has a strong effect on the low dimensional visualization of the data. For a more detailed explanation of the inner workings of UMAP, the reader is referred to [10].

2.9 Hierarchical Clustering and Dendrogram

Clustering methods are used to find subgroups within a dataset [5]. The main idea is to partition the observations into subgroups where observation in the same group is similar to each other and each group is different from the other groups. One such clustering method is hierarchical clustering and is illustrated in Figure 4. The method starts with computing the dissimilarity between all datapoints where each datapoint is viewed as its own cluster. The most similar clusters (or data points) are grouped together into a larger cluster. This is marked with a red circle in the top row of the figure. Correspondingly, the grouping of data is illustrated with a dendrogram in the lower row of the figure. The dendrogram has all the data points lined up on the x-axis. Then they are connected in a tree-like structure until all data points are connected. The vertical height of the line in the dendrogram displays the dissimilarity between the clusters (or data points) being merged. When all data points are grouped together, one decides where to cut the dendrogram. A cut higher up in the tree results in fewer and larger groups, while a cut further down in the tree results in smaller subgroups. The cut in the bottom right corner of the figure results in the three clusters displayed in the top right corner of the figure.

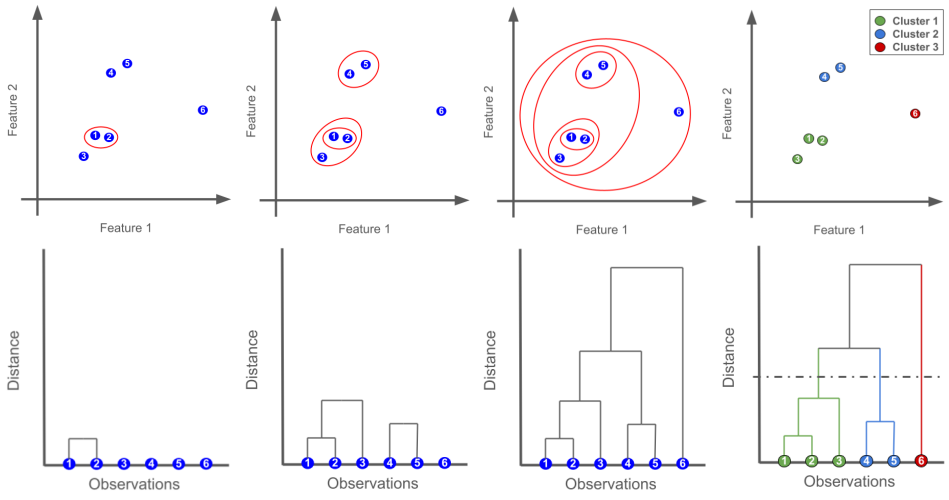


Figure 4: The upper row of figures displays the feature space of the dataset. Going from left to right, the most similar data points or clusters are grouped together (marked with the red circles). The grouping of data is illustrated with dendrograms, for each corresponding step, in the lower row of figures. The vertical height of the lines in the dendrogram represents the dissimilarity between the clusters and datapoints for each merger. The final dendrogram to the right displays a cut with a dashed line, resulting in the three clusters colored in the figure above.

When performing the hierarchical clustering and creating the dendrograms, one needs to decide how to measure the similarities between data points and clusters [5]. For data points the euclidean distance is used. When computing the similarity between a data point with a cluster or a cluster with another cluster one uses the so-called linkage. There are several different linkages to choose between, here $Ward$ linkage is used which minimizes the sum of squared differences within all clusters.

3 Method

The method section consists of four main parts. The first part, Section 3.1, is the data preprocessing to make the dataset compatible with the machine learning methods. The second part, Section 3.2, consists of a short statistical comparison of the decedents labeled with heart failure as cause of death with the rest of the dataset. The third part, Section 3.3, consists of dimension reduction and clustering methods, with UMAP and hierarchical clustering. The fourth part, Section 3.4, consists of the classification of heart failure as cause of death. The classification methods used are logistic regression, random forest, neural network and TabPFN.

3.1 Data Preprocessing

The data originates from historical records of autopsy reports from the Section of Forensic Pathology, Department of Forensic Medicine in Copenhagen. In total, there are about 23 000 individual cases stored as electronic text documents. Through a custom python script, a combined dataset has been created, and stored as comma separated values (CSV-file) [3]. The data includes variables such as sex, age, body height, body weight, organ weight, organ dimensions, degree of putrefaction, listed cause of death, medical history and scene description. The way the script works is that it searches for specific words or phrases in the autopsy reports and then it stores the data under different columns. For example, a search for the word "Age" followed by a number, the numeric age is stored under a column named "Age". Sometimes the extracted data is a descriptive text as a string or a boolean statement stored as true or false. In total, the construction of the original dataset resulted in about 80 variables. Even if 80 variables sounds like a lot, only a fraction of the available data in the autopsy reports has been extracted into the dataset.

Before the dataset could be used by the machine learning methods, some data preprocessing was needed to make the dataset and machine learning methods compatible. The machine learning methods used in the project need numeric representations of the variables. Numeric, boolean, ordinal and categorical variables were kept, and variables containing long descriptive texts were dropped. Ordinal variables such as degree of putrefaction were given numeric representations as 1, 2, 3, etc. Binary variables such as sex were set to 0 and 1. Boolean variables such as cause of death were set to 0 for false and 1 for true. In order to make the autopsy date numerically comparable, it was changed from date given as (yyyy-mm-dd), to days since first of January 1970. Some variables had corresponding unit variables. For example the variable weight would store the numeric weight, while the weight unit variable would store the unit as g or kg. These variables had to be scaled such that they have a uniform unit in the numeric variable. It was also discovered that the organ measurements were given in different orders in different autopsy reports. For example, one observation was given as height, width, depth and the next as width, depth, height etc. Hence when the dataset was constructed, the organ dimensions of height, width and depth got mixed between the columns. A decision was made to sort the organ dimension such that height is the largest, width the second largest and depths the smallest measurement for each organ and decedent. The dataset was then filtered based on age to only keep the decedents with an age of 18 years or older. Last, some particularly large values were assumed to be errors from writing down the values in the autopsy reports. These values were set to "missing", including an age over 500 years, an autopsy date going back 1800 years, a height above 16 meters, two right kidney heights above two meters, two left kidney heights above a meter and six liver heights above 1.6 meters. Last, some causes of death were updated after secondary analyses.

The data preprocessing resulted in 32 autopsy variables and 22 causes of death variables. In Table 1 a full list of the autopsy variables is presented with variable descriptions. In Figure 2, a full list of the cause of death variables with variable descriptions is presented. The variable names were originally created with danish variable names. In the Appendix A the original danish variable names are listed together with the respective english translation used for the project. Table A1 shows the autopsy variable translations and Table A2 shows the cause of death variable translations.

Table 1: The left column lists the autopsy variables and the right column lists the corresponding variable descriptions.

Variable	**Description**
Autopsy Date	Date when the autopsy was performed given in days since 1970-01-01.
Age	The age of the decedent in years.
Sex	The sex of the decedent, 0 for female and 1 for male.
Height	The height of the decedent in cm.
Weight	The weight of the decedent in kg.
Putrefaction level	The degree of putrefaction as none, discreet, moderate or pronounced.
Brain weight	The weight of the brain in g.
Spleen weight	The weight of spleen in g.
Right lung weight	The weight of the right lung in g.
Left lung weight	The weight of the left lung in g.
Heart weight	The weight of the heart in g.
Heart height	The height of the heart in cm.
Heart width	The width of the heart in cm.
Heart depth	The depth of the heart in cm.
Right cardiac wall thickness	The thickness of the right cardiac wall in mm.
Left cardiac wall thickness	The thickness of the left cardiac wall in mm.
Cardiac septum width	The width of the cardiac septum in mm.
Right kidney weight	The weight of the right kidney in g.
Left kidney weight	The weight of the left kidney in g.
Right kidney height	The height of the right kidney in cm.
Right kidney width	The width of the right kidney in cm.
Right kidney depth	The depth of the right kidney in cm.
Left kidney height	The height of the left kidney in cm.
Left kidney width	The width of the left kidney in cm.
Left kidney depth	The depth of the left kidney in cm.
Liver weight	The weight of the liver in g.
Liver height	The height of the liver in cm.
Liver width	The width of the liver in cm.
Liver depth	The depth of the liver in cm.
Right pleural cavity fluid	The amount of fluid in the right pleural cavity in ml.
Left pleural cavity fluid	The amount of fluid in the left pleural cavity in ml.
Abdominal cavity fluid	The amount of fluid in the abdominal cavity in ml.

Table 2: The left column lists the cause of death variables and the right column lists the corresponding variable descriptions.

Variable	Description
Heart failure	Death caused by hearth failure.
Poisoning	Substance-related deaths. For example drugs, alcohol and toxins.
Coronary Artery	Death caused by a problem in the coronary arteries.
Inflammation	Death due to widespread inflammation in the body.
Disease	Death caused by disease.
Suffocation	Death cause by suffocation.
Pneumonia	Death caused by severe inflammation in the lungs.
Exsanguination	Death due to rapid and excessive loss of blood.
Cancer	Death resulting from cancer.
Drowning	Death caused by drowning.
Inner bleeding	Death due to internal blood loss within the body.
Pulmonary Embolism	Death caused by sudden blockage in the arteries of the lungs.
Burns	Death resulting from severe burns.
Cardiac Hypertrophy	The heart muscle has become abnormally enlarged.
Inner airway obstruction	Death caused by blockage of the internal airways preventing breathing.
Injuries	Death due to physical trauma or bodily harm.
Stabs and slashes	Death caused by sharp-force injuries such as stabbing or cutting.
Gunshot Wounds	Death resulting from gunshot wounds.
Hanging	Death caused by hanging.
Cold	Death due to prolonged exposure to extremely low temperatures.
Subdural hematoma	Death caused by blood between the brain and its outer covering.
Undisclosed	The cause of death is undisclosed.
Missing	Cause of death label missing.

3.2 Heart Failure Statistics

To get a first insight to the decedents labeled with heart failure, each autopsy variable is reviewed. The distributions of the autopsy variables of the decedents labeled with heart failure, are compared with the distributions of the autopsy variables of the rest of the dataset. The comparison is made through a Wilcoxon rank test [12]. Variables with distributions significantly different to each other are plotted together for a visual comparison.

3.3 Dimension Reduction and Clustering

The third step consists of dimension reduction and clustering methods, with the purpose of determining if there exists groups within the dataset. Clustering methods try to cluster decedents similar in the autopsy data, without considering their labeled cause of death. The potential clusters are then analysed in order to see how they correspond to the causes of deaths and how the autopsy variable distributions vary between the groups.

3.3.1 kNN-imputation

For the dimension reduction and clustering methods used in this project, the dataset needs to be complete. To fill in all the missing values, kNN-imputation is used with the package *KNNImputer* from scikit-learn [13] with the parameter values in Table 3. See Section 2.1 for explanation of how kNN-imputation works.

Table 3: Selected parameter values for kNN-imputation.

Parameter	Selected Value
missing_values	np.nan
n_neighbors	5
weights	distance
metric	nan_euclidean
copy	True
add_indicator	False
keep_empty_features	False

3.3.2 UMAP

The dimension reduction method used is UMAP, which makes a two-dimensional projection of the high-dimensional dataset, where geometric relations are preserved. Meaning data similar in the autopsy variables in the high-dimensional space, appears closer together in the two-dimensional projection. This makes it possible for a visualisation of the autopsy data and see if any clusters are formed. This is done with the package *umap* [10] with the parameters in Table 4. See Section 2.8 for explanation of how UMAP works.

Table 4: Selected parameter values for the UMAP.

Parameter	Selected Value
n_neighbors	25
min_dist	0.8
n_components	2

The data in the UMAP is then colored based on cause of death. This is done to see how any potential clusters correspond to the labeled causes of death and if potential subgroups exist within a particular cause of death.

3.3.3 Hierarchical clusters

The clustering method used is hierarchical clustering. The method is implemented with the package *cluster.hierarchy.linkage* from scikit-learn [13] with the parameter *method* selected as *ward*. See Section 2.9 for explanation of how hierarchical clustering works. Instead of a 2D projection of the data set as in UMAP, hierarchical clustering is visualized through a dendrogram. The dendrogram is created using the package called *dendrogram* from SciPy [14] with the parameter Z selected as the result from *cluster.hierarchy.linkage*. The dendrogram trees are then cut resulting in the clusters. See Section 2.9 for explanation of how a dendrogram works. The clusters found with hierarchical clustering are first used to color the data in the UMAP. This is done to compare if the clusters found from hierarchical clustering correspond with the clusters found through UMAP. Then the clusters found through hierarchical clustering are further analysed. For each cluster, the percentage of decedents labeled with each cause of death is calculated. Clusters with high concentration of heart failure are further analysed by comparing the autopsy variables in the cluster with the rest of the data with the Wilcoxon rank test.

3.4 Classification

The classification of cause of death is performed with four different methods, including logistic regression, random forest, neural network and TabPFN. The four models are tasked with classifying heart failure as cause of death based on the autopsy data. The first step is to split the dataset into training and test data. The dataset is split randomly into 80% training data and 20% test data. For all random processes, the seed 42 is used. Random forest and TabPFN can handle missing values, while logistic regression and the neural network uses the kNN-imputated version of the dataset, explained in Section 3.3.1. TabPFN is ready to be used off the shelf, while the hyperparameters need to be selected for logistic regression, random forest and the neural network. Randomised grid search

is used to select the hyperparameters, explained in Section 2.5. The best set of hyperparameters is decided by the highest ROC-AUC score. In order to evaluate the performance, k-fold cross-validation is used, explained in Section 2.6. The threshold for classification is selected based on the highest f1 score. The four different models are then compared based on several different metrics, including ROC-AUC, f1, accuracy, precision and recall. SHAP values are also calculated to see what variables were the most influential for the model's decision making and in which way the variables affected the predictions [15]. The following four subsections describe how the four classification models were implemented and which packages and variables were used.

3.4.1 Logistic Regression

The logistic regression model was implemented by the *LogisticRegression* package from scikit-learn [13] with model parameters and tested values used in grid search presented in Table 5.

Table 5: Model parameters and tested values used in grid search for logistic regression.

Parameter	**Value**
solver	lbfgs
max_iter	2000
C	[0.01, 0.1, 1, 10]

3.4.2 Random Forest

The random forest model was implemented by the $RandomForestClassifier$ package from scikit-learn [13] with model parameters and tested values used in randomized grid search presented in Table 6. The random grid search was performed for 300 iterations with 5-fold cross validation, choosing the best set of parameters based on highest ROC-AUC score.

Table 6: Model parameters and tested values used in randomized grid search for random forest.

Parameter	**Value**
n_estimators	500
criterion	gini
bootstrap	True
max_depth	[10, 20, 30, None]
min_samples_split	[2, 5, 10]
min_samples_leaf	[1, 2, 4, 8]
max_features	[sqrt, log2, 0.5]
class_weight	[balanced, balanced_subsample]

3.4.3 Neural Network

The neural network model was implemented by the $MLPClassifier$ package from scikit-learn [13] with model parameters and tested values used in randomized grid search presented in Table 7. The random grid search was performed for 300 iterations with 5-fold cross validation, choosing the best set of parameters based on highest ROC-AUC score.

Table 7: Model parameters and tested values used in randomized grid search for the neural network.

Parameter	**Value**
activation	relu
solver	adam
max_iter	2000
early_stopping	False
hidden_layer_sizes	[(2,), (4,), (8,), (16,), (32,)]
alpha	[1e-5, 1e-4, 1e-3, 1e-2, 1e-1]
learning_rate_init	[1e-5, 1e-4, 1e-3, 1e-2, 1e-1]
batch_size	[32, 64, 128, 256]

3.4.4 TabPFN

The TabPFN model was implemented by the $TabPFNClassifier$ package from TabPFN [9]. No model parameters need to be selected.

3.5 Use of Generative AI Tools

When formatting figures in python, it was often several subfigures plotted at once with labels, titles, text boxes and different colors to separate what was displayed. Generative AI tools such as ChatGPT were used as an assisting tool for the formatting, to present the result in a clear and structured way. Everything was reviewed and verified such that it showed the intended result.

4 Results

The result is divided up into three main parts. The first part, in Section 4.1, is a short statistical overview over how the autopsy data of the decedents labeled with heart failure compared to the rest of the dataset. The second part, in Section 4.2, consists of the dimension reduction and clustering methods of UMAP and hierarchical clustering. The third part, in Section 4.3, is the result from the four classification models used for classifying heart failure as cause of death. The four models are Logistic Regression, Random Forest, Neural Network and TabPFN.

4.1 Heart Failure Statistics

The data preprocessing explained in Section 3.1 resulted in 32 autopsy variables and 22 cause of death variables. Figure 5 displays an example of nine of the autopsy variables distributions and statistics. The distributions show that around 500 autopsies are performed each year. The age of the decedents is normally distributed with a median age of 51 years. Around two thirds of the decedents are male and one third is female and around 80% show none or discrete level of putrefaction. Several of the variables have distributions with both small and large outliers such as height and weight. For more autopsy variable statistics, see the Appendix Table A3. The table lists the number of observations, number of missing values, the median and the first and third quantile of all the continuous autopsy variables.

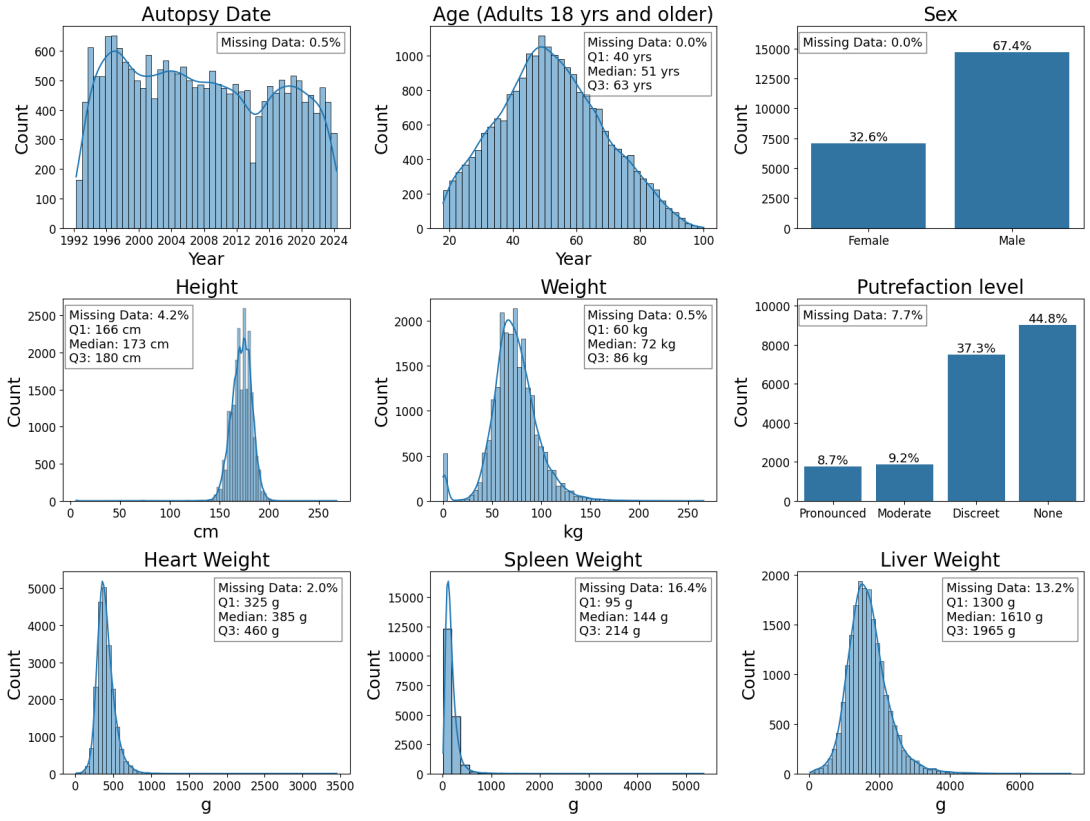


Figure 5: Each subfigure displays the distribution of one autopsy variable each, with count on the y-axis.

Figure 6, displays the ten most common causes of deaths labeled in the dataset. The most common labeled cause of death is poisoning, which makes up 26.5% of the causes of death. This includes all substance-related deaths, for example drugs, alcohol and toxins. Heart failure, that is the focus of this project, is the third most common labeled cause of death and makes up 10.5% of the causes of death. Most other causes of death make up single digit percents of the total dataset. See Table A4 in the Appendix for a list of the count and percentage of all labeled causes of deaths.

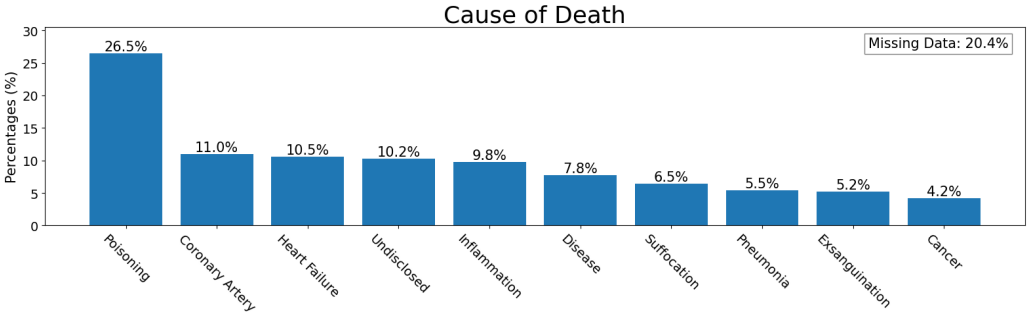


Figure 6: The ten most common causes of deaths in the dataset are ordered from most common to least common. The y-axis displays the percentage of decedents labeled with the causes of death.

Figure 7 displays a statistical comparison between the decedents labeled with heart failure as cause of death compared to the rest of the dataset. The dataset contains 2298 decedents labeled with heart failure, which constitutes 10.5% of the dataset. The autopsy variables presented in the figure all had significantly different distributions according to the Wilcoxon rank test. The figure shows a visual comparison of the autopsy variable densities for heart failure compared to the rest of the dataset.

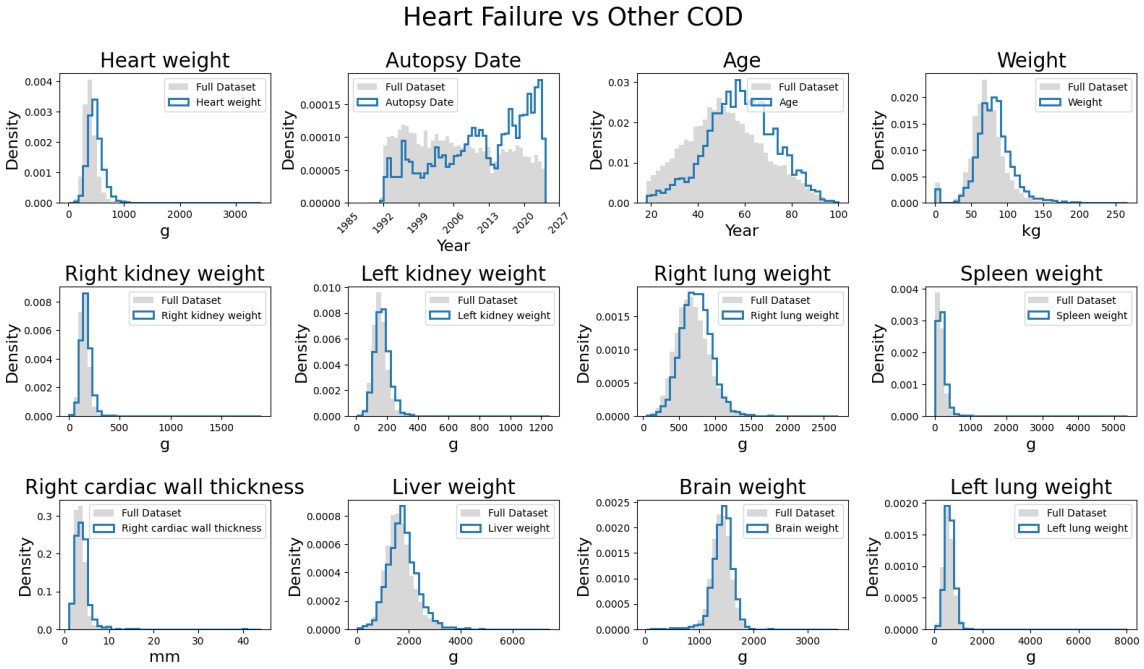


Figure 7: Each subfigure displays an autopsy variable whose density is significantly different for decedents labeled with heart failure compared to the rest of the dataset according to the Wilcoxon rank test. The blue line shows the density of the decedents labeled with heart failure as cause of death. The grey area shows the density of the rest of the dataset.

Figure 7 shows a steady increase in the amount of labeled heart failures in the past decades. The decedent are generally older and heavier and with heavier hearts. Other examples of organs that generally are heavier include the kidneys, the lungs, the spleen, the liver and the brain. Also the right cardiac wall thickness is thicker.

4.2 Dimension Reduction and Clustering

The following section presents the result from the dimension reduction and clustering methods. Section 4.2.1 displays the UMAP result and Section 4.2.2 displays the hierarchical clustering result.

4.2.1 UMAP

The result of UMAP on the whole autopsy dataset, excluding the variable sex, is displayed in Figure 8. The data is colored by sex, where blue represents males and orange represents females. The UMAP displays data of the same sex generally closer together, resulting in one orange and one blue area in the figure.

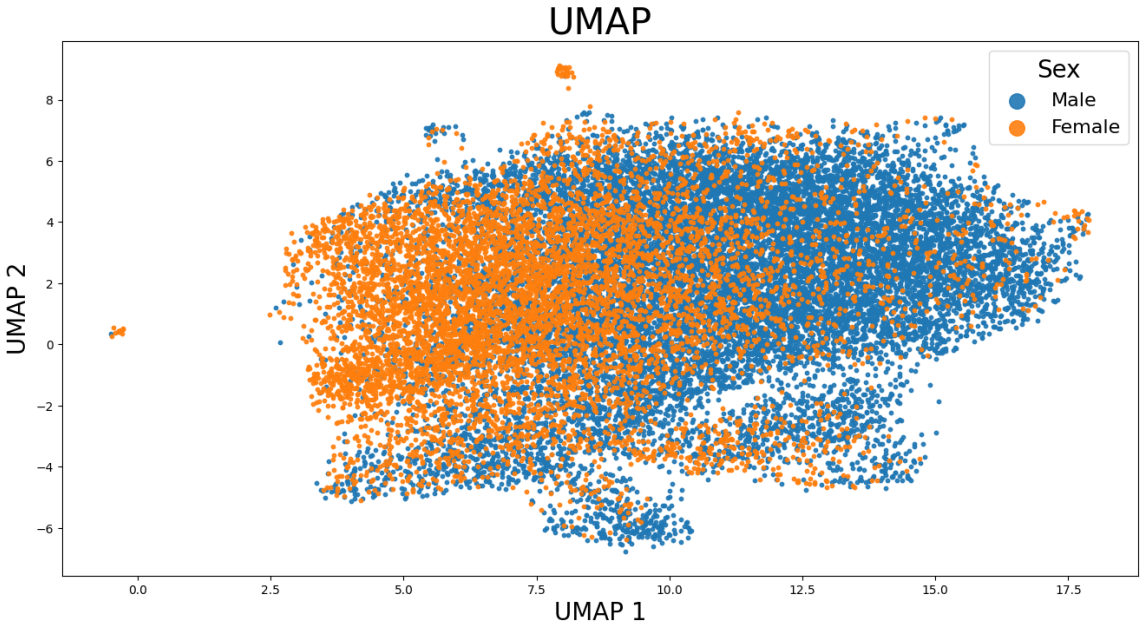


Figure 8: The UMAP displays the two dimensional projection of the autopsy dataset, excluding the variables sex. The data is colored by sex, where blue represents male decedents and orange female decedents.

The difference in males and females seen from the UMAP led to the decision to split the dataset into male and female and perform the clustering separately. Figure 9 displays the UMAP constructed on the male autopsy dataset with the data colored by labeled cause of death. The UMAP displays one large cluster in the middle, with one small cluster above and two small clusters underneath with almost no separation between the clusters. Most causes of deaths are uniformly distributed over the UMAP. One can see some higher concentrations in the top half of the UMAP for heart related causes of deaths, including heart failure, coronary artery and cardiac hypertrophy. There are also higher concentrations in the bottom half of the UMAP for trauma related causes of death, such as gunshot wounds, hangings, stabs and slashes.

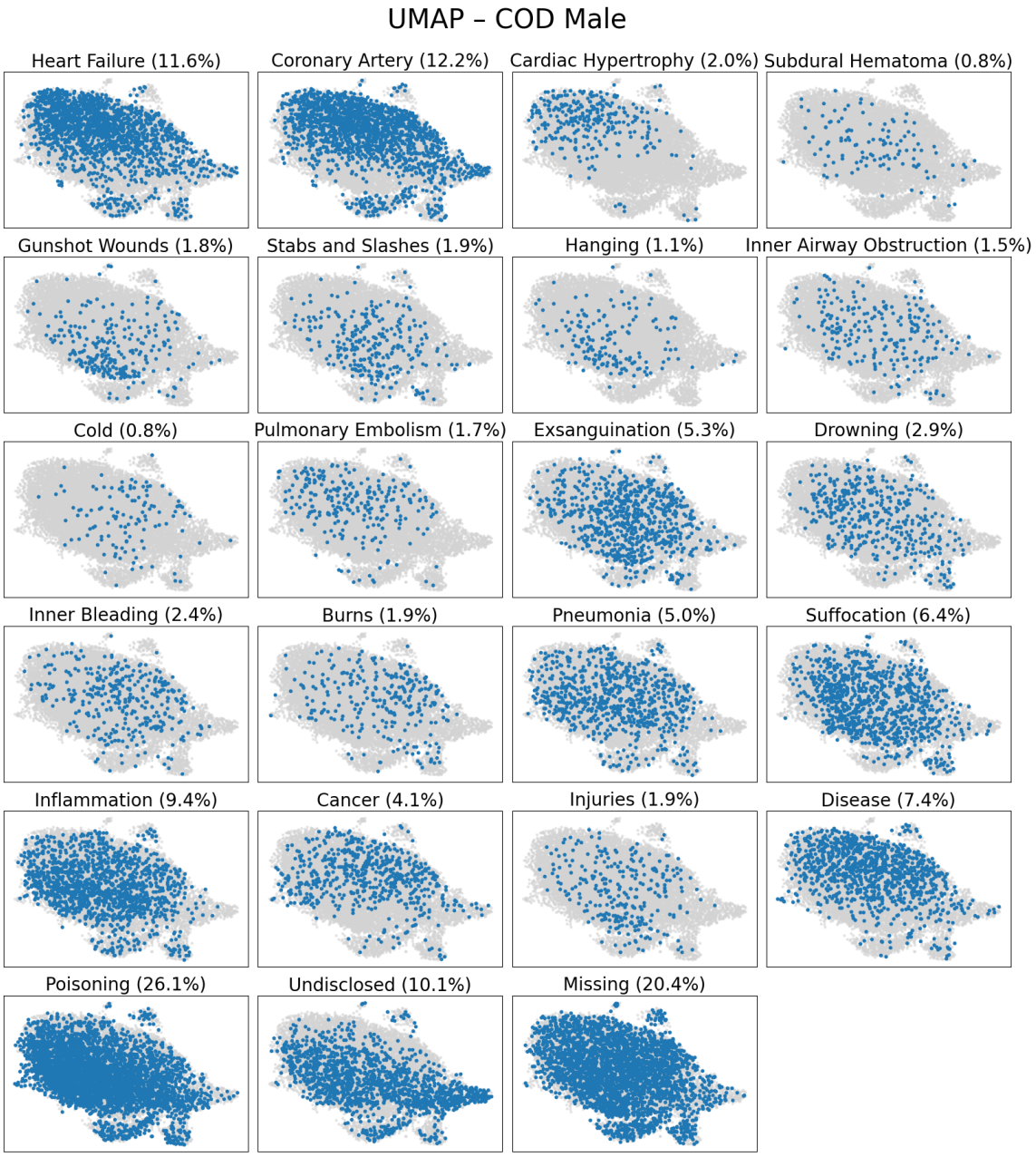


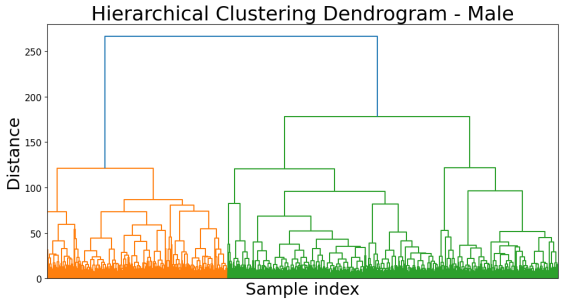
Figure 9: The UMAP is constructed with the male autopsy dataset. Each subfigure has the data colored by one cause of death each. The percentage of males labeled with the particular cause of death is written in the subfigures titles.

Figure 10 displays the UMAP constructed on the female autopsy dataset with the data colored by labeled cause of death. Similarly to the Male UMAP, the female UMAP displays one large cluster, with indications of smaller clusters formed at the edge of the main cluster. The left side of the UMAP has higher concentrations of heart related causes of deaths, including heart failure, coronary artery and cardiac hypertrophy. The right side of the UMAP has higher concentrations of trauma related causes of death, such as gunshot wounds, hangings, stabs and slashes.

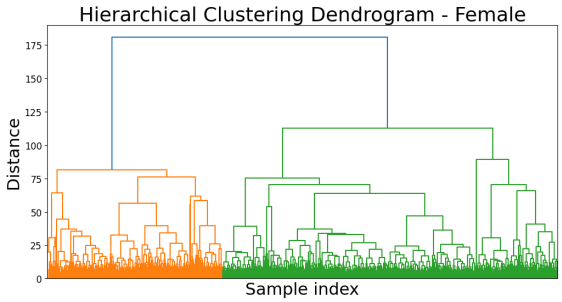
Figure 10: The UMAP is constructed with the female autopsy dataset. Each subfigure has the data colored by one cause of death each. The percentage of females labeled with the particular cause of death is written in the subfigures titles.

4.2.2 Hierarchical Clustering

The hierarchical clustering was performed on the male and female dataset separately. Figure 11a displays the resulting dendrogram from the hierarchical clustering on the male autopsy dataset. The male dendrogram shows relatively large distances between the clusters above the distance of 90, compared to the splits further down in the tree. The male dendrogram was cut at 90, resulting in eight clusters. The cluster sizes are tabularized in Table A5. In the same way, the hierarchical clustering of the female autopsy dataset is displayed with a dendrogram in Figure 11b. The female dendrogram shows relatively large distances between the clusters above the distance of 70, compared to the splits further down in the tree. The female dendrogram was cut at 70, resulting in 8 clusters. The cluster sizes are tabularized in Table A6.



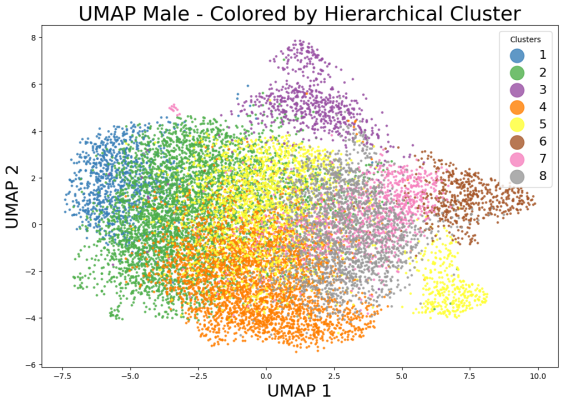
(a) The dendrogram displays the clustering process of the hierarchical clustering for the male decedents.



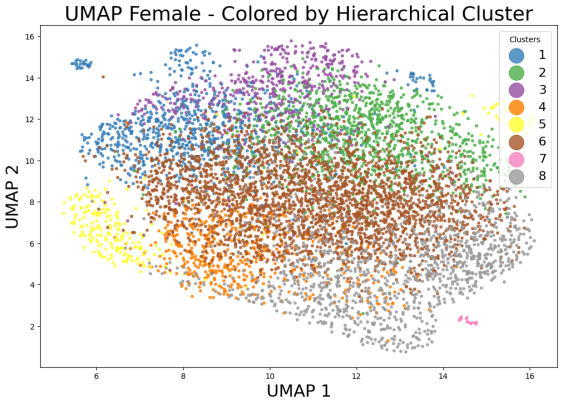
(b) The dendrogram displays the clustering process of the hierarchical clustering for the female decedents.

Figure 11: Figure (a) and Figure (b) displays the dendrograms of the hierarchical clustering of the male respective female decedents. The male dendrogram was cut at 90 and resulted in eight male clusters. The female dendrogram was cut at 70 and resulted in eight female clusters.

The clusters found through the hierarchical clustering process were used to color the data in the male and female UMAPs. Figure 12a displays the male UMAP collared by the eight male hierarchical clusters. Figure 12b displays the female UMAP collared by the eight female hierarchical clusters. The data belonging to the same hierarchy clusters also appears close together in the UMAP.



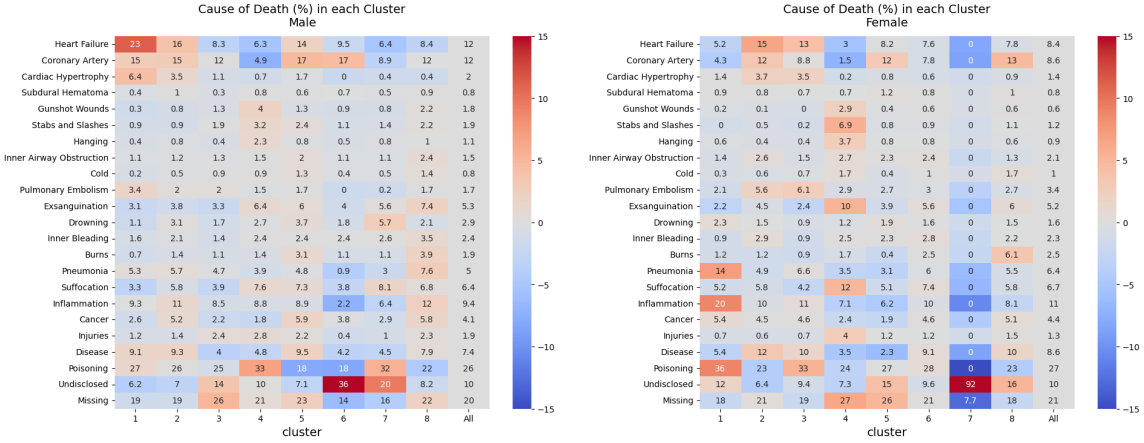
(a) The data in the male UMAP is colored by the eight clusters produced from the hierarchical clustering on the male dataset.



(b) The data in the female UMAP is colored by the eight clusters produced from the hierarchical clustering on the female dataset.

Figure 12: Figure (a) and Figure (b) displays the male respectively female UMAPs with the data collared by the clusters produced from hierarchical clustering.

For the clusters found through hierarchical clustering, Figure 13 displays the distributions of causes of deaths within each cluster. For the male hierarchical clusters, cluster 1 has the highest percentage of decedents labeled with heart failure. Male cluster 1 is located in the upper left part of the UMAP in Figure 12a, which is the same area where heart related causes of death were located in Figure 9. For the female hierarchical clusters, cluster 2 has the highest concentration of heart failures. This cluster is located in the top right part of the UMAP in Figure 12b. Which is not the same region with the highest concentrations of heart failures in Figure 10.



(a) Percentages of causes of deaths in the male hierarchical clusters.

(b) Percentages of causes of deaths in the female hierarchical clusters.

Figure 13: Figure (a) is based on the male dataset and Figure (b) is based on the female dataset. The columns represent the hierarchical clusters and the rows the causes of deaths. The columns show the percentage of each cause of death in each cluster. The column furthest to the right shows the percentage in the whole male respective female dataset. The colors are row normalized, indicating if the cluster has a higher or lower percentage of each cause of death, compared to the full male respective female dataset.

According to the Wilcoxon rank test, all autopsy variable distributions in cluster 1 are significantly different compared to the rest of the male dataset. Figure 14 displays 12 of the most different variable distributions. The same thing is displayed in Figure 15 for the female cluster 2. All the autopsy variable distributions for female cluster 2 were also significantly different compared to the rest of the female dataset. Male cluster 1 shows generally heavier organs such as the liver, the kidneys, the heart, the spleen, the lungs and the brain. The decedents are generally heavier, taller and has a wider cardiac septum width and left cardiac wall thickness. The female cluster 2 generally has heavier organs such as the heart, the kidneys, the liver, the spleen and the lungs. The decedents are generally heavier, taller and older. The cardiac septum width and the left and right cardiac wall thickness are thicker. The significantly different distributions in the clusters are generally the same variables that were the most different in the statistical comparison of heart failures in Figure 7. The distributions in the clusters are shifted even more to the right.

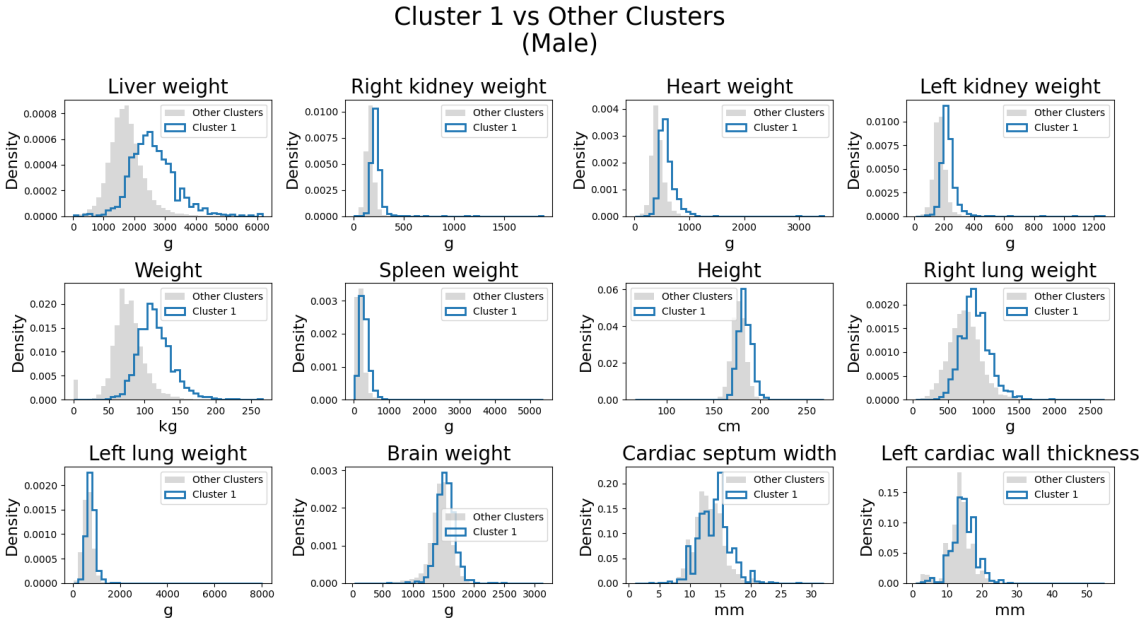


Figure 14: Each subfigure displays an autopsy variable whose density is significantly different in the male cluster 1, compared to the rest of the male dataset, according to the Wilcocon rank test. The blue line shows the density of the decedents in cluster 1. The grey area shows the density of the rest of the male dataset.

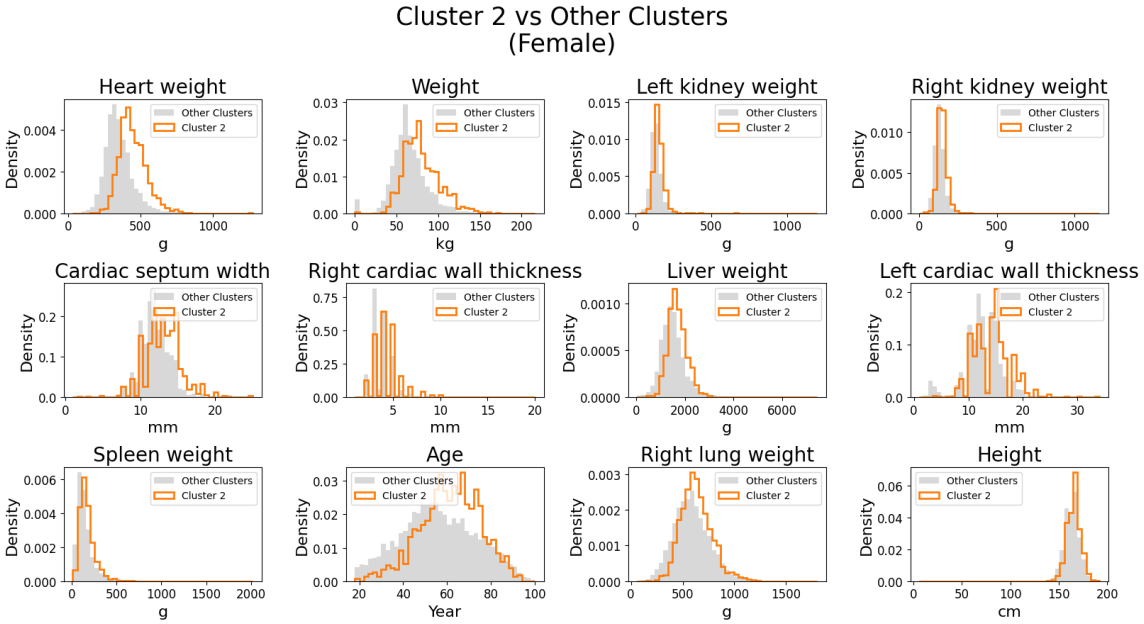


Figure 15: Each subfigure displays an autopsy variable whose density is significantly different in the female cluster 2, compared to the rest of the female dataset, according to the Wilcoxon rank test. The blue line shows the density of the decedents in cluster 2. The grey area shows the density of the rest of the female dataset.

4.3 Classification

The result from the classification models contains the selected hyperparameters, the ROC-AUC curves, confusion matrices from the test data, the performance scores on the validation and test

data and the SHAP values for the models. The hyperparameters selected for logistic regression are presented in Table 8. Table 9 lists the hyperparameters for random forest and Table 10 lists the hyperparameters for the neural network. TabPFN does not require selection of hyperparameters by the user, hence only the performance of the model is presented.

Table 8: The first column displays the hyperparameters for the logistic regression model. The second column states the selected parameter values.

Model Parameter	**Selected Value**
solver	lbfgs
max_iter	2000
C	1

Table 9: The first column displays the hyperparameters for the random forest model. The second column states the selected parameter values.

Model Parameter	**Selected Value**
n_estimators	100
criterion	gini
bootstrap	True
max_depth	20
min_samples_split	2
min_samples_leaf	8
max_features	log2
class_weight	balanced

Table 10: The first column displays the hyperparameters for the neural network model. The second column states the selected parameter values.

Parameter	**Value**
activation	relu
solver	adam
max_iter	2000
early_stopping	False
hidden_layer_sizes	(32,)
alpha	1e-1
learning_rate_init	1e-4
batch_size	64

The hyperparameters were selected based on the highest ROC-AUC score. The left column in Figure 16 presents the ROC curve and AUC score for the selected models on the test data. The middle column presents the confusion matrices on the test data and the right column presents the row-normalized confusion matrices on the test data for all models.

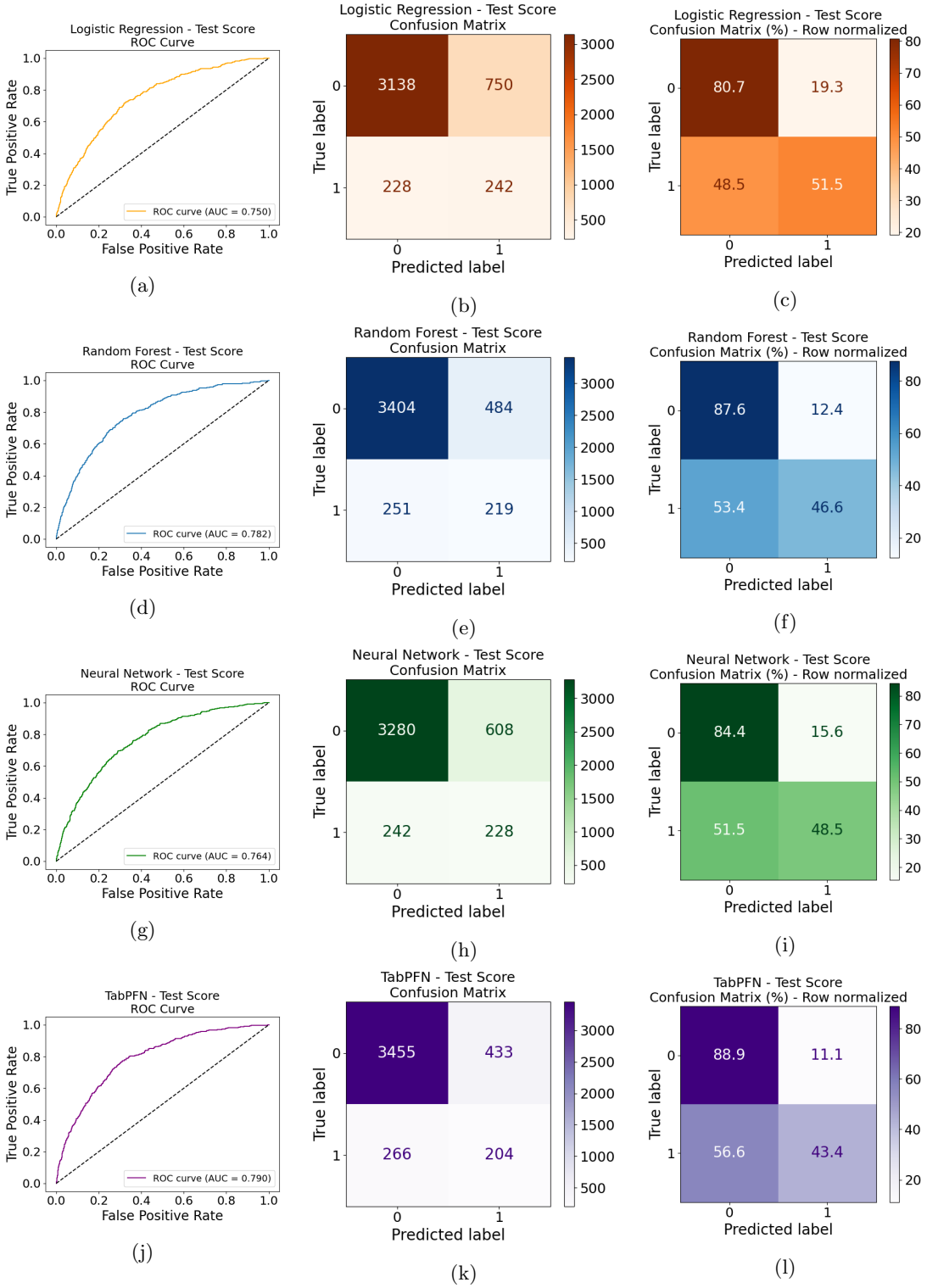


Figure 16: Each row is the result from one of the four classification models. The first row is logistic regression, the second row is random forest, the third row is the neural network model and the last row is the TabFPN model. The left column displays the ROC-AUC curves from the test data. The middle column displays confusion matrices on the test data and the right column displays a row-normalized version of the confusion matrices.

Figure 16 shown that all the four classifiers scored between 0.75-0.79 on the ROC-AUC score on the test data. Out of the four models TabPFN performed the highest ROC-AUC score, followed

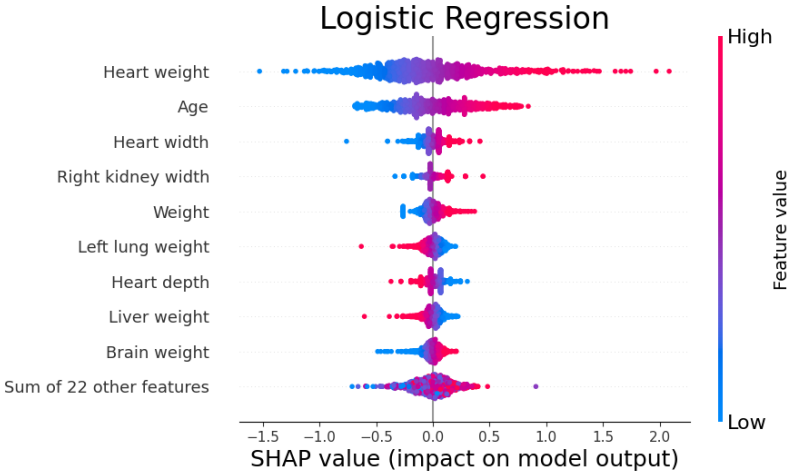
by random forest, then the neural network and last the logistic regression model. Out of the 4358 test data points there are 10.8% (470) decedents labeled with heart failure. Across the four models, between 204-242 of these positive cases were correctly classified, corresponding to 43-52% of all heart failures. The remaining 3888 samples consisted of the negative cases, where the models correctly classified between 3138-3455 samples, corresponding to 81-89% of the non heart failures. The logistic regression model classified the largest number of positive cases correctly, but it also produced the highest number of false positives. In contrast, TabPFN correctly classified the largest number of negative cases while identifying the fewest positive cases. The neural network achieved the second-highest number of correctly classified positive cases and the second-highest number of false positives. Random forest classified the second most positive cases incorrectly and the second most negative cases correctly.

Continuing to Table 11, it shows that all the test scores are slightly higher than the validation scores for all the four classifiers. On the test data, TabPFN performed the best ROC-AUC score, then random forest, followed by the neural network and last logistic regression. For f1 score, TabPFN and random forest both scored 0.37 on the test data, followed by the neural network with 0.35 and last logistic regression with 0.33. The overall accuracy of the models are between 78-84%, with highest score by TabPFN, then random forest, then the neural network and last logistic regression. The highest precision is scored by TabPFN, then random forest, then the neural network and last logistic regression. The highest recall is scored by the logistic regression model, then the neural network, followed by random forest and last TabPFN.

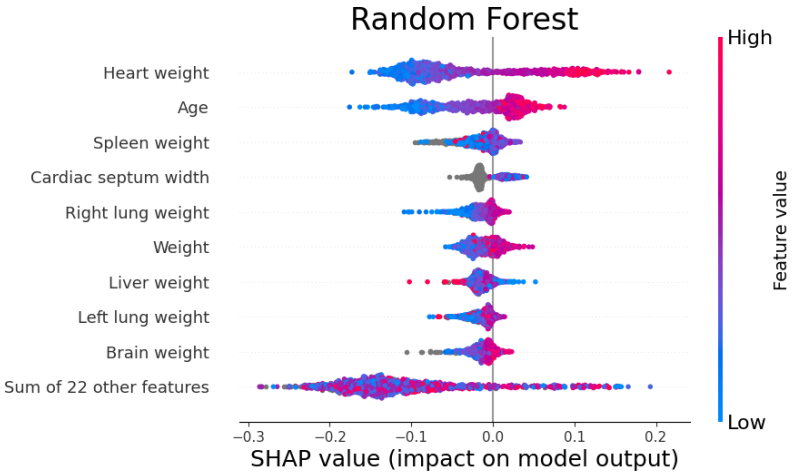
Table 11: Model performance comparison of the four classification models on validation and test data.

Dataset	Model	ROC-AUC	F1	Acc	Prec	Recall
Validation	Logistic Regression	0.72	0.31	0.77	0.23	0.49
	Random Forest	0.77	0.35	0.83	0.30	0.43
	Neural Network	0.74	0.33	0.80	0.26	0.48
	TabPFN	0.77	0.36	0.84	0.30	0.43
Test	Logistic Regression	0.75	0.33	0.78	0.24	0.51
	Random Forest	0.78	0.37	0.83	0.31	0.47
	Neural Network	0.76	0.35	0.80	0.27	0.49
	TabPFN	0.79	0.37	0.84	0.32	0.43

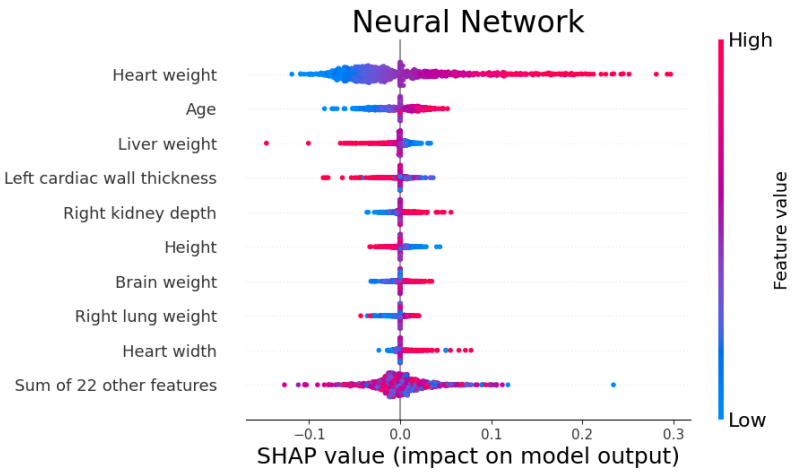
Figure 17 displays the SHAP values for the logistic regression model, the random forest model and the neural network model. SHAP value compatibility with the TabPFN model is under development and was hence not used in the project. The SHAP values indicate which variables have been the most influential for each model's predictions and in which way they have affected the predictions [15]. The figures display the nine most influential variables for each model. Heart weight and age are the most influential variables for predicting heart failure for all three models. A large heart weight and a high age pushes the prediction towards predicting heart failure, while a low age and low heart weight pushes the model towards predicting non heart failure. In Figure 7, heart weight and age are also shown to be the autopsy distributions that are the most significantly different with distributions pushed towards higher values. Liver weight is also shown to have distributions pushed towards higher values for decedents labeled with heart failure, but according to the SHAP values a high liver weight pushes the model away from predicting heart failure as cause of death. Spleen weight and kidney weight are also shown to have distributions pushed towards higher values for heart failures. It is however only random forest that has spleen weight in its top nine most inferential variables and it is only random forest that doesn't have kidney width in its top nine most influential variables. All the models also have brain weight in its top nine most influential variables. In total there are 14 different variables in the three models' top nine. Five of the variables are heart related, including heart weight, heart width, heart depth, cardiac septum width and left cardiac wall thickness.



(a) SHAP values - Logistic Regression



(b) SHAP values - Random Forest



(c) SHAP values - Neural Network

Figure 17: The three subfigures display the SHAP values for the logistic regression model, the random forest model and the neural network model. The variables listed on the left are the nine most influential variables for the model predictions. Red color indicates a high feature value and blue color a low feature value. A positive SHAP value pushes the model towards predicting heart failure, while a negative SHAP value pushes the model away from predicting heart failure.

5 Discussion

The discussion will cover the three main parts of this project, including the heart failure statistics, the clustering methods and the classification models. The discussion will relate this to the purpose and goal of the project, being to evaluate whether machine learning algorithms can classify manner and cause of death and predict cause of death from forensic autopsy data.

5.1 Heart Failure Statistics

Starting with the statistical comparison between decedents labeled with heart failure and the rest of the dataset in Figure 7, a steady increase in heart failures over the past decades is displayed. There could be several possible explanations for this trend, including a change in lifestyle leading to more deaths due to heart failure. It could also be due to a change in directives for when to include heart failure as the cause of death in the autopsy report. Or it could also be due to a change in policy for whom to send to an autopsy, which now includes more deaths due to heart failure. Some of the distributions showed outliers with particularly small values for weights and measurements in the dataset. These values are necessarily not errors. It could instead be explained by putrefaction, burns or trauma, only leaving fractions of the body or organ for the autopsy. The comparison between decedents labeled with heart failures with the rest of the dataset showed differences in the autopsy variable distributions. Even if the autopsy variable distributions showed that decedents labeled with heart failure are generally older, heavier and with heavier organs, there are still large overlaps with the distributions with other causes of deaths. Meaning that just because there is a difference seen in the distributions, it does not tell if it is possible to distinguish a heart failure from a non heart failure in specific cases. Since there could still be old and heavy decedents with heavy organs without heart failure and young and light decedent with lighter organs labeled with heart failure.

5.2 Dimension Reduction and Clustering

The dimension reduction with UMAP on the whole dataset in Figure 8 displayed two areas of higher concentrations of males and females respectively. This means that males appear more similar in the autopsy data to other males and females appear more similar to other females. By dividing the analysis by sex, it hopefully captures patterns of non sex related differences better. From the male and female UMAPs in Figure 9 and Figure 10, both UMAPs show one large main cluster each. This means that the dimension reduction of UMAP does not capture distinctly different groups in the autopsy data for either males or females. Data points appearing closer together in the UMAP are still more similar to each other than data points appearing further away. These differences do however not correspond well with the characteristics for the different causes of deaths, since most labeled causes of deaths are uniformly distributed in the UMAPs. Even if decedents labeled with heart failure do not appear in separate clusters, there are still areas with higher concentrations of heart failures in the UMAPs. These areas are not only occupied by heart failures, but mixed with other labeled causes of deaths.

Continuing to the dendrograms from the hierarchical clustering in Figure 11, the male and female dataset both cluster into eight clusters each with relatively large differences between the clusters. This means that the hierarchical clustering method does find different groups of decedents similar in the autopsy data. The hierarchical clustering and the UMAP seem to deem the same decedents as similar, since decedents within the same hierarchical cluster appear close together in the UMAPs in Figure 12. From the cause of death distributions within each cluster in Figure 13 it shows that most clusters have similar distributions of causes of deaths as the distributions in the full male and female datasets. Meaning the clusters do not correspond to specific causes of deaths. Even the hierarchical cluster with the highest percentage of heart failures still has three quarters made up of decedents labeled with other causes of deaths. The autopsy variable distributions of the clusters with the highest concentrations of heart failures in Figure 14 and Figure 15 differ in a similar way as the statistical comparison of heart failures in Figure 7. The clusters display distributions shifted even more to the right for weight and organ weights such as heart weight, liver weight and kidney weights. This means that the hierarchical clusters do find groups of decedents with generally much higher weights and organ weights. But these differences do not seem to only correspond to heart

failures as cause of death. There are several decedents similar in the autopsy data labeled with other causes of deaths.

5.3 Classification

For the classifiers, the SHAP values in Figure 17 displays the nine most influential variables for predicting heart failure for the logistic recession model, the random forest model and the neural network model. If the models display the same variables as the most influential variables, it means they agree on which variables are the most influential for predicting heart failure in the autopsy data. If they display different variables it means that the models are influenced more by different variables when predicting heart failure. The SHAP values in Figure 17 show that the three models agree that heart weight, age, liver weight and brain weight are in the nine most influential variables for predicting heart failure. They do however disagree on spleen weight and kidney weight. Only random forest views spleen weight to be in the nine most influential variables while random forest is the only one that doesn't have kidney weight in its top nine. This means that the models are influenced more by different autopsy variables when predicting heart failure. The result from the distributions of decedents labeled with heart failure in Figure 7 showed that decedents labeled with heart failure have liver weight distributions shifted towards higher values compared to the non heart failures. All the three models also showed that liver weight was in their top nine most influential variables for predicting heart failure. The SHAP values however showed for all the three models that high liver weights pushes the models away from predicting heart failure. A possible explanation could be that there exists other causes of deaths better corresponding with high liver weights. As seen in male cluster 1 in Figure 14. The cluster has a liver weight distribution significantly shifted towards higher values, but only one quarter of the decedents are actually labeled with heart failure.

Continue with the model performances in Table 11. Usually one expects the validation performance to be better than the test performance, since the model parameters are selected based on the best validation score. The reason that the test performance was slightly better in this case is most likely just a coincidence from the test split that led to a slightly easier test set for predicting heart failure. All the models performed quite similarly to each other on both the validation and test data. Since they all got an AUC score above 0.5, all the four models are better than randomly guessing. Which model is the best depends on what one values the most. If one values model interoperability, one might choose logistic regression or random forest. If one values the model with the highest performance, TabPFN scored the highest AUC, f1, accuracy and precision. TabPFN does however appear like a black box and currently has no SHAP values. If one values classifying the most heart failures correctly, logistic regression has the highest recall score. It does however also classify the most non heart failures incorrectly as heart failures. If one values that the percentage of classified heart failures is as high as possible, TabPFN has the highest precision, but it also classified the fewest number of heart failures as heart failures. If one values both, meaning one wants to find as many heart failures as possible, but still want to classify as few non heart failures incorrectly as possible, TabPFN and Random forest have the highest f1 score. Random forest could be a good trade off, since it offers more interoperability than TabPFN and correctly classified more heart failures than TabPFN and still misclassifies fewer non heart failures than logistic regression.

5.4 Conclusion

In summary, the dimension reduction of UMAP did not show strong differences between groups within the dataset. The hierarchical clusters showed stronger indications of groups within the dataset, but neither showed groups corresponding to the different causes of deaths. The largest differences found were groups of heavier decedents with heavier organs. These groups were however not exclusive to heart failures. Meaning, other autopsy variables are needed to better distinguish between different causes of deaths. The classifiers performed quite similar to each other and all performed better than randomly guessing with AUC scores between 0.75-0.79. In conclusion, the project works as a proof of concept for classifying cause of death based on autopsy data. In order to increase the performance of the classifiers, other autopsy variables better capturing differences between the causes of death are needed.

5.5 Future Work

With a continuation of the project, there are several things that could be improved upon. The most important one is to extract more autopsy variables from the autopsy reports. The increase in variables could possibly capture important differences needed to distinguish decedent from different causes of death. During this project, the focus has been on classifying heart failure as cause of death. If more causes of deaths are able to be classified, the classification models could be changed such that it can classify between several causes of deaths simultaneously. This would be more similar to the end product of a tool that can be implemented to the forensic practice. If time would have allowed it, the clustering would have been performed on the decedents of specific causes of death separately. For example, do clustering on only the decedents labeled with heart failure. This could potentially reveal subgroups within each cause of death.

A Appendix

Table A1: The left column lists the autopsy variables with the english names used in the project. The right column lists the corresponding original Danish variable names from the dataset.

Variable (English)	**Variable (Danish)**
Autopsy Date	Autopsy Date
Age	Age
Sex	Sex
Putrefaction level	Putre_level
Height	Højde
Weight	Vægt
Brain weight	hjernen
Right lung weight	højre lunge
Left lung weight	venstre lunge
Spleen weight	milt
Heart weight	hjerte
Heart height	Hjertet_højde
Heart width	Hjertet_bredde
Heart depth	Hjertet_dybde
Right cardiac wall thickness	højre hjertekammer
Left cardiac wall thickness	venstre hjertekammer
Cardiac septum width	hjerteskille
Liver weight	leveren
Liver height	Leveren_højde
Liver width	Leveren_bredde
Liver depth	Leveren_dybde
Right kidney weight	højre nyre
Left kidney weight	venstre nyre
Right kidney height	Højre nyre_højde
Right kidney width	Højre nyre_bredde
Right kidney depth	Højre nyre_dybde
Left kidney height	Venstre nyre_højde
Left kidney width	Venstre nyre_bredde
Left kidney depth	Venstre nyre_dybde
Right pleural cavity fluid	højre
Left pleural cavity fluid	venstre
Abdominal cavity fluid	bughule

Table A2: The left column lists the cause of death variables with the english names used in the project. The right column lists the corresponding original Danish variable names from the dataset.

Variable (English)	**Variable (Danish)**
Undisclosed	uoplyst
Drowning	drukning
Heart failure	hjertesvigt
Poisoning	forgiftning
Hanging	hængning
Gunshot Wounds	skud
Stabs and slashes	stik snit
Exsanguination	forblødning
Burns	forbrænding
Pneumonia	lungebetændelse
Suffocation	kvælning
Coronary Artery	coronary art
Cardiac Hypertrophy	cardiac hypertrofi
Inflammation	inflammation
Inner airway obstruction	indr sprring
Cancer	cancer
Subdural	hematoma SDH
Pulmonary Embolism	lungeemboli
Inner bleading	indre forbld
Cold	kulde
Injuries	kvæstelse(r)
Disease	sygdom

Table A3: The first column lists the autopsy variables. It is followed by the unit, the first quantile, the median and the third quantile. Then the count, number of missing values and the percentage of missing values is listed.

Variable	**Unit**	**Q1**	**Median**	**Q3**	**Count**	**Missing**	**% missing**
Autopsy Date	Days	10821	13550	16709	21678	110	1
Age	Years	40	51	63	21787	1	0
Height	cm	166	173	180	20884	904	4
Weight	kg	60	72	86	21688	100	0
Brain weight	g	1295	1408	1520	20968	820	4
Spleen weight	g	95	144	214	18222	3566	16
Right lung weight	g	530	672	830	19180	2608	12
Left lung weight	g	456	585	728	18781	3007	14
Heart weight	g	325	385	460	21350	438	2
Heart height	cm	11	12	13	5928	15860	73
Heart width	cm	10	10	11	5928	15860	73
Heart depth	cm	4	5	5	5928	15860	73
Right cardiac wall thickness	mm	3	4	5	19515	2273	10
Left cardiac wall thickness	mm	11	13	15	19931	1857	9
Cardiac septum width	mm	10	12	14	6271	15517	71
Right kidney weight	g	122	150	180	14225	7563	35
Left kidney weight	g	126	153	183	13832	7956	37
Right kidney height	cm	11	12	13	9644	12144	56
Right kidney width	cm	5	6	6	9646	12142	56
Right kidney depth	cm	3	3	4	9646	12142	56
Left kidney height	cm	11	12	13	9409	12379	57
Left kidney width	cm	5	6	6	9411	12377	57
Left kidney depth	cm	3	3	4	9411	12377	57
Liver weight	g	1300	1610	1965	18904	2884	13
Liver height	cm	25	27	29	19316	2472	11
Liver width	cm	17	19	21	19322	2466	11
Liver depth	cm	6	7	9	19322	2466	11
Right pleural cavity fluid	ml	100	200	400	3058	18730	86
Left pleural cavity fluid	ml	100	200	450	3165	18623	85
Abdominal cavity fluid	ml	100	300	900	1421	20367	93

Table A4: The left column lists the causes of deaths. The middle column is the number of decedents labeled with the cause of death. The last column is the percentage of decedents labeled with the cause of death.

Variable	**Count**	**% COD**
Poisoning	5776	26.5
Missing	4455	20.4
Coronary Artery	2401	11.0
Heart failure	2298	10.5
Undisclosed	2233	10.2
Inflammation	2138	9.8
Disease	1694	7.8
Suffocation	1407	6.5
Pneumonia	1188	5.5
Exsanguination	1141	5.2
Cancer	914	4.2
Drowning	536	2.5
Inner bleading	515	2.4
Pulmonary Embolism	497	2.3
Burns	451	2.1
Cardiac Hypertrophy	385	1.8
Inner airway obstruction	371	1.7
Injuries	371	1.7
Stabs and slashes	364	1.7
Gunshot Wounds	306	1.4
Hanging	224	1.0
Cold	188	0.9
Subdural hematoma	173	0.8

Table A5: The first row indicates the eight male clusters found through hierarchical clustering. The second row is the number of observations in each cluster. The third row is the percentage of observation belonging to the cluster. The last column is the result on the whole male dataset.

Male Cluster	1	2	3	4	5	6	7	8	All
Count	972	4217	697	3325	2029	546	947	1933	14676
Percentage	6.6	28.7	4.7	22.7	13.9	3.7	6.5	13.2	100

Table A6: The first row indicates the eight female clusters found through hierarchical clustering. The second row is the number of observations in each cluster. The third row is the percentage of observation belonging to the cluster. The last column is the result on the whole female dataset.

Female Cluster	1	2	3	4	5	6	7	8	All
Count	867	1122	456	593	257	2678	13	1122	7108
Percentage	12.2	15.8	6.4	8.3	3.6	37.7	0.2	15.8	100

References

[1] R. Menezes and F. Monteiro, "Forensic autopsy," *StatPearls*, 2023.

[2] P. Saukko and B. Knight, *Knight's forensic pathology*. CRC press, 2015.

[3] J. R. Busch and C. J. Wingren, "Automated data extraction from autopsy reports using a custom python script," *Forensic Science International*, p. 112756, 2025.

[4] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[5] G. James, "An introduction to statistical learning with applications in r," 2013.

[6] A. Lindholm, N. Wahlström, F. Lindsten, and T. B. Schön, *Machine learning: a first course for engineers and scientists*. Cambridge University Press, 2022.

[7] A. V. Joshi, *Machine learning and artificial intelligence*, vol. 261. Springer, 2020.

[8] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization.," *Journal of machine learning research*, vol. 13, no. 2, 2012.

[9] N. Hollmann, S. Müller, L. Purucker, A. Krishnakumar, M. Körfer, S. B. Hoo, R. T. Schirrmeister, and F. Hutter, "Accurate predictions on small data with a tabular foundation model," *Nature*, vol. 637, no. 8045, pp. 319–326, 2025.

[10] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[11] J. Healy and L. McInnes, "Uniform manifold approximation and projection," *Nature Reviews Methods Primers*, vol. 4, no. 1, p. 82, 2024.

[12] W. J. Conover, *Practical nonparametric statistics*. john wiley & sons, 1999.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[14] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[15] A. V. Ponce-Bobadilla, V. Schmitt, C. S. Maier, S. Mensing, and S. Stodtmann, "Practical guide to shap analysis: Explaining supervised machine learning model predictions in drug development," *Clinical and translational science*, vol. 17, no. 11, p. e70056, 2024.

Master's Theses in Mathematical Sciences 2026:E65
ISSN 1404-6342

LUNFBV-3013-2026

Computational Science
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/