Master Research Internship

Master Thesis

Interpretation of laboratory results and elaboration of a clinical diagnosis in blood coagulation domain

Domain: Clinical Decision Support - Artificial Intelligence - Data Mining - Classification

Author:
Alban Delamarre

Supervisor:
Cédric Buche
IHSEV - Lab-STICC - France

Academic years 2015 - 2016

Abstract

During a haemorrhage, a phenomenon called coagulation stops the blood effusion. It can be measured with coagulation curves known as thrombogram. In some individuals affected with haemophilia this phenomenon is disturbed. If some precautions are not taken their life could be put in danger while prescribing a treatment or performing a surgery. Thus it is important for clinicians to establish an accurate diagnosis. However, providing such a diagnosis is a complicated task as a lot of parameters have to be taken into account. The goal of this study is the realisation of a clinical decision support system that would help health professionals during the process of decision making. Machine learning techniques are used to classify thrombograms and therefore, to detect haemophilia, its type, A or B, and its severity. This paper presents a comparison of different classification methods and the selection criteria identified to choose the best method. On a dataset of 14 000 artificially generated thrombograms, results show 98% recognition for haemophilia and 93% recognition for the different types and severities of haemophilia. It proves that our system is a viable solution and that machine learning techniques can efficiently detect haemophilia and provide a reliable complete diagnosis.

Contents

1 Introduction

Coagulation is the phenomenon that takes place during a haemorrhage, a blood clot forms and stops the flow of blood. Some pathologies can disturb this phenomenon, like haemophilia. Prescribing a treatment that is incompatible with these kind of diseases could be dangerous for the patient's life. Moreover, before a surgery, a clinician must be totally aware of the clinical history of his patient to avoid any risks. That's why the diagnosis of this kind of pathology must be established without any errors. However, establishing such diagnosis is really complicated, clinicians encounter difficulties identifying the type of illness as there are a lots of parameters to take into account.

Thrombograms are temporal series that represent the evolution of the rate of thrombin during a haemorrhage. In other words, these curves are a measure of coagulation. Analysis of these kind of curves is a difficult task, as they take into account many parameters. We propose to undertake a clinical decision support based on machine learning approach to detect the type and the severity of haemophilia using thrombograms.

Based on a database composed of artificially generated thrombograms, we have studied the ability of machine learning techniques to learn and classify different natures of coagulation curves. As artificial intelligence methods had never been used on these kind of curves, various were tested to discover the more effective ones.

Given the clinical context, evaluation classification must fulfil some requirements, quantity of data for example is an important factor in this domain where data collection is under various constraints. This paper also presents different ways to evaluate the classification of the different methods. As shown by the results of our experiments, our system is a viable solution to help clinicians analysing thrombograms.

1.1 Problem specification

As analysing coagulation curves is a very complicated task and can be source of serious consequences if badly interpreted, therefore the objective is to create a clinical decision support which, given a thrombogram as an input, returns whether or not this curves belongs to a healthy patient or a haemophiliac patient. And if so, the type and severity of haemophilia this patient has.

Given a space $\mathcal{V}$ of unlabelled data and $\mathcal{Y}$ a finite set of labels, we have $\mathcal{X} = \mathcal{V} \times \mathcal{Y}$ the space of labelled samples. Let $\mathcal{D} = \{x_1, x_2, ..., x_n\}$ be a dataset composed of n labelled instances, where $x_i = \langle v_i \in \mathcal{V}, y_i \in \mathcal{Y} \rangle$ and v_i a vector representing a time series of length m such as $v_i = \{t_1, t_2, ..., t_m\}$. The objective is to find the best classifier $\mathcal{C}$ which for a given time series v associates a label y such as $\mathcal{C}(v) = y$ with $\langle v, y \rangle \in \mathcal{X}$.

1.2 Overview of the approach

The achievement of the clinical decision support depends on the ability of machine learning techniques to accurately classify thrombograms. The different steps of our approach are briefly described below:

- **1**. First of all we need to constitute a dataset with the different categories of haemophilia we want our clinical decision support system to detect. We must generate a sufficient amount of thrombograms to be able to perform an efficient training for each classification method that will be tested.
- **2**. After dataset creation, to obtain the most accurate results, we have to find the best configuration for each method and for each classification that can be perform on the different categories. This is done by identifying parameters which have the most influence and determine their values for each classification. A grid search is used for this purpose.

- **3**. Once we have the parameters, we can isolate the best method for the different classifications depending on the evaluation criteria established according to the medical context: quantity of data, precision, recall, we can determine the most valuable classification technique.
- **4**. Since some methods perform better on a specific classification, we developed a cascade classification technique. At each step, a classifier is used to separate the overall dataset into two sub-datasets according to the different classes. The different types of haemophilia are identified step by step. The advantage of a cascade, is that we can use a specific classifier for the categories we want to analyse. Results obtained using this technique are then compared to the ones of single methods.

The paper is organized as follows. Section 2 presents a state of the art of classification techniques and summary of the different types of clinical decision support systems. It also describes thrombograms and how they permit detection of blood illness like haemophilia. Section 3 explains in detail the approach set up in our context. In Section 4 evaluation criteria are identified and described according to the context, classification results are then presented. Section 5 contains an analysis of the results obtained. We finally discuss the issues of this work and the directions for future work in section 6.

2 State of the art

This section presents a state of the art of clinical decision support system, classification method and extraction technique that will be used to identify thrombograms nature.

2.1 Thrombograms

During a haemorrhage, a phenomenon called coagulation can be observed. The blood fluidity is regulated by activators and inhibitors to form a blood clot and stop the flow. However, in some pathologies, like haemophilia, there is a deficit of these factors, thus it influences coagulation.

Thrombin is one of the key factors that will determine the blood coagulability. Its concentration can be measured by calibrated automated thrombography [10]. In [31], Young and al. shows that this measurement process provides coagulation curves which permit to identify haemophilia. Figure 1 shows a sample of diverse coagulation curves. We can clearly notice the difference between each curve: The one with the highest peak represents the healthy patient, others are from haemophiliac patients, and the most severe haemophilia have the lowest peak. Thus, as explained by Young and al. it is possible indeed to determine which curves represent haemophiliac patients.

But, figure 2 on the contrary, shows 5 undistinguishable thrombograms. In this figure, curves with the highest peaks describe the coagulation of three different pathologies: a healthy patient, an A mild haemophiliac and a B mild haemophiliac. Curves with the two lowest one shows different types of severe haemophilia (A and B). As we can see, identifying the true nature of thrombograms becomes a complicated task. Even experts encounter difficulties facing those kind of curves. That's why we propose to develop a clinical decision support system.

2.2 Clinical Decision Support

In the clinical domain, diagnosis mistakes can have disastrous consequences. Clinical Decision Support, CDS, is a system which brings advice to clinicians during the process of decision making and thus reduce diagnostic error rates. In a given context, a CDS links observations with knowledge and match them to the patient characteristics to influence the health professional. The aim of the CDS is certainly not to make a decision, but only to provide suggestions that the clinician can use to produce a diagnosis. There are two types of CDS:

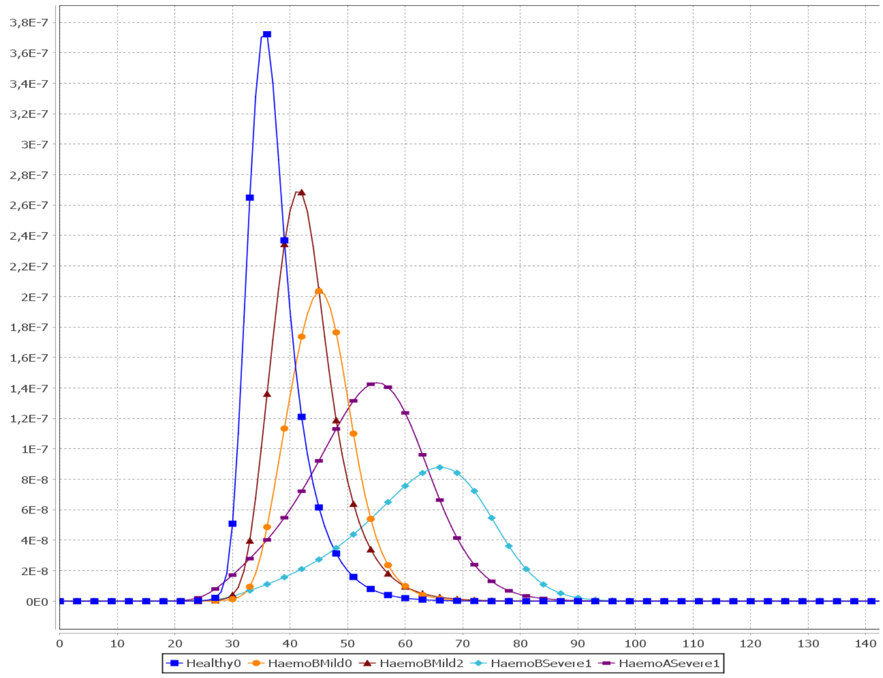

Figure 1: Samples of distinguishable thrombogram types

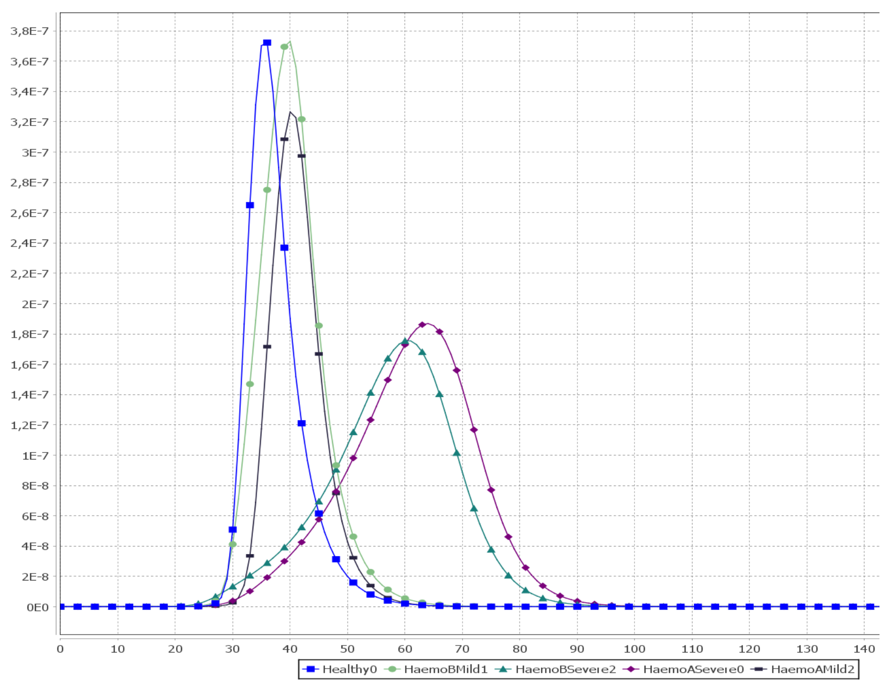

Figure 2: Samples of undistinguishable thrombogram types

- Knowledge-based
- Non-knowledge-based

Knowledge-based CDS is composed of rules and associations established by experts, an inference engine and an interface. The interface permits to set the inputs and obesrve results. The inference engine associates inputs, patients' data, with rules and associations to produce a diagnosis suggestion. This type of CDS needs an expert knowledge of the context. In MYCIN [26] expert knowledge are represented as *If-Then* rules, Seixas and al.[24] created a CDS to detect Alzheimer and dementia based on a Bayesian Network modelled using expert knowledge. A huge variety of knowledge-based CDS exists and tools to developp them as well. The following are semantic-based languages for the development and the execution of clinical guidelines: Arden Syntax[22], GLIF3 [3], PROforma [29]. MET3 [23] is a multi agent system that supports the entire decision making process.

For non-knowledge-based CDS, no knowledge expert or inference rules are needed as machine learning techniques learn from a clinical dataset and associate patterns found in the data to pathologies. Shin and al.[25] used a non-knowledge-based CDS to detect cancer using mass spectrometry. A study of neonatal sepsis prediction [15] also use machine learning in a clinical decision support.

Clinicians encounter difficulties to correctly analyse thrombograms, therefore CDS is a viable solution for this problem. As there are no rules to identify thrombograms natures we will develop a non-knowledge-based CDS. As shown in figure 3, given a dataset, machine learning algorithms for classification will be trained to analyse and find discriminative patterns in coagulation curves to accurately classify them and provide a diagnosis suggestion.

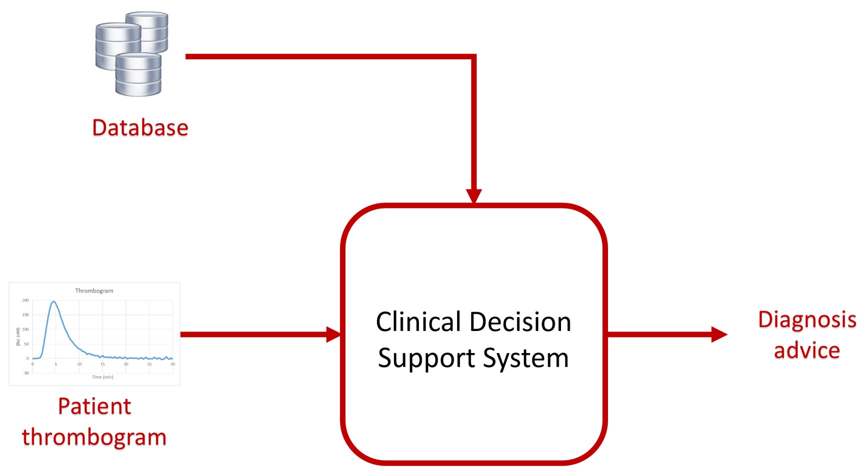

Figure 3: Clinical Decision Support

2.3 Classification Methods

To realize the most effective CDS we will compare various machine learning classification techniques and observe how accurate they are with thrombograms. These techniques are presented in this section.

2.3.1 KNN - K-Nearest Neighbours

The k-nearest neighbours method establish the label of an unlabelled sample datum by comparing it with a set of k data from an established dataset. This set is settle by choosing the "closest" neighbours of this sample data and provide a class suggestion based on their labels. Therefore this technique needs two parameters: k the number of neighbours and the distance, or similarity, measure between two samples. KNN has been used in various domains and also with times series [6]. To establish the distance between two samples it calculates different distance functions like the Euclidean distance or the Mahalanobis distance.

Applying KNN to thrombograms may have long computation time. For each sample, the distance with the others needs to be computed. The longer the time series the longer will be the computation time depending on the distance function. A feature extraction might sharply reduce this computation time.

2.3.2 FFNN - Feed Forward Neural Network

The Feed Forward Neural Network FFNN is used in this paper as a multi-layer perceptron. Each layer is composed of various neurons, and each neurons of a layer has outputs to every neurons on the next layer. The input layer, has described in [16], may take a different quantity of variables: features, each points of a time series, statistical properties etc... In a FFNN, the information only goes forward. A neuron in a layer receives the values of its inputs neurons, calculates its activation via a sigmoid activation function and sends its value to every neurons in the next layer. Diverse structures have been tested to determine the most efficient one on each classification.

The larger the dataset, the better training of a FFNN. Nanopoulos and al. [16] have shown that this technique directly uses the whole time series as an input, but the use of extracted features obtain equivalent results for a much shorter computation time. Moreover structure of the FFNN needs to be determined.

2.3.3 SVM - Support Vector Machine

The support vector machine, SVM, permits a binary linear classification of the data. By representing the training data in a high dimensional space, the method finds the hyperplane that best separates the two classes. Non-linear classification is possible by mapping inputs data into high dimensional feature spaces, this is done by using kernel functions [21]. If the dataset contains more than two labels, like our dataset, classification can still be performed by using the one-versus-all technique or the one-versus-one technique.

Using SVM seems like a good option as its accuracy depends on kernel functions, our dataset contains more than two labels. Gudmundsson and al.[9] had shown that its performance was equal to the one of a KNN, however Zhang and al.[32] obtained much better results.

2.3.4 Adaboost

Boosting methods is the association of various "weak" classifiers (with a probability to correctly classify a sample higher than the random choice). These classifiers are then weighted according to their discriminative power. During the training phase, classifier weights are adjusted depending on the classification results they obtained. While on the test stage, the output is computed based on weighted votes of each classifier. For example, we know that the peak value of a thrombogram is a discriminative feature between healthy patients and haemophiliac ones, so during a classification between this two categories the peak feature will have a good weight and its vote will be more important. As the Adaboost method is a binary classifier, a cascade classification has been used to handle multiclass datasets. Alonso and al. [2] used Adaboost technique for the classification of time series, represented as literals depending on their behaviour (increasing, decreasing, stable etc).

2.3.5 Decision Tree

In decision trees, labels are represented by leaves and branches by input parameters leading to the leaves. By identifying parameters which are the most discriminative, the algorithm will builds the tree structure. Inputs may be time series features, statistical characteristics or segment values. Decision trees are widely used for their simplicity and the fact that the user can easily interpret them. In this study, we used the C4.5 algorithm[18], which builds decision trees top-down.

When applying decision trees to thrombograms, it seems obvious that giving the whole curve as input might have poor results. But the use of extracted values should produce good classification results.

2.4 Extraction Method

Some classification techniques may need extracted values to improve their results. As giving the whole time series might be heavy, the idea is to extract features that could have a discriminative function. This section presents different extraction techniques and their characteristics.

2.4.1 Piecewise Linear Approximation - PLA

The Piecewise Linear Approximation [12] (figure 4) technique approximates time series with linear functions. The curve is divided into segments and each segment is represented by a straight line. Segmentation can be uniform, segments of equal length, or non-uniform. The second one best fit the shape of the time series. To compute the straight line various techniques exist, by giving the right end and left end points or by applying the least square method.

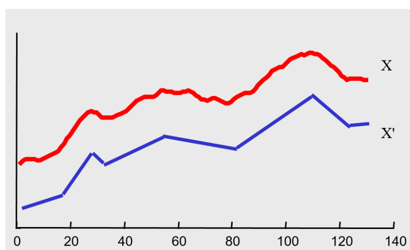

Figure 4: Example of a PLA [20]

2.4.2 Piecewise Aggregate Approximation - PAA

The Piecewise Aggregate Approximation [7] consists in segmenting the original time series into pieces of the same size and calculate the mean value on each segment (figure 5). The succession of mean values form the curve approximation. This method offers a dimensionality reduction depending on the segmentation, however information about the curve is lost during the PAA process. For example, on a given segment, it is impossible to determine if values are increasing, decreasing or if there is a peak. Thus, two really different segments can have similar mean value.

2.4.3 Adaptive Piecewise Constant Approximation - APCA

Adaptive Piecewise Constant Approximation, as described in [4], is a PAA where segment length is not arbitrarily fixed. The approximation adapts itself to the curve. In areas of high activity the APCA computes various small segments, few big ones are calculated for area of low activity (figure 6).

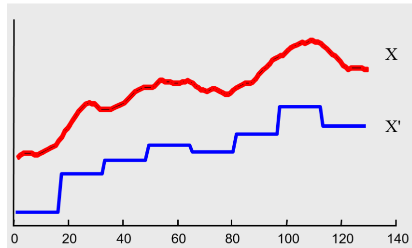

Figure 5: Example of a PAA [20]

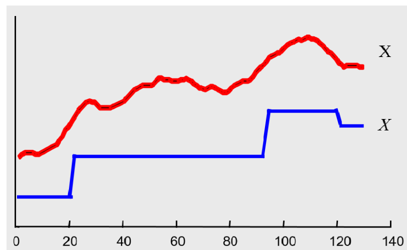

Figure 6: Example of a APCA [20]

2.4.4 Symbolic Aggregate approXimation - SAX

Symbolic Aggregate Approximation [7] represents a time series as a character string (figure 6). A PAA is applied, and then each mean value is traduced into a character. To establish the scale value/character, breakpoints, under a Gaussian curve, are computed depending and the alphabet size, determined by the user. A letter is assigned to each space between these breakpoints and thus the mean value obtains its letter. Disadvantages are the same as the PAA, curve trends are not represented.

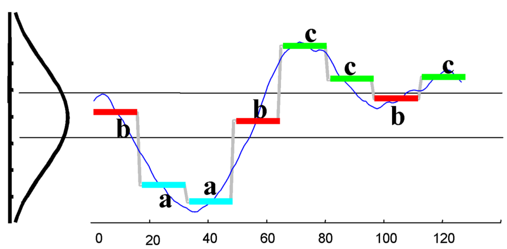

Figure 7: Example of a SAX

2.4.5 Piecewise Trend Approximation - PTA

As the SAX uses the PAA to determine the string, PTA uses the PLA. Each straight line is represented in this method by a letter corresponding to the slope. Using the least square technique, letters are attributed depending on the type of slope: increasing, stable, decreasing (figure 8). Esmael and al.[7] used this method combined with a SAX to obtain a representation based and values and slope.

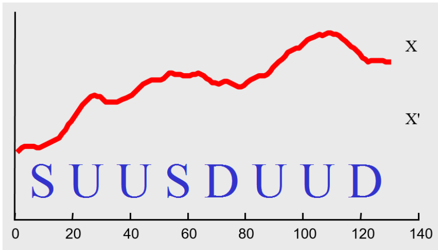

Figure 8: Example of a PTA

2.4.6 Discrete Wavelet Transform - DWT

Discrete Wavelet Transform [5] decomposes a signal into a combination of wavelets. There are different kinds of wavelets. The Haar wavelet transform, pairs up input values and computes the difference to create coefficients. Recursively repeating this process on the obtained coefficients offers a dimensionality reduction but also time and frequency information. The input must be a list of 2^n numbers. Figure 9 shows the signal reconstruction from the DWT coefficients.

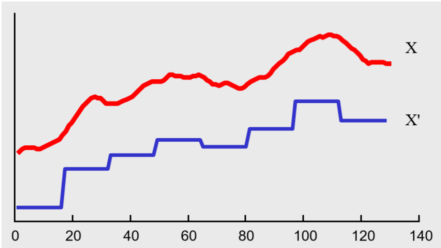

Figure 9: Example of a DWT [20]

2.4.7 Discrete Fourrier Transform - DFT

DFT computes a time series into a sequence of coefficients ordered by their frequency. The transformation converts the curve from the time domain to the frequency one. The idea is that thrombograms may have frequency particularities that could be discriminative for classification. Wu and al. [30] obtained results with both DFT and DWT while detecting similarities between two signals.

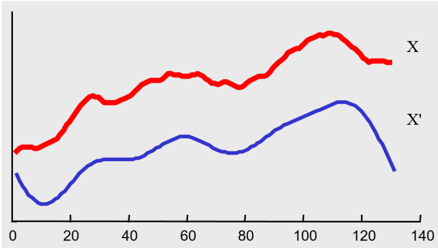

Figure 10: Example of a DFT [20]

2.4.8 Methods comparison

Among the described methods above and their characteristics, we can establish a selection based on the representation of thrombograms they can provide. First of all, we can separate the extraction techniques in three categories

- Piecewise approximation: PLA, PAA and APCA
- Symbolic approximation: SAX, PTA
- Coefficient representation: DWT, DFT

Representations based on piecewise approximation are generally fast to compute (cf Table1). PLA gives a linear approximation, it conserves values and shapes. Otherwise, with PAA, time series represented by their mean values lose part of the information. The adaptivity of APCA offers a good trade-off between dimensionality reduction and conservation of curve values, however shapes of time series are not conserved. For experimenting piecewise approximations with thrombograms, PLA is sufficient, as this technique provide an efficient dimensionality reduction and brings more information than the two other techniques.

Method	Shape approximation	Computation	Values	Parameters
PLA	yes	$O(n)$	Curve values	3
PAA	no	$O(n)$	Average	2
APCA	no	$O(nLog(n))$	Average	2

Table 1: Characteristic comparison of piecewise approximation techniques

Symbolic approximation gives a simple representation of time series which is an advantage for classification. However SAX and PTA while traducing a curve to a character string lose information, the slope for the SAX method and values for the PTA method. An efficient technique is to combine both methods. The resulting string is twice as long but the representation of the curve contains twice as much information for an equivalent computation complexity. Table 2 summarizes the different characteristics of each technique.

Method	Shape approximation	Computation	Values	Parameters
SAX	no	$O(n)$	Average	2
PTA	no	$O(n)$	Slope	2
SAX/PTA	no	$O(n)$	Average Slope	3

Table 2: Characteristic comparison of symbolic approximation techniques

Coefficient based representation, otherwise, doesn't offer an approximation based on values or slopes, but a different approach. DFT, for example, offers a representation of time series in the frequency domain. It would be interesting to see if some frequencies can be discriminative for thrombograms classification. DWT not only gives frequency information but also temporal contents. When treating time series these methods are frequently used and generally obtain good results [30] [1] [17]. Characteristics are summarised in table 3.

Method	Time series size	Computation	Values	Parameters
DWT	n^2	$O(n)$	Wavelets	2
DFT	n^2	$O(n)$	Sine Cosine	3

Table 3: Characteristic comparison of coefficient representation techniques

2.5 Synthesis

To assist the decision making process while diagnosing a thrombogram nature, a clinical decision support CDS is implemented. As no explicit rules exist to clearly identify this kind of curve, machine learning are used to conceive a non-knowledge-based CDS. Five classification techniques (KNN,FFNN,SVM,Adaboost,decision tree) will be compared to find the one that best performs on coagulation curve. As time series might be difficult to handle, a dimensionality reduction based on feature extraction method (PLA,SAX/PTA,DFT,DWT) will be tested. A tool, ORION [27], containing all these methods is used to compare their performances. Figure 11 summarizes this state of the art.

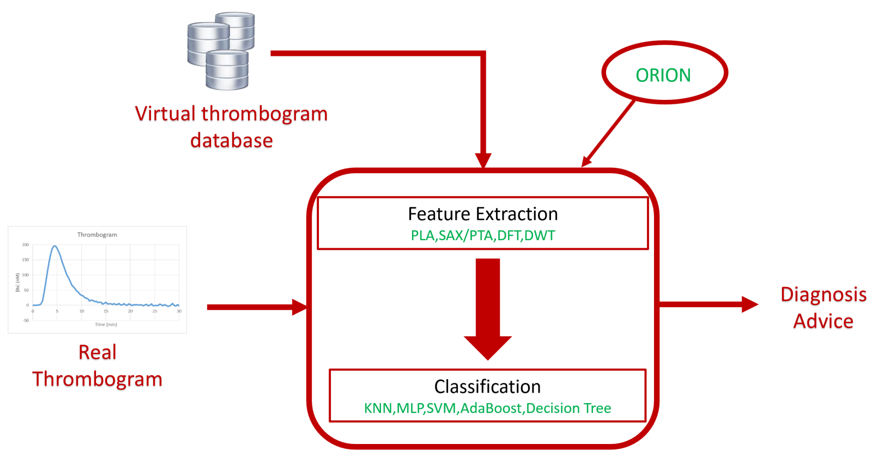

Figure 11: CDS representation with classification methods and extraction techniques

3 Approach

The objective is to find the most accurate classification method. To obtain comparable results for the diverse techniques described in the previous section, a workflow was followed. It is detailed in the second part of this section. The first part is dedicated to the dataset generation.

3.1 Datasets

Thrombograms were generated using the numerical model defined in [13]. Every protein of coagulation (coagulation factors, inhibitors) and their interactions (biochemical reactions) are represented by as a system of differential equations. This system describes the process of coagulation over time, and particularly, the thrombin generation. Initial proteins concentrations are, for each thrombogram, randomly generated according to a Gaussian distribution centred on averaged concentrations and with a standard deviation computed using the 2.5 and 97.5

percentiles of normal values. In the haemophiliac A category, for the three different severities mild, moderate and severe, standard deviation were computed with a concentration in factor VIII respectively, between 40% and 5%, 5% and 1%, 1% and 0%. Concerning haemophiliac B mild, moderate and severe, standard deviation were calculated using factor IX concentration respectively, between 40% and 5%, 5% and 1%, 1% and 0%. Once initial concentration of proteins in coagulation are fixed, the differential equation system is resolved and thrombograms can be calculated.

Resort to simulated data provides two advantages in the clinical domain. Firstly, data acquisition is complicated and very expensive. Patients with the illness corresponding to the study have to be found, samples have to be taken in the same conditions etc. Secondly, the simulation can provide a large amount of data and in machine learning data is gold.

The dataset was fill with the following thrombograms:

Table 4: Composition of the dataset

Labels	Quantity
Healthy	5 000
Haemophiliac A Mild	1 500
Haemophiliac A Moderate	1 500
Haemophiliac A Severe	1 500
Haemophiliac B Mild	1 500
Haemophiliac B Moderate	1 500
Haemophiliac B Severe	1 500

So we have a dataset composed of 5 000 Healthy patient and 9 000 Haemophiliac, this repartition is efficient as it offers a good training to determine whether or not a patient is haemophiliac, it also provides a good quantity for each type and severity of haemophilia.

Our dataset D is composed of 14 000 thromobgrams. Table 4 shows its composition, the set of label $\mathcal{Y}$ is made of 7 categories such as $\mathcal{Y} = \{$*Healthy, Haemophiliac A mild, Haemophiliac A moderate, Haemophiliac A severe, Haemophiliac B mild, Haemophiliac B moderate, Haemophiliac B severe*$\}$. Therefore, we will be able to process different kinds of classification, they are all listed below in table 5.

Table 5: Different classifications

Classifications
Healthy / Haemophiliac
Haemophiliac A / Haemophiliac B
Haemophiliac Severity
All Categories

Thus, it allowed us to observe the result of different classification techniques on different datasets. The idea was that some methods might be more effective on detecting if a patient is haemophiliac or not than detecting the nature of the haemophilia and conversely.

3.2 Workflow

3.2.1 Overview

On the basis of a database composed of various types of thrombograms, a feature extraction step will produce new datasets. They will serve for the training of a classification method. As we apply supervised learning, inputs for this method are: the class to which samples and thrombograms belong or extracted values. Due to the lack of real data, this learning phase will be held on artificially generated thrombograms (Figure 12). The goal is to determine which techniques is the most accurate. Given the medical context, some criteria will be taken into account to evaluate this method (see section Evaluation criteria). For example, diagnosing a haemophiliac patient as healthy may have calamitous consequences.

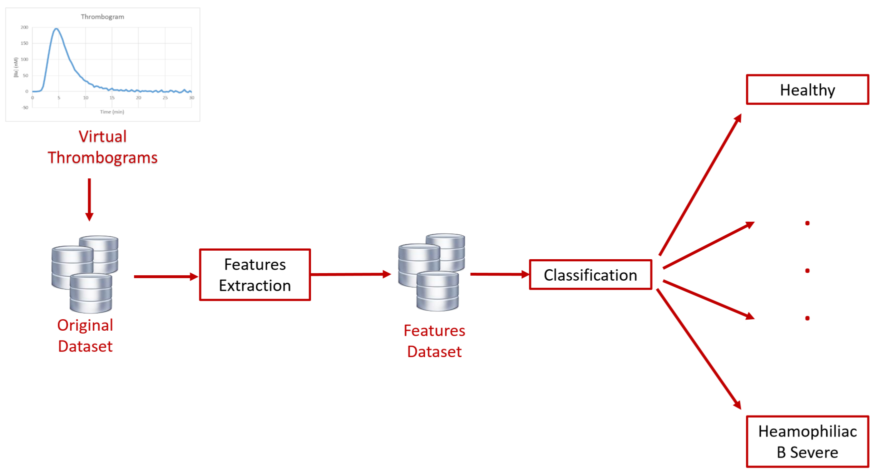

Figure 12: Training step of the CDS

The testing phase (figure 13) of the CDS consists in performing classification on virtual data that weren't used for training with the previously trained method. This step will validate the training, and will serve to compare each method.

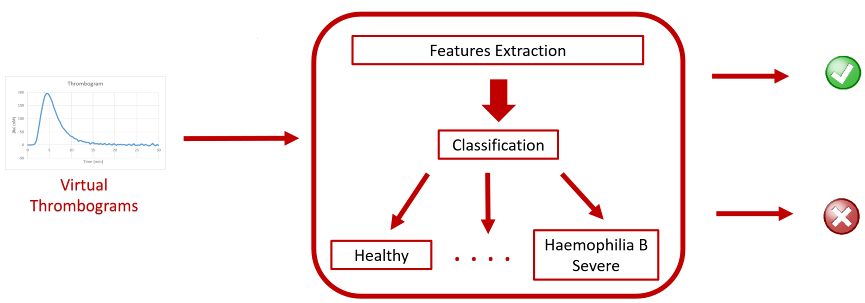

Figure 13: Testing step of the CDS

Once the method has been chosen and trained, the CDS will be tested on real data. This final step will validate the used method. Moreover, if results are accurate the thrombogram generation model will also be validated.

3.2.2 Features extraction

We have identified different extraction techniques during the state of the art. Thus based on the thrombogram dataset, we generated different versions listed below:

- Feature-based
- PLA-based
- APCA-based
- SAX/PTA-based
- DFT-based
- DWT-based

The feature-based dataset hasn't been realized using the feature extraction techniques described in the state of the art. It is based on features that are considered the most significant for thrombograms. Figure 14 shows the characteristics taken into account: Time to peak, peak, lag time and etp (Endogenous thrombin potential = area under curve).

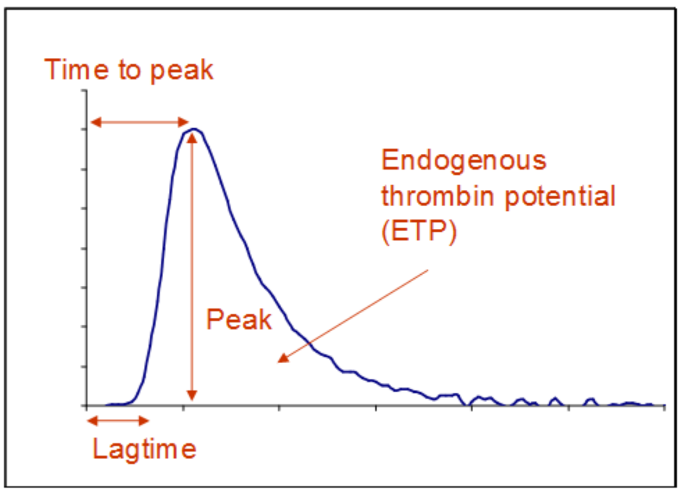

Figure 14: Thrombogram features

To implement the PLA-based dataset we segmented the thrombograms into 10 segments. Then, using the mean square method we computed the straight line representing each segment. The dimensionality reduction was performed by only using three points for each segment, left-end point, middle point and right-end point. Therefore in the PLA-based dataset each sample will be represented by vectors of 30 values.

In the SAX/PTA-based dataset, curves are divided into 15 pieces. For the SAX part, we have chosen an alphabet size of 7 lower-case letters (a to g). For the PTA part, the alphabet is composed of 8 upper-case letters (U to Z). This choice was made to visually distinguish letters for mean values and letters for slopes. This extraction represents each thrombogram by a string of 30 characters.

For the DFT-based dataset, a Fast Fourrier Transform was performed on each coagulation curve. Thrombograms we disposed are made of 118 points, so, to obtain a time series size of 2^n zero-padding was performed. It consists in adding 0 at the end of the time series until the adequate size is reached, 256 here. A sample in this dataset is composed of the 30 first Fourrier coefficients.

We choose to apply a Haar wavelet transform to create the DWT-based dataset. Likewise the DFT technique, DWT also required a time series length of 2^n, therefore zero-padding was also performed here. But after the application of this extraction technique we obtained 127 DWT coefficients, and thus, a missed dimensionality reduction. However, we decided to keep this dataset for comparison.

3.2.3 Parameter search

Before comparing the accuracy of the different classification techniques for the different datasets, we performed a parameter research.

The first step was an exploratory one. By testing diverse values on them, we identified those which are most likely to affect classification results (Marked with a $*$ in table 6).

Table 6: Parameters for each method

Methods	Parameters
KNN	k* Distance function*
SVM	SVM type kernel type degree gamma* coef0 nu C* eps p shrinking probability
FFNN	Layers* Learning rate* momentum Iteration*

As there are few parameters that significantly influence results, we decided to perform a grid search to find the best values for theses parameters. The grid search consists in fixing all parameters to a given value except one which will change according to a finite set of values. If there are more than one parameter to change, we evaluate the method by doing all the possible combinations. For example, with the KNN method, we executed the training phase and testing phase with all the different values of k(1,3,5,10,25,50) for each distance function [28] (Euclidean,Normalized euclidean, Manhattan, Minkowski(2.0) and Mahalanobis). That to say 6 values of k for 6 distance functions, that gives us 36 different combinations to perform.

3.2.4 Classification

Depending on the dataset, inputs are vectors of various dimensions, 4 dimensions for the feature based dataset and 118 for the time series based. The output is a class name (Healthy, Haemophiliac A Severe etc...), as it is a known finite set we apply supervised learning during the training stage. This permits to associate a kind of entry to a specific output, by generalising the known data treated during learning. In this paper we used the hold out method, also called test sample estimation. The dataset is split into two parts, it is common to use 2/3 of the dataset for training and 1/3 for testing [14]. The test stage consists of only giving inputs and observe the prediction made by the trained technique. Output of these samples are known, so we can evaluate the classification rate of this method.

3.2.5 Cascade

A single method can be used on a complete dataset to identify each categories in the same time, but it is also possible to perform a cascade classification method to determine labels one by one. The idea is that some categories can be pulled out of the dataset at a different step of the classification process, and thus applies the best classification method for the remaining categories. This technique, illustrated in figure 15 with SVM, allows a specialized classifier to perform on a sub-dataset.

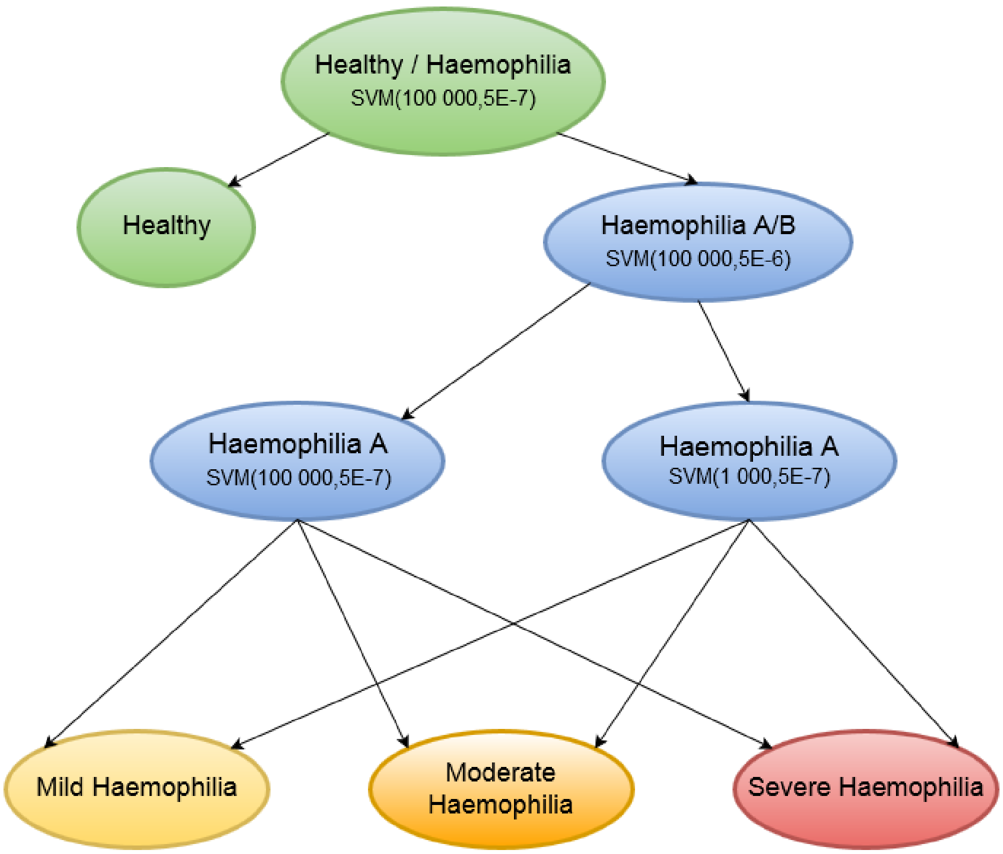

Figure 15: Cascade diagram developed for the all categories classification using SVM

The dataset is separated in two parts at the beginning. The training part (70% of the dataset) and the testing part (30% remaining). A given classifier is trained. The test set is then divided following the classification performed at each step by the trained method. We repeat this process until every classes are separated. It means that, after the first step for example, some data labelled as healthy will be in the haemophiliac group, therefore the classification error rate at each following step will be affected.

4 Evaluation

To measure the efficiency of our system, and given the clinical context, criteria have to be taken into account, they are presented in this section. Performance measures of these criterion is also detailed. Results obtained are shown in a third part.

4.1 Performance measure

To illustrate performance measures, let's take an example of a healthy/haemophiliac classification. The machine learning technique identify some data as healthy and some data as haemophiliac. We can distinguish four different cases:

- Healthy sample classified as healthy (True positive, TP).
- Haemophiliac sample classified as haemophiliac (True negative, TN).
- Healthy sample classified as haemophiliac (False negative, FN).
- Haemophiliac sample classified as healthy (False positive, FP).

Therefore, the total number of positives nP and the total number of negatives nN are calculated as follow:

$$nP = TP + FN \qquad\qquad nN = TN + FP$$

Precision measures the proportion of correctly classified data for a given label. In the example, how many data were healthy into data classified as such. *Recall* expresses the proportion of positives which are correctly classified as such. In the example above, for all healthy data how many were identified healthy.

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

However, *Precision* doesn't inform about the data belonging to a given label but that was not classified as such. *Recall* doesn't take into account data which don't belong to but were classified as a given category. To evenly combine both measures the $F - measure$ provides a good trade-off between them. It is calculated as follows:

$$F - measure = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$$

False positive rate FPR, is the proportionate negative data identified as positive on the ensemble of negative data. In the example, how many patients were classified as healthy while they were haemophiliac?

$$FPR = \frac{FP}{nN}$$

The receiver operating characteristic, or ROC Curve described in [8], illustrates the classifier accuracy. It is created plotting the true positive rate TPR, or *recall*, against the false positive rate FPR. Both are explained above. Therefore, a ROC curve shows the trade-off between true positives, in the example healthy patient recognized as healthy, and false positives haemophiliac patients identified healthy. The best method will have the higher true positive rate and the lower false positive rate, it is represented by the upper left corner on the ROC plot. But in some cases it may be more advantageous to choose a classifier which detects a lot of positive cases at the cost of having a lot of false alarm, i.e. false positive cases.

In the medical domain, data acquisition is a constrained task. Patients able to give all the different categories have to be found, in our case: Haemophiliac A and B with the different levels of severity. Moreover, samples have to be taken in the same conditions etc. Thus, an important criterion is the classification rate against the quantity of samples. Given this important fact, the learning curve has been computed. These curves express the increase of learning with experience. As best parameters are already identified, experience, here, is represented as the number of samples. To establish needed amount of data, we tested different sizes of dataset. Increasing the dataset size at each step by 500 samples, size goes from 500 to 14 000. To compute the classification rate for each different size, we used the cross validation technique [14]. It consists in dividing the dataset into equal parts, testing on one and training on the others. The method is trained and tested as many times as the quantity of parts. Classification rate is then averaged according to the obtained results. In this study we proceed to a 5-fold cross validation, we divided our dataset in 5 parts.

4.2 Evaluation criteria

The most important purpose in this study is haemophilia detection. Once a patient is diagnosed as haemophiliac, tests will be made to determine the kind of haemophilia and its severity. Nevertheless, detecting haemophilia type is still an advantage, as the CDS can suggest an order to perform all the tests. For example, if a patient is classified as haemophiliac A, the CDS could recommend performing the haemophilia A test at first. Severity classification produce a complete diagnosis. The fact is that each haemophilia severity needs different types and quantity of medication.

The CDS goal is to avoid medical errors. The worst error that could be made in our case is diagnosing a patient that is healthy while he or she truly is haemophiliac. Some medications are incompatible with haemophilia, an error of diagnosis could have disastrous consequences, and endangered a patient. We can measure these misclassifications with FPR. Therefore we have to find the method which minimize it.

Secondly, if a patient is detected as haemophiliac, lots of different tests will be lead to confirm the haemophilia. Performing all these tests is very expensive, so, another objective of our classification is to maximise the $Recall$, which represents the proportion of healthy patients classified as healthy.

The aim is to find a good trade-off between our 2 first criterions. A method which never classifies a haemophiliac as healthy and which finds the largest quantity of healthy patients to reduce cost.

Once methods that best satisfy these requirements are found, a third objective regarding the reduction of data amount has to be considered. In the medical domain, data acquisition is a complex and expensive task. Reducing the needed quantity of data for training is also an important criterion.

4.3 Results

We have established comparison measurements of the different classification techniques for thrombogram analysis. The next section presents these measures for each step of our workflow.

4.3.1 Parameter search

A parameter search was performed for KNN, SVM and FFNN. For each different classification on each dataset, diverse values were tested. Figure 16 shows learning curves of a SVM depending on C and $Gamma$ parameters. The classification is made with all category on the time series-based dataset. We can notice that best classification rates are reached for $Gamma = 5.00e-6$.

Higher C values appears to obtain better classification rates, in fact, the higher the C, the higher the cost of misclassification.

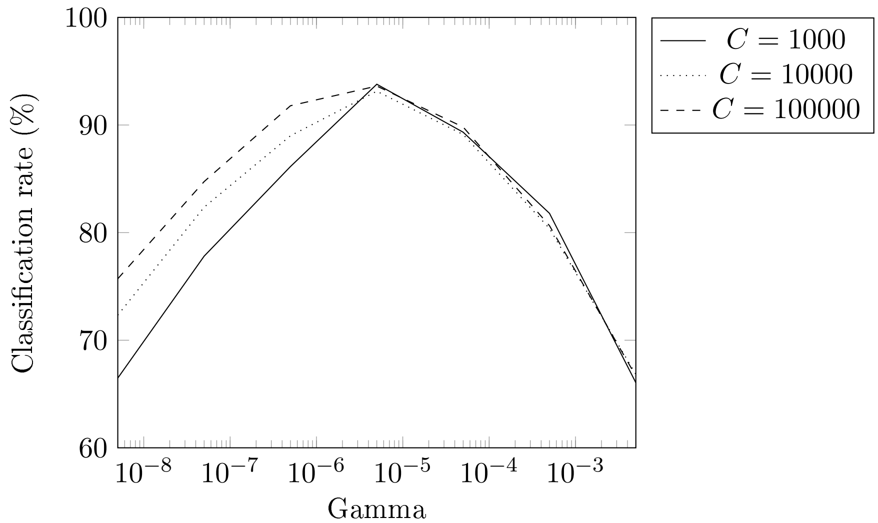

Figure 16: Classification rate against SVM *Gamma* parameter for each C values, performing all categories classification on the time series-based dataset

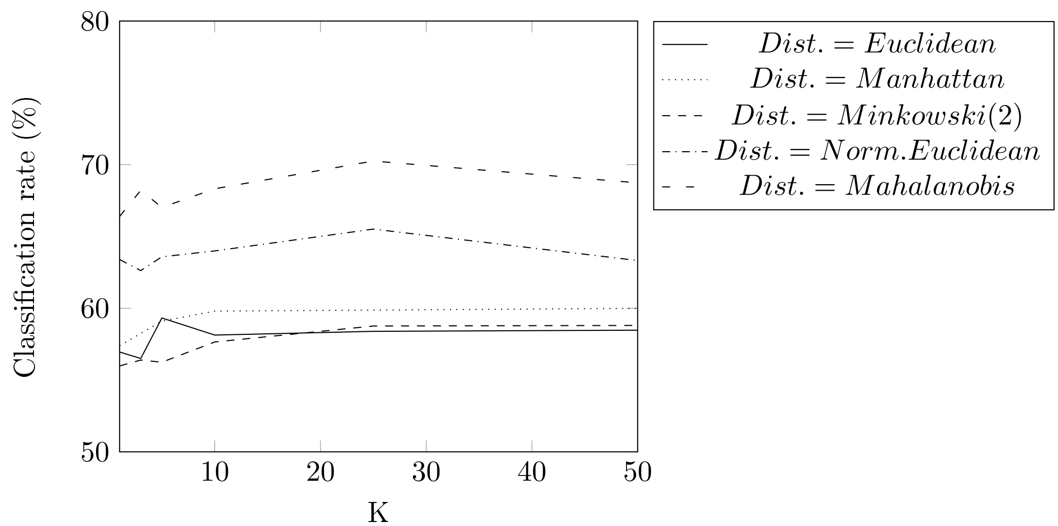

Figure 17: Classification rate against KNN K parameter for each distance function, performing haemophiliac A/B classification on the features-based dataset

Grid search results, for a haemophiliac A/B classification performed on the features-based dataset, are plotted in figure 17. We can see that, different values of k don't influence results, however, distance functions play a major role in increasing accuracy. In this case, the Mahalanobis distance measure is the best one. On other datasets, we also noticed that distance function is the most influential parameter for a KNN when applied to thrombograms.

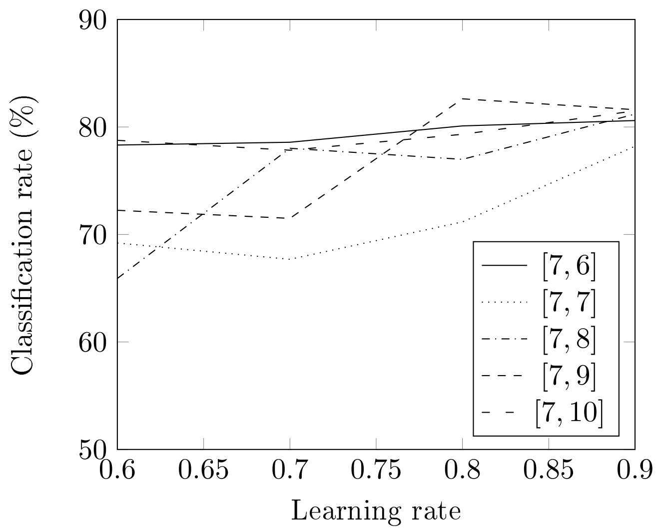

Figure 18: Classification rate against FFNN *learning rate* parameter for different neural network structures with 100 iterations, performing haemophiliac A/B classification on the features-based dataset

Concerning FFNN, we tested various structures, learning rate values and quantity of iterations. Globally, results between different structures are equals. Increasing the learning rate slightly increase accuracy, as shown by figure 18. Finally, computation time for a high amount of iterations become rapidly non reasonable.

4.3.2 Classification

Results obtained on the features-based dataset are shown in table 7. First of all, we can notice that SVM is the best method for each kind of dataset, it outclasses every other methods. Secondly, every method performed very well while identifying whether a patient is haemophiliac or not, ignoring the severity. Excluding the decision tree classification method, severity is also well recognized, classification rate tops to 95.33% with SVM.

Classifications with the worst accuracies are the all-categories and the haemophiliac A/B.

While performing a classification on the time series-based dataset, results are shown in table 8, we observed similarities with classifications on the features-based dataset. SVM, again, outperform all others machine learning techniques. SVM classification rate surpasses the 90% on every different kind of datasets. We can notice that again, datasets with the best results are the healthy/haemophiliac and haemophilia severity.

FFNN performs really poorly on time series-based datasets. This can be explained by the quantity of input values, 118 with this dataset. An explanation could be the lack of training samples. As the number of neurons increase in the neural network, the training needs to be on a larger dataset. A solution to this problem would be to use the same training data various times. As well as considerably increasing the training time, it could also lead to overfitting.

Results obtained using the other extraction techniques are shown in annexe A.

Table 7: Performance of each method on feature-based datasets. Highest classification rate, recall, precision and F-measure and lowest FPR are shown in bold

Dataset	Method	Classification rate	Recall	Precision	F-Measure	FPR
All Categories	Decision Tree	45.70	34.87	34.79	34.81	65.21
	Adaboost	54.58	39.57	35.54	37.45	60.43
	KNN	76.21	69.03	69.04	68.81	30.96
	SVM	**85.81**	**80.95**	**80.81**	**80.80**	**19.19**
	FFNN	67.74	57.71	56.92	49.76	43.08
Healthy / Haemophiliac	Decision Tree	90.92	90.12	89.98	90.05	10.02
	Adaboost	87.92	89.15	86.33	87.19	13.67
	KNN	98.54	98.40	98.38	98.39	1.62
	SVM	**98.83**	**98.60**	**98.90**	**98.75**	**1.10**
	FFNN	98.71	98.46	98.75	98.60	1.25
Haemophilia A/B	Decision Tree	52.35	52.30	52.29	52.29	47.71
	Adaboost	54.39	54.16	54.20	54.12	45.80
	KNN	70.24	70.27	70.25	70.24	29.75
	SVM	**83.62**	**83.59**	**83.62**	**83.60**	**16.38**
	FFNN	57.76	57.44	57.92	56.96	42.08
Haemophilia Severity	Decision Tree	65.76	65.97	65.78	65.87	34.22
	Adaboost	66.95	66.94	66.06	66.49	33.06
	KNN	93.59	93.59	93.79	93.65	6.21
	SVM	**95.33**	**95.33**	**95.37**	**95.32**	**4.63**
	FFNN	94.44	94.38	94.37	94.37	5.63

Table 8: Performance of each method on time series-based datasets. Highest classification rate, recall, precision and F-measure and lowest FPR are shown in bold

Dataset	Method	Classification rate	Recall	Precision	F-Measure	FPR
All Categories	Decision Tree	75.00	68.16	67.94	67.97	32.06
	Adaboost	56.60	43.14	41.16	42.13	56.86
	KNN	78.95	72.68	72.53	72.51	27.47
	SVM	**93.61**	**92.44**	**91.98**	**92.15**	**8.02**
	FFNN	48.40	32.28	19.18	22.46	37.97
Healthy / Haemophiliac	Decision Tree	97.88	97.78	97.60	97.69	2.40
	Adaboost	88.07	87.42	86.95	87.17	13.05
	KNN	98.21	97.98	98.13	98.05	1.87
	SVM	**98.90**	**98.74**	**98.88**	**98.81**	**1.12**
	FFNN	91.28	91.30	90.07	90.61	9.93
Haemophilia A/B	Decision Tree	69.63	69.62	69.64	69.62	30.36
	Adaboost	52.85	52.86	52.87	52.81	47.13
	KNN	74.37	74.41	74.45	74.37	25.55
	SVM	**96.48**	**96.50**	**96.46**	**96.48**	**3.54**
	FFNN	59.85	60.20	62.51	58.06	37.49
Haemophilia Severity	Decision Tree	91.66	91.65	91.67	91.66	8.33
	Adaboost	84.04	83.87	84.14	84.01	16.13
	KNN	93.22	93.21	93.27	93.17	6.73
	SVM	**95.59**	**95.50**	**95.76**	**95.59**	**4.24**
	FFNN	70.59	71.26	74.41	70.18	25.59

4.3.3 Cascade

In an attempt to improve classification rates on the all categories classification and given the results with other classifications, we performed another approach. Cascade technique has been implemented using the best classification methods we have identified for each sub-classification, i.e. healthy/haemophiliac, haemophiliac A/B and haemophiliac severity. Table 9 shows overall results and results reached at each step of cascade on the time series based dataset.

Table 9: Cascade classification rates for the time series-based dataset

	Classificaction rate	Recall	Precision	F-Measure	FPR
Healthy/ Haemophiliac	99,07	98,14	99,17	98,65	0,98
Haemophiliac A/B	94,24	94,06	94,53	94,29	6,05
Haemophiliac Severity	84,29	84,40	84,59	84,49	15.40
Overall	93,51	91,87	92,08	91,97	7.92

We can see that the overall classification rate obtained with time series surpasses the 90%. However, classification rates at each step, except for the Healthy/Haemophiliac step, are less accurate than those with the single method. This can be explain by the fact that after each classification, some data are in the wrong class group and it will be impossible to classify them correctly. For example, after the Healthy/Haemophiliac classification step, some data labelled as Healthy will be in the haemophiliac group, next steps won't be able to identify them correctly. Therefore error rates rise for these steps.

Results obtained with extracted features using cascade classification are shown in annexe B.

5 Analysis

In this section we compare the performance of each classification method and analyse the use of extraction techniques. Finally, we present the most accurate method according to the identified criteria.

5.1 Classification comparison

Figures 19, 20 and 21, show ROC curves of the methods for the different classifications. The healthy haemophiliac one is not shown. As results obtained are all very accurate, the graph readability is affected. ROC curves of decision tree classification technique can't be plotted in the graph. This is due to the discrete classification realized by this technique, which doesn't provide a classification threshold for each sample.

On figure 19, we clearly see that SVM outperform all other methods on the all categories classification, by nearly reaching the upper left corner, which represent a perfect classifier. Thus SVM provide an excellent way to distinguish the nature of thrombograms.

As shown by figure 20, we can see that SVM are much more accurate than the other techniques for the classification of haemophiliac A and B curves. For this classification we can also notice that Adaboost method performs really poorly. Approaching the ROC diagonal space signifies that the classifier performs as well as the random choice. FFNN technique also encounters difficulties identifying the type of haemophilia.

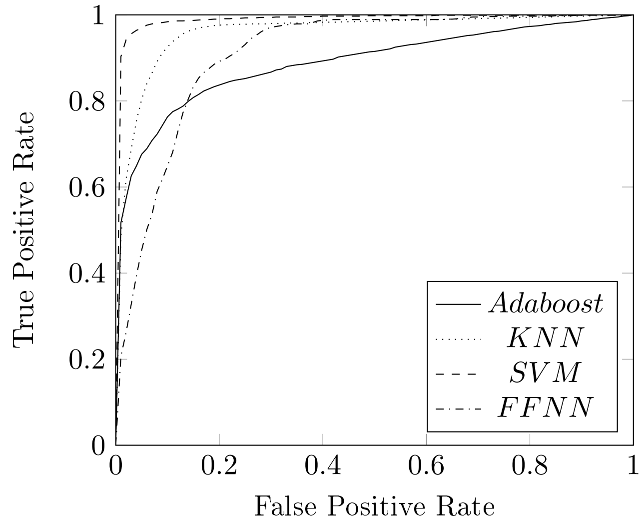

Figure 19: ROC curves of the different classification techniques on the all categories classification using the time series-based dataset

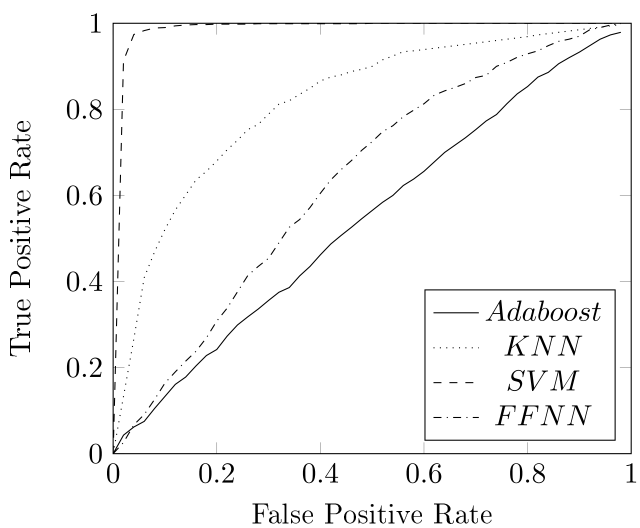

Figure 20: ROC curves of the different classification techniques on the haemophiliac A/B classification using the time series-based dataset

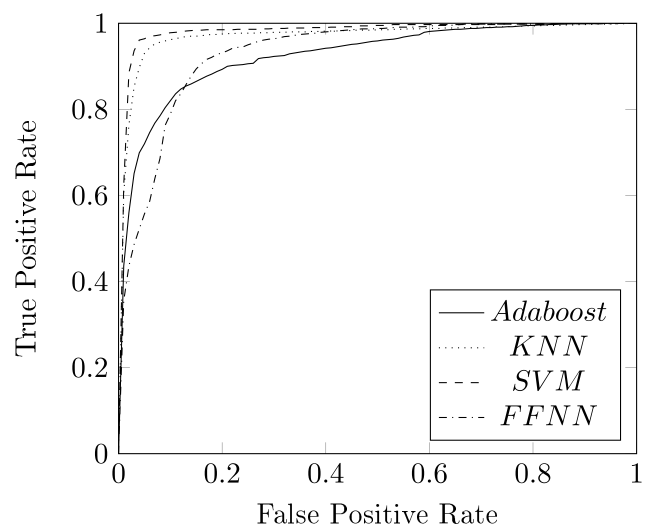

Figure 21: ROC curves of the different classification techniques on the haemophiliac severity classification using the time series-based dataset

The severity is easier to detect than the type. In general, methods perform much better on the severity classification than on the type classification as shown by the figure 21. These two graphs explain why SVM obtains better results while classifying all the categories. This technique, compared to the others tested in this study, has succeeded in identifying the different severities and the different types of haemophilia. Thus producing a complete diagnosis is possible and accurate using SVM.

Results obtained for the different types of classification using the extraction methods are shown in table 10. This table shows the classification rate of the method which performs the best to compare the efficiency of each extraction technique.

Table 10: Best classification rates obtained using extraction techniques on the different classifications

Datasets	All categories	Healthy/ Haemophiliac	Haemophiliac A/B	Haemophiliac Severity
Time series based	93.61	98.90	96.48	95.59
Features based	85.81	98.83	83.62	95.33
DFT based	87.95	98.93	90.30	93.33
DWT based	93.29	98.93	95.56	94.11
PLA based	92.59	99.07	94.85	95.52
SAXPTA based	69.86	95.19	64.48	91.15

The feature based dataset represent thrombograms with 4 distinct characteristics, described as the most discriminative ones by experts, i.e. time to peak, peak, etp and lag time. Results obtained are correct for the healthy/haemophiliac classification and for the haemophiliac severity, accuracies equal the ones obtained with time series. We can deduce that differences between the different levels of severity and between a healthy patient and a haemophiliac patient are present in these 4 characteristics.

Representing thrombograms as a vector of Fourrier coefficients offers dimensionality reduction. Samples are composed of the 30 first coefficients. As this extraction technique permits to equal the results obtained with the whole time series for the healthy/haemophiliac classification, other classifications are less accurate. The objective of using the DFT was to reveal discriminative frequencies for each different type of thrombograms. Results for the all categories classification shows, that these frequencies do exist, but they are not sufficient to provide a complete diagnosis.

The DWT extraction technique presents accurate results, 93% of classification rate, which equals the one obtained with time series. But the dimensionality reduction attempt failed with the thrombograms. DWT based dataset samples are composed of vectors with 127 values, while time series are only composed of 118 values. Classifying haemophiliac A and B thrombograms or analysing the severity of the haemophilia is slightly less efficient than using time series.

On the other hand PLA extraction technique shows an efficient ratio classification result/vector dimension. By only presenting vectors composed of 40 values, classification rate only decreases by 1% for each classification except the healthy/haemophiliac classification, which equals the one obtained with time series.

The originality of the SAXPTA technique is to present the time series as a character string.

However, results are not satisfying. This is due to information loss induced by the technique. Representing a thrombogram only using averaged values and slopes seems to be insufficient to produce an accurate classification.

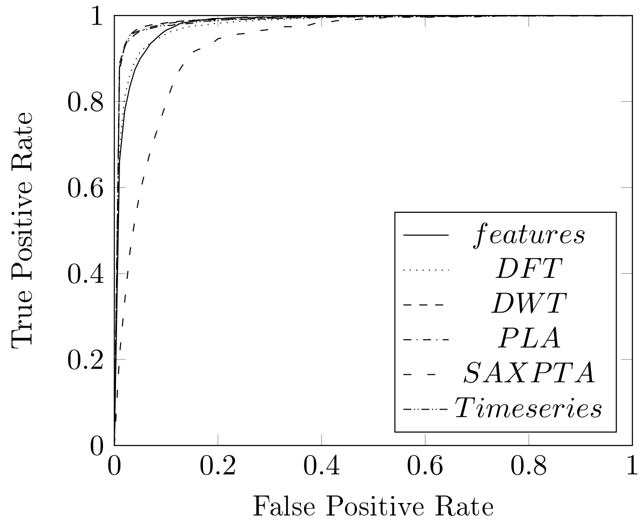

Figure 22: ROC curves of the all categories classification on the time series-based dataset for the different extraction techniques

Extraction techniques provides less accurate results that those obtained with the whole time series. However, this study was lead with artificially generated curves. It means that these curves are clean of any noise that can appear during the recording of real thrombograms. As PLA technique also reduces the noise of real curves, it would be interesting to include this technique for analysing real coagulation curves.

5.2 Cascade classification comparison

Table 11 compares results obtained with cascades and the ones obtained using single methods. We can observe that cascades never obtain better results than single method. Accuracies are equal using time series or PLA extraction, and lower with the other extraction techniques.

Table 11: Classification rates using single methods or cascade classification technique

Dataset	Single	Cascade
Features based	85.81	83.86
Time series based	93.61	93.51
DFT based	87.95	89.50
DWT based	93.29	91.38
PLA based	92.59	92.73
SAXPTA based	69.86	67.76

5.3 Performance of haemophilia detection

The first objective of this study is to detect haemophilia using thrombograms. Table 12 shows results obtained while classifying healthy and haemophiliac patients. Decision Tree, KNN and SVM obtained very good results on this dataset. SVM method is slightly above the others. Comparatively, Adaboost and FFNN perform relatively badly.

As identified in the performance evaluation section, detecting a haemophiliac patient as healthy could have serious consequences. The SVM technique outperforms all other techniques by having a *False Positive rate* under 1%. The second criterion, i.e. having a maximum of healthy patient classified as such, is also satisfied. *Recall* tops to 98%. Thus, a CDS composed of a SVM classification technique is able to fulfil the two main criterion, avoiding clinical errors and reducing costs.

Table 12: Performance of the different methods performing a healthy/haemophiliac classification. Best classification rate, FPR and Recall are shown in bold

Methods	Classification rate	No. haemophiliac classified as healthy	FPR	No. heatlhy classified as healthy	Recall
Decision Tree	97.88	51/2710	1.88	1452/1490	97.45
Adaboost	88.07	274/2679	10.23	1294/1521	85.08
KNN	98.21	32/2698	1.19	1487/1515	98.15
SVM	**98.90**	18/2685	**0.67**	1487/1515	**98.15**
FFNN	91.28	238/2717	8.76	1355/1483	91.37

5.4 Complete diagnosis performance

To produce a complete diagnosis, specifying the type of haemophilia and its severity, we consider the all categories classification and the cascade technique. Table 13 shows results obtained by the best single method in this classification and the best cascade. We can see that they obtained equivalent results. Those two techniques have classification rates over 90%.

Table 13: Performance of single method and cascade technique on the all categories classification

	Classificaction rate	Recall	Precision	F-Measure	FPR
Single Method	93.61	92.44	91.98	92.15	8.02
Cascade Method	93,51	91,87	92,08	91,97	7.92

5.5 Learning curves

To establish the best ratio accuracy/amount of data, learning curves were computed for classification techniques which perform best. Figure 23 shows the quantity of thrombograms needed to perform an efficient healthy/haemophiliac-classification on the time series-based dataset. To compute this curve a 5-fold cross validation was performed. SVM method with 500 coagulation curves, can clearly distinguish healthy patients from haemophiliac ones. Moreover, this figure shows the FPR and *Recall* for healthy patient. As we can see, FPR remains low, and *Recall* remains high regardless of the size of the dataset. Therefore, to fulfil the security requirement and the low costs requirement, i.e. a low FPR and a high *Recall*, few patients are needed.

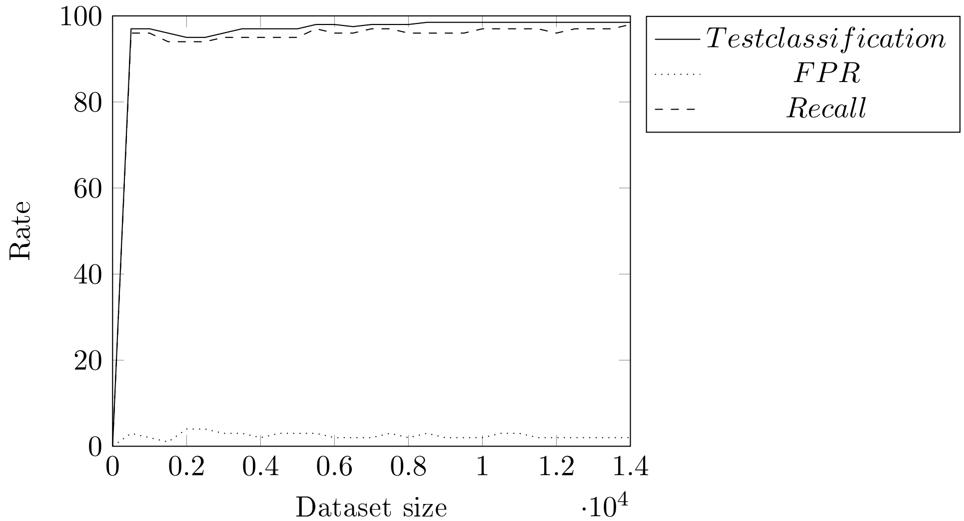

Figure 23: Classification rates versus samples quantity for a SVM performing Healthy/Haemophiliac classification on the time series-based dataset

We can see on figure 24 that the SVM method, with only 500 samples, has an accuracy of 80%. The final classification rate, 93% is reached for a dataset of 6 000 thrombograms. Proportions of the original dataset are conserved. A dataset of 6 000 coagulation curves contains:

- 2 150 Healthy
- 650 haemophiliac for each category

On the previous graph we have seen that a few samples were needed to satisfy the criteria. Meanwhile, producing an accurate complete diagnosis require more thrombograms samples.

As we can see in figures 25 and 26, classification rates of the haemophiliac A/B classification and severity classification respectively, accuracies surpass 90% correctness with few training samples.

6 Discussion

This final section brings to light the issues of CDS to detect haemophilia and suggests ways for further research.

6.1 Issues of this work

We suggested a clinical decision support system to help clinicians for haemophilia diagnosis. Our system is able to provide an accurate diagnosis using thrombograms. Support Vector Machine, which obtained the best results according to our evaluation criteria, provides an accurate classification. Moreover, this machine learning technique is able to detect the kind of haemophilia and its severity.

Providing such a diagnosis provides various advantages. With a low *False Positive Rate*, it offers a good way to avoid medical errors, and thus reduce the risk of putting patient's life in danger. It also reduces costs by correctly identifying healthy patients, and so reduces the

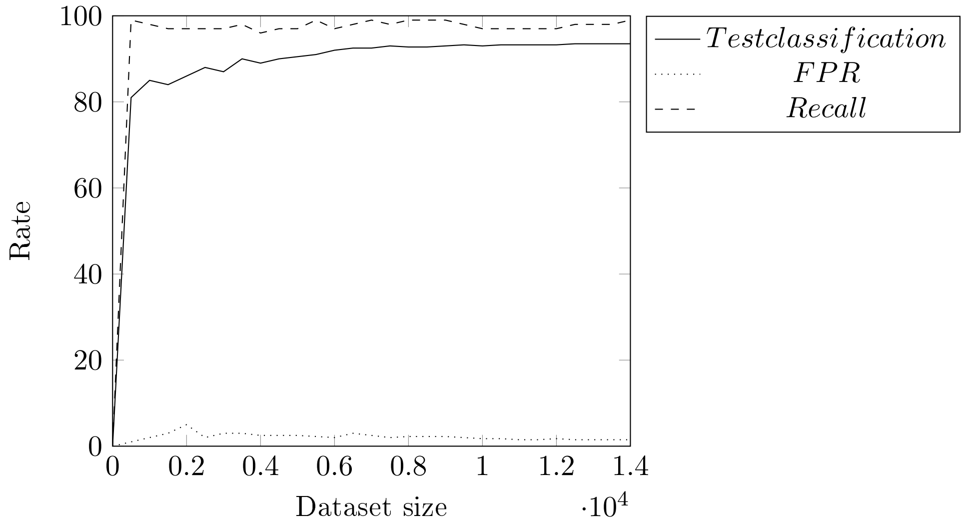

Figure 24: Classification rates versus samples quantity for a SVM performing the all categories classification on the time series-based dataset

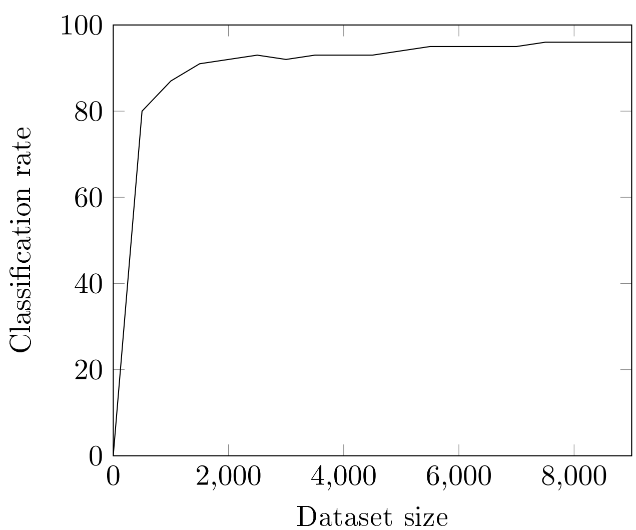

Figure 25: Classification rate versus samples quantity for the Haemophiliac A/B classification on the time series-based dataset

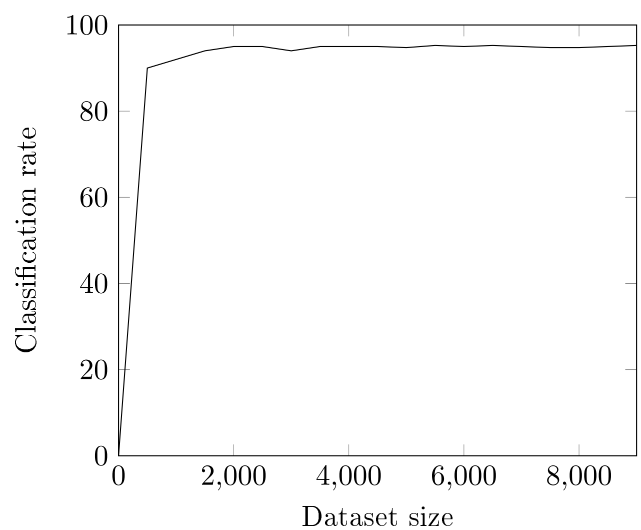

Figure 26: Classification rate versus samples quantity for the Haemophiliac severity classification on the time series-based dataset

quantity of tests needed. Finally, the fact that our CDS provides a complete diagnosis, type of haemophilia and severity, also permits the management of test quantity by suggesting an order in the application of the different tests.

6.2 Future work

There are several avenues for this work. First is the validation phase. Both the learning phase and the test phase were carried out on artificially generated thrombograms. This generation can produce a huge quantity of data for every possible type of haemophilia, allowing an effective learning phase. However, if the thrombograms generation model is erroneous, classifying real coagulation curves with methods trained on an artificial dataset will have poor results. Moreover, generated data are not polluted with noise, and noise may greatly affect classification results. Some techniques might be used to reduce the impact of noise. A collect campaign will be lead to acquire real data. Collecting medical data is a complicated and constrained task. Thus the learning curves obtained in this study give an estimation of the amount needed for each categories and thus optimizing data acquisition. Only after this validation step, the CDS can be completed.

Secondly, other machine learning techniques might obtain accurate classification results on this dataset. Hidden Markov Models [33], HMM, generally performs well with time series, they are extensively used in time series prediction. Testing a HMM on the thrombograms dataset and comparing the obtained results will be a great contribution to this work. Moreover, Dynamic Time Warping [19], DTW, hasn't been taken into account in this study. DTW evaluates the distance between two time series. They are widely used in speech recognition, but it is possible to apply it to any time series. The use of DTW may reveal a good way to identify the different categories of thrombogram.

Another approach would be to test different types of neural network, backpropagation [11] or deep convolutional [34] neural network. This kind of neural network can be trained by re-using the training set of data (epochs) and thus reducing the number of samples needed while optimizing the training.

This study has shown the ability of machine learning techniques to detect haemophilia and determine its nature. Results obtained with haemophilia and thrombin generation open doors for other clinical application in the domain of blood illness. If illnesses can be detected by some measurements, like thrombogram in the case of haemophilia, artificial intelligence methods can be used to help during the decision making process.

References

[1] R. Agrawal, C. Faloutsos, and A. Swami. Efficient similarity search in sequence databases. *Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms (FODO '93)*, 8958546:69–84, 1993.

[2] C. Alonso and J. Rodriguez. Time Series Classification by Boosting Interval Based literals. *Inteligencia Artificial*, 4(11):2–11, 2000.

[3] A. A. Boxwala, M. Peleg, S. Tu, O. Ogunyemi, Q. T. Zeng, D. Wang, V. L. Patel, R. A. Greenes, and E. H. Shortliffe. GLIF3: A representation format for sharable computer-interpretable clinical practice guidelines, 2004.

[4] K. Chakrabarti, E. Keogh, S. Mehrotra, and M. Pazzani. Locally adaptive dimensionality reduction for indexing large time series databases. *ACM Transactions on Database Systems*, 27(2):188–228, 2002.

[5] P. Chaovalit, A. Gangopadhyay, G. Karabatis, and Z. Chen. Discrete wavelet transform-based time series analysis and mining. *ACM Computing Surveys*, 43(2):1–37, 2011.

[6] W. A. Chaovalitwongse, Y.-J. Fan, and R. C. Sachdeo. On the Time Series K-Nearest Neighbor Classification of Abnormal Brain Activity. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 37(6):1005–1016, 2007.

[7] B. Esmael, A. Arnaout, R. K. Fruhwirth, and G. Thonhauser. Multivariate time series classification by combining trend-based and value-based approximations. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7336 LNCS(PART 4):392–403, 2012.

[8] J. Fan, S. Upadhye, and A. Worster. Understanding receiver operating characteristic (ROC) curves. *CJEM: Canadian Journal of Emergency Medicine*, 8(1):19–20, 2006.

[9] S. Gudmundsson, T. P. Runarsson, and S. Sigurdsson. Support vector machines and dynamic time warping for time series. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 2772–2776, 2008.

[10] H. C. Hemker, S. Kerdelo, and R. M. W. Kremers. Is there value in kinetic modeling of thrombin generation ? No (unless ...). *Journal of thrombosis and haemostasis*, 10:1470–1477, 2012.

[11] I. Kaastra and M. Boyd. Designing a neural network for forecasting financial and economic time series. *Neurocomputing*, 10(3):215–236, 1996.

[12] E. Keogh, S. Chu, D. Hart, and M. Pazzani. Segmenting Time Series: A Survey and Novel Approach. *Data Mining in Time Series Databases*, pages 1–21, 2003.

[13] Kerdelo. *Méthodes informatiques pour l'expérimentation in virtuo de la cinétique biochimique. Application à la coagulation du sang. Modélisation et simulation.* Thèse, Université Rennes 1, 2006.

[14] R. Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence*, 14(12):1137–1143, 1995.

[15] S. Mani, A. Ozdas, C. Aliferis, H. A. Varol, Q. Chen, R. Carnevale, Y. Chen, J. Romano-Keeler, H. Nian, and J.-H. Weitkamp. Medical decision support using machine learning for early detection of late-onset neonatal sepsis. *Journal of the American Medical Informatics Association : JAMIA*, 21(2):326–36, 2014.

[16] A. Nanopoulos, R. Alcock, and Y. Manolopoulos. Feature-based classification of time-series data. *Information processing and Management*, 0056:49–61, 2001.

[17] I. Popivanov and R. Miller. Similarity search over time-series data using wavelets. In *Proceedings 18th International Conference on Data Engineering*, pages 212–221. IEEE Comput. Soc, 2002.

[18] J. R. Quinlan. *C4.5: Programs for Machine Learning*, volume 1. 1992.

[19] C. Ratanamahatana and E. Keogh. Everything you know about dynamic time warping is wrong. *Third Workshop on Mining Temporal and Sequential Data*, pages 22–25, 2004.

[20] C. A. Ratanamahatana, J. Lin, D. Gunopulos, E. Keogh, M. Vlachos, and G. Das. Mining Time Series Data. *Data Mining and Knowledge Discovery Handbook*, pages 1049–1077, 2010.

[21] S. Rüping. SVM Kernels for Time Series Analysis. *Time*, page 8, 2001.

[22] M. Samwald, K. Fehre, J. de Bruin, and K. P. Adlassnig. The Arden Syntax standard for clinical decision support: Experiences and directions. *Journal of Biomedical Informatics*, 45(4):711–718, 2012.

[23] J. Sayyad Shirabad, S. Wilk, W. Michalowski, and K. Farion. Implementing an integrative multi-agent clinical decision support system with open source software. *Journal of Medical Systems*, 36(1):123–137, 2012.

[24] F. L. Seixas, B. Zadrozny, J. Laks, A. Conci, and D. C. Muchaluat Saade. A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer's disease and mild cognitive impairment. *Computers in Biology & Medicine*, 51:140–158, 2014.

[25] H. Shin and M. K. Markey. A machine learning perspective on the development of clinical decision support systems utilizing mass spectra of blood samples, 2006.

[26] E. H. Shortliffe, D. a.B. Lindberg, and E.H.S. *Computer-Based Medical Consultations: Mycin*. Elsevier, 1976.

[27] J. Soler. *ORION A Generic Model for Data Mining Application to Video Games*. Thèse, Université de Bretagne occidentale, 2015.

[28] J. Soler, F. Tencé, L. Gaubert, and C. Buche. Data Clustering and Similarity. *Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference*, pages 492 – 495, 2013.

[29] D. R. Sutton and J. Fox. The Syntax and Semantics of the PROforma Guideline Modeling Language. *Journal of the American Medical Informatics Association*, 10(5):433–443, 2003.

[30] Y.-L. Wu, D. Agrawal, and A. El Abbadi. A Comparison of DFT and DWT Based Similarity Search in Time-series Databases. *Proceedings of the 9th International Conference on Information and Knowledge Management (CIKM)*, 35(2000-08):488–495, 2000.

[31] G. Young, B. Sørensen, Y. Dargaud, C. Negrier, K. Brummel-Ziedins, and N. S. Key. Thrombin generation and whole blood viscoelastic assays in the management of hemophilia: current state of art and future perspectives. *Blood*, 121(11):1944–50, 2013.

[32] D. Zhang, W. Zuo, D. Zhang, and H. Zhang. Time series classification using support vector machine with Gaussian elastic metric kernel. *Proceedings - International Conference on Pattern Recognition*, pages 29–32, 2010.

[33] Y. Zhang. Prediction of Financial Time Series with Hidden Markov Models. *The Magazine Of The Fine Arts*, B. Eng(May):102, 2004.

[34] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao. Time series classification using multi-channels deep convolutional neural networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8485 LNCS, pages 298–310. Springer Verlag, 2014.

A Classification results

Table 14: DFT results

Dataset	Method	Classification Rate	Recall	Precision	FMeasure	FPR
All categories	Decision Tree	78.31	71.98	72.24	72.11	28.02
	Adaboost	44.64	27.30	16.11	20.27	72.70
	KNN	76.57	70.32	70.36	70.34	29.68
	SVM	87.95	84.45	85.10	84.77	15.55
	FFNN	37.57	14.29	5.37	7.80	85.71
Healthy Haemophiliac	Decision Tree	98.24	98.10	98.04	98.07	1.90
	Adaboost	96.88	96.86	96.49	96.67	3.14
	KNN	98.60	98.44	98.42	98.43	1.56
	SVM	98.93	98.75	98.91	98.83	1.25
	FFNN	66.12	50.00	33.06	39.80	50.00
Haemophiliac A/B	Decision Tree	79.22	79.21	79.23	79.22	20.79
	Adaboost	60.52	60.53	60.63	60.58	39.47
	KNN	71.67	71.71	71.67	71.69	28.29
	SVM	90.30	90.29	90.39	90.34	9.71
	FFNN	52.41	50.00	26.20	34.39	50.00
Haemophiliac severity	Decision Tree	88.15	88.27	88.27	88.27	11.73
	Adaboost	56.19	55.40	42.45	48.07	44.60
	KNN	90.78	90.74	90.93	90.83	9.26
	SVM	93.33	93.26	93.60	93.43	6.74
	FFNN	92.78	92.85	92.80	92.83	7.15

Table 15: DWT Results

Dataset	Method	Classification Rate	Recall	Precision	FMeasure	FPR
All categories	Decision Tree	9.19	73.02	73.08	73.05	26.98
	Adaboost	72.98	62.87	67.85	65.26	37.13
	KNN	82.71	77.45	77.39	77.42	22.55
	SVM	93.29	91.65	91.45	91.55	8.35
	FFNN	65.17	54.47	56.81	55.61	45.53
Healthy Haemophiliac	Decision Tree	98.64	98.57	98.50	98.53	1.43
	Adaboost	98.31	98.46	97.90	98.18	1.54
	KNN	98.79	98.54	98.78	98.66	1.46
	SVM	98.93	98.62	99.02	98.82	1.38
	FFNN	98.02	97.51	98.15	97.83	2.49
Haemophiliac A/B	Decision Tree	76.04	76.10	76.11	76.11	23.90
	Adaboost	66.19	66.08	66.32	66.20	33.92
	KNN	80.93	80.97	80.93	80.95	19.03
	SVM	95.56	95.61	95.53	95.57	4.39
	FFNN	62.96	62.11	65.34	63.68	37.89
Haemophiliac severity	Decision Tree	91.15	91.16	91.20	91.18	8.84
	Adaboost	91.93	91.91	91.96	91.93	8.09
	KNN	94.11	94.05	94.18	94.12	5.95
	SVM	92.81	92.76	92.93	92.85	7.24
	FFNN	90.33	90.08	90.73	90.40	9.92

Table 16: PLA Results

Dataset	Method	Classification Rate	Recall	Precision	FMeasure	FPR
All categories	Decision Tree	73.82	66.53	66.76	66.65	33.47
	Adaboost	67.58	57.48	62.72	59.99	42.52
	KNN	79.20	72.76	72.88	72.82	27.24
	SVM	92.59	91.03	90.74	90.88	8.97
	FFNN	76.25	69.03	77.23	72.90	30.97
Healthy Haemophiliac	Decision Tree	98.28	98.19	98.04	98.11	1.81
	Adaboost	98.19	98.17	97.93	98.05	1.83
	KNN	98.52	98.44	98.36	98.40	1.56
	SVM	99.07	98.93	99.03	98.98	1.07
	FFNN	98.79	98.74	98.61	98.68	1.26
Haemophiliac A/B	Decision Tree	65.99	66.06	66.03	66.04	33.94
	Adaboost	59.80	59.80	59.80	59.80	40.20
	KNN	74.95	74.99	75.01	75.00	25.01
	SVM	94.85	94.84	94.87	94.85	5.16
	FFNN	74.99	75.24	75.46	75.35	24.76
Haemophiliac severity	Decision Tree	91.22	91.21	91.21	91.21	8.79
	Adaboost	89.70	89.52	89.48	89.50	10.48
	KNN	93.55	93.48	93.65	93.57	6.52
	SVM	94.74	94.76	94.84	94.80	5.24
	FFNN	95.52	95.40	95.67	95.54	4.60

Table 17: SAXPTA Results

Dataset	Method	Classification Rate	Recall	Precision	FMeasure	FPR
All categories	Decision Tree	-	-	-	-	-
	Adaboost	16.50	23.16	7.12	10.89	76.84
	KNN	60.79	47.27	56.96	51.66	52.73
	SVM	69.86	62.01	62.32	62.17	37.99
	FFNN	64.40	54.60	56.95	55.75	45.40
Healthy Haemophiliac	Decision Tree	-	-	-	-	-
	Adaboost	82.19	85.59	83.04	84.29	14.41
	KNN	90.71	92.49	89.98	91.22	7.51
	SVM	94.81	94.60	94.24	94.42	5.40
	FFNN	95.19	94.80	94.75	94.77	5.20
Haemophiliac A/B	Decision Tree	-	-	-	-	-
	Adaboost	50.41	50.56	50.72	50.64	49.44
	KNN	54.70	52.74	67.33	59.15	47.26
	SVM	64.48	64.48	64.48	64.48	35.52
	FFNN	61.52	61.53	61.56	61.54	38.47
Haemophiliac severity	Decision Tree	-	-	-	-	-
	Adaboost	73.04	72.56	73.70	73.13	27.44
	KNN	75.85	75.40	82.39	78.74	24.60
	SVM	91.07	91.03	91.21	91.12	8.97
	FFNN	91.15	91.23	91.32	91.28	8.77

B Cascade results

Table 18: DFT Cascade Results

	Classification Rate	Recall	Precision	FMeasure	FPR
Healthy Haemophiliac	98.79	98.51	98.88	98.69	1.49
HaemophiliacA/B	88.00	87.98	87.97	87.98	12.02
Haemophiliac Severity	74.61	74.67	74.57	74.62	25.32
Overall	89.50	86.64	86.31	86.47	13.36

Table 19: DWT Cascade Results

	Classification Rate	Recall	Precision	FMeasure	FPR
Healthy Haemophiliac	98.62	98.22	98.80	98.51	1.78
Haemophiliac A/B	92.54	92.53	92.54	92.53	7.47
Haemophiliac Severity	78.45	79.00	78.70	78.85	20.99
Overall	91.38	89.54	89.05	89.29	10.46

Table 20: PLA Cascade Results

	Classification Rate	Recall	Precision	FMeasure	FPR
Healthy Haemophiliac	98.81	98.70	98.73	98.72	1.30
Haemophiliac A/B	93.06	93.07	93.04	93.05	6.93
Haemophiliac Severity	82.97	83.10	83.11	83.11	16.89
Overall	92.73	90.63	90.69	90.66	9.37

Table 21: SAXPTA Cascade Results

	Classification Rate	Recall	Precision	FMeasure	FPR
Healthy Haemophiliac	93.31	92.51	92.76	92.64	7.49
Haemophiliac A/B	61.91	61.98	61.99	61.99	38.02
Haemophiliac Severity	22.51	25.08	22.59	23.74	74.92
Overall	67.76	60.39	60.49	60.44	39.61