Bachelor's thesis

Information Technology

NINFOS13

2017

Núria A. Vilarrasa

VISUAL EXPLORATION OF CLINICAL DATA WITH SELF-ORGANIZING MAPS

– A study case of atrial fibrillation

TURKU AMK
TURKU UNIVERSITY OF APPLIED SCIENCES

BACHELOR'S THESIS | ABSTRACT

TURKU UNIVERSITY OF APPLIED SCIENCES

Degree in Information Technology

December 2017 | 56 pages

Núria A. Vilarrasa

VISUAL EXPLORATION OF CLINICAL DATA WITH SELF-ORGANIZING MAPS

- A study case of atrial fibrillation

This Bachelor's thesis presents machine learning as a research tool by delving into Self-Organizing Maps for the exploration of clinical data. The purpose of this thesis was to provide hands-on experience on data analysis methods and the availability of real-world data confers additional value to this objective.

Self-Organizing Maps have been chosen for the simplistic visual approach they provide at a glance to explain the relationships within the data. All the implementations were carried out using MATLAB. This research was also aided by statistical methods such as principal component and correlation analyses.

The data contains attributes from patients suffering from atrial fibrillation, a heart condition that can lead to heart failure, stroke and other cardiac complications. The research data was provided by Turku University Hospital and was gathered prior to the start of this work.

The main question was to assess the bleeding and stroke risks for the patient depending on the duration of an onset. Answering such a specific question was proven too challenging and a general analysis of the data was performed instead. Alternatives are also proposed, such as determining the relationship between related medical scores and the research data.

KEYWORDS:

Self-organizing Maps; machine learning; atrial fibrillation

CONTENTS

APPENDICES

FIGURES

TABLES

LIST OF ABBREVIATIONS

AF	Atrial Fibrillation
ANN	Artificial Neural Network
ASA	Acetylsalicylic Acid
BMU	Best Matching Unit
CSV	Comma-Separated Value
ECG	Electrocardiogram, electrocardiography
IQR	Interquartile Range
ML	Machine Learning
NaN	Not a Number
NHLBI	National Health, Lung and Blood Institute
NSAID	Nonsteroidal Anti-Inflammatory Drug
PC	Principal Component
PCA	Principal Component Analysis
SOM	Self-Organizing Map
TIA	Transient Ischemic Attack

1 INTRODUCTION

In the last decade, machine learning has become one of the most prominently developing fields in the IT sector. Although the term was first coined more than half a century ago, it has not been until the recent years when it has started to appear in the spotlight. From robotics to finance, passing through biology and healthcare, a huge scope of applications has been derived thanks to this field.

There are two key factors that explain this occurrence. On the one hand, the sudden increase in data availability, promoted by the spread of smartphones and Internet of Things, requires data to be analysed in order to draw better conclusions based on underlying patterns. On the other hand, the improvements in computing hardware have made possible to build more complex and powerful machine learning models which were not feasible only a decade ago due to the computation complexity.

Among many available techniques, Self-Organizing Maps have been chosen as the main tool for the work on this thesis. They provide a method for visualising high-dimensional data on a two-dimensional representation, commonly referred as a map. This map is built on a lattice of nodes, and each of these nodes contains a weight vector that finds the closest data vector to itself. It is an intuitive approach to visualise high-dimensional data due to the concept that similar items are represented close together on the map.

The main motivation for this thesis was to study an innovative research area such as machine learning. There was a variety of publicly available datasets to base the actual work on, but thanks to Tapani Ojanperä, supervisor of this work, and Tuija Vasankari, a doctor from the Cardiology Unit at Turku University Hospital, a real-world problem has been proposed. The working dataset contains data of patients from Turku area who suffer from atrial fibrillation disorder and have received cardioversion. Therefore, the purpose of this thesis is to analyse the data provided by Turku University Hospital with the proposed methods in order to find underlying patterns. Overall, the real-world scope of the data has proved to be a great encouragement to pursue this work.

This work contains 5 additional chapters. Chapter 2 is an overview of machine learning and a general introduction to neural networks and to the statistical methods used in this thesis. Chapter 3 mainly focuses on the concept of Self-Organizing Maps and the algorithm that defines them as well as visualisations with the Iris dataset for an easy

approach to understanding this technique. Chapter 4 covers details on atrial fibrillation, a cardiac disorder and the domain of this thesis, for which TYKS has provided a dataset, also detailed in this section. Chapter 5 mainly focuses on the empirical analyses carried out in this work, including the data pre-treatment, the statistical results obtained as well as the main focus, the visualisations obtained through the use of Self-Organizing Maps. The closing chapter summarises the work and provides the final conclusions.

2 METHODS

The current chapter delves into the more theoretical aspects of the tools used for this work and is divided in two subsections. The first one starts offering an introductory perspective of the basics of Machine Learning and Neural Networks and discusses the choice of software tools selected for the development of the work. The subsequent section, on the other hand, explains the tools of statistical nature that have been chosen to further explore the given data before proceeding with the main methods.

2.1 Machine Learning

Machine learning (ML) was first introduced around the 1950s as a subfield of Artificial Intelligence but it was not until 1990s when it started to develop as an independent area. Back in 1959, Arthur Samuel (1959, 210) defined ML as

> «the field of study that gives computers the ability to learn without being explicitly programmed».

Decades later, in 1997, Tom M. Mitchell (1997, 2) provided a more formal and scientific definition:

> «A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E».

For a better understanding of Mitchell's definition, let us use a typical learning problem example of this field, which is the e-mail spam filter problem: A user receives dozens of e-mails daily but part of them are spam. Sticking to the definition, the task T would be filtering whether an e-mail is spam or not, experience E could be previously obtained from a known spam filter list, and performance P would measure the fraction of correctly classified e-mails.

The algorithms used in ML are essentially divided into two different categories: supervised learning and unsupervised learning.

Supervised and unsupervised learning

In supervised learning, a problem is approached by analysing first some available data in order to infer a general rule that can be applied to the data of the problem. The data used to derive this rule or function from is called *training set* and it consists of a representative collection of labelled examples containing an input vector and output value pair in each entry. Once this data has been generalised from and has produced an inferred function, the learning algorithm is then applied to the data of the problem, also referred to as *test set*, which contains only a set of input vectors, in order to obtain some label or output, for each object (Bishop 2006, 3).

In short, the training set that contains correct answers provides a rule that is later applied to the test set in order to estimate a correct solution for the problem. Among various supervised learning algorithms, two notable models are classification and regression tasks, being the difference between the form of the output; while the former considers two or more different discrete labels per output, serving as an example the previously introduced spam filter problem (spam/no spam), the latter produces continuous output values, such as a temperature prediction problem.

Unsupervised learning, on the contrary, does not depend on a labelled training set to infer a general rule for the learning algorithm. Instead, the algorithm itself finds structures and patterns inside the unlabelled data to discover hidden properties or features (Bishop 2006, p. 3). That implies there is no indicator to weigh the obtained result. A well-known type of unsupervised learning models are clustering algorithms, whose purpose is to find common patterns in the data in order to group in a common cluster different elements that were previously unrelated. A particular example is the algorithm used by Google News to crawl the web and group similar articles together.

Artificial Neural Networks

Artificial neural networks (ANN) are computational tools which seek to simulate the architecture and internal operations of the human brain and its nervous system. The goal is to imitate the human ability to learn, generalise and process large quantities of information rapidly and in parallel. These networks can be trained in order to determine

an output from some data input by reproducing the basic characteristics of biological neurons.

An ANN consists of processing elements, called neurons, connected to one another in a net fashion, each receiving multiple inputs and issue one single output. A graphical example can be seen in Figure 1.

Their procedure is similar to that of a biological neuron; if the sum of inputs exceeds a threshold value, the neuron emits a particular output to the neighbouring neurons. Reproducing the learning and adapting mechanisms of the network is established through mathematical calculations which define the parameters (e.g. number of neurons, number of layers) of the model.

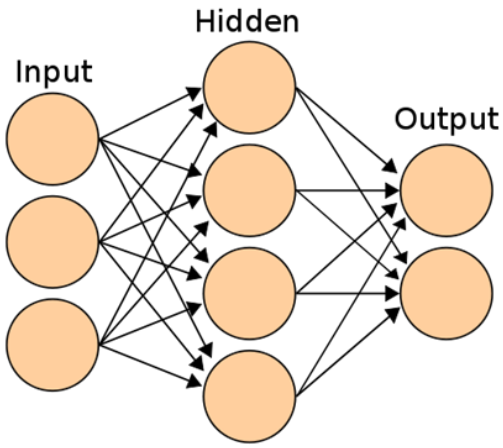


Figure 1. Graphical schema of an ANN.

The main difference between ANN and traditional interpolation programs is that the latter are based on known rules or equations that provide a clear definition of the problem. Then, the program defines step by step what has to be done to obtain the desired result. This system is useful if the rules of a particular behaviour are known. Neural networks can be used when these behavioural rules are hard to determine.

On the other hand, ANN can generalise from the given examples, even when the data is incomplete, not clear or missing values. The data can be theoretical, experimental, or empirical based on past experiences, or a combination of these types. The network will approximate to a rule or function, even though it will still be based on the experience obtained from training and does not contain explicit rules. Neural networks neither

provide explanations nor justify the obtained results, furthermore, they cannot apply the solution to another problem out of their training domain.

2.1.1 Software tools proposed for this method

In the first instance, multiple tools were presented to perform the present work. There were three main choices: R language, Octave and MATLAB. For convenience, MATLAB was selected along the SOM Toolbox v2.1, developed by the Laboratory of Computer and Information Science at Helsinki University of Technology, currently Aalto University (Vesanto et al. 1999; Vatanen et al. 2015, 35 – 40). The reason for opting for this particular implementation was due to the fact that it has been emphasised by Kohonen and his team (2013) to rely on the SOM Toolbox for research, as well as the availability of MATLAB on the university's computers.

This work also used Microsoft Office Excel 2016 for minor cases of the work.

2.2 Statistical methods

This section introduces the statistical methods used in this work. The first approach has been by performing a correlation analysis on the whole dataset by using Pearson r in order to perceive the correlations between attributes. After observing the results, Principal Component Analysis has been chosen to reduce the dataset's dimensionality.

2.2.1 Pearson correlation analysis

In order to find more relationships in the data, a correlation matrix is computed with the remaining attributes. The chosen method was Pearson r correlation.

Pearson's r correlation is a statistical measure to find how two variables such as X and Y are related to each other by measuring the strength of the linear dependence between them. There can be three outcomes: Positive correlation, negative correlation, or no correlation. The formula for Pearson correlation is

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(\mathrm{n} - 1)s_x s_y} \quad (1)$$

where n is the number of units, x_i and y_i are one value, $\bar{x}$ and $\bar{y}$ are the sample mean of each dataset and s_x and s_y are the standard deviations for each variable X and Y.

The correlation is represented by the range $[-1, 1]$ and the closer to the edge, the more intense the correlation is. On a relationship scale, it can be interpreted as:

- 0.00 – 0.19: Very weak
- 0.20 – 0.39: Weak
- 0.40 – 0.59: Moderate
- 0.60 – 0.79: Strong
- 0.80 – 1.00: Very strong

However, due to Pearson r correlation being a linear measurement, a value of $r = 0$ does not imply the target attributes are not correlated but rather a lack of linear correlation between them. This is important as data can present other types of correlations, such as quadratic, which this method would not recognise. A visual example is pictured in Figure 2, where both plots display a positive correlation, but the second portrays a quadratic relationship.

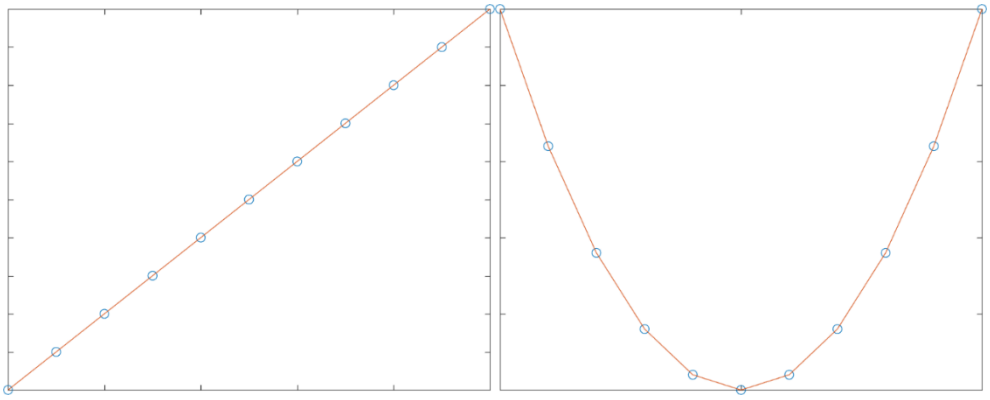


Figure 2. Plots representing positive linear and positive quadratic correlation.

Phi coefficient

The phi coefficient is a measure of association particularly used when observing binary variables, i.e. variables which values only have two states, 1 and 0. Examples are features such as alive or dead, male or female, pass or fail. This measure is comparable to Pearson r correlation coefficient in the interpretation and given a 2x2 contingency table, such as the one presented in Table 1:

Table 1. Basic contingency table for phi coefficient.

		Attribute 1	
		Yes	**No**
Attribute 2	**Yes**	a	b
	No	c	d

Where a, b, c and d represent the observation frequency of each attribute, ϕ then follows

$$\phi = \frac{a\,d - bc}{\sqrt{\{(a+b)(c+d)(a+c)(b+d)\}}} \tag{2}$$

The correlation is positive if most data falls in the diagonal cells. On the contrary, if most data falls off the diagonal, the correlation is negative.

An example is proposed in order to understand the phi coefficient, where 50 people, divided in two groups by gender and by type of diet, are presented:

Table 2. Example for phi coefficient.

	Male	**Female**
Vegetarian	10	5
Non-vegetarian	10	25

By applying Equation 2, a value of ϕ = 0.356 is obtained from the values f the previous table, which can be interpreted as a weak positive correlation between these variables (Chedzoy 2006).

MATLAB's function for computing correlation recognises whether the data is binary and automatically applies phi-coefficient if required.

2.2.2 Principal Component Analysis

Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction meaning its aim is to represent a high-dimensional dataset, i.e., data containing a high number of attributes, into a set with a reduced number of dimensions while keeping most of its original information. It is useful when dealing with data with redundant

attributes which do not provide new information or to improve the visualisation of noisy data (Silipo et al. 2014, 11).

The requirement on the data in order to carry a successful transformation is to contain linearly correlated features, or in other words, features that can be expressed as a result of a linear combination of other features from the dataset. As an example, given a dataset $X = \{x_1, x_2\}$ containing the sizes of different rectangular papers, with attributes x_1 as length and x_2 as height, a correlated attribute would be x_3 as the perimeter of the paper, since it can be expressed as $x_3 = 2x_1 + 2x_2$, and hence it can be reduced from the dataset due not providing new information. In the case of a fourth attribute x_4 representing the area of the paper, it would not be recognised as redundant since the attribute itself is a result of the non-linear operation $x_4 = x_1 \cdot x_2$.

So the principal components (PC) are uncorrelated normalised linear combinations of the original features, ranked by an information basis represented by the variance of the data. Therefore, for n number of attributes in a dataset there can be n PC where the first PC, or PC1, is the linear combination that carries the largest amount of variance from the data, PC2 will be the is the linear combination that carries the largest amount of variance with the constraint of being uncorrelated with PC1, and so forth up to the last PC, or PCn, which is the linear combination of the attributes with the minimum variance.

These PC are derived from the data, contained in matrix form, by calculating the eigenvalues and eigenvectors related to it. An eigenvector v is a non-zero vector that does not change its direction when multiplied by a scalar multiple of itself referred to as an eigenvalue λ, hence the linear transformation $T(v) = \lambda v$.

Eigenvectors represent the PC of the data. Eigenvalues for a given PC measure the variance in all the variables related to that PC. Therefore, eigenvalues represent the importance of a PC with respect to the variables. If a component has a low eigenvalue, then it has very low influence in explaining the variances in the variables.

Eigenvalues measure the amount of variation in the data related to each component. This variation is defined by the variance of the data itself. Variance is a statistical measurement used to indicate the spread of the data points from their mean. The variance for any given population, often represented as σ^2, follows the formula

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N} \tag{3}$$

where X is the values of the population data points, μ is the mean of the data and N is the number of data points.

In this work, PCA has been applied to reduce the dimensionality of the last section of the data, mostly containing information about medications.

3 SELF-ORGANIZING MAPS

Unsupervised learning has been previously introduced as a type of algorithm to learn without relying on more data than the available in the dataset. Unsupervised learning is used to find patterns in unlabelled datasets. An interesting kind of unsupervised learning is competitive learning, which is based on the output neurons competing amongst themselves to be activated. The result is that only one activates at a time, and this neuron is called the “winner-takes-all” neuron. Afterwards, the neurons are forced to organise themselves, and thus this network is known as Self-Organizing Map, or SOM for short.

3.1 Overview on Self-Organizing Maps

The SOM algorithm was first described in 1981 by the Finnish professor Teuvo Kohonen (1981), who gives the name to the most well-known type of SOM, the Kohonen networks. At the early stages, SOM emerged from neural network models centred around associative memory and adaptive learning fields in order to provide an explanation of the spatial organisation of the functions of the brain, observed particularly in the cerebral cortex (Kohonen 2014, 16). Although he developed it to map the distribution of metric vectors, such as measurement values or statistical features, it can actually be used to define a visualisation of any data items that have pairwise common distances between its items. As a consequence, an SOM can work on both numerical and categorical data, but the former requires pre-treatment to numerical equivalent values.

Given a set of high-dimensional data items, an SOM represents them as a lower-dimensional image, typically two or three-dimensional, which maps every item into one node, or neuron. This image is referred as a map and is built on a lattice, usually with hexagonal form (Figure 3). The distances between nodes in the low-dimensional visualisation reflect the similarities between the items in the high-dimensional space. This similarity preservation is what differentiates SOM from other ANN models (Kohonen 2014, 2).

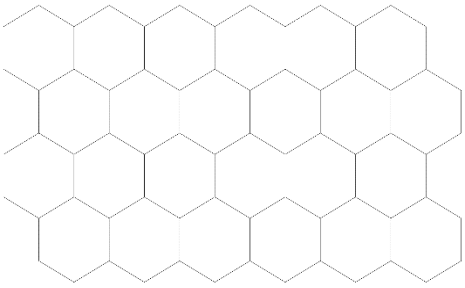


Figure 3. Empty SOM lattice with size 6 x 4 nodes, represented as hexagons.

Starting from an initial state of complete disorder, the SOM algorithm will gradually lead to an organised representation of activation patterns drawn from the input space. The initialisation of the nodes can take place by simply selecting random vectors, but Kohonen demonstrated that a much faster convergence follows if the initial values are selected as a

> «regular, two-dimensional sequence of vectors taken along a hyperplane spanned by the two largest principal components of x»

namely, the PC associated with the two highest eigenvalues of the input data (Kohonen 2014, 22 – 23). This strategy is considered in the situation where the distances between the input data use the Euclidean metric.

The SOM learning principle is that

> «every input data vector shall select the model that matches best with it, and this "model", called the winner, as well as a subset of "models" that are its spatial neighbors in the array, shall be modified for better matching.» (Kohonen 2014, 2)

This principle is followed by the SOM algorithms, and two of these algorithms, both proposed by Kohonen, are briefly introduced below.

3.2 Stepwise recursive algorithm

The stepwise algorithm is mainly theoretical with no application in the current work. It was proposed heuristically by Kohonen (1990) when trying to materialise the general learning principle listed above. The initialisation algorithm follows the previously introduced approach where the initial weights are randomly selected along the vectors taken from the main eigenvectors of the data.

Let the input data items be considered $\{x(t)\}$ on a particular step in the sequence t. Let $\{m_i(t)\}$ be another sequence representing the computed approximations of the weight m_i, where i is the index of the node's spatial location on the grid following

$$m_i(t+1) = m_i(t) + \mathrm{h}_{ci}(t)[x(t) - m_i(t)] \quad (4)$$

where $h_{ci}(t)$ is the neighbourhood function, c represents the index of the node in $m(t)$ has the smallest distance from $x(t)$, namely, the winner node, also referred as best matching unit (BMU), represented by $m_c(t)$.

Subsequently, a recursive step where the input data item $x(t)$ selects the BMU in the grid and the BMU's weight, as well as the weights of the nodes in the BMU's neighbourhood, are updated. The basis of the update is to direct the nodes' weights in order to match better with eventual similar inputs. The modification rate per node depends on the mathematical form (of the function $h_{ci}(t)$, such as)

$$\mathrm{h}_{ci}(t) = \alpha(t) \exp[-sqdist\ (c, \mathrm{i})/2\,\sigma^2(t)] \quad (5)$$

where $\alpha(t)$ is a decreasing function of t, $sqdist(c,i)$ is the square of the distance between the nodes c and i in the grid, and $\sigma(t)$ is another decreasing function of t. The topographic order of the units on the grid is established by this iteration.

Every model must be updated often enough in order to achieve sufficient statistical accuracy (Kohonen 2013).

Summarising the stepwise algorithm can be done in five steps:

1. Initialization: The initial weight vectors for the nodes of the map are initialised by choosing random values.
2. Sampling: A sample input vector is selected from the input space.
3. Matching: Every map node is checked to calculate which weight is most similar to the input vector. This winner node is the BMU.
4. Updating: The closest nodes to the BMU are calculated and updated.
5. Convergence: Repeat from step 2 until the feature map does not change anymore.

3.3 Batch algorithm

As mentioned before, the stepwise algorithm has only theoretical interest. In practice, the batch computation is preferred since it contains fewer parameters and the convergence time is much faster. Although not included in this work, this algorithm can also be used to generalise non-vectorial data (Kohonen 2001; Kohonen 2013, 37).

For each node i, the algorithm associates a weight m_i and a list containing copies of certain input vectors $x(t)$. The initialization of the weight values follows the same principle as in the stepwise algorithm, by choosing regular two-dimensional values per node selected from the two-dimensional hyperplane originated from the two largest principal components of x.

Let the input data items be considered a set $\{x(t)\}$, having t as a particular integer-valued index of a vector. Drawing a parallel between this concept and the dataset of this work, $x(t)$ would be a patient. Each $x(t)$ is compared with all the weights of the nodes, and the node that is more similar to $x(t)$ is selected. Thus, $x(t)$ is copied into a sublist associated with this node, the BMU, and this is repeated for each node in the input data set.

Being N_i the neighbourhood set, the mean of all $x(t)$ for each node is computed and is copied into the union of all sublists in N_i. A similar mean is computed over the neighbourhoods of all the nodes. Updating the weights of each node means replacing the respective means and is done in one updating operation cycle for all nodes of the grid (Kohonen 2014, 37).

Hence, the batch algorithm can be summarised by the following steps:

1. Initialization: The initial weight vectors for the map's nodes are initialised.
2. Sampling: From the input set of vectors, each sample input is selected.
3. Matching: Every map node is checked to calculate which weight is most similar to the input vector. This winner node is the BMU. When found, the sample input vector is associated in the sublist of the winner.
4. Repetition: For each vector in the input set of vectors, repeat steps 2 and 3.
5. Updating: The nodes are updated by calculating the mean from the vectors stored in each sublist of a node of the neighbourhood. All the nodes of the grid are updated at once.
6. Convergence: The cycle is repeated until the map does not change anymore.

3.4 Visualising the Iris flower dataset

The Iris flower data set is, perhaps, the most well-known data set in ML. Collected by Edgar Anderson (1935) and first introduced by Ronald Fisher the following year (1936), their work is still referenced to this day. The dataset can be found in the Machine Learning Repository from the University of California, Irvine (Lichman 2013). The database contains 150 instances representing a type of the iris plant with its corresponding attributes, which include sepal length, sepal width, petal length and petal width and the corresponding class: *Iris setosa*, *Iris versicolor* or *Iris virginica*.

Being a popular and well-studied data set with properties that can be easily represented, it is an ideal tool to provide an introduction to SOM visualisation. Therefore, a more accessible perspective to the current method is provided in order to approach its practical aspects in a clearer style.

Distribution visualisation

Before proceeding with the training of the SOM, it is a useful step, when the data allows it, to start by visualising the distribution of the data along the variables. In this case, since the data contains values in float numbers, it is a good idea to perform histograms and scatter plots to get a general overview. This can be observed in the following figure.

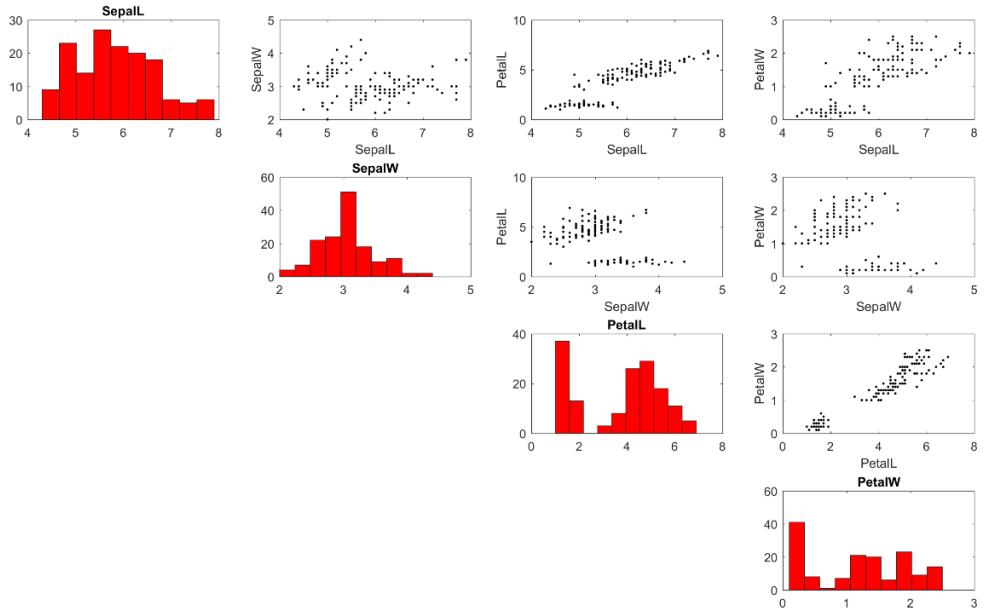


Figure 4. Visualisation of the Iris dataset by feature frequency and pair correlation.

In this representation, the histograms can be observed on the diagonal while the scatter plots are on the off-diagonal. It can be observed how all the scatter plots detail two clearly distinct groups, with approximately one-third of the data represented in the less crowded cluster. This could be understood as a relationship by the fact that the data contains three different classes of Iris flower; hence maybe two are more closely similar while the third one has more distant features.

Unified distance matrix

It has been mentioned in previous sections that the position of the nodes on the map represents how similar the dataset points are between themselves or, in other words, the distance they share. Note that being more similar denotes the distance between data points is lower, therefore the concept of similarity is opposite to distance. Hence, two points which have fairly similar features will likely be close together on the map; their distance is low. Likewise, two points which have notably dissimilar attributes will be more distant from each other; their distance is higher (Kohonen 2013, 11).

This property is displayed in the unified distance matrix or U-matrix for short. Once the SOM is trained, the U-matrix represents the distance vectors of the nodes for all the attributes at the same time. These distances can be visualised through the colours of the image. In this particular example, Figure 5 shows how the Iris dataset is distributed along the nodes of its map.

At this point, it is important to mention that the distance values of the nodes are stored in a matrix referred as codebook (Kohonen 1990). This codebook matrix is randomly initialised at the beginning of the computation and is tuned with each iteration of the algorithm. Each cell of the matrix contains the similarity value for a particular node and a particular attribute. Therefore, the matrix contains as many rows as cells and as many columns as attributes of the input data. By default, the U-matrix illustrates the distance while accounting for all the input attributes, often represented them with a greyscale colour scheme.

The lateral colour bar displays the maximum, minimum and midpoint distance. The lighter the colour, the closer the points are, representing where the similar units cluster. On the other hand, the darkest areas represent the most dissimilar data points and, therefore, can be considered as the borderlines between clusters.

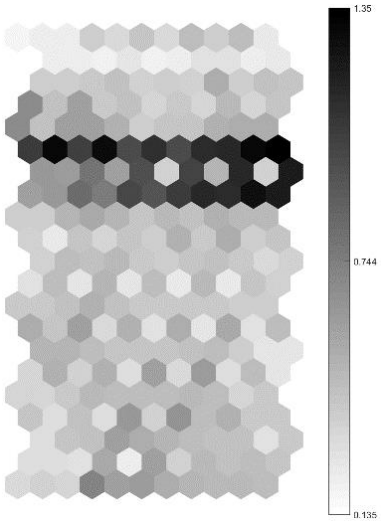


Figure 5. U-matrix of the Iris dataset.

Component planes

Another useful visualisation tool is, together with the U-matrix, the representation of the components on the map where the components are the different attributes the data has. Using the current example, the iris dataset contains four different components, namely the attributes sepal length (SepalL), sepal width (SepalW), petal length (PetalL) and petal width (PetalW). Figure 6 shows the representation of these components in the greyscale colourmap.

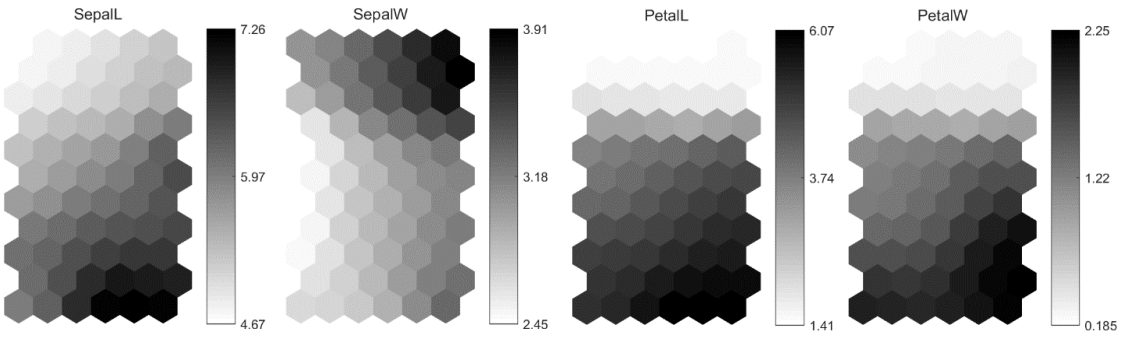


Figure 6. Component planes for each attribute of the Iris dataset.

By comparing each of the components, it can be observed how related they are in respect of the others. Clearly, it can be said that the petal width of a sample is related to its petal length, while there is a secondary relationship in regard to the sepal length,

seemingly close to the petal size. On the contrary, sepal width seems to present higher values on those elements where the previously mentioned values of the attributes are lower.

Hit histograms

It is important to note that one element is represented in the same position for all representations of a map. In Figure 7 it can be seen how a single data point (represented in blue) is in the same position for both, the U-matrix and the component planes representing sepal width and length. The individual elements represented on the map are often referred as “hits” (Kohonen 2013).

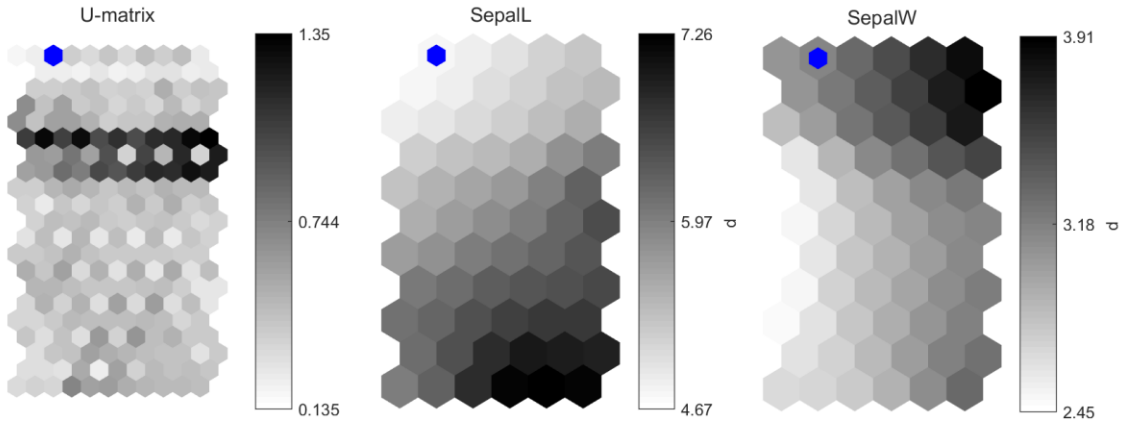


Figure 7. Visualisation of the position of a particular element on three different representations.

When dealing with classified data, this is a useful feature to provide more insight on the map. The Iris data set contains four attributes per each element, and a fifth column listing the class of that element as the name of the iris flower species it represents. A total of 150 elements under 3 classes are equally portrayed and a particular visualisation can be achieved by plotting the hits on the map.

In the following page, Figure 8 depicts the U-matrix of the Iris dataset plotted with all the hits, coloured in yellow. The larger the hit point, the more points there are in that area, hence can be confirmed that the darker areas contain a lesser amount of elements while the lighter ones contain most of the data points.

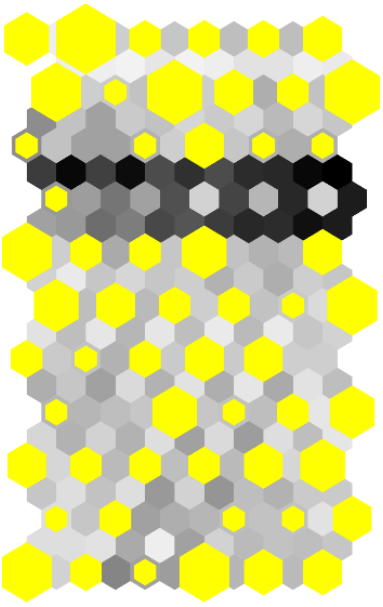


Figure 8. Iris data set hit histogram over the U-matrix.

In Figure 9, by contrast, the hits have been coloured according to their class and are represented over an empty lattice thus providing a visual reference concerning how the elements have been distributed by the SOM. Red colour indicates *I. setosa*, green colour *I. versicolor* and blue colour *I. virginica*.

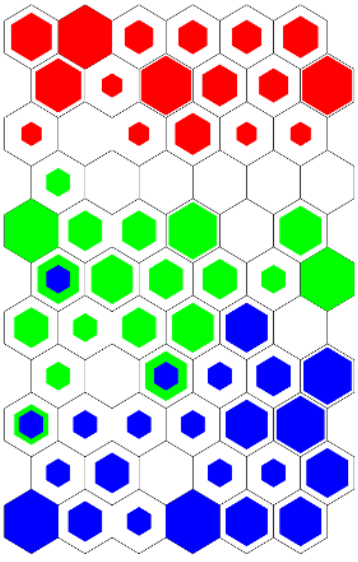


Figure 9. Hit histogram over empty lattice, coloured by class.

4 CASE STUDY

This chapter serves as an overview of the medical field to which this work is related. It first introduces the studied disorder and proceeds with the introduction of the dataset this thesis has used.

4.1 Atrial fibrillation

Atrial fibrillation, or AF for short, is a cardiac disorder caused by an abnormal heart rhythm, or arrhythmia that can cause blood clots, stroke or heart failure among other complications.

In a healthy heart, the electrical impulse that characterises the heartbeat is generated in the sinoatrial node (SA node) located in the right upper chamber (atria) of the heart. From there, it moves through the atria causing them to contract and release the blood to the lower chambers to the atrioventricular node (AV node) which is located right above the ventricles. When the ventricles are filled with blood, the impulse travels through them causing a contraction to pump the blood to the rest of the body (NHLBI 2011).

When AF is present, impulses are generated faster and more chaotically. When they travel through the atria, these start to flutter releasing the blood irregularly to the ventricles while the signals reach the AV node. The node processes the impulses through the ventricles, which in turn contract and release blood, but at a slower rate than the atria so both upper and lower chambers beat with an irregular rhythm (NHLBI 2014).

The most common noticeable symptom is a rapid and irregular heart rate, which in turn may cause palpitations, chest pain or shortness of breath, among other possibilities. It is often detected during the course of a transient ischemic attack (TIA) or stroke, although it is not unusual for the patient not to present clear symptoms and therefore being unaware of the condition until a routine examination or an electrocardiography (ECG) are performed (NHLBI 2014, Diagnosis). On the left side, Figure 10 locates the previously mentioned parts of the heart, while on the right side it illustrates on an ECG the difference between a healthy heart with normal heart rhythm, also called sinus rhythm, and a heart suffering an AF arrhythmia. On the left, there is the representation

of the electrical signals through the heart, from the SA node through the AV node and to the ventricles.

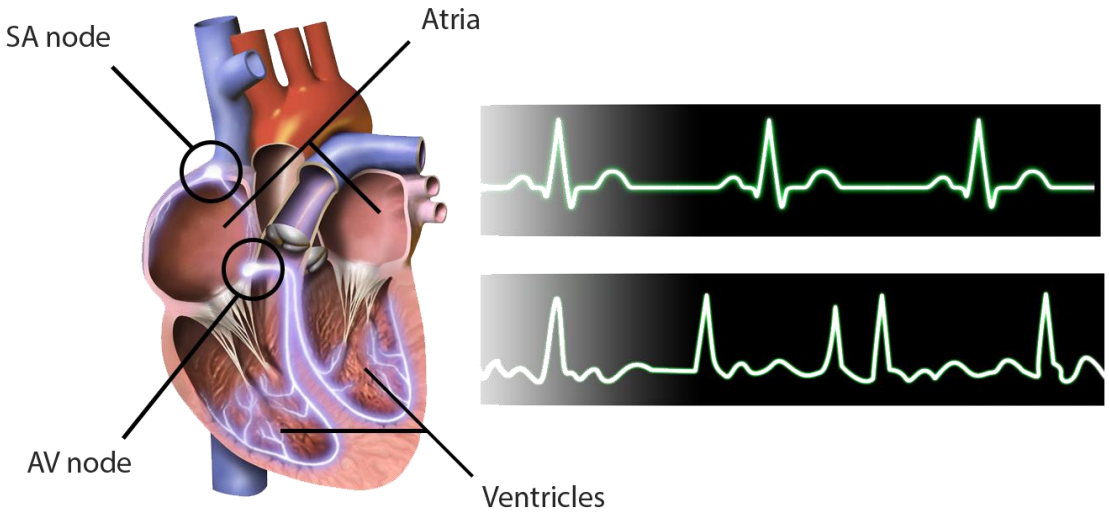


Figure 10. Heart diagram (left). ECG with sinus rhythm (top) and AF arrhythmia (bottom).

The predominant factor is age; the older the person is the higher the risk is to suffer AF. Other major risk factors, listed by the United States NHLBI (2014, Who is at risk), include:

- High blood pressure
- Coronary heart disease (CHD)
- Heart failure
- Congenital or structural heart defects
- Other heart conditions

A common complication is the formation of blood clots, which affects the treatment procedure. When AF interferes with the normal heart rhythm, blood can pool in the left atrium which can consequently clot. This highly influences the treatment procedure, because performing a cardioversion restarts the heart rhythm, it can lead potential clots to be released into the blood flow which could cause serious damage upon reaching other vital organs. The risk increases the longer the AF episode, and the guidelines recommend a limit of 48 hours after the onset to perform a cardioversion without oral anticoagulants (OAC). Otherwise, OAC are administered for a period of 3 to 4 weeks prior to and following a cardioversion. (Nuotio et al. 2014)

However, even though there is no definitive cure for this condition, and although the most common approach is performing an electrical cardioversion, it can also be treated with drug therapy or surgery (NHLBI 2014, Treatments). Concerning prevention and risk reduction, the recommendations go along the lines of following a healthy lifestyle and keeping the circulatory system in the best condition by doing physical activity regularly, eating a heart-healthy diet, controlling alcohol consumption and not smoking (NHLBI 2014, Prevention).

Presently, TYKS uses two medical scores for patients suffering AF: CHA_2DS_2-VASc for stroke risk and HAS-BLED for bleeding risk.

CHA_2DS_2-VASc

CHA_2DS_2-VASc is a stratification schema used to predict the stroke risk in AF patients. The index ranges between 0 – 9 and is derived from 7 parameters weighing between 0 – 2. The parameters are divided into definitive risk factors (i.e. Age ≥ 75 and stroke, TIA or thromboembolism (TE), score 2) and combination risk factors (i.e. the rest, score 1).

Patients can be categorised into high-risk patients, with one definitive or two or more combination risk factors, intermediate-risk, with one combination risk factor, and low-risk, with no risk factors present. (Lip et al. 2010)

The stroke risk assessment parameters are:

- **C**ongestive heart failure/Left Ventricular dysfunction
- **H**ypertension
- **A**ge ≥ 75 years old
- **D**iabetes mellitus
- **S**troke/TIA/TE
- **V**ascular disease
- **A**ge 65 – 74 years old
- **S**ex **c**ategory (i.e. female gender)

CHA_2DS_2-VASc supersedes a previously established scoring system, $CHADS_2$, as it provides a better assessment for low-risk patients by including the distinction of definitive risk factors as well as considering risk factors related to TE. (Odum et al. 2012)

HAS-BLED

HAS-BLED is a scoring system introduced in 2010 to provide an assessment of the bleeding risk factor for AF patients during a 1-year period. The score's range is between 0 – 9 and is calculated from 7 parameters weighing between 0 – 1. A score of 9 means a high risk of major bleeding while a score of 0 means no risk.

The bleeding score parameters are:

- **H**ypertension
- **A**bnormal renal and/or liver function
- **S**troke, prior history
- **B**leeding, prior history or predisposition
- **L**abile international normalised ratio (INR)
- **E**lderly (> 65 years)
- **D**rugs and/or alcohol, usage history

The score was developed to determine whether the use of OAC is optimal for a patient. OAC are used to reduce the stroke risk in AF patients. However, stroke risk is often associated with a bleeding risk that OAC can aggravate. (Pisters et al. 2010)

In the working dataset, some OAC drug therapies include Marevan, Arixtra, Clexane, ASA, clopidogrel, Efient and Fragmin.

4.2 Working dataset

The work on this thesis has been carried out using a dataset containing 7,727 records and 107 attributes of patients with AF disorder. The dataset is in the Finnish language. The data has not been processed prior to the obtaining, hence being considered a "raw data set", meaning it has only been extracted from a source but has not received any treatment.

The dataset is in Comma-Separated Value (CSV) format and contains mixed data types per features, most of them being categorical. This poses a limitation for this work as an SOM provides better results when trained with numerical attributes where the distance between the values is clearly defined. Nevertheless, it is possible to compute an SOM

using categorical data, but a further study based on the domain of the dataset must be considered in order to derive the most information out of the parameters it defines.

The entries on the dataset represent patients who suffer from AF and have been treated in TYKS cardiology department. Each patient has one record per diagnose meaning a patient can appear more than once if multiple diagnoses have been carried out.

The 107 attributes are divided into four main sections:

- Personal information: data that identifies each entry
- Current diagnosis: data recorded during the diagnose for which the entry was created
- Previous medical records: medical history prior the diagnose
- Medication: medications provided for current treatment

The study centres on data from the last three sections.

5 EMPIRICAL ANALYSIS

This section contains the implementation of the methods previously described as well as the results obtained. It concludes with the discussion of the results.

5.1 Data pre-treatment

The first treatment done to the data consists of string detection and substitution from all value fields, excluding the dataset headers. This step is required due to the inability to manipulate mixed data files on MATLAB and also because the data provided in these cells is too diverse to establish a logical connection. Given the dimensionality of the dataset, this action has been carried out in a semi-automated fashion using Microsoft Excel software.

In particular, the most important change has been the balancing of features. The apparent problem is that some string-only values have occupied more than one cell during the data input phase, creating a displacement in the whole dataset where some records had a total number of features larger than expected by few cells. The finding of the affected attributes has been carried out by finding which records had more features than intended and then detecting and removing the problematic features. Hence, after this step data has been reduced to 107 attributes for all records.

Subsequently, some attributes have been deleted to maintain the dataset with numerical-only features. These include all the features obtained from the string-only fields in the questionnaire as well as date-based values. Additionally, four other variables have been deleted due to not providing relevant information. These include parameters such as the initials or the number of the patient. After this step, the dimensionality of the data has been reduced to 88 numerical attributes.

On a closer look at the data, various entries containing string values on numerical attributes have been revealed along with other cells containing Not a Number (NaN) values, neither being suitable for analysis. The remaining string values are converted to NaN value all the same, albeit not performing well with most data analysis tools. At this point, the options are to delete the affected records and features or substitute them for numbers.

The options are not complementary so both are considered. First, a thorough analysis of the data types was carried by searching for values out of the expected range. Most of the features should have accepted only binary attributes, but due to them not being mandatory, leaving them blank would cause this issue. Features containing a majority of NaN values have been discarded automatically, i.e. all the data from the last section of the questionnaire, “Paralysis/At the time of bleeding”, since the answering rate is the lowest compared to the rest of the attributes so it does not provide enough information. The response/completion rate of these attributes can be observed in Table 3.

Table 3. Response rate proportion with respect to the whole dataset.

Marevan	Clexane	Fragmin	Arixtra	ASA	Clopidogrel	Efient
1,70% (132)	1,64% (125)	1,63% (126)	0,36% (28)	1,68% (130)	1,74% (135)	1,59% (123)

For simplicity, the binary format was chosen for as many attributes as possible. The work uses MATLAB from this point onward.

At the same time, features containing numerical values out of the range 0 – 1 previously established were modified. In the case of attributes such as “last rhythm” which contains values from 1 – 3 and “duration of current AF” which ranges from 1 – 4, new features were created for each answer choice in order to convert them to multiple binary attributes. Variables such as “gender” or “cardioversion” were kept as dichotomous by changing the value 2 to 0 while maintaining the value 1. In these particular cases, “gender” 1 represents the female patients (0 for male) while “cardioversion” 1 represents the patients who received electrical cardioversion (0 for pharmacological cardioversion).

At that point, the dataset contained 7,681 entries with 86 attributes.

5.2 Implementation of statistical methods

A search for far outliers, values isolated from other values found outside of the accepted range, were carried out by visualising boxplots from 2 of the non-binary features. Boxplots were computed by using the four quartiles of the data, which visualise where most of the data is located and how spread it is. From this information, the interquartile range (IQR), which is represented by the main box of the diagram, can be formulated as $IQR = Q3 - Q1$. This information is used to calculate the range for outliers, values

under the minimum and over the maximum accepted values, calculated by $min = Q1 - 1.5 \cdot IQR$ and $max = Q3 + 1.5 \cdot IQR$. The boxplots for two attributes are presented in Figure 11.

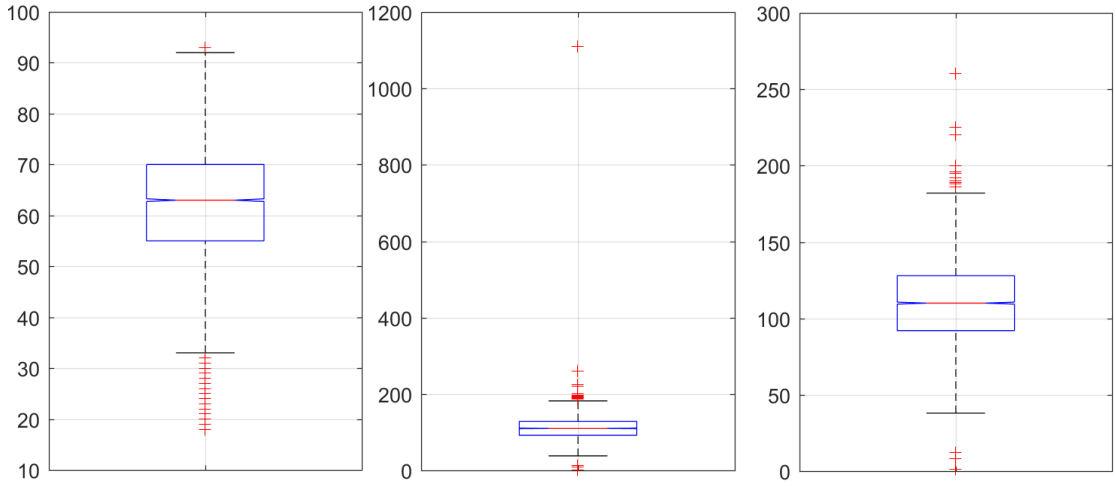


Figure 11. Boxplots for ”age” (left) and ”ECG” value (centre and right).

In the first boxplot it can be seen that the age median is around 64 years. Half of the patients fall within the range of 55 to 70 years, represented by the box. The remaining population go from as young as 33 and up to 92, displayed by the "whiskers" above and below the box. Still, there are some outliers, one above 92 and many below 33, to as young as 18, represented by the red crosses.

The second boxplot represents the value for ECG which represents the electrical activity of the heart. The normal sinus rhythm for an adult is between 60 and 100 beats per minute (bpm), with the average at 82 bpm. However, it can be seen there is a far outlier at the 1100 mark. This value is clearly a typing error at the moment of inputting the data so this data entry has been deleted. Due to this outlier, the boxplot was hard to read.

Hence, the third boxplot represents the same ECG attribute without that far outlier and in this case, it is possible to read it fairly well. The median value is around the 120 bpm mark with the maximum around 185 bpm and the minimum at 40 bpm. Yet, most of the data can be found between the marks from 90 to 130 bpm and most outliers are between 180 and 250 bpm.

The next step was to perform a significance test to decide whether there is evidence that the population has a linear correlation. Statistical significance is achieved when the probability value, usually referred as ρ-value, is less than the significance level, referred

as α. The ρ-value represents the probability of obtaining at least the same results given the condition that the null hypothesis is true. Hence, a null hypothesis must be proposed. A value of ρ = 0 is chosen to represent the null hypothesis, assuming there is no linear correlation in the population, against an alternative hypothesis of ρ = 1, opposing the previous statement meaning the population is linearly correlated.

Next, a significance level α was chosen in order to accept or reject the null hypothesis. If the probability of a larger difference is less than or equal to the significance level, then the null hypothesis is rejected and the result is said to be statistically significant. A value $\alpha = 0.05$ is chosen.

The sample values for the data were computed by using the corr function from the Statistics Toolbox in MATLAB and are plotted as an 86x86 matrix (Figure 13). The test is two-tailed and set to ignore NaN values. Using the same function, the ρ-values were computed and those for the variables with moderate or stronger correlation ($r > 0.4$) were visualised. Coincidentally, for all $r > 0.4$, the probability values were found to be $\rho = 0$. This implies there is a very strong evidence to refute the null hypothesis and thus believe there is some correlation. The obtained results are as follows.

There is a very strong positive correlation (0.819) between patients who suffered stroke or TIA and those who suffered TE, stroke or death during the last month which is an expected outcome. There is also a strong relationship (0.705 and 0.592) between those patients who took nonsteroidal anti-inflammatory drugs (NSAIDs) and acetylsalicylic acid (ASA). Besides, albeit predictable, there is a strong correlation between the medications taken while admitted at the hospital and those prescribed at home. Unfortunately, some of the listed drugs have too little representation in the current population that the connection cannot be considered reliable although both, ρ-value and r value, gave proper results. Lastly, a relation can be observed between Marevan users and a three week INR value (0.758 and 0.693), which is consistent with the facts as INR is used to monitor patients who are on OAC drug therapy, such as Marevan.

On the other hand, a moderate negative correlation between electric cardioversion and “flecainide” can be appreciated. After a brief research, it were found that flecainide is a medication used to control cardiac arrhythmias, hence the relationship (Drugs.com 2017). A correlation between Marevan users at home and ASA users at home can also be noticed (-0.424). Exclusive features such “last heart rhythm” (AF, SR or Not Available) or the “AF episode duration” in hours (< 12, 12 - 24, 24 - 48 or > 48) are ignored.

Various weaker relationships can be observed such as age with hypertension (0.287) or other coronary artery diseases (0.256), or other vascular problems with AF number (0.201) or heart failure (0.209).

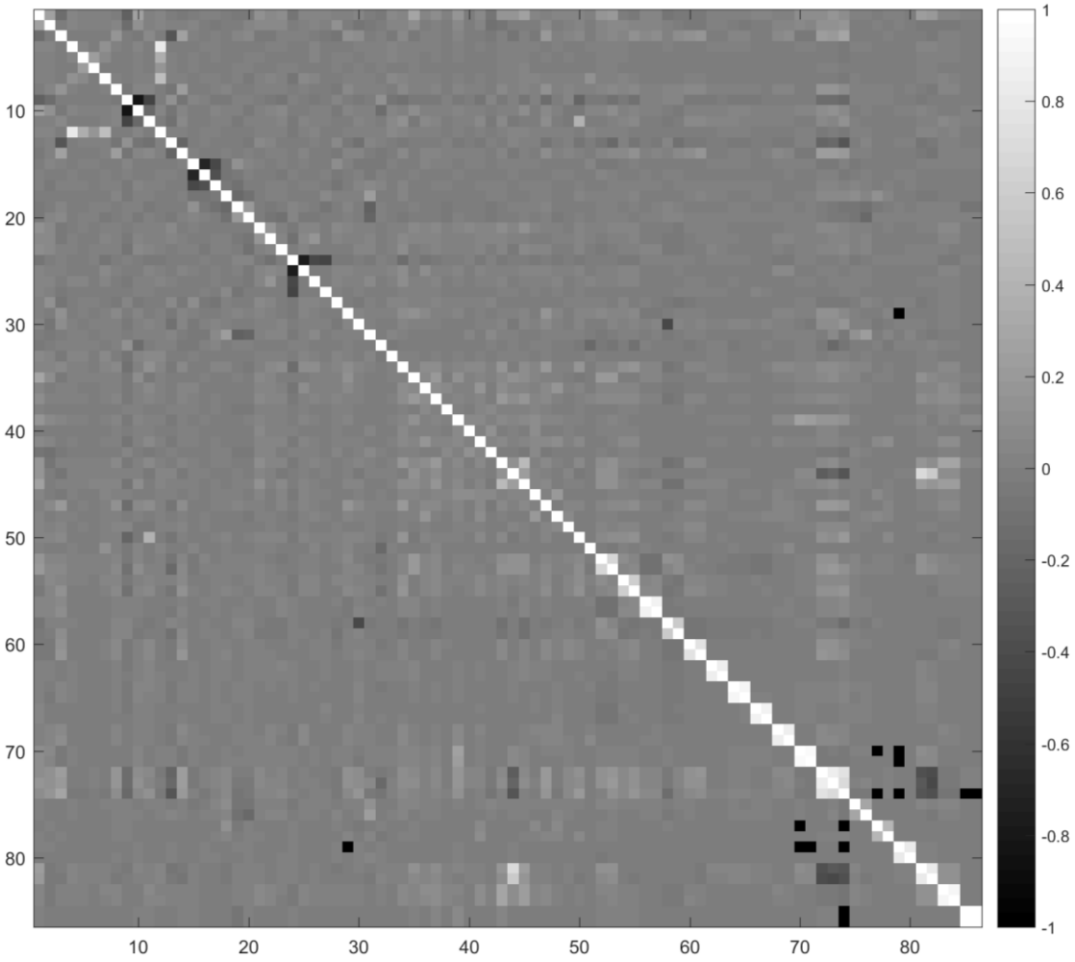


Figure 12. Pearson's matrix correlation on a greyscale visual representation.

Figure 12 presents the correlation table results in a visual representation with grey colourmap. The brighter areas represent positive correlation, as it can be observed in the diagonal, while darker areas picture negative correlation. Grey portrays areas where the correlation is close to 0.

Due to the evident correlation of the pharmacological section of the data, PCA has been applied to reduce its dimensionality. The number of variables was settled at 68.

5.3 Visualisations with Self-Organizing Maps

The data visualisations with SOM are introduced in this section. The main question that this work was trying to address is whether after the AF onset 48 hours is the safest choice as a limit to proceed with a cardioversion. Also, by observing the maps, there is also the intent to consider relationships that were not apparent through the previous analyses as well as reaffirming or refute the preceding conclusions. The presented results by no means hold any clinical validity due to the absence of medical background.

As previously mentioned, the SOM algorithm distributes the data units in regard to the concept of “similarity”. However, quite often the attributes are not presented in the same units since they describe different properties of the data. For example, in the working dataset, the parameter “age” is entered in years while “gender” is binary, representing 0 male and 1 female. This can lead to a biased outcome towards “age” since it has a larger variance, resulting in this attribute highly influencing the “similarity” of the units with respect to each other. As a result, data must be standardised prior to the SOM computation. In this work, the data has been normalised by scaling its features to range 0 – 1.

Atrial fibrillation duration visualised as hits

Due to the importance of the period between the AF onset until the cardioversion holds in this study case, the first consideration is to observe this attribute's distribution across the SOM. In the first place, in order to visualise the situation of this attribute, the hits of this parameter were coloured to represent each possible state: Red depicts "below 12 hours"; green, "between 12 and 24 hours"; blue, "between 24 and 48 hours", and yellow, "over 48 hours". The distribution of the attribute over an empty lattice can be seen in Figure 13.

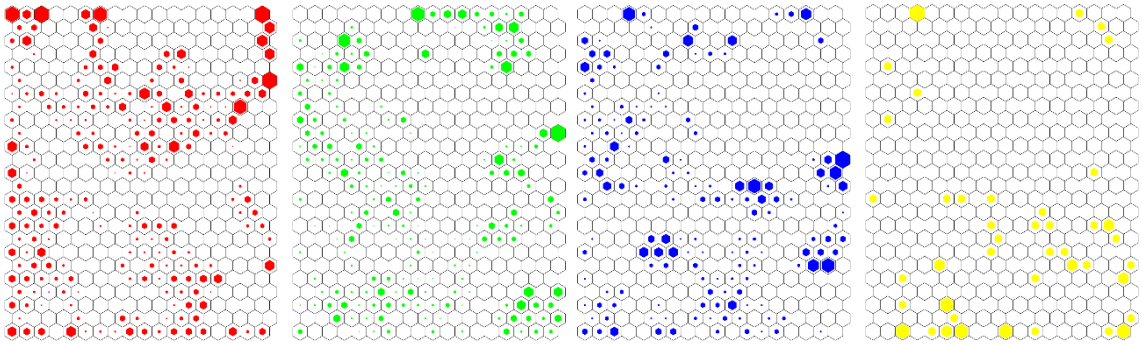


Figure 13. Data hit representation by ”AF duration”.

It can be observed at first sight that the hits are complementary across the lattice, barely overlapping, meaning the cells contain mostly one category of patients. This behaviour is explained by the nature of the SOM which utilises all entered parameters to determine the similarity between units, consequently placing them across the map. In other words, since “AF duration” has been included in the parameters for the SOM computation, it has also been used to better distinguish the clusters of patients.

Although this can be seen as an advantage, it is, in fact, considered counter-productive. As depicted in the previous figure, it is possible to highlight a particular attribute in the form of hits and further distinguish them depending on their state. Therefore, also using this information to model the SOM biases the result.

Thus, a new SOM has been computed with the same data but shifting the “AF duration” from an attribute to be the label of the units, hence reducing the number of parameters to compute from 68 to 64. The result, presented in Figure 14, does not use the target attribute to organise the SOM but instead it clusters the units by the remaining parameters.

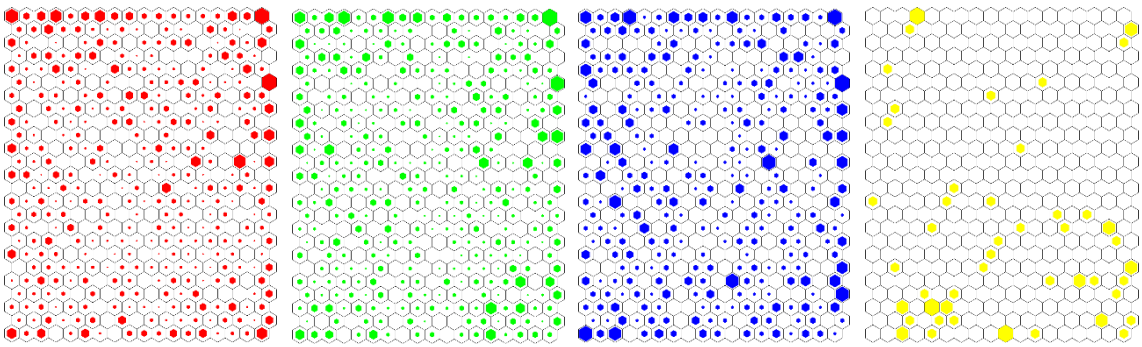


Figure 14. Data hit representation using ”AF duration” as the units’ label, excluding it from the computation.

In this second figure, it can be noticed how the units are distributed across the whole lattice overlapping with units with different labels. As previously mentioned, the total number of patients is 7,681. From these, 3,520 (45.83%) are in the red group, 2.757 (35.98%) in the green, 1.344 (17.50%) in the blue and 60 (0.78%) in the yellow. Hence, almost half of the patients fall under the first group, and only one-fifth are in the third or fourth group. The low representation of the fourth group, below 1%, is to be expected since it is not common to cardioversion a patient after 48 hours.

Moreover, another detail can be noticed. Since they are histograms, the size of the hits is proportional to the number of units portrayed in each category. Otherwise, the first and second groups would populate most of the lattice while the fourth group would be barely visible. Although some clusters can be apparently appreciated, displaying the total hits over the U-matrix will help unveil how the unit vectors have been distributed across the neurons to shape the map. The visualisation is presented in Figure 15.

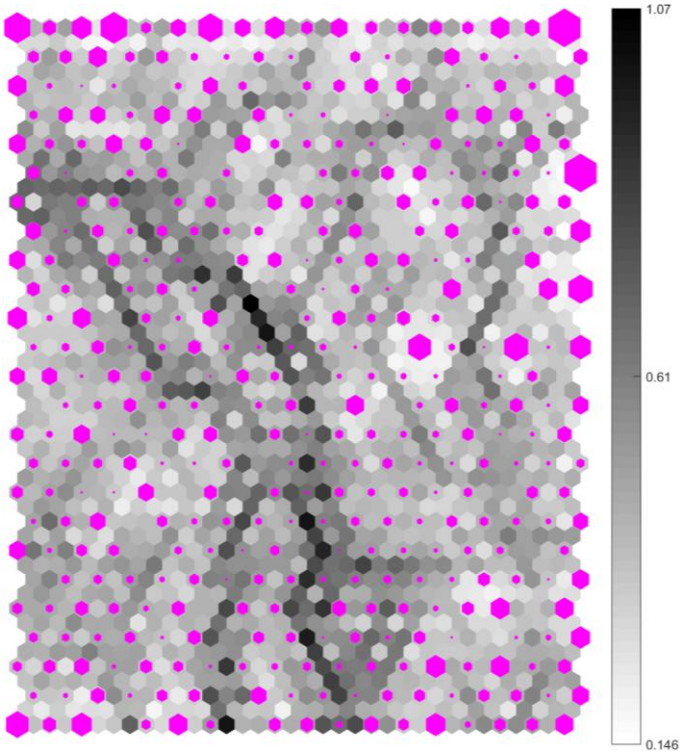


Figure 15. Data units represented as hits over the U-matrix.

Recalling the function of the U-matrix introduced in previous chapters, light coloured areas display similarity between units while darker ones convey dissimilarity. With this in mind, one clear borderline can be appreciated through the middle of the y-axis of the figure, separating the cells at the lower-left from the rest. Right beside that border, a

second one can also be observed with a similar shape. Hence, two main clusters can be deduced: The left side, at the left side of the dark strip, and the right side of the image, on the right.

By representing all the units as hits of the same colour, the relevance of the clusters is more clearly outlined by the size of each hit. In this case, it is noticeable how most of the points are adjacent to the edges and, especially, on the upper side, where both the left and the right corner present high concentrations. There are also some hits on the lower side, mostly on the lower-right. The lower-left corner also displays few large hits, which are more isolated. As it can be observed, the places where the hits are largest coincide with the lighter areas, while the cells with darker colours have smaller or no hits at all.

Component analysis

The distribution of the U-matrix varies depending on the midpoint value, or centroid, of each attribute. The centroid of a particular attribute is determined from the maximum and minimum values in the codebook matrix for that particular component. Table 4 displays the top ten components by midpoint value in both, descendant and ascendant order.

Table 4. Top ten attributes presented in descendant (left) and ascendant (right) order by codebook mid-point value.

Attribute	Value
lr FA	0.500
lr SR	0.500
NSAIDS	0.500
3 week INR	0.500
hypertension	0.500
first AF	0.499
marevan	0.495
ASA	0.495
gender	0.494
lr NA	0.481

Attribute	Value
arixtra	0.004
bleeding compl	0.004
cirrhosis	0.006
embol oth	0.008
fragmin	0.009
ECG	0.011
death	0.012
efient	0.012
stroke TIA	0.018
hd prob NaN	0.020

It is presumed that the components with higher centroid values are those that more clearly influenced the final shape of the U-matrix, or, in other words, the final disposal of

the nodes in the SOM. In order to test this hypothesis, an image that plots the U-matrix together with the components with higher centroid values is generated. The result is presented in Figure 16.

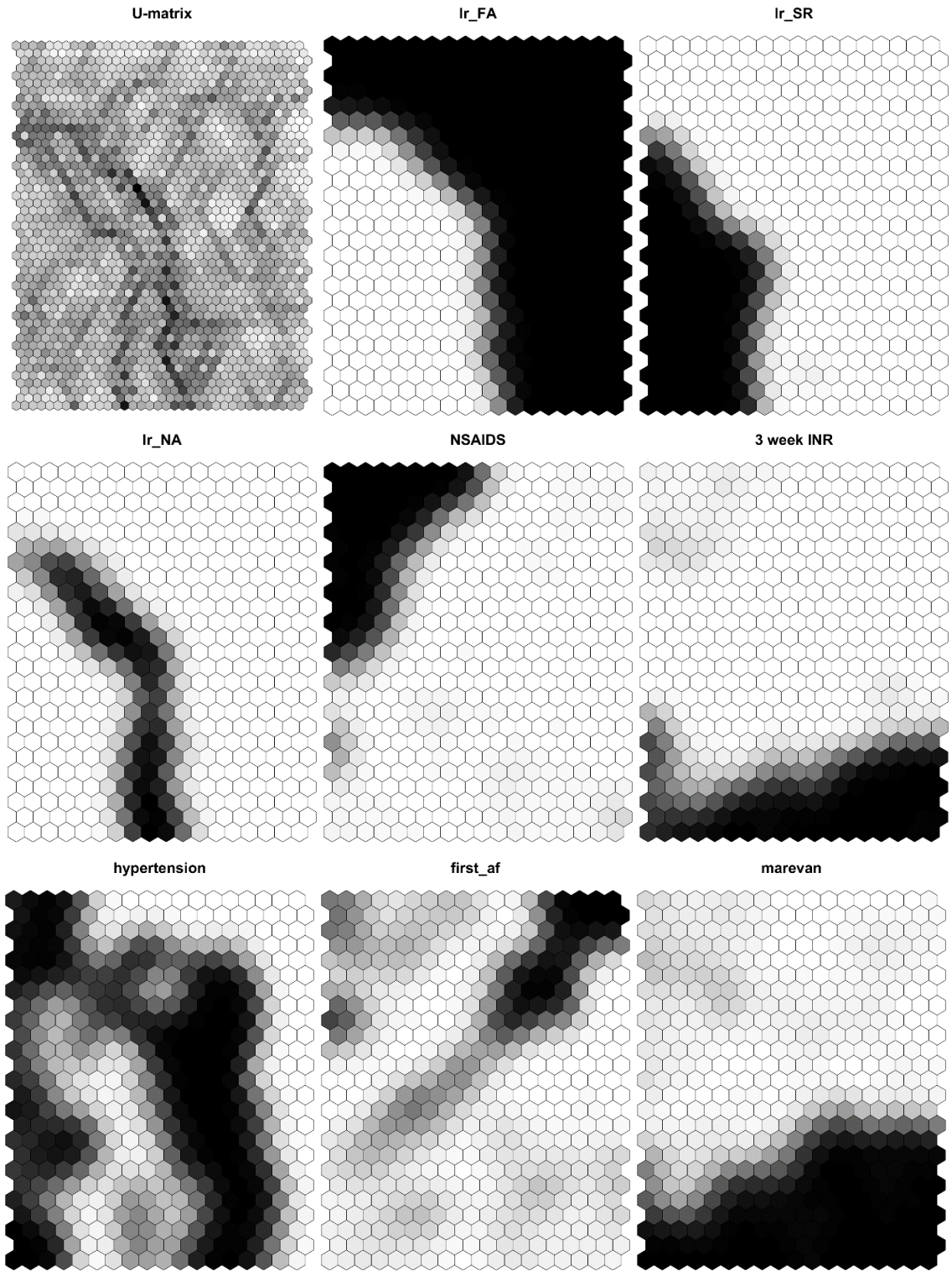


Figure 16. Component plots with the highest influence plotted with the U-matrix.

The first three components displayed represent the last recorded cardiac rhythm of the patient, which can have three states: "Atrial Fibrillation", "Sinus Rhythm" and "Not Available". Upon closer inspection, it can be deduced that having this information divided in three different attributes adds a layer of redundancy to the distribution of the nodes. Since a person cannot have two different cardiac rhythms at the same time, the fact that the last rhythm is reported as AF discards the other options. There are 715 entries which report the last rhythm as "Not Available", which his roughly 8% of the total.

Dataset rearrangement

Therefore, the data needs to be reassembled in order to avoid redundant attributes from biasing the distribution. Therefore, the previous choice of dividing some other attributes into binary attributes was not the best approach to prepare the data to use with the SOM algorithm. A new dataset is created by transforming the redundant binary attributes into one attribute with different states. Target attributes include “last rhythm”, “haemodynamic problems” and “AF duration”.

In the first instance, the attribute now explains whether the patient had the cardiac rhythm altered due to an AF onset or not. In the case of the second attribute, which was previously divided into four mutually exclusive features, have been compressed into one attribute which includes all four states. With regards to the duration, the component has been handled by first including the few cases of patients who received cardioversion after 48 hours into the previous group, converting the group into the subset of patients who received treatment over 24 hours later. After this aggregation, the remaining features have been combined into one attribute to explain them all. Data entries that were previously listed as “Not Available” have been replaced by NaN. The list of attributes with their description can be found in Appendix 1.

The top midpoint values of the attributes have been presented again in Table 5 in order to compare them with the previous table's results.

At first sight, there are many attributes that appear in both as most and least influential, such as “NSAIDS”, “3 week INR” and “hypertension” on the top or “bleeding complications”, “cirrhosis” or “other embolisms” on the bottom. However, the first and expected detail appears as the influence of the “last rhythm” parameter. In Table 4, it held the 1st, 2nd and 10th positions on the list while in Table 5 it is listed as 9th. This

helps explain the biased distribution of the previous SOM since, even though the algorithm still regards it as highly relevant, in the previous computation the evident redundancy made it hold threefold the influence when it came to dispersing the units across the map neurons while diminishing the impact of the remaining attributes, hence revealing a main strong border in the U-matrix.

Table 5. Top ten attributes after redistributing the variables in descendant (left) and ascendant (right) order.

Attribute	Value
3 week INR	0.500
NSAIDS	0.500
ASA	0.500
hypertension	0.499
first AF	0.496
beta-blocker	0.496
marevan	0.495
clexane	0.495
lr AF	0.486
digoxin	0.483

Attribute	Value
bleeding comp	0.004
ECG	0.004
embol oth	0.006
death	0.008
cirrhosis	0.009
arixtra	0.011
hd prob	0.012
labile INR	0.012
stroke TIA	0.018
TE stroke death 31d	0.020

On a closer inspection, some other observations have been noticed after reading the whole table. “Clexane” has seen a significant increase from the lower half of the table (50th) to the 8th most influential. “Digoxin” has also seen a notorious increase, not as relevant as Clexane but still noteworthy, by going up from the 27th to the 10th position. Clexane, generic name enoxaparin, is an OAC drug while digoxin is an antiarrhythmic used to help make the heartbeats more strong and regular (Drugs.com 2017) (Drugs.com 2015). Some other medications have also seen an increment in relevance after this transformation.

The U-matrix with the highest scoring components obtained after rearranging the dataset is presented in Figure 17. Comparing the results with the image from the previous dataset, it can be observed how, in the current figure, the attributes with higher mid-point value equally reflect upon the distribution of the neurons through the map in contrast to the former, where the three redundant attributes intensified two particular borderlines while minimising the other features' influence on the distribution. As a result, the overall

lattice is coloured with a darker shade of grey, which leads to thinking the components that are regarded as mildly relevant now have more impact than in the previous dataset.

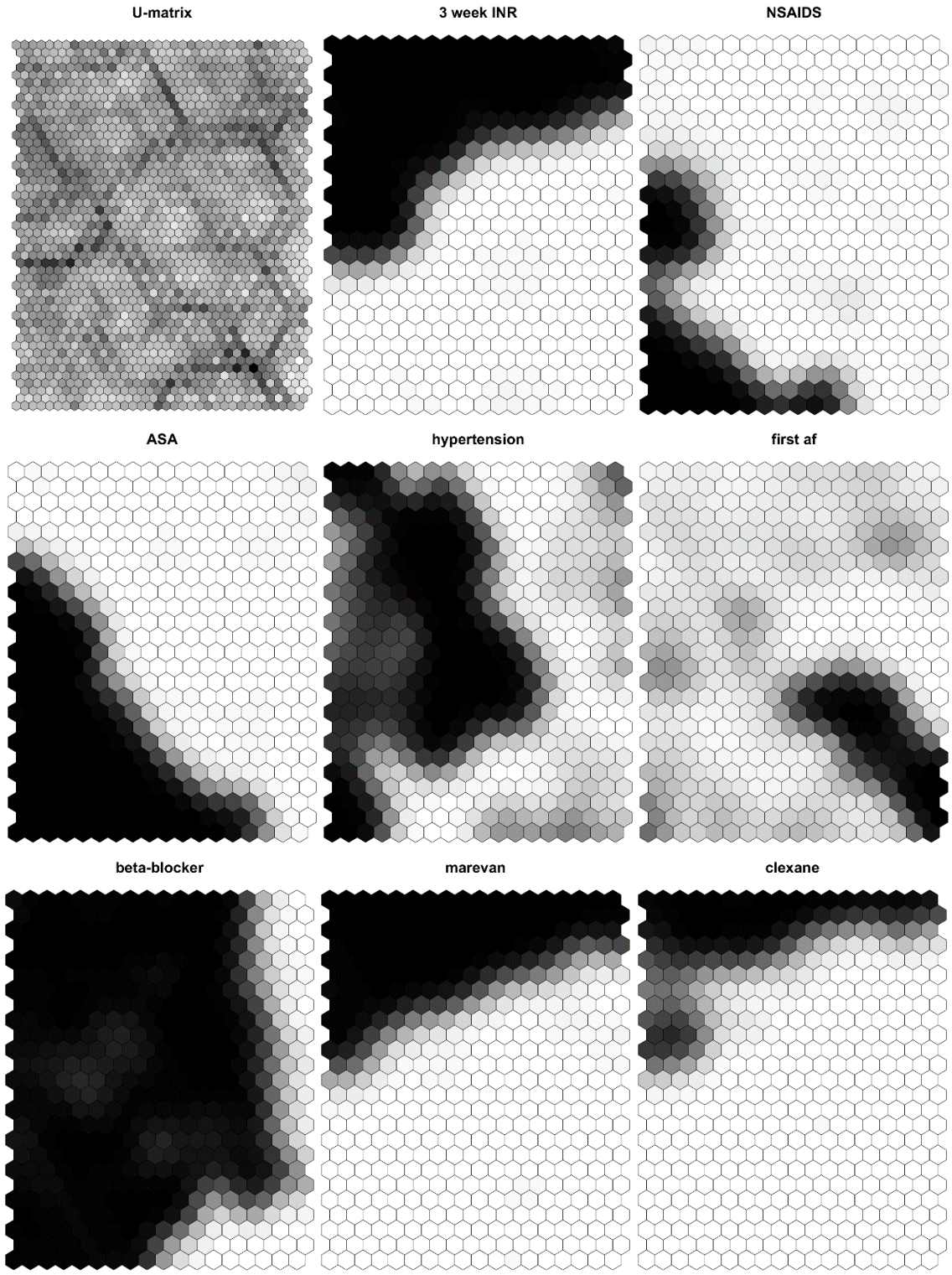


Figure 17. U-matrix with the top eight most relevant attributes.

Visual inspection of component planes

After confirming that the data rearrangement proved relevant, the component planes were plotted over again. The aim was to demonstrate that it is possible to derive consistent conclusions by observing the distribution of the units in the SOM. Subsequently, some hypotheses are presented, which are based only on speculation.

The first subset of images is presented in Figure 18. It can be seen how there is a correlation between the duration of the arrhythmia "AF duration", whether this duration is confirmed "duration sure" and whether the patients were aware of the onset of the episode "AF aware". The component informing whether it was the first registered arrhythmia "first AF" is included in the fourth plot.

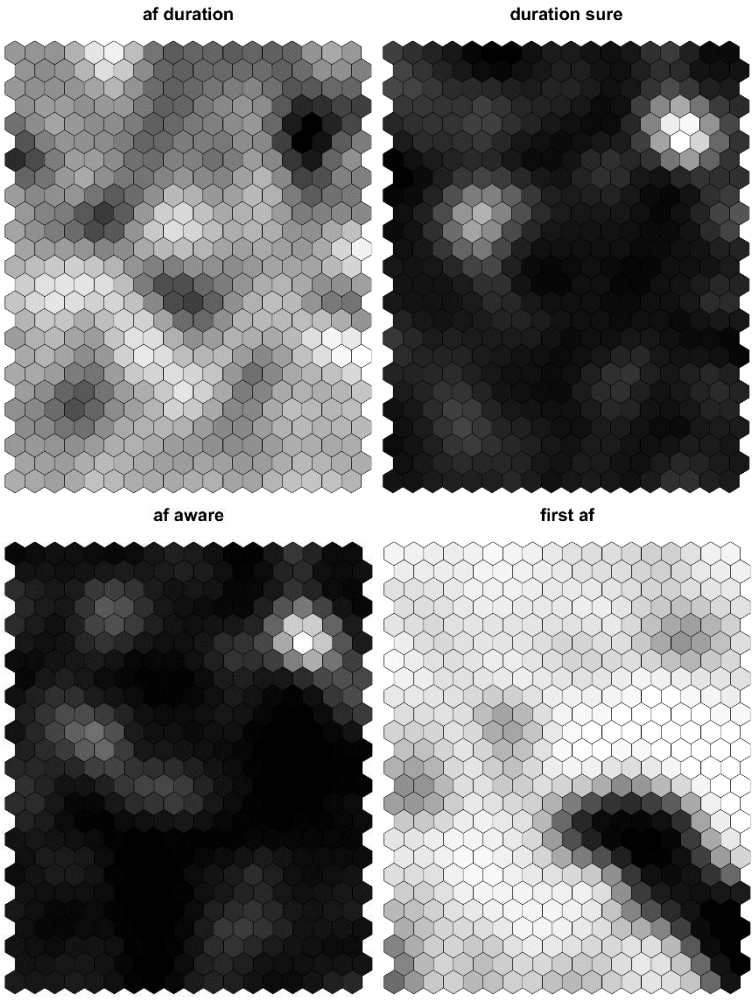


Figure 18. Comparison between arrhythmia's duration, awareness and prevalence.

As displayed in the figure, two main defined clusters can be distinguished. The first is found close to the right side, on the fifth row, while the second, around the left, on the tenth row. These clusters could be interpreted as containing patients who were not aware of the AF, and thus the onset could not be clearly established. Due to the safety risks involved with the uncertainty, the doctor categorised them into the group "over 24 hours since the onset of the arrhythmia".

It can be seen in the second subplot of the second row how there is a shade of grey in the same spot for the patients who registered the first episode. Therefore, most of them were not likely first-timers.

The second group of images, displayed in Figure 19, comprises the attributes regarding the relapse of the patients "AF relapse" and the patients who suffered an arrhythmia during the last 30 days "AF last month".

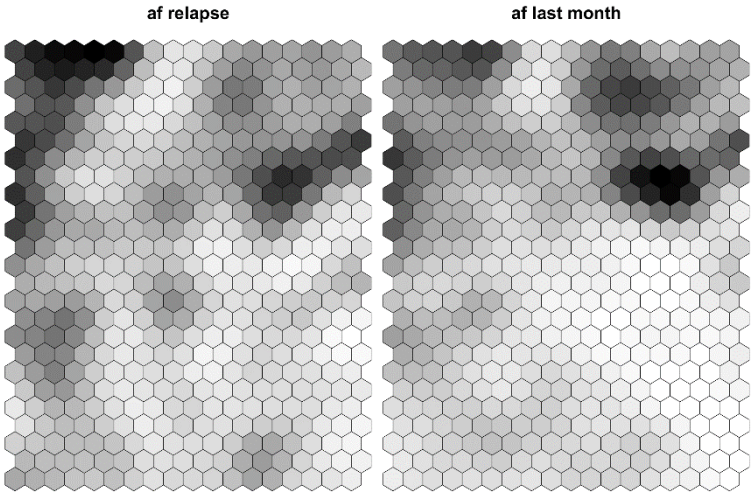


Figure 19. Comparison between relapsing rate and incidence during last 30 days.

It can be observed how the two main clusters represented in "AF relapse" are visible in the monthly component, albeit the cluster located at the right side, around the eight row, is more clearly defined in the later. The consideration at this point is what makes this particular cluster different from the other. So the rest of the components will be consulted in an attempt to find commonalities.

About this group, some inferred characteristics include an area comprised mostly of male patients, of an age presumably younger than the average, with multiple AF registered. There are no common clusters with other conditions or symptoms reported during current

diagnosis, and the "hypertension" component, relatively prevalent overall, displays a light grey shade, hence most of the patients in this area do not suffer from this condition either. The attribute "bleeding diathesis" portrays four clusters, and one of these matches the aforementioned attributes. Still, this is a feature that did not hold a particular impact on the placement of the nodes and its density isn't remarkable either, with roughly 1.69% of the patients reporting it. Therefore, the clusters delimited by "bleeding diathesis" attribute are likely to be low populated.

Complementing the previous observations, it is noticed that this area is comprised of the border between areas in other significant attributes, such as "beta-blocker" or "digoxin". Therefore, the conclusion is that this subgroup found at the right side of both "AF relapse" and "AF last month" is a scarcely populated area that has some common features, such as male patients of a younger age, suffering from bleeding diathesis and without hypertension.

The last example consists of the components linked to the kind of cardioversion received "cardioversion" and two antiarrhythmic medications, "flecainide" and "amiodarone".

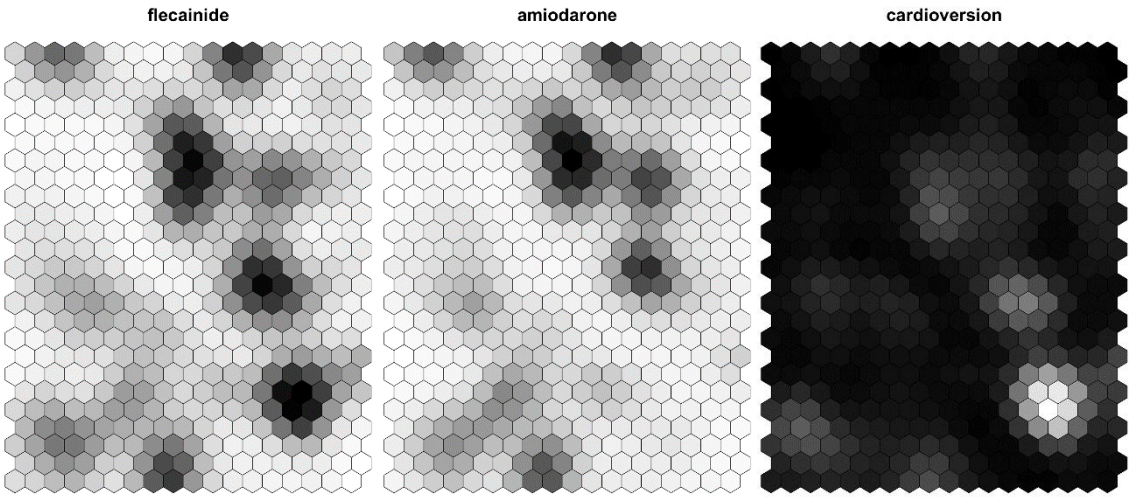


Figure 20. Comparison between two antiarrhythmic drugs and cardioversion type.

The first noteworthy consideration is that both, "flecainide" and "amiodarone", display a considerably similar distribution. When compared to "cardioversion", it can be discerned how the clusters of the first two components mostly reflect upon the last. This coincidence strengthens the previous hypothesis derived from the statistical correlation implementation that flecainide is a drug used for pharmacological cardioversion. Given the similitude with the first drug, "amiodarone" could be a medication which is

administrated to patients who display a similar clinical profile to the ones who take flecainide, maybe also serving to perform chemical cardioversion. ”Flecainide” and ”amiodarone” have in common all but one cluster, located on the right side of the ”flecainide” component, around the fifth row counting from below, which can also be distinguished in the "cardioversion" component.

This cluster seems to be formed mostly by non-elderly male patients that report the first AF and don't suffer from hypertension. Alcoholism and cirrhosis are marginally present around the same area. When comparing it to cardioversion, it can be noticed how the whitest cluster corresponds to this particular group of nodes. Therefore, the interpretation is that most of the patients who are treated with pharmacological cardioversion are located in this cluster, while the rest of the light areas in the "cardioversion" component are less populated in comparison. Hence, the patient profile that receives flecainide is mostly uniform, while amiodarone is a secondary drug administered to a smaller amount of patients.

Other clear similarities have been found across the 60 components of the dataset. These include, for example, the connection between patients who suffer from diabetes mellitus and have had a heart failure episode or the relationship between those who reported a stable INR value during three weeks and also took Marevan. Given the scope of this work, the current considerations are regarded as sufficient. The aforementioned components can be found in Appendix 2.

5.4 Further considerations

Up to this point, the work has been focused on the explorative nature of the SOM rather than its predictive ability. A proposal to explore this theory is explained. By using the knowledge behind the introduced medical scores, CHA_2DS_2-VASc and HAS-BLED, displaying the distribution of the patients in regard to their parameters should be feasible. By portraying a U-matrix that only depicts the similarity of the nodes based on the reduced subset of components that explain the scores, the resulting figure should display the units divided into clusters representing these scores.

A list of few components from the current dataset that could be considered for each score can be found in Table 6.

Table 6. Possible components related to medical score estimation.

CHA_2DS_2-VASc	**HAS-BLED**
coronary heart disease	hypertension
age	age
gender	bleeding diathesis
heart failure	labile INR
previous MI	renal failure
…	…

Unfortunately, two main constraints were found; First, the lack of an attribute to verify the validity of the clusters displayed in the U-matrix by these components. If the dataset contained an additional attribute explaining the CHA_2DS_2-VASc or the HAS-BLED value registered to each patient by the doctor, a fair comparison could be carried out. A solution could be to derive the score value for each patient entry, but the validity would be less than ideal and could also entail redundancy issues.

Secondly, although the parameters of the medical scores are concisely detailed, in the current dataset they might be explained by a variety of attributes, some of them not even considered. Also, some parameters hold more relevance than others at delimiting the associated risks that the medical score explains. Therefore, in order to proceed with this approach and better reflect the outcome, a transformation of the dataset would be suggested.

All in all, this is regarded as a hypothesis that could be further developed under other circumstances, but for the associated drawbacks and considering the scope of this work, it is not deemed to be necessary.

5.5 Discussion

Lastly, the conclusions derived from the SOM implementation for the visualisation of medical data are presented in this section.

Advantages

The first advantage is the ability of the images to represent high-dimensional data in a low-dimensional space. That is because each node of the lattice has a weight which is assigned to represent how similar it is to the neighbouring nodes. This weight is assigned by iteration after learning from the data itself.

Another advantage is that in order to interpret the visual representations there is no need to comprehend the algorithm that generates them. In other words, any person can easily understand how to read the images and draw conclusions out of them by referring to the colour shades. It has been easy to present the results to a wider audience who had no knowledge of the underlying nature of the SOM.

Limitations

On the other hand, every SOM is different and can display different relationships between the vectors by only seeing trivial rearrangements in the data. Needing to modify the data to better reflect the situation resulted in completely different results and thus, presenting them proved challenging.

Also, the data must be prepared in order to avoid redundancy of the attributes. Since the SOM uses similarity to distribute the units across the nodes, the presence of redundant attributes can easily bias the computation. Not considering the real connections the data features shared outside of the dataset led to requiring modifications.

Finally, since the SOM relies on distance, data transformations must be handled with precaution, such as applying dimensionality reduction techniques. NaN values, although computable with the SOM Toolbox, should also be avoided when possible. The transformations carried out at some points in the work have later proved counter-productive.

6 CONCLUSION

The aim of using computers with ML algorithms is to draw improved conclusions out of the data although not every dataset fits every learning model. For that reason, it is of utmost importance to have the right knowledge about the context surrounding the data as well as the available methods that best suit each situation. Since the beginning, this work has taken into consideration the advantages, as well as the drawbacks, associated with the chosen model, the SOM.

Initially, TYKS proposed questions to answer regarding the data but it has proven difficult to comply. The reasons could be due to the methods proposed not being suitable enough for the data or the lack of experience in optimally implementing them to yield the most convenient outcome. Nevertheless, in the presentation of results, the method was considered as a potential tool to explore the medical data from an alternative point-of-view. The author believes the SOM proffer a highly valuable asset in presenting the connections the data has to offer in a neat and informative approach that any users can become acquainted with ease.

Finally, working hands-on real-world data has helped me confront a variety of situations from diverse nature. Given other circumstances, the author would have unlikely had the chance to address and hence learn from them. This experience ranges from learning about the context of the data, the diverse implementations performed and the mistakes associated with the information retrieval.

The author believes this work could be expanded by contrasting it with other ML methods or performing improved implementations of the SOM. Overall, the knowledge gained from working on this thesis is invaluable and therefore, has served the main purpose for which was proposed.

REFERENCES

Anderson, E. 1935. The irises of the Gaspe Peninsula. Bulletin of the American Iris Society (Volume: 59, 2 – 5)

Berthold, M. 2010. Guide to intelligent data analysis. London, Springer.

Bishop, C. 2006. Pattern Recognition and Machine Learning. New York, Springer.

Chedzoy, O. B. 2006. Phi-coefficient. Encyclopedia of Statistical Sciences. John Willey & Sons, Inc.

Drugs.com. August 2015. Digoxin Uses, Dosage & Side Effects. Consulted on 02.05.2017 https://www.drugs.com/digoxin.html

Drugs.com. May 2017. Enoxaparin medical facts. Consulted on 02.05.2017 https://www.drugs.com/cdi/enoxaparin.html

Drugs.com. May 2017. Flecainide: Indications, Side Effects, Warnings. Consulted on 02.05.2017 https://www.drugs.com/cdi/flecainide.html

Fisher, R. 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics (Volume: 7, Part II, 179 – 188)

Kohonen, T. 1981. Construction of similarity diagrams for phonemes by a self-organizing algorithm. Report TKK-F-A463, Helsinki University of Technology, Espoo, Finland

Kohonen, T. 1990. The self organizing map. Proceedings of the IEEE (Volume: 78, Issue: 9)

Kohonen T. 2001. Self-Organizing Maps. 3rd edition. Berlin Heidelberg, Springer

Kohonen, T. 2013. Essentials of the self-organizing map. Elsevier, Journal of Neural Networks (Volume: 37, Issue: January 2013 52 – 65)

Kohonen, T. 2014. MATLAB Implementations and Applications of the Self-Organizing Map. Unigrafia Oy, Helsinki, Finland

Lichman, M. 2013. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. Consulted 29.08.2016 https://archive.ics.uci.edu/ml/datasets/Iris

Lip, GY et al. 2010. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest Journal (Volume: 137, Issue: February (2) 263 – 272)

Mitchell, T. 1997. Machine Learning. McGraw-Hill, Inc. New York, NY, USA

National Heart, Lung and Blood Institute. September 2014. What Is Atrial Fibrillation? Consulted on 09.2016 https://www.nhlbi.nih.gov/health/health-topics/topics/af

National Heart, Lung and Blood Institute. November 2011. Your Heart's Electricla System. Consulted on 09.05.2016 https://www.nhlbi.nih.gov/health/health-topics/topics/hhw/electrical

Nuotio I. et al. 2014. Time to Cardioversion for Acute Atrial Fibrillation and Thromboembolic Complications. JAMA (Volume: 312, Issue: August (6) 647 – 649)

Odum, LE et al. 2012. The $CHADS_2$ versus the new CHA2DS2-VASc scoring systems for guiding antithrombotic treatment of patients with atrial fibrillation: review of the literature and recommendations for use. Pharmacotherapy (Volume: 32, Issue: 3 (March) 285 – 296)

Pisters, R et al. 2010. A novel user-friendly score (HAS-BLED) to assess 1-year risk of major bleeding in patients with atrial fibrillation: the Euro Heart Survey. Chest Journal (Volume: 138, Issue: 5 (November) 1093 – 1100)

Romanazzi, M. February 2014. Componenti Principali. Consulted on 10.10.2016 http://venus.unive.it/romanaz/edami/disp_pca.pdf

Samuel, A. L. 1959. Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development (Volume: 3, Issue: 3)

Silipo, R. et al. 2014. Seven Techniques for Dimensionality Reduction. KNIME, Open for Innovation

Ultsch, A. and Siemon, H. P. 1990. Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. Widrow, Bernard; Angeniol, Bernard. Proceedings of the International Neural Network Conference (INNC-90), Paris, France, July 9–13, 1990. 1. Dordrecht, Netherlands: Kluwer. pp. 305–308. ISBN 978-0-7923-0831-7.

Vatanen et al. 2015. Self-organization and missing values in SOM and GTM. Elsivier, Neurocomputing (Volume: 147, Issue: 5 January 60 – 70).

Vesanto, J. et al. 1999. Self-Organizing Map in MATLAB: the SOM Toolbox. In Proceedings of the Matlab DSP Conference. November 16 – 17, Espoo, Finland

APPENDIX 1. Table of the attributes with description

Table 7. List of transformed dataset attributes

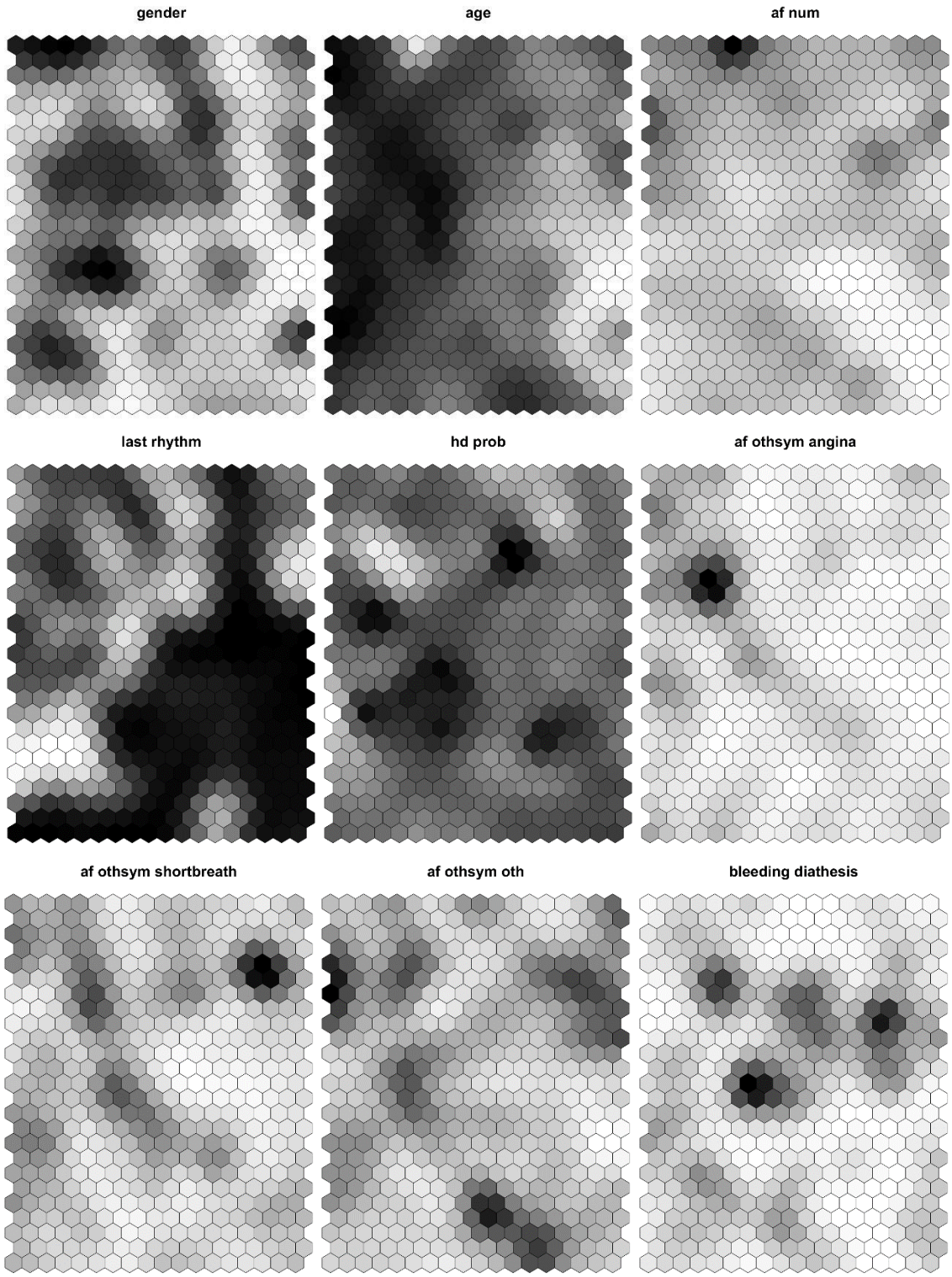
Attribute	Description
age	Age at the moment of admission.
gender	Female gender (0 male, 1 female).
AF num	Number of AF arrhythmias registered.
stroke TIA	Stroke or TIA at the moment of admission.
embol oth	Other embolism at the moment of admission.
bleeding comp	Bleeding complications at the moment of admission.
death	Death at the moment of admission.
AF relapse	Current arrhythmia is a relapse.
last rhythm	Registered cardiac rhythm at the moment of admission.
TE stroke death 31d	TE, stroke or death during the last month.
first AF	First registered AF.
AF last month	Suffered an AF during the last month.
AF duration	Duration since the onset of the current AF. Less than 12 hours, 12 to 24 hours or more than 24 hours.
duration sure	Doctor is sure of the chosen “AF duration” range.
AF aware	Patient was aware of the AF arrhythmia.
AF othsym angina	Angina symptoms presented with AF.
AF othsym shortbreath	Short breath symptoms presented with AF.
AF othsym oth	Other symptomps presented with AF.
hd prob	Haemodynamic problems (-1 hypotension, 0 none, 1 pulmonary oedema).
ECG	Electrocardiogram value.
LVH	Left-Ventricular Hypertrophy condition present.
cardioversion	Cardioversion type (0 electrical, 1 pharmacological).
esophagus us	Transoesophageal echocardiogram.
cv managed	Cardioversion restored sinus rhythm.
acute compl	Acute complications during cardioversion.
heart failure	Heart failure registered prior to cardioversion.
hypertension	Hypertension condition.
renal failure	Renal failure registered prior to cardioversion.
DM	Diabetes Mellitus present.

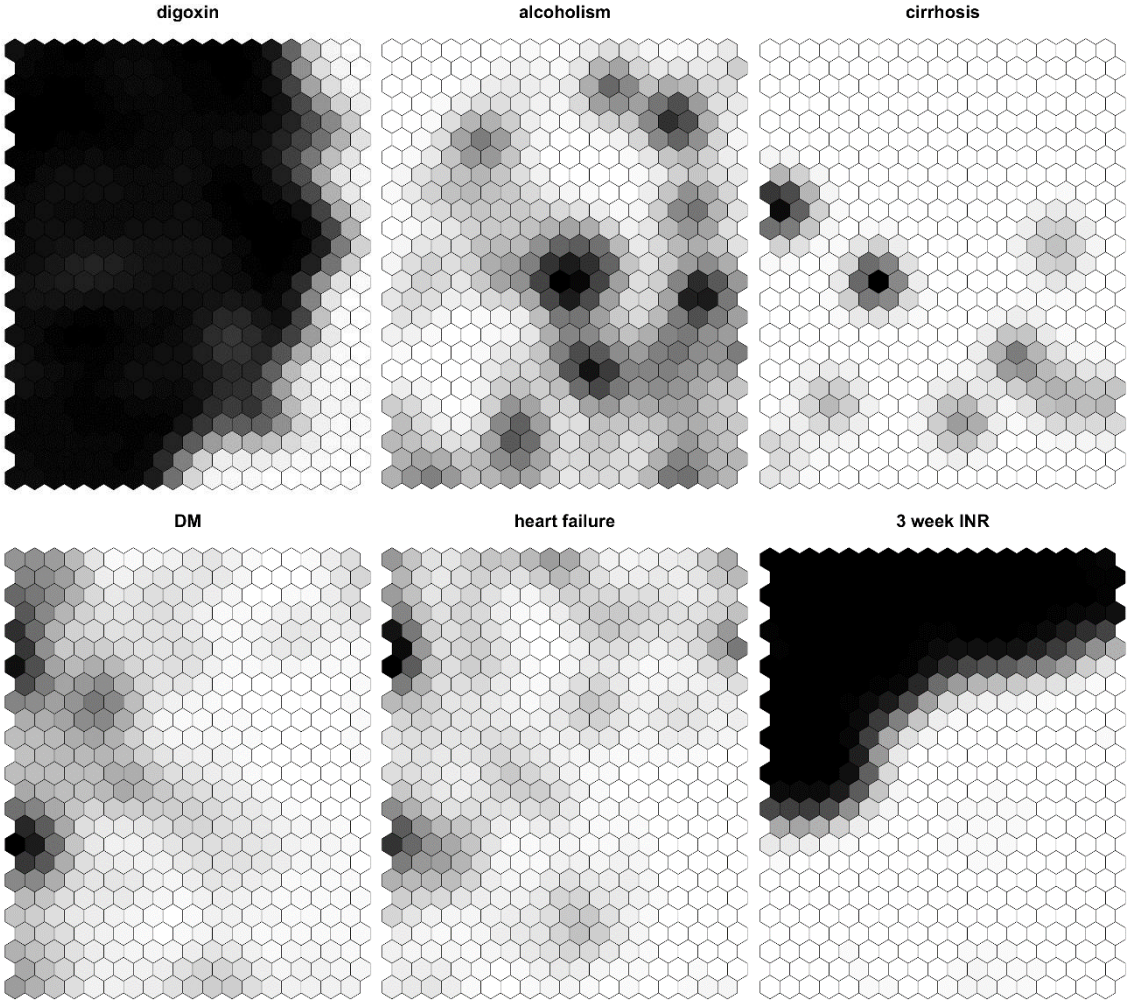
(continue)

Table 7 (continue).

Attribute	Description
cirrhosis	Cirrhosis present.
paralysis TIA	Paralysis or TIA prior to cardioversion.
labile INR	Unstable or high INR value.
oth embol diagn	Other embolism diagnosed.
alcoholism	Alcoholism.
previous MI	Previous Myocardial Infarction
NSAIDS	Nonsteroidal Anti-Inflammatory medications.
oth coronary artery disease	Other coronary artery diseases.
prev GI bleed	Previous gastrointestinal bleeding.
oth vascular disease	Other vascular disease.
bleeding diathesis	Bleeding diathesis condition present.
coagulopathy	Coagulopathy condition present.
pacemaker	Pacemaker user.
beta-blocker	Beta-blocker medication prescribed.
digoxin	Digoxin medication prescribed.
verapamil	Verapamil medication prescribed.
flecainide	Flecainide medication prescribed.
amiodarone	Amiodarone medication prescribed.
sotalol	Sotalol medication prescribed.
propafenone	Propafenone medication prescribed.
quindine disopyramide	Quindine disopyramide medication prescribed.
dronedarone	Dronedarone medication prescribed.
dipyridamole	Dipyridamole medication prescribed.
marevan	Marevan medication prescribed.
3 week INR	3 weeks in stable INR range.
clexane	Clexane medication prescribed.
fragmin	Fragmin medication prescribed.
arixtra	Arixtra medication prescribed.
ASA	Acid Acetylsalicylic medication prescribed.
clopidogrel	Clopidogrel medication prescribed.
efient	Efient medication prescribed.

Appendix 2. Additional component visualisations




digoxin
alcoholism
cirrhosis
DM
heart failure
3 week INR