الجمهورية الجزائرية الديمقراطية الشعبية
REPUBLIC ALGERIAN DEMOCRATIC AND POPULAR
وزارة التعليم العالي و البحث العلمي
Ministry of Higher Education and Scientific Research
جامعة أبي بكر بلقايد – تلمسان –
Aboubakr Belkaïd – Tlemcen – University

TECHNOLOGY Faculty

THESIS

This Thesis is submitted for the degree of ***Doctor of Sciences***

Speciality : BIOMEDICAL ELECTRONICS

By : *Omar BEHADADA*

Subject

Knowledge fusion from the human expert and database: Medical application

The 01 / 06 / 2017

Committee Member :

Mr ABDERAHIME Amine	Senior Lecturer	Univ. Tlemcen	Chair Comittee
Mr CHIKH M.Amine	Professor	Univ. Tlemcen	Supervisor
Mr TROVATI Marcello	Senior Lecturer	Univ. Edge Hill	Co- Supervisor
Mr BENYETOU Abdelkader	Professor	Univ. USTO. Oran	Member Comittee
Mr MERAD Lotfi	Senior Lecturer	E.S.S.A Tlemcen	Member Comittee

Academic year : 2016-2017

الجمهورية الجزائرية الديمقراطية الشعبية
REPUBLIC ALGERIAN DEMOCRATIC AND POPULAR
وزارة التعليم العالي و البحث العلمي
Ministry of Higher Education and Scientific Research
جامعة أبي بكر بلقايد – تلمسان –
Aboubakr Belkaïd – Tlemcen – University

TECHNOLOGY Faculty

THESIS

This Thesis is submitted for the degree of ***Doctor of Sciences***

Speciality : BIOMEDICAL ELECTRONICS

By : *Omar BEHADADA*

Subject

Knowledge fusion from the human expert and database: Medical application

The 01 / 06 / 2017 ,

Committee Member :

Mr ABDERAHIME Amine	Senior Lecturer	Univ. Tlemcen	Chair Comittee
Mr CHIKH M.Amine	Professor	Univ. Tlemcen	Supervisor
Mr TROVATI Marcello	Senior Lecturer	Univ. Edge Hill	Co- Supervisor
Mr BENYETOU Abdelkader	Professor	Univ. USTO. Oran	Member Comittee
Mr MERAD Lotfi	Senior Lecturer	E.S.S.A Tlemcen	Member Comittee

Academic year : 2016-2017

I would like to dedicate this thesis to my both loving parents, my wife and her family, my sisters and brothers and their families .....

ACKNOWLEDGEMENTS

I would like to thank my principal supervisor, CHIKH Amine for guiding and supporting me over the years. You have set an example of excellence as a researcher, mentor, and instructor. The good advice, support and friendship of my second supervisor Dr Marcello TROVATI has been invaluable on both academic and personal level for which I am extremely grateful.

I would like to thank my committee chair Dr ABDERAHIME Amine, for serving on my committee and taking time to talk with me on many occasions.

I would like to thank my thesis committee members Pr BENYETOU Abdelkader and Dr MERAD Lotfi for all their guidance through this process; your discussion, ideas, and feedback have been invaluable.

I'd like to thank the students, research technicians, collaborators, friends and collegues and every person who contributed to this research.

...Thank you for your constant enthusiasm and encouragement,

I am grateful to all of you…

Table of contents

List of Figures:

List of tables:

Chapter 1

1.1 Introduction

Knowledge based systems – based on a reasoning mechanism and a knowledge base (KB) – are very important, therefore they need to be interpretable, comprehensible and intelligible, in order to be used with human interaction in such fields like medicine for decision support systems.

People would not feel confident and would not accept suggestions unless KBs of such systems are really comprehensible. While expert knowledge and expert systems are comprehensible, expert knowledge acquisition keeps being 'a hard and critical task'. However, the formalization of some parts of expert knowledge is hard to be done and remains at unconscious level.

Experimental data can constitute another source of information to these systems and provide a clear interaction image between variables. Besides, there are algorithms and machine learning techniques that affect the induction of knowledge based systems. Since there are two types of knowledge, their interpretability level would differ. On the one hand, the expert knowledge is usually a general knowledge related to the most influential variables and the global system behavior. On the other hand, the induced knowledge from data is always a specific knowledge related to the situations described in the training data set. The two types of knowledge are complementary and they would give rise to high performance compact systems. One of the several ways of cooperation is to make of the induced knowledge a posteriori expert validation, whereas Guiraud-Carrier [1] proposed the a priori knowledge to guide the induction process and/or to reduce the computational effort in the ML process. The simplest one consists in making a posteriori expert validation of induced knowledge. Alternatively, other authors proposed the use of expert knowledge as a priori knowledge to guide the induction process and/or to reduce the computational effort in the ML process. Induced knowledge can also be used in order to complete the expert knowledge. In some cases, both kinds of knowledge are combined in the KB generation [2, 3, 4, 5]. As different sources of information may include

some contradictions and/or redundancies, the integration process should be made carefully and a consistency analysis is needed.

Cardiovascular diseases are one of the most worrying health issues and the largest cause of mortality in the world, based on the World Health Report 2017 **[6].** Thus, low-cost and high-quality cardiac assessment offers a very valuable challenge to this problem. Furthermore, the availability of a huge amount of information created by the continuous development of big data methods and techniques provides new challenges as well as new opportunities in this field. The detection of cardiac arrhythmia is a very interesting area, because premature ventricular contraction (PVC) is an effective predictor of sudden death. Over the past decade, several studies have focused on methods and algorithms that detect and signify cardiac arrhythmias, aiming to achieve a good classification rate. More specifically, many classification methods have been used in the field such as Bayesian classifiers, decision trees and neural and rule-based learners **[7].** However, the classification methods with good classification rates usually have a low degree of interpretability preventing the user (such as a cardiologist) from fully taking advantage of such a method. The interpretability of any knowledge-based system is crucial especially when dealing with big data applications **[8].** In fact, this ensures an effective decision-making progress by producing interpretable knowledge, which can easily be maintained and assessed. The acquisition of expert knowledge is a complex yet essential task because of its inherent accuracy, if carried out by human intervention. However, this tends to be inefficient when dealing with big datasets and, furthermore, such knowledge contains an unconscious component that is hard to formalise **[9].** Alternatively, knowledge can also be defined by analysing information extracted from experimental data, which is likely to provide an accurate insight into the different parameters. In particular, there is a wealth of algorithms and machine-learning techniques for model identification that are based on the properties of accuracy indices, and which can be applied to the knowledge induction process **[10].** Another valuable source of knowledge is based on articles and texts published in scientific journals which includes the most critical knowledge and through which many experts share their results, analysis and experiences. However, because such textual information is typically very large if not huge, scientists are faced with an overwhelming amount of information which poses a big computational and implementation challenge. In modern medicine large amounts of data are

generated but there is a widening gap between data acquisition and data comprehension. It is often impossible to process all of the data available and to make a rational decision on basic trends. Thus, there is a growing need for intelligent data analysis such as, data mining to facilitate the creation of knowledge in order to support clinicians in decision-making. In fact, data mining approaches can be used in such databases to improve classification tasks. In this thesis, we introduce a novel method to semi-automatically identify fuzzy partition rules applied to cardiac arrhythmia detection which combines an automated information extraction from textual sources with expert elicitation to create a robust, scalable and accurate knowledge-based system which provides a crucial insight into arrhythmia detections from big data information sources.

1.2 Medical context

An electrocardiogram (ECG) reflects the activity of the central blood circulatory system which can provide extensive information on the normal and pathological physiology of heart activity. As a consequence, it is an important non-invasive clinical tool for the diagnosis of heart diseases **[11].** Early and quick detection and classification of ECG arrhythmia are important especially for the treatment of patients in the intensive care unit **[12].** For over four decades, computer-aided diagnostic (CAD) systems have been used in the classification of the ECG resulting in a huge variety of techniques. Included in these techniques are multivariate statistics, decision trees, fuzzy logic, expert systems and hybrid approaches **[12].** In designing an effective and efficient CAD system the most important step is the integration of suitable feature extractor and pattern classifier so that they can operate in coordination **[13].**

The medical sector has always generated large amounts of data based on patient record, regulatory requirements and so on… **[14].** All the information created by the aforementioned data provides a valuable opportunity to improve further the state-of-the-art tools available to clinicians as well as to provide scalable, robust and efficient methods to extract, assess and manage such informations. Furthermore, the digitalisation of medical datasets with the implementation of big data techniques has created further benefits. These include the detection

of diseases at earlier stages resulting in more effective and successful treatments, as well as the management of specific individual and population medical information. Specific scenarios can also be predicted and estimated based on large amounts of historical data, combined with real-time information to determine and assess their crucial properties **[15].**

1.3 Main goals

- The main objective of this thesis is to define a new fuzzy modelling process that is able to achieve a good interpretability-accuracy trade-off through the cooperation between expert and induced knowledge. In order to achieve this goal, we have set the following intermediate targets:

- Analyse the fusion of knowledge from different sources.
- Ensure that the interpretability concept will be maintained in all the process
- Set a knowledge system where it will be possible to achieve high accuracy while also maintaining high interpretability.
- Propose a new methodology of knowledge fusion for building a common linguistic knowledge base.
- Test the new methodology in classification and automatics diagnosis of arrhythmias such as PVC.

1.4 Contributions

The method we are proposing and in particular the overall structure of the extraction process from expert knowledge, data and textual information is depicted in *Figure 1.1*. More specifically, fuzzy logic as a modelling platform allows us to merge and manage those three types of knowledge, where the fuzzy partition design aims to define the most influential variables according to the aforementioned knowledge. An important part of this process is the rule base (RB) definition and integration, where the expert is invited to describe the system behaviour, expressing his/her system knowledge as linguistic rules (expert rules). Furthermore, rules are induced from data (induced rules) according to the common universe of fuzzy

partition. Both types of rules use the same linguistic terms defined by the same fuzzy sets. As a consequence, rule comparison can be performed at the linguistic level and subsequently both types of rules are merged into a unique knowledge base.

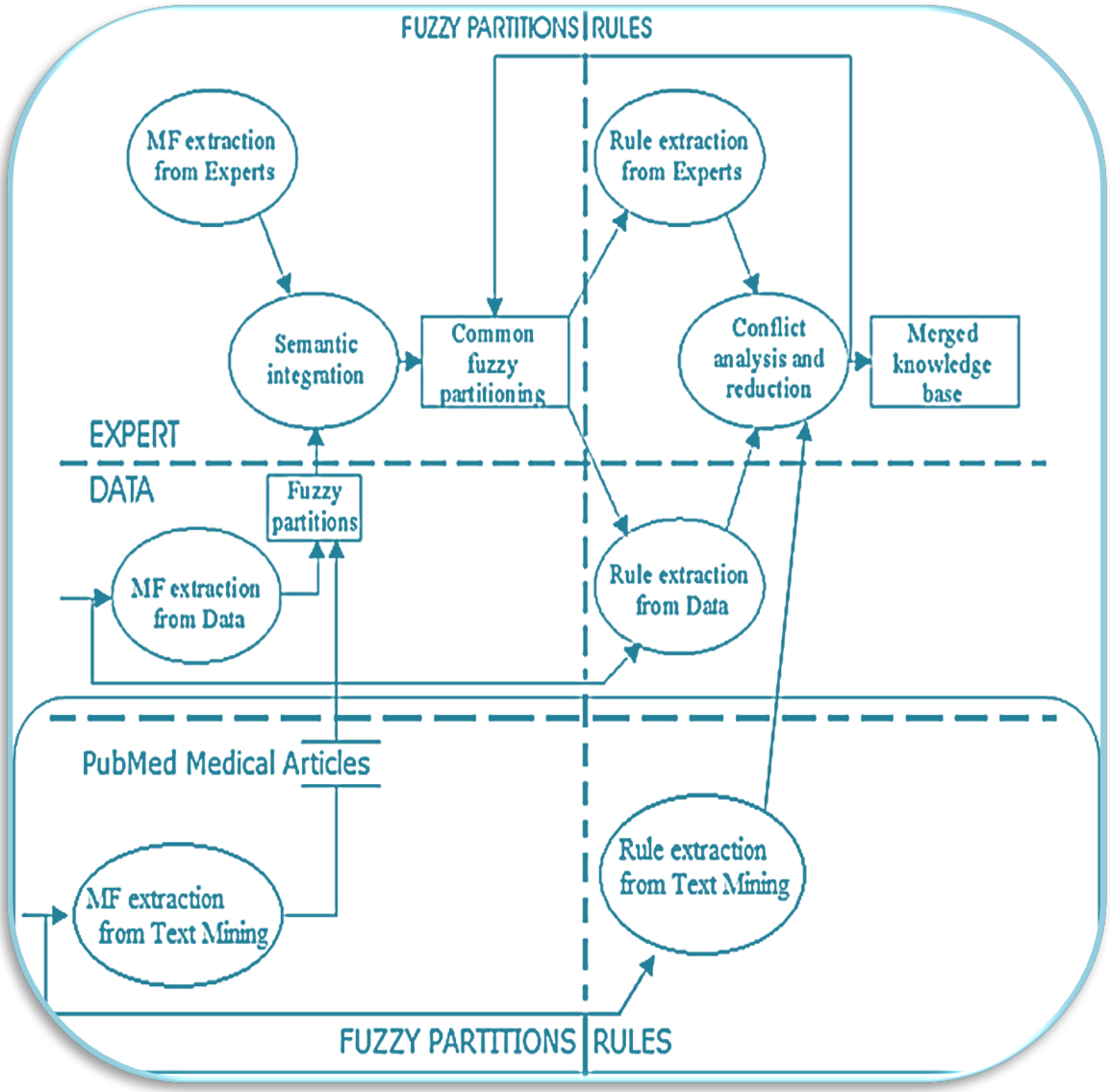


Figure 1.1 Overall structure of the knowledge extraction process.

As part of the process the expert can provide complete or partial information about the linguistic variables. Additionally, several algorithms can be used to create fuzzy partitions from data, or

induced partitions, and linguistic constraints are superimposed to the fuzzy partition definition in order to ensure their interpretability. The result is the definition of a common universe for each of the variables, according to both expert knowledge and data distribution.

1.5 Thesis structure

The remainder of this thesis is structured as follows:

- Chapter 1: we introduce our system and motivation and we define interpretability and the different ways to fuse knowledge from different sources.
- Chapter 2: we describe the clinical electrocardiography and highlight arrhythmias such as premature, junctional and auricular contraction and introduce also the cardiologist diagnosis.
- Chapter 3: we present all the methods and techniques used in our proposed system such as text mining and fuzzy logics.
- Chapter 4: we explain the experiments and the obtained results of classification problems which have been treated and we compare our results with those gained by popular methodologies.
- Chapter 5: we offer some conclusions and suggest options for further researches.

Chapter 2

2.1 INTRODUCTION TO CLINICAL ELECTROCARDIOGRAPHY:

The electrocardiogram is used as a diagnostic test, as well as the clinician's desire to determine cardiac abnormalities from the body surface potentials. As a rough framework [16], it is worth thinking of the heart as three separate systems: a functional electrical system, a functional system of coronary (or cardiac) arteries to channel nourishing blood to every cell of the myocardium, and as a culmination of these, an effective mechanical pump.

First, we consider how an ECG is used to assess electrical abnormalities of the heart: The surface ECG has inherent limitations as a diagnostic tool, since given a distribution of body surface potentials, we cannot precisely specify the detailed electro physiologic behaviour of the source since this inverse problem does not have a unique solution (as demonstrated in 1853 by Hermann von Helmholtz). It is not in general, possible to specify uniquely the characteristics of a current generator from the external potential measurements alone. Therefore, an exacting assessment of the electrical activity of the heart involves an invasive electrode study. Despite these inherent limitations, the surface ECG is extremely useful in clinical assessments of electrical pathologies and an invasive electro physiologic study is indicated in only a small fraction of cases.

To a first approximation, electrical problems come in two forms: those that make the heart pump too slowly or infrequently (bradycardias) and those that make the heart pump too quickly (tachycardias). If the pumping is too slow, the cardiac output of life-sustaining blood can be dangerously low. If too quick, the cardiac output can also be too low since the heart does not have time to fill and also because the heart can suffer damage (e.g., demand ischaemia) when it tries to pump too rapidly.

2.1.1 THE NORMAL DETERMINANTS OF HEART RATE: THE AUTONOMIC NERVOUS SYSTEM

One class of heart rate abnormalities arises from abnormal function of the heart rate control system. There are specialised cells in the SA node whose function is to act as the heart's pacemaker, rhythmically generating action potentials and triggering depolarisation for the rest of the heart (recall that once any portion of the heart depolarises, the wave front tends to propagate throughout the entire myocardium). The SA node has an intrinsic rate of firing but ordinarily this is modified by the central nervous system, specifically, the autonomic nervous system (ANS). The decision-making for autonomic functions occurs in the medulla in the brain stem and the hypothalamus. Instructions from these centres are communicated via nerves that connect the brain to the heart. There are two main sets of nerves serving the sympathetic and the parasympathetic portions of the ANS which both innervate the heart. The sympathetic nervous system is activated during stressful times and increases the rate of SA node firing (hence raising heart rate) and innervates the myocardium itself, increasing the propagation speed of the depolarisation wave front primarily through the AV node and increasing the strength of mechanical contractions. These effects are all consequences of changes to ion channels and gates that occur when the cells are exposed to the messenger chemical from the nerves. The time necessary for the sympathetic nervous system to actuate these effects is approximately 15 seconds.

The sympathetic system works in tandem with the parasympathetic system. For the body as a whole, this system controls quiet-time functions like food digestion. The nerve through which the parasympathetic system communicates with the heart is named the vagus. The parasympathetic branch's major effect is on heart rate and the velocity of propagation of the action potential through the AV node. Furthermore, in contrast with the sympathetic system, the parasympathetic nerves act quickly, decreasing the velocity through the AV node and slowing the heart rate within a second when they activate. Most organs are

innervated by both the sympathetic and the parasympathetic branches of the ANS and the balance between these competing effects determines function.

The sympathetic and parasympathetic systems are rarely totally off or on; instead, the body adjusts their levels of activation, known as tone as appropriate to its needs. If a medication that inactivates the sympathetic system (e.g., propranolol) is used on a healthy resting subject with a heart rate of 60 bpm, the classic response is to slow the heart rate to about 50 bpm. If a medication that inactivates the parasympathetic system (e.g., atropine) is used, the classic response is an elevation of the heart rate to about 120 bpm. If you administer both medications and inactivate both systems (parasympathetic and sympathetic withdrawal), the heart rate rises to 100 bpm. Therefore, in this instance for normal subjects at rest, the effects of the heart rate's "brake" are greater than the effects of the "accelerator", although it is the balance of both systems that dictates the heart rate. The body's normal reaction when the vagal tone is increased (the brake) is to reduce sympathetic tone (the accelerator) simultaneously. Similarly, when sympathetic tone is increased, the parasympathetic tone is usually withdrawn. Indeed, if a person is suddenly startled, the earliest increase in heart rate will simply be due to parasympathetic withdrawal rather than the slower-acting sympathetic activation.

So on, what basis does the autonomic system make heart rate adjustments? There are a series of sensors throughout the body that send information back to the brain known as afferent nerves, which bring information to the central nervous system. Those parameters sensed by afferent nerves include the blood pressure in the arteries (baroreceptors), the acid-base conditions in the blood (chemoreceptors) and the pressure within the heart's walls (mechanoreceptors). Based on this feedback the brain unconsciously adjusts heart rate.

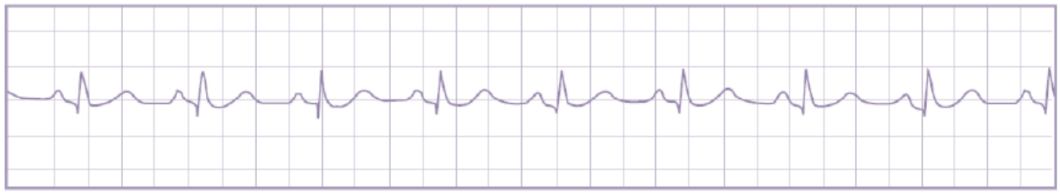


Figure 2.1 Normal sinus rhythm [17]

The system is predicated on the fact that as heart rate increases, cardiac pumping and blood output should also increase, in turning raising arterial blood pressure, blood flow and oxygen delivery to the peripheral tissues, carbon dioxide clearance from the peripheral tissues, and so on.

When the heart rate is controlled by the SA node's rate of firing, the sequence of beats is known as a sinus rhythm (see Fig. 2.1). When the SA node fires more quickly than usual (for instance, as a normal physiologic response to fear or an abnormal response due to cocaine intoxication), the rhythm is termed sinus tachycardia (see Fig. 2.2). When the SA node fires more slowly than usual (for instance, either as a normal physiologic response in a very well-conditioned athlete or an abnormal response in an older patient taking too much heart-slowing medication), the rhythm is known as sinus bradycardia (see Fig. 2.3). There may be cyclic variations in heart rate due to breathing, known as sinus arrhythmia (see Fig. 2.4). This non-pathologic pattern is caused by activity of the parasympathetic system (the sympathetic system responds too slowly to alter heart rate on this time scale), which is responding to subtle changes in arterial blood pressure, cardiac filling pressure and the lungs themselves during the respiratory cycle.

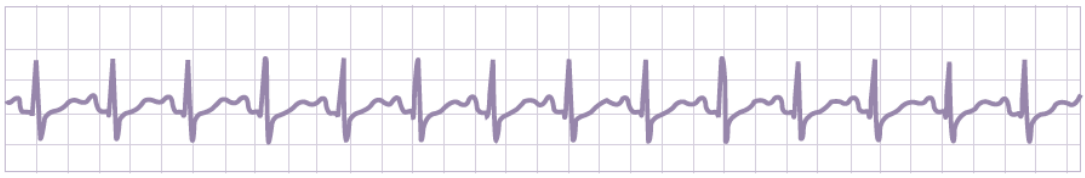


Figure 2.2 Sinus tachycardia [17]

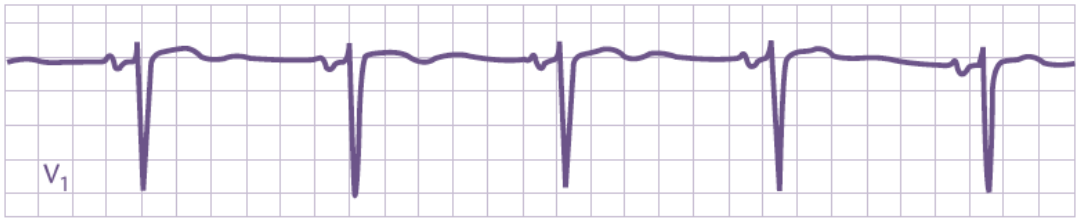


Figure 2.3 Sinus bradycardia [17]

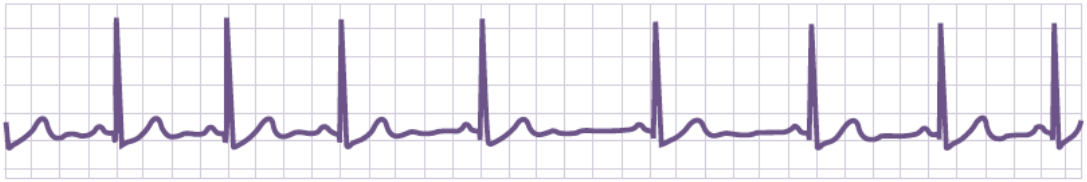


Figure 2.4 Sinus arrhythmia [17]

2.1.2 Ectopy, Tachycardia, and Fibrillation

An arrhythmia is any abnormal cardiac rhythm. One category of arrhythmias occurs when the trigger to depolarise originates outside of the SA node in another part of the myocardium, known as ectopic depolarisation leading to ectopic beats. Common causes of ectopy include a drug effect (e.g., caffeine) or a viral infection of the myocardium or other inflammation or damage of part of the heart (e.g., ischaemia). When the ectopic beat originates in the atria, it leads to a premature atrial beat also known, as an atrial premature contraction (APC) (see Fig. 2.5). When it originates in the ventricles, it leads to a premature ventricular beat or ventricular premature contraction (VPC) (see Fig. 2.6). Note in Figure 2.6 that the ectopic ventricular beat looks very different from the other sinus beats. The spread of the wave front for a VPC can be backwards, such as when the action potential starts at the apex of the heart rather than the septum. The depolarisation wave front can move in very different directions than the typical sinus-driven heart vector.

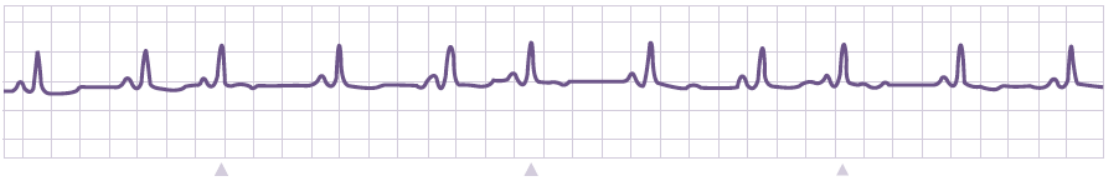


Figure 2.5 Atrial premature contractions (indicated by arrowheads) [17]

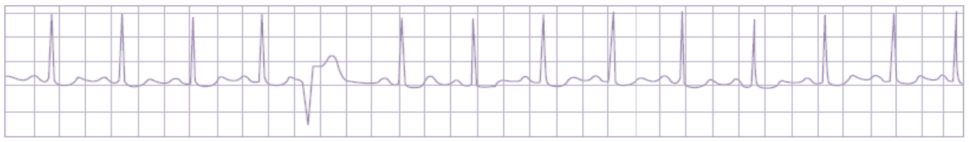


Figure 2.6 Ventricular premature contraction [17]

Furthermore, the ectopic beat is typically wider because its wave front propagates slowly through the myocardium rather than through the high-speed Purkinji system.

After an ectopic wave front has propagated to all portions of the heart, the myocardium is left temporarily depolarised. After a pause, the tissue repolarises and the regular sinus mechanism can instigate a subsequent beat. The conditions that caused the ectopic beat might persist (e.g., still too much caffeine leading to an excitable myocardium) but the ectopic beat itself is left behind in the past. Sometimes, a semi-stable pattern of sinus beats and ectopic beats develops. For instance, a repeating pattern of "sinus beat – VPC – sinus beat – VPC – and so forth" can occur, termed ventricular bigeminy (see Fig. 2.8). There can be so many ectopic beats ongoing that the overall heart rate is driven much higher than normal.

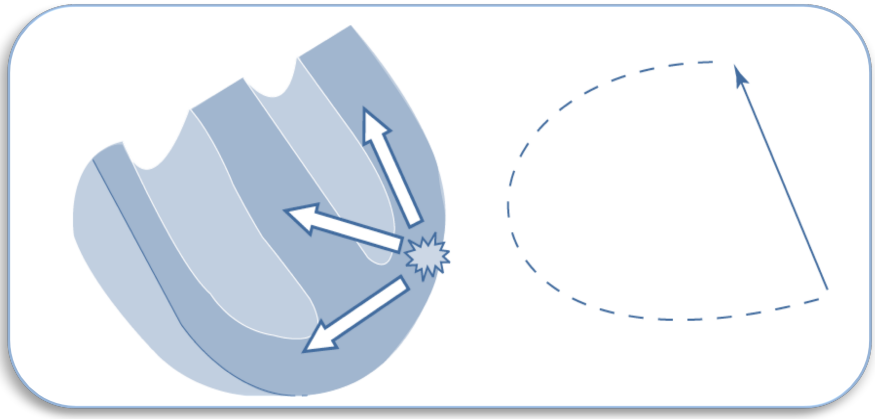


Figure 2.7 Wavefront trajectory in a ventricular premature contraction [17]

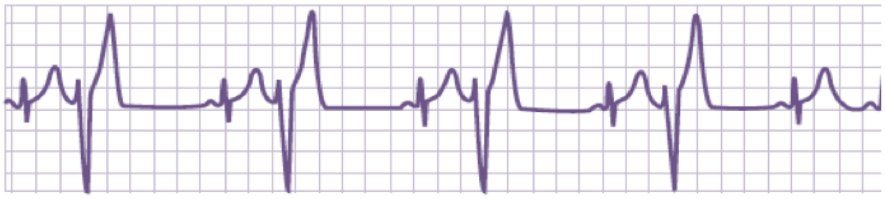


Figure 2.8 Ventricular bigeminy [17]

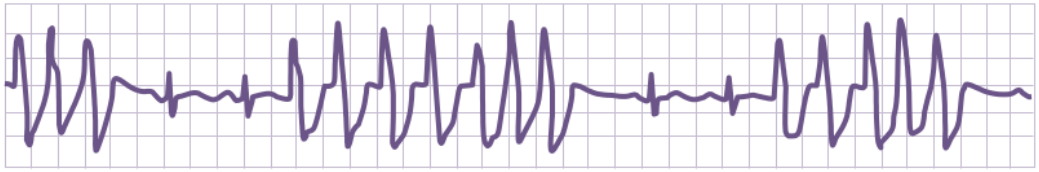


Figure 2.9 Three episodes of non-sustained ventricular tachycardia [17]

The potential for serious problems is higher for those less common conditions in which a wave front continues to propagate in a quasi-stable state, circulating repeatedly through the heart leading to repeated waves of tissue depolarisation at an abnormally high rate [15] [16]. When this cyclic, quasi-stable phenomenon occurs in the heart, it is called a re-entrant arrhythmia. A classic example often caused by a re-entrant pattern is ventricular tachycardia (VT) (see Fig. 2.9). These states are extremely pathologic and can be rapidly fatal because the rate of depolarisation can be incompatible with effective cardiac pumping. At one extreme (rapid VT in an older frail heart), VT can be fatal in seconds to minutes. At the other extreme (slow VT in a younger healthy heart) the cardiac output can remain at a life- sustaining level.

The ECG criteria for VT are three or more consecutive ectopic ventricular beats at a rate over 100 bpm. If the VT terminates within 15 seconds (or 30 seconds, by some conventions) it is known as no sustained VT; otherwise it is sustained VT.

Because of the grave medical consequences of VT and related re-entrant arrhythmias, they have been well investigated experimentally, clinically and theoretically.

Figure 2.10 Atrial fibrillation-------two examples.

In some cases, the unified wave front of depolarisation can break down into countless smaller wave fronts, which circulate quasi-randomly over the myocardium. This leads to a total breakdown of coordinated contraction and the myocardium will appear to quiver, which is termed fibrillation. In atrial fibrillation (see Fig. 2.10), the AV node will still act as a gatekeeper for these disorganised atrial wave fronts, maintaining organised ventricular depolarisation distal to the AV node with normal QRS complexes. The ventricular rhythm is generally quite irregular and the rate will often be elevated. Atrial fibrillation is frequently well tolerated, provided the consequent ventricular rate is not excessive. AF can lead to a minor impairment in cardiac output due to reduced ventricular filling. In the long term, there can be regions in a fibrillating atrium where, because of the absence of contractions, the blood sits in stasis and this can lead to blood clot formation within the heart. These clots can re-enter the circulation and cause acute arterial blockages (e.g., cerebrovascular strokes) and therefore patients with atrial fibrillation are often

anticoagulated. In contrast to atrial fibrillation, untreated ventricular fibrillation (see Fig. 2.11) is fatal within seconds to minutes. The appearance of fibrillating ventricles has been likened to a "bag of worms" and this causes circulatory arrest which is the termination of blood flow through the cardiovascular circuit.

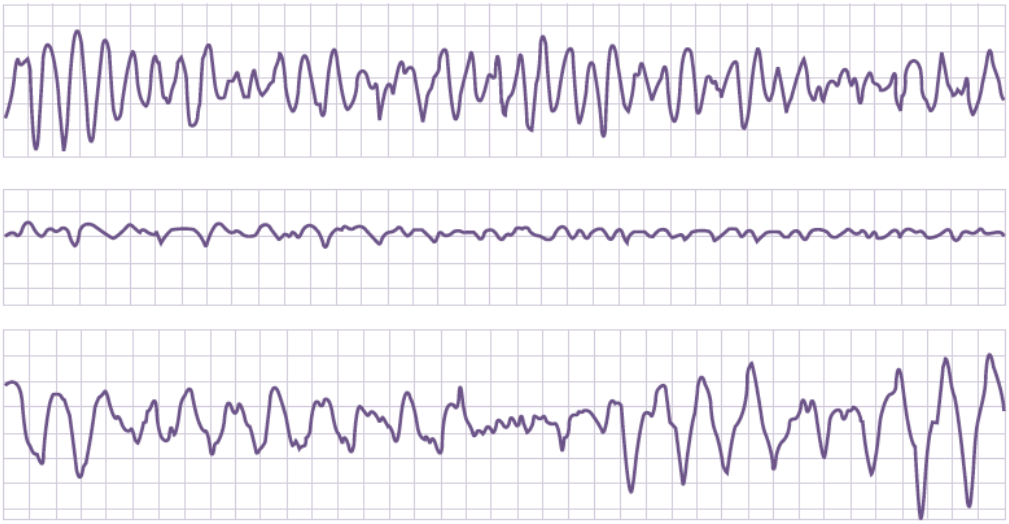


Figure 2.11 Ventricular fibrillation --three examples.

2.1.3 Conduction Blocks, Bradycardia, and Escape Rhythms

The other category of arrhythmias is related to excessively slow rhythms and abnormal blockages of wave front propagation. For instance, an aging SA node may pace the heart too slowly, leading to low blood pressure and weakness or fainting. Assuming this is not a result of excessive medication, these symptoms might require that an artificial pacemaker be implanted. The AV node may also develop conduction pathologies that slow the ventricular heart rate, it may fail to conduct some atrial wave fronts (termed second-degree AV block, see Fig. 2.12), or it may fail to conduct all atrial wave fronts (termed third-

degree AV block, see Fig. 2.13). Assuming this is not a result of excessive medication, most third-degree and some second-degree AV blocks require pacemaker therapy. In first-degree AV block, the AV node conducts atrial wave fronts with an abnormal delay but because the atrial wave fronts are eventually propagated to the ventricles, first-degree AV block does not slow the overall ventricular rate.

Sections of the specialised conducting fibres in the ventricles can also fail, such that depolarisation waves must reach some portions of the ventricles via slower routes.

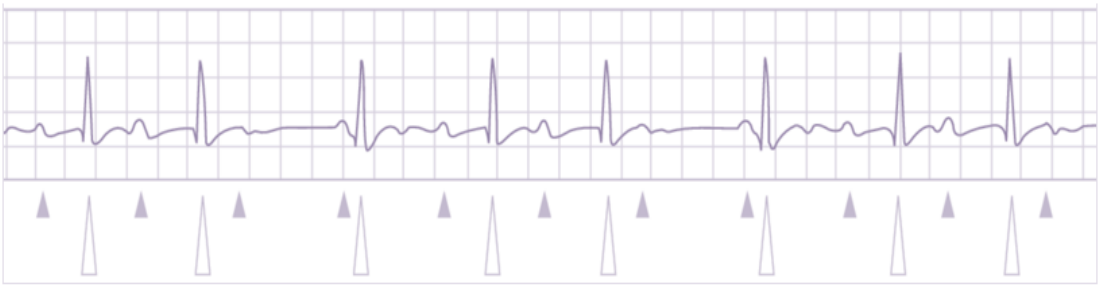


Figure 2.12 Second-degree AV block. In this subtype of second-degree AV block, termed Wenckebach or Mobitz Type I, there is a characteristic lengthening of the delay between the atrial P wave and the ventricular QRS and ultimately there is a failure to conduct a P wave. Then this cycle repeats. In the example illustrated, there are three P waves (small purple arrowheads) followed by ventricular beats (large white arrowheads), and then the AV node fails to conduct the fourth P wave in each cycle (small purple arrowheads without subsequent large white arrowheads).

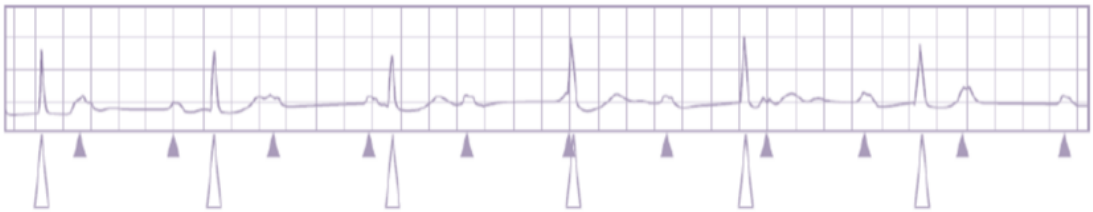


Figure 2.13 Third-degree AV block. There is a failure of the AV node to conduct any wave fronts from the atria to the ventricles. The ventricular beats are escape beats, originating electrically from the specialised conducting fibres just below the AV node. The ability to generate escape beats is the heart's fail-safe mechanism for what would otherwise cause fatal cardiac (e.g., ventricular) arrest. Notice there is no relationship between the atrial P waves (small purple arrowheads) and the junctional escape beats (large white arrowheads). See Figure 2.15 for an example of a ventricular escape beat. [17]

muscle-to-muscle propagation. There are a classic set of changes associated with failures of different conduction bundles (e.g., right bundle branch block and left bundle branch block, see Fig. 2.14). These blocks usually have a minimal effect on pumping efficacy. However, they can dramatically change the cardiac vector's trajectory and hence the surface ECG and can mask other ECG changes indicative of disease (e.g., ischaemia). In some cases, these conduction abnormalities indicate some other underlying pathology of great importance (for instance, a pulmonary embolism can cause a new right bundle branch block and acute anterior ischaemia can cause a new left bundle branch block).

The topic of bradyarrhythmias and heart blocks leads to the topic of escape beats (see Figs. 2.13 and 2.15). An escape beat is similar to an ectopic beat, in that it is the initiation of a depolarisation wavefront outside of the SA node. However, the difference is that the escape beat is a normal and compensatory response, a normal fail-safe functionality of the heart.

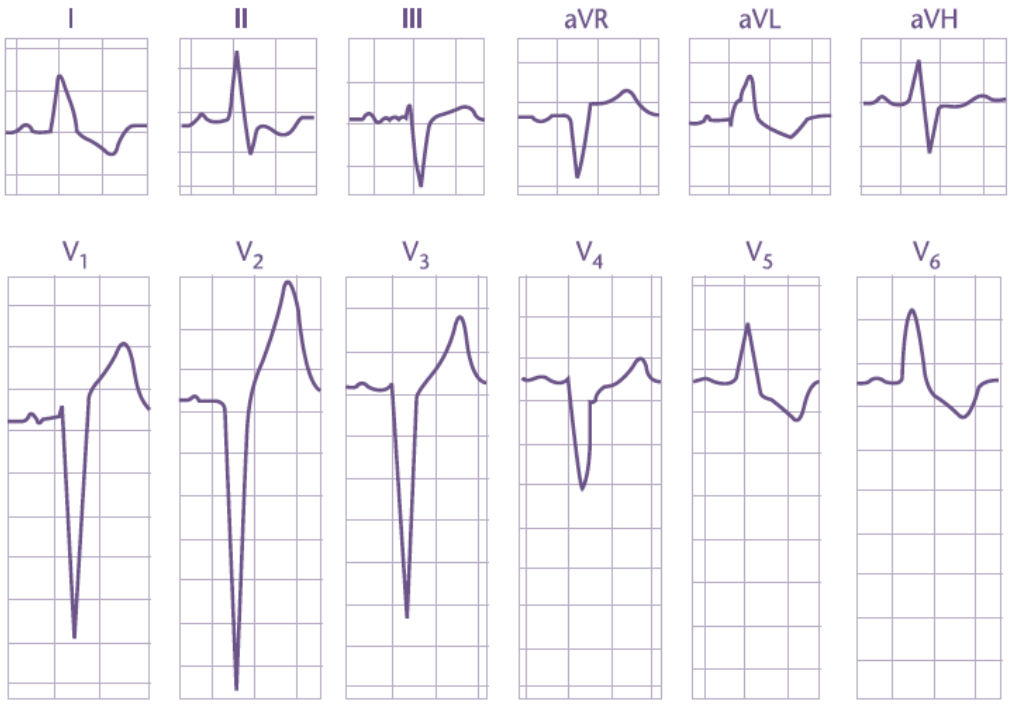


Figure 2.14 Classic ECG pattern of left bundle branch block

There is a network of cardiac cells able to initiate heart beats, so that a key life-sustaining function (e.g., pacing the heart) is not exclusively relegated to a microscopic collection of cells in the SA node. The cells in the backup system have intrinsic rates of firing that are slower than the SA node. So while the latter is given the opportunity to pace the heart, the other regions will initiate heart beats if there has been an excessively long pause in the SA node. Therefore, a ventricular beat that occurs in the setting of a third-degree heart block is termed a ventricular escape beat and it represents an appropriate and healthy response to the lack of any other pacemaker trigger. When non-sinus beats appear on the ECG, it is important to differentiate ectopic beats (where the pathology is aberrant automaticity) from escape beats (where the pathology is abnormal conduction) because the treatments are quite different. Note that ectopic beats are generally premature whereas escape beats terminate a prolonged RR interval.

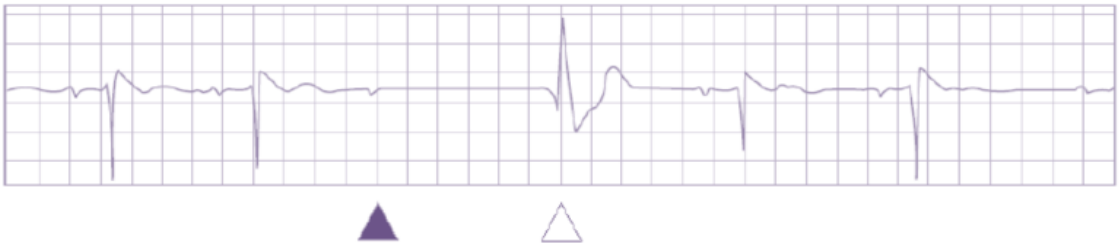


Figure 2.15 Ventricular escape beat. Note the atrial P wave (purple arrowhead) followed by an evident pause, indicating a failure to conduct through the AV node. The ventricular escape beat (white arrowhead) is a fail-safe mechanism so that con duction blocks do not cause ventricular cardiac arrest. See Figure 2.6 for more information.

2.1.4 Cardiac Ischaemia, Other Metabolic Disturbances, and Structural Abnormalities

The ECG can reveal metabolic abnormalities of the myocardium. The most medically significant abnormality is ischaemia, which ultimately progresses to myocardial cell death and which occurs when part of the myocardium is not receiving enough blood flow, often caused by disease of the coronary arteries. Recording a 12-lead ECG is a standard-of-care diagnostic test in any evaluation of possible cardiac ischaemia. Ischaemia often changes the appearance of the T wave and the ST interval because of a current of injury between the ischaemic and non-ischaemic myocardium which alters the main cardiac vector. There are classic patterns of ischaemia, which are seen in only a minority of ischaemic events (see Fig. 2.16). In most such events there are "nonspecific" ECG changes such as, changes in the T wave that may or may not be caused by ischaemia. In a small percentage of cases, there can be ischaemia without any grossly evident ECG changes. Overall, the ECG is not highly sensitive nor specific for cardiac ischaemia, but larger regions of cardiac ischaemia are associated with more substantial ECG changes (as well as a higher mortality rate), so this information is very useful in decision-making. For example, a patient with classic changes of acute ischaemia is often a good candidate for powerful thrombolytic (clot dissolving) medication, whereas patients without those classic changes are not considered good candidates because on average, the risks of the drug therapy, including intracranial haemorrhage and internal bleeding outweigh the benefits.

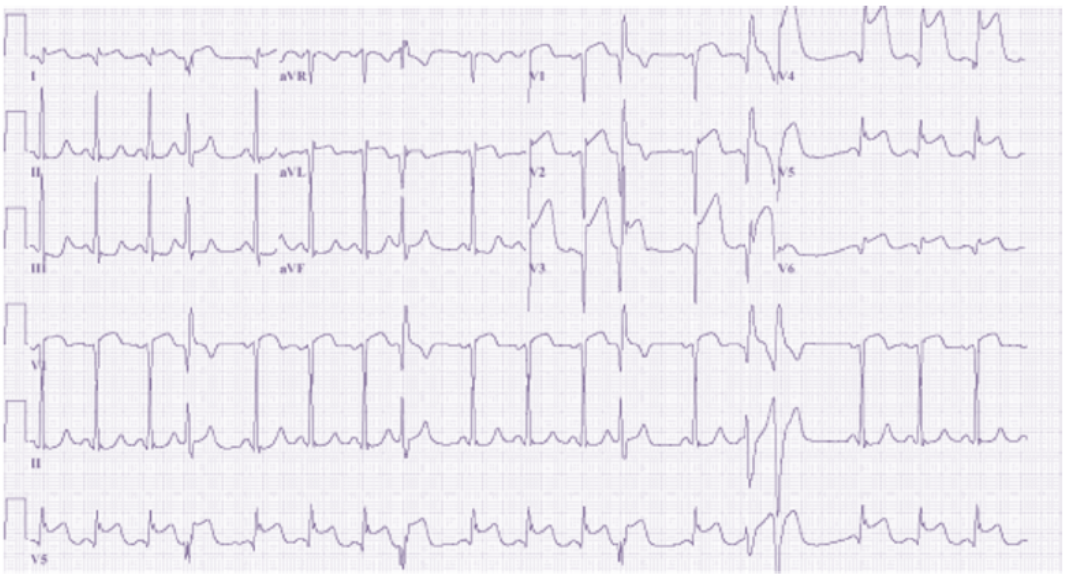


Figure 2.16 Acute myocardial infarction. Large areas of ischaemic anterior myocardium often produce ST-segment elevation in multiple contiguous precordial ECG leads (or in all the precordial leads, in this dramatic example). Also note there is minor ST-segment depression in the inferior lead III, which in this context is referred to as a ''reciprocal change".

Other metabolic abnormalities that cause characteristic changes in the ECG include electrolyte abnormalities. The classic indicators of high serum potassium levels (hyperkalemia) include a high pointed T wave and ultimately, the loss of the P wave and distortion of the QRS (see Fig. 2.17). Hypokalemia causes an undulation after the T wave called a U wave. Calcium and magnesium disturbances, and also extreme cold body temperature are other causes of ST–T wave abnormalities.

Figure 2.17 Hyperkalemia (moderate/severe). The K^+ was 10.5 mEq/L in a patient with renal failure. Note the loss of P waves and the widening of the QRS complex. There are numerous classic ECG morphologies associated with hyperkalemia. This example shows what Marriott has termed a "dumping pattern" because "it looks as though a rotund body has been dumped in the hammock of the ST segment […] making the ST segment sag horizontally […]and verticalising the proximal limb of the upright T wave.

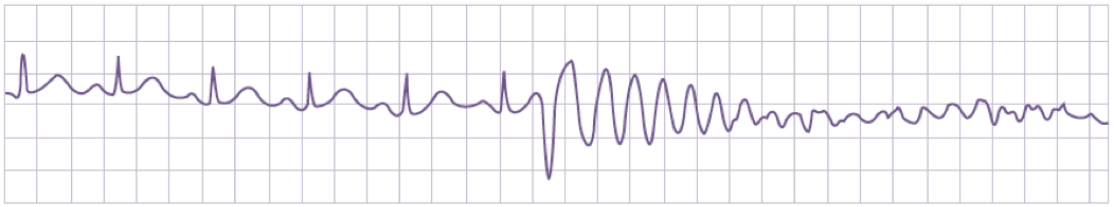


Figure 2.18 An R-on-T ventricular premature beat initiates polymorphic ventricular tachycardia

Therapeutic drugs can also alter the appearance of the ECG. One alteration of significance is a delay in the T wave relative to the QRS, the so-called QT prolongation. A prolonged 20 QT indicates a myocardium at risk for triggered activity in which a cardiac cell will rapidly and repeatedly depolarise, associated with a kind of dangerous ventricular tachycardia called *torsades des pointes*. In the so-called R-on-T phenomenon, the heart depolarises in the midst of its repolarisation which can trigger a life-threatening tachyarrhythmia (see Fig. 2.18). R-on-T can occur because the QT interval is abnormally long. It can also occur because of a mechanical blow to the chest during repolarisation (e.g., a young hockey player struck in the chest) and relevant to signal processors, even it can occur if a patient receives electrical cardioversion in the middle of repolarisation when the T wave occurs. Hence, these devices are imbued with a synchronisation mode in which the QRS complex is automatically identified and the device is prevented from discharging in the temporal vicinity of the following T wave.

The ECG can reveal abnormalities in the geometry of the heart or pathologies in which part of the heart has become enlarged (too much myocardium) or has undergone cell death and scarring (electrically absent myocardium). Non-invasive echocardiography is now the standard method by which such abnormalities are diagnosed and so in general the ECG has been relegated to a convenient if imperfect screening test for structural abnormalities of the heart. Examples of conditions that are often apparent in the surface ECG include:

- Thickening of the ventricular walls caused by years of beating against high arterial blood pressure (hypertrophy);
- Ballooning of the ventricular walls caused by accommodating large volumes of blood regurgitated from upstream when there is a leaky, incompetent valve;
- Ventricular wall aneurysms that form after heart attacks;
- Scars in the heart (after heart attacks) that cause part of the heart to be electrically silent;

- Abnormal fluid collecting around the heart (termed an effusion) that can impair the heart's ability to fill and hence pump effectively (cardiac tamponade).

These conditions change the trajectory and/or magnitude of the normal heart vector, which can distort the normal ECG morphology. A review of all these conditions and their resultant ECG appearance is beyond the scope of this chapter and introductory or advanced textbooks focused solely on the clinical interpretation of the ECG are plentiful [17]. However, in the following section a brief overview of how clinicians analyse the ECG is presented. This should not be taken as a definitive guide but as a general strategy that may prove useful in designing an ECG analysis algorithm.

2.1.5 A Basic Approach to ECG Analysis

In analysing the clinical electrocardiogram, it is important to use a systematic approach. The following overview which illustrates a clinical approach, should not be considered completely thorough but simply as a guide to understanding how clinicians identify abnormalities in the ECG.

1. ***Identify the QRS complexes.*** The following observations should be made:

- What is the ventricular rate?
- Are the QRS complexes spaced at regular intervals? If not, what is the nature of the irregularity?
- Are the QRS complexes identical in shape in a given lead? Are they of normal size and morphology?

2. ***Identify the P waves.*** In some cases, this will require careful observation and more than one lead axis may be necessary. The following questions should be explored :

- Is there a one-to-one relationship between P-waves and QRS complexes? If not, is there a definable pattern ?
- Is the PR interval of normal duration?
- What is the atrial rate?
- Are the P waves identical in shape in a given lead? Are they of normal size and shape?

Based on the above analysis, it should be possible to identify the mechanism of the rhythm in most cases.

3. ***Examine the QRS complex in each lead.*** Is the QRS axis normal? Overall, are the QRS widths and amplitudes normal? Often, the QRS complexes are viewed in several "groups" that are specific to a particular region of the heart. The waveform patterns should also be checked

for signs of intraventricular conduction block, significant amplitude Q waves, and precordial R-wave pattern normality.

4. ***Examine the ST-T segments.*** Are there abnormalities such as elevation or depression, and is the abnormality suggestive of ischaemia, infarction or hypothermia?

5. Examine the T waves. Are their shapes normal? In each lead, are they oriented in the same direction as the QRS complex? If not, is it suggestive of ischaemia, ventricular conduction abnormalitiesor a potassium abnormality such as hyperkalemia?

6. ***Examine the QT interval.*** Is it over half the RR interval?

Once an abnormality is identified, there are often several potential explanations, many of which lead to several ECG pathologies, and it may not be possible to determine the significance of the abnormality with certainty. To confirm a potential diagnosis from the ECG, other characteristic abnormalities are often sought. For a given individual, comparing a new ECG with a prior ECG provides an invaluable reference, particularly if trying to ascertain when ECG abnormalities are acute or not. For instance, for a patient with chest pain, abnormal ST-T wave patterns are much more concerning for ischaemia when they are new (e.g., not present in an ECG recorded one year earlier). If the ST-T wave patterns existed long before the new symptoms, one may deduce that these patterns are not directly related to the acute pathology.

Chapter 3

3.1: Bayesian Belief Network Extraction from Text

3.1.1 Introduction

As graphical models, Bayesian Belief networks (BBNs) capture independence relationships among random variables and are a popular modelling framework in risk and decision analysis BBNs have been employed in a variety of applications such as safety assessment of nuclear power plants, risk evaluation of a supply chain and medical decision support tools [26]. In BBN modelling, the strong statements are not about dependencies but rather about independences (i.e. absence of edges at the graph level) as it is always possible to capture them through the conditional probability tables when an edge is present, even though the reverse is not true. The construction of a BBN can be done either through data or when unavailable, literature reviews or expert elicitation. While the first approach can be automated, the other two require a significant amount of manual work that can make them impractical on a large scale.

There is extensive research on BBNs and in particular, their extraction from text corpora is increasingly attracting attention. In [27], the authors suggest a domain-independent method for acquiring text causal knowledge to generate BBNs. Their approach is based on a classification of lexico-syntactic patterns that refer to the causation described in [23] where an automatic detection of causal patterns and semi-validation of their ambiguity is carried out. Similarly, in [29] a supervised method for the detection and extraction of causal relations from open domain texts is presented. The authors provide an in-depth analysis of verbs, cue phrases that encode causality and to a lesser extent, influence. A thorough investigation of text patterns and cue phrases is described in [30], providing good accuracy in relation to extraction. In [31], the authors discuss BBN extraction from text based on [32] and also include an investigation of probability extraction and discuss its assessment by considering a variety of measures to assess probabilistic information. In [32], a light and fast causal relation tool based on an architecture to extract causal relationships automatically is described. This approach is aimed to produce

Bayesian network fragments by improving the existing causal relation identification through text patterns. However, the authors do not attempt any verb disambiguation, which leads to a low precision percentage. In this chapter we will describe a framework based on natural language processing (NLP) techniques to build Bayesian Networks from text corpora within the biomedical domain, providing decision makers with a method to generate a skeleton automatically of a Bayesian network based on a text corpus. To this end, we have widened the parameters used for the relation extraction by including an influence that is often regarded as a generalisation of causality. In fact, we believe that by considering only causal relations it is not possible to address fully the BBN extraction from text. In particular, influence relation extraction has been investigated significantly less often in comparison with causality, mainly due to the former's intrinsic ambiguity and vagueness. One of the main advantages of this approach is that more general information can be extracted from the text, which facilitates the generation of more accurate BBNs.

3.1.2 Bayesian Belief Networks

3.1.2.1 Identification of Dependence Relations

While BBNs represent dependence relationships between random variables, an intuitive (and commonly used) way of building a Bayesian network is to rely on cause-and-effect relationships among the variables [27, 28]. However, any BBN model is based on a wider range of relations that also include causality. Thus, an important aspect of this work is that it addresses the need for more comprehensive lexical rules to capture influence as well as causality. In [29], the authors provide a good description of the semantic of influence, pointing out that the distinction between causation and influence is that the latter only affects the manner and the intensity of the referent, whereas the former also affects its occurrence. However, influence terms typically carry a higher level of ambiguity since their semantic range is clearly wider than those only expressing causality.

Another important point in the generation of BBNs from text is that in general there is some level of inconsistency among the information extracted. When a large corpus of documents is analysed, the likelihood of inconsistencies increases significantly. These inconsistencies, which

can include cycles or contradicting extraction results, can arise from lexical properties, for example when dependence and independence relations are extracted between two concepts within the same sentence or when the information obtained from the text analysis is aggregated at a later stage.

3.1.2.2 Dependence Relations

Usually in BBN literature, dependence relations are defined over mathematical domains. However, this current study suggest that causal and influence relations have linguistic domains. Thus, to avoid any confusion, in this part of chapter we will refer to a dependence relation as a linguistic entity containing both causal and influence relations between concepts in text corpora.

We manually assessed several verbs from the biomedical domain to identify dependence relations that may be associated with a dependence relation between two or more concepts. These were then classified as causal or influence verbs and further divided into subsets according to their level of ambiguity by considering their frequency of usage and the number of senses that refer to non-relation cases. Cue phrases were found in a similar way and subsequently classified according to their ambiguity. In [28], the authors analyse some causal verbs and propose a classification based on their ambiguity and usage. However, we used WordNet in our analysis, which is appropriate to both the study of linguistics and BBN modellers, in order assessing specifically the lexical and semantic properties of the verbs and cue phrases we identified.

3.1.2.3 Independence Relations

It can be a complex and challenging task identifying the independence relations correctly, as they often appear in implicit ways. Thus, the use of text patterns to identify independence is less successful than when applied to causal and influence extraction. In this chapter, we focus on explicit independence relation extraction as follows:

- **Explicit text patterns and cue terms.**
- **Negations.** depicts the negation terms which, when used with causal and influence verbs, are likely to indicate independence relations.

3.1.2.4 Assessment of Relations

The use of the above text patterns may produce results containing inconsistencies, especially when the extraction is carried out over medium/large sized text corpora. Therefore, in order to generate meaningful information for obtaining a BBN, we need to assess the relations extracted. We propose the following method:

> Let a and b be two concepts from a text corpus, and let C (a, b), F (a, b), and I (a, b) be a causal, influence, and independence between a and b, respectively.

Suppose we have both F (a, b) (or C (a, b)) and I (a, b) for the two concepts a and b (i.e., they are both dependent and independent). We will assume we have I (a, b). The reason is that when both a negation and a dependence verb (or a cue term) are present in a sentence, this may result in the extraction of both a dependence and an independence relation. As an example, in the case of smoking does not cause cancer, we may have both C (smoking, cancer) and I (smoking, cancer).

Suppose we have both F (a, b) and C (a, b). In such case, we assume that two concepts are linked by F (a, b), since influence tends to have a wider linguistic range as discussed above.

Suppose there are two causal relations C1(a, b) and C2(a, b) and that the ambiguity level of the latter is greater than the former. Then we will take C1(a, b) (i.e. lower ambiguity wins over a greater one). The same applies for influence.

3.1.3 Algorithm

This is a grammar-based extraction, defined by text patterns. These are rules, which identify sentences with a specific structure.

The pattern we considered was of the form (NP, verb, NP) where NP refers to the noun phrases (which contains nouns, proper names, etc) and verb to the linking verb. For example, a sentence such as "PVCs can be related to electrolytes" would be captured by the above pattern, where

- 'PVC' is the first NP
- 'can be related' is part of verb
- 'electrolytes' is the second NP

The implementation of the above text pattern is via the following algorithm:

1. Let T be a text
2. Tokenise and parse T to create a syntactic hierarchical structure of each sentence in T. In other words, each sentence is analysed and each word is given its syntactic role such as noun, verb, punctuation, etc.
3. Finally, via regular expressions, only sentences where two biomedical words are linked by a verb are identified.
4. Each sentence is then analysed via sentiment analysis, so that it is given a number between -1 and 1, depending on how 'positive' it is. A sentence that describes something bad (e.g. 'I hate wind') will be given a number close to -1, whereas positive sentences ('I love sunny days') will be given a number close to 1. This is based on a simple algorithm, which is defined by keywords expressing an emotion. Keywords such as 'hate', 'bad', 'ugly', etc. express a negative emotion, whereas keywords such as 'love', 'like', 'beautiful', etc. express a positive emotion
5. The output is a list of three-tuples of the form [[noun 1, noun 2, polarity], ….], i.e. couples of words from the biomedical domain, which are joined by a verb, with the polarity (the sentiment analysis measure between -1 and 1, as described above)

We then applied the above algorithm to texts from PubMed, so that the names below

- Premature ventricular contractions, or PVCs,
- Premature ventricular complexes,

- Ventricular premature beats,
- Extrasystoles,

were parts of the NPs.

3.2 Fuzzy Inference Systems

3.2.1 Fuzzy partition design

The readability of fuzzy partitioning is a pre-requisite condition in building an interpretable rule base. The use of linguistic variables favours the readability and each system variable is described by a set of linguistic terms, modelled as fuzzy sets that form Strong Fuzzy Partitions (SFP). Figure 3.1 shows an example of SFP with five terms. It can include membership functions of several shapes, although triangular, trapezoidal, and semi-trapezoidal only in the edges [9]. This kind of partition satisfies all the semantic constraints (distinguishability, normalisation, coverage, overlapping, etc.) with a coverage level of $\epsilon = 0{,}5$ and the next equation:

$$\forall_x \in U, \sum_{i=1}^{M} \mu_{A_i}(x) = 1 \qquad 3.2.1$$

where U=[Ul, Uu] is the universe of discourse, M is the number of linguistic terms, and $\mu_{A_i}(x)$ is the membership degree of x to the A_i fuzzy set.

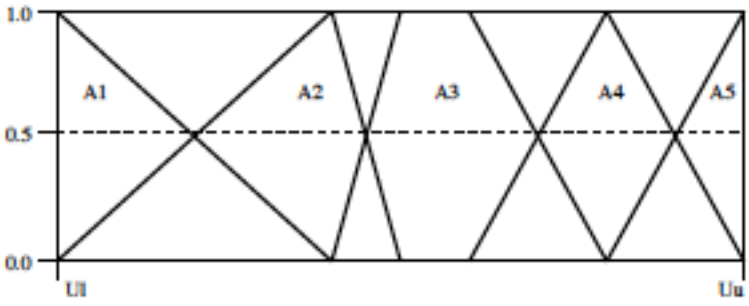


Figure 3.1 A strong fuzzy partition

As illustrated in Figure 3.1, each fuzzy set Ai is defined by four characteristic break points (Sleft, Cleft, Cright, Sright), which determine the limits of the support (S) and the core (C) for each set. If only three parameters are needed, then Cleft = Cright = C. Three contiguous fuzzy sets (Ai-1, Ai, Ai+1) have boundaries such as:

$$\begin{cases} A_i(S_{left}) = A_{i-1}(C_{right}) \\ A_i(C_{left}) = A_{i-1}(S_{right}) \\ A_i(C_{right}) = A_{i+1}(C_{left}) \\ A_i(S_{right}) = A_{i+1}(C_{left}) \end{cases}$$

When working with fuzzy sets in software applications, it is usual to consider parameterised membership functions 2D and it is important to choose functions reducing the computational cost. A trapezoidal membership function, for instance picture (a) appearing in Figure 3.2, is usually preferred because it can be defined by using only four 2D parameters (and in most cases, only four real parameters as shown in the picture). These kinds of curves are also interesting because they may degenerate into semi-trapezoidal and triangular functions, which are defined by only three parameters [37], [40].

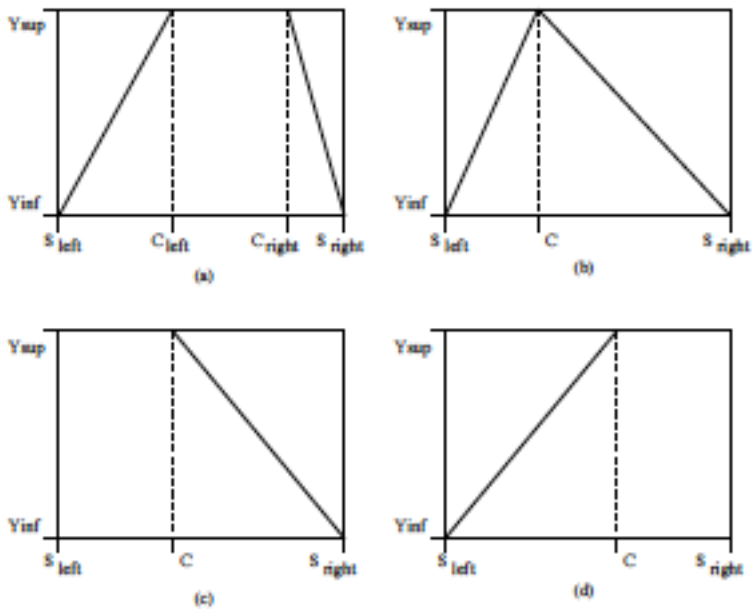


Figure 3.2 Membership functions

However, their main advantage against other well-known curves like Gaussian functions is their comprehensibility. Trapezoidal functions have a finite support while Gaussian ones are infinite which means an overlapping of all the fuzzy sets. The use of SFP with trapezoidal functions guarantees that for each value x in U, only two fuzzy sets have $\mu_{A_i}(x)$ bigger than zero but the use of Gaussian functions implies that this happens with all the fuzzy sets. Thus, the use of trapezoidal membership functions to characterise the fuzzy sets, favours the interpretability and reduces the computational cost.

The objective is to design a highly interpretable kind of partition by combining expert knowledge and data distribution. Regarding skilled knowledge, the expert should be able to provide us with at least a minimal knowledge of the linguistic variables of the system. For a given variable, maximum information from the expert, in membership function definition, is desirable: The definition of the universes of discourses. The first thing is to set a domain of interest, included in a physical range.

The number of linguistic terms (labels). It would be valuable for experts to provide us with the number of linguistic terms that they need to express their reasoning.

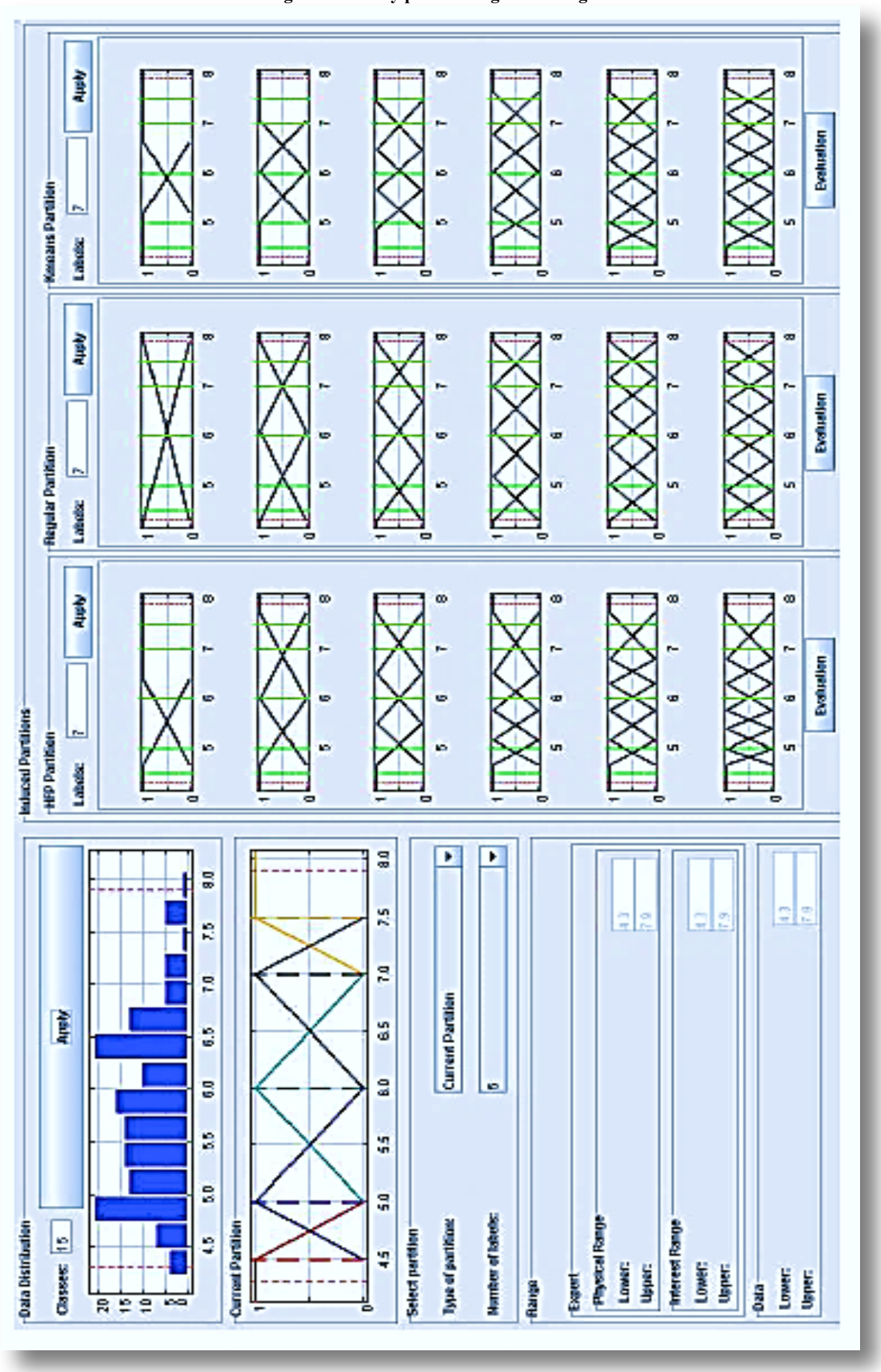
The prototypes of some linguistic labels (modal points, i.e., the most significant values of the fuzzy set centres). For instance, the normal charge of a battery is 12 V.

This information can also be extracted from experimental data. If the universe is unknown, then it can be extracted from the whole data set. When the expert does not know the suitable number of terms, a default value is considered. Five seems to be a good number because it is the intermediate odd number between two and nine. According to [41], the use of odd numbers makes the partition interpretation around a central label easier. Finally, if the expert does not provide us with modal points, then a uniform partition is built in the range of interest with a number of fuzzy sets equal to the number of linguistic terms given by the expert. The automatic generation of the fuzzy partitions from data involves defining the most appropriate shapes for the membership functions, determining the optimum number of linguistic terms in the fuzzy partitions (i.e., the granularity) and/or locating the fuzzy sets into the universe of the variable. Many algorithms can be found in the specialised literature for generating partitions from data but we use the following ones: K-means [42] and Hierarchical Fuzzy Partitioning (HFP) [43].

These methods, as well as those selected to induce rules, are implemented in Fispro [44] and are also used by KBCT [45]. Figure 3.3 illustrates an example of fuzzy partitioning through a histogram of the data distribution, comparing the current partition with all the induced ones with a different number of terms. The green vertical lines represent the modal points defined by the expert. Three kinds of fuzzy partitions are considered :

- **K-means Partition:** uses the well-known clustering method of the same name. There is no *a priori* relationship for a given variable between the modal points related to partitions of different sizes [42].
- **HFP Partition:** corresponds to an ascending procedure. At each step for each given variable, two fuzzy sets are merged. The computational time can be high and essentially depends on the initial number of fuzzy sets. There are two ways to build the initial partition: The first one consists of considering the values in the learning set as identical if they differ by less than a threshold (given as a percentage of the input range). Each group of values, the number of groups depending on that threshold, then corresponds to one fuzzy set. The second way sets the number of groups and calculates the group centres by applying a k-means clustering with a large number of groups [43].
- **Regular Partition:** a uniform fuzzy partition defined in the range derived from the data distribution.

figure 3.3 Fuzzy partitioning with Gauge



3.2.3 Rule-based definition

The considered rules were introduced by [41] and they are of the following form:

$$If\ X_a\ is\ A_a^i\ \ and\ \dots and\ X_z\ is\ A_z^j\ \ ,then\ Y\ is\ C^n$$

Rule premises are made up of tuples (input variable, linguistic term) where X_a is the input variable name ***a***, while A_a^i represents the label ***i*** of such a variable. Notice that the absence of an input variable in a rule means that the variable is not considered in the evaluation of that rule. The rest of this section is dedicated to explaining how build these rules from both expert knowledge and data while maintaining the rule base interpretability.

3.2.3.1 Induced rules

The process of generating rules from data is called induction and it aims to produce general statements, expressed as fuzzy rules in our case, valid for the whole set from partial observations. The observed output for each sample item, is part of the training set allowing supervised training procedures. Many methods are available in the fuzzy logic literature ([47]; [48]) but we are only interested in those one generating rules sharing the same fuzzy sets. Thus, we choose the two following which can be run with previously defined partitioning. They are implemented in Fispro [44] and are used by KBCT [45].

- **Wang and Mendel (WM).** The used method is adapted from the original algorithm proposed by [49]. Given the fuzzy partitions, it starts by generating one rule for each data pair of the training set. The $\mathbf{i_{th}}$ pair is written as :

  If $\mathbf{X}_1$ **is** $\mathbf{A}_{i\ 1}$ and $\mathbf{X}_2$ **is** $\mathbf{A}_{j\ 2}$ . . . and $\mathbf{X}_z$ **is** $\mathbf{A}_{k\ z,}$ then **Y** is $\mathbf{C}_n$.

  The fuzzy sets $\mathbf{A}_{i\ m}$ are those for which the matching degree of $\mathbf{x}_{im}$ is maximum for each input variable $\mathbf{X}_m$ from pair **i**. The fuzzy set $\mathbf{C}_n$ is the one for which the matching degree of the observed output, $\mathbf{y}_i$ is maximum. A degree is assigned to each rule. For a given rule, it is equal to the rule firestrength for the considered pair. In case of identical premises for two rules, only the one with the higher degree is kept. This procedure is

very easy to use since it does not require any parameter. Moreover, it allows the rule base to be adaptive, i.e., new rules may compete with existing ones. WM generates complete rules that are defined considering all the available variables and are likely to be simplified.

- **Fuzzy Decision Trees (FDT).** [50], [51] and [52] proposed the first decision trees and (FDT) are an extension of these classical decision trees. The used method is based on the fuzzy implementation of Quinlan's algorithm. The tree is made up of one root node, which is the tree top or starting point and a series of other nodes. Terminal nodes are called leaf nodes, or leaves and each node corresponds to a split on the values of one input variable. This variable is chosen to maximise the information gain. The goal is to reach a maximum homogeneity among the examples that belong to the node, relative to the output variable which is equivalent to minimising the entropy. The paths from the root node towards the leaf nodes are easy to interpret as decision rules. Thus, the tree represents a subspace of all possible rules. The tree induction is an iterative process. At each step, a new node is added. A number of sub-nodes are generated equal to the number of fuzzy sets of the selected variable. The process is repeated until all leaves contain elements belonging to the same given class. The advantage of decision trees is twofold: first, they generate incomplete rules, and second, the input variables are sorted per their importance, which could be useful information to the expert. Figure 3.4 shows an example of a decision tree built using the FDT method. The associated rule base turns up in Table 3.1.

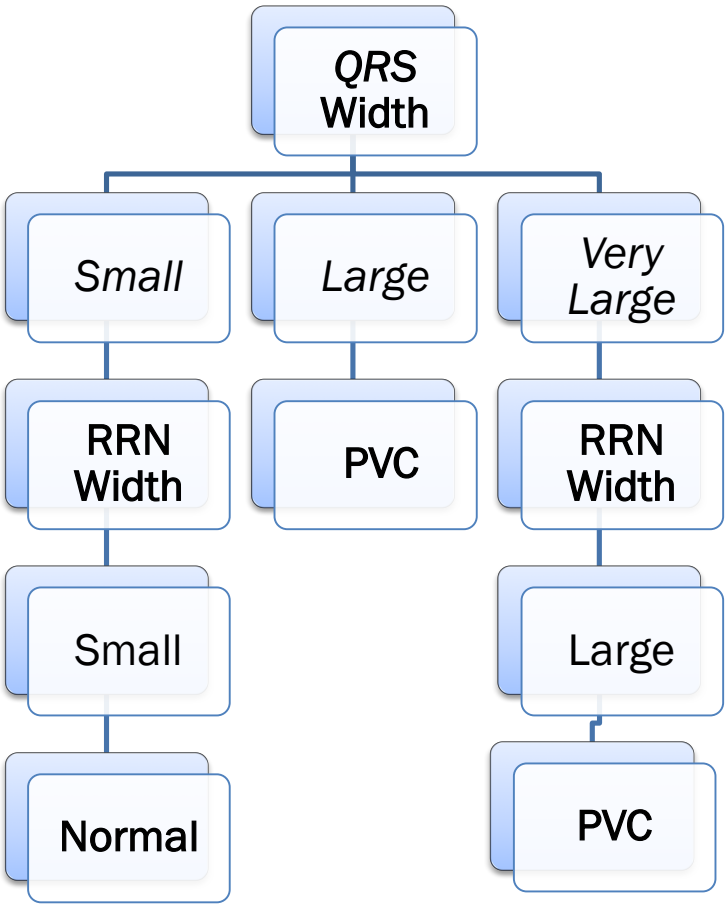


Figure 3.4 Fuzzy decision tree

Rule	If QRS Width	And RRN Width	Then Arrhythmias class
(R1)	Small	Small	Normal
(R1)	Large		PVC
(R1)	Very Large	Large	PVC

Table 3.1 Linguistic rules derived from the tree illustrated in Figure 3.6

Chapter 4

4.1 presentation of the experimentation

In this part, we present our method to define semi-automatically fuzzy partition rules providing a powerful and accurate insight into cardiac arrhythmia. We define a text mining approach applied to a large dataset consisting of the freely available scientific papers provided by PubMed. The information extracted is then integrated with expert knowledge as well as experimental data, to provide a robust, scalable and accurate system, which can successfully address the challenges posed by the management and assessment of big data in the medical sector. We use as platform of rules induction the GUAJE environment (see annex page 102-105).

4.3 Automated extraction of fuzzy partition rules from text

Text mining (TM) [57] is a branch of computer science aiming to accurately extract, identify and analyse information and semantic properties from text sources. Although, there has been steady and successful progress in addressing the challenges, TM research is still very much expanding to provide further state-of-the-art tools to improve accuracy, scalability and flexibility. The extraction of information from textual sources is typically a complex task because of the ambiguous nature of human language. In fact, depending on the general context and the given semantic information, a variety of TM techniques which in general depend on the type of data and their structure can be used [57]. In this thesis, we apply a grammar-based text extraction, based on text patterns, which relies on a set of rules identifying sentences with a determined structure. More specifically, we consider text patterns of the form (NP, verb, NP), where NP refers to the noun phrases and verb to the linking verb [57]. For example, sentences such as 'PVCs can be related to electrolytes' are identified to extract a relationship between PVCs and electrolytes. The effectiveness of this approach is fully exploited when syntactic properties of a sentence are investigated, by using suitable parsing technology [58]. The syntactic roles of the different phrasal components are essential in extracting the relevant information and they can contribute towards a full understanding of the type of relationship.

Furthermore, we also apply basic sentiment analysis, which aims to identify the mood described by text fragments based on specific keywords [59]. Table 4.1 shows a small selection of such keywords used in our approach. We mined all the articles in journals freely available from PubMed [60], a very large database containing biomedical literature, as follows:

We identified articles from the journals containing the following keywords:

- Premature ventricular contractions, or PVCs
- Premature ventricular complexes
- Ventricular premature beats
- Extrasystoles

The identified articles were first lexically and syntactically analysed via the Stanford Parser [58]. Subsequently, a grammar-based extraction identifies the relevant information based on the aforementioned keywords as well as on sentiment analysis [58]. More specifically, only sentences with one or more of the keywords, including those in Table 4.1, in the NPs will be extracted.

Negative keywords	**Positive keywords**	**Uncertain keywords**
Bad	Satisfactory	Unpredictable
Negative	Enhancing	Possible
Underestimate	Advantage	Somewhat
Unsafe	Benefit	Precautions
Unwelcome	Good	Speculative
Tragic	Excellent	Confusing
Problematic	Great	Fluctuation

Table 4.1. A selection of keywords used.

4.3.1 Text mining extraction results

The output of the extraction consists of keywords set found in each text fragment (usually a sentence) which was also extracted, see Table 4.2 as example. A full assessment of this Information extraction type, from text goes beyond the scope of this paper because it specifically addresses issues that are not directly relevant to this context.

Keywords in relation extraction	**Sentences identified**
'PVC', 'PVCs', 'imbalances'	PVCs can be related to electrolyte or other metabolic imbalances'
'PVC', 'death'	'70 mmol/L and T2DM significantly increases risk of PVC and sudden cardiac death; the association between sMg and PVC may be modified by diabetic status'.
'Premature ventricular complexes', 'PVC', 'PVCs', 'atrial', 'ventricular', 'beat', 'missed', 'premature'	'The system recognises ventricular escape beats, premature ventricular complexes (PVC), premature supraventricular complexes, pauses of one or two missed beats, ventricular bigeminy, ventricular couplets (two PVCs), ventricular runs (> two PVCs), ventricular tachycardia, atrial fibrillation/flutter, ventricular fibrillation and asystole', width of the QRS complex, QRS large, QRS complex morphology, Heart rate Irregular. QRS complex (≥ 120 ms).

Table 4.2. Example of relation extraction.

However, we considered a small evaluation consisting of two randomly chosen papers [61], [62] from those identified earlier. The automatic extraction was then compared with a manual one, which produced a recall of 71% and a precision of 84%. In future research, we will investigate more specialised keyword set, as well as larger set of text patterns compared with those used in this paper to improve the mentioned measures.

4.4Data preparation

4.4.1 MIT-BIH data base:

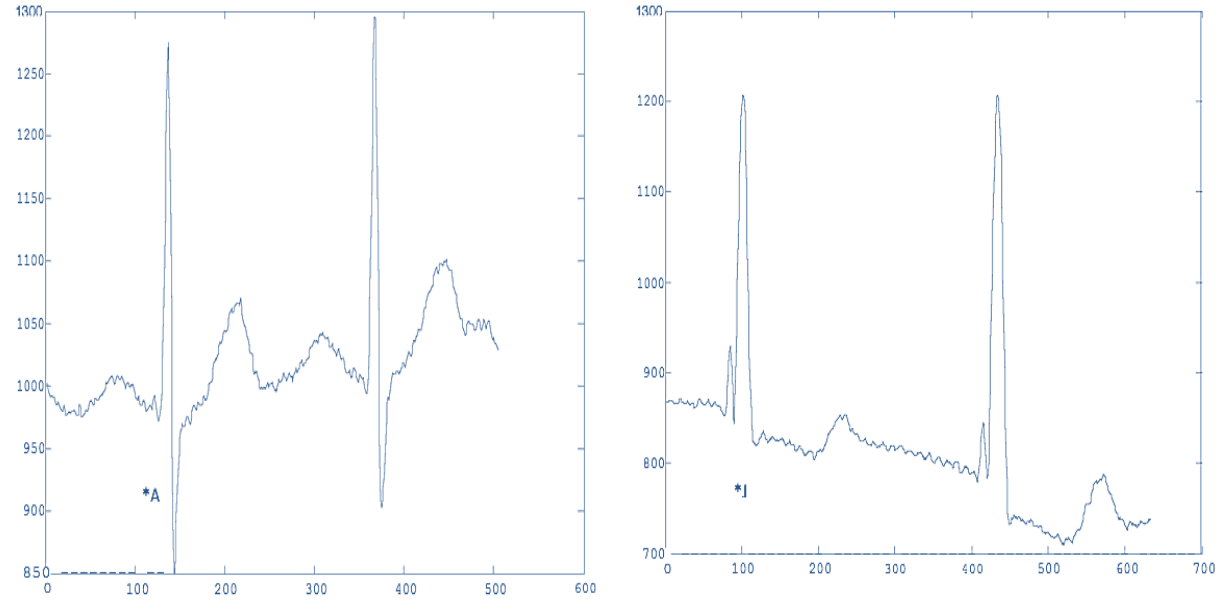


Figure 4.1 Electrocardiogram recording (ECG) type of heart beats: AVC (* A) and JVC (* J)

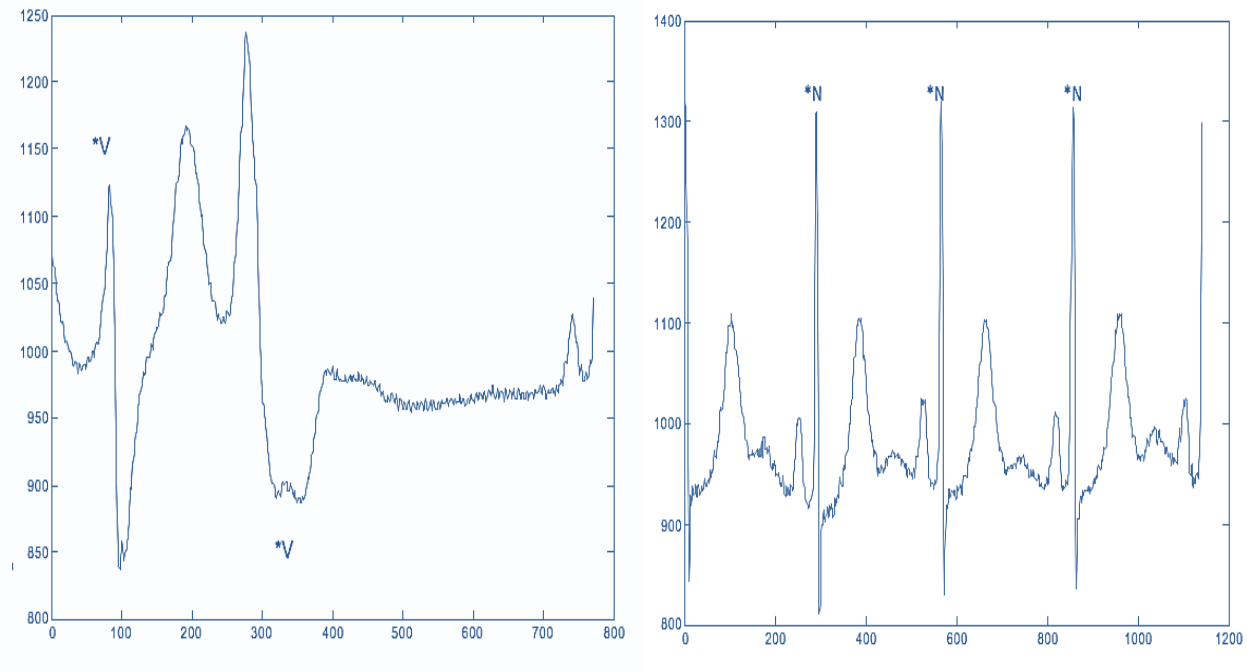


Figure 4.2 Electrocardiogram recording (ECG) type of heartbeats: PVC (* V) and normal (* N)

4.4.2 Data collection

The patients who have been considered in the experiments, taken from MIT-BIH [63], are shown in Table III. The R peaks of the ECG signals were detected using the Tompkins algorithm [64], which is an online real-time QRS detection algorithm. This algorithm reliably detects QRS complex using slop, amplitude and width information. This algorithm automatically adjusts thresholds and parameters periodically to the standard 24 h MIT-BIH arrhythmia database; this algorithm correctly detects 99.3% of QRS complex. From patients with cardiac arrhythmia, taken from MIT-BIH database, we chose only patients with three conditions, namely PVC beats, premature arterial contraction (PAC) beats and premature junctional contraction (PJC) beats, because they provide the best quality of records, and more specifically, PVC is a predictive element of the Cardiac Arrhythmia (CA) sudden death.

1. Dataset:

2. Class	3. Normal	4. PVC	5. PAC	6. PJC
7. Number of samples	8. 60 190	9. 6709	10. 2130	11. 83

Table 4.3 dataset

4.4.3 Feature selection

The feature vector, which is used for recognition of beats, has been selected as follows:

- The R–R interval of the beat RRp (calculated as the difference between the QRS peaks of the pres-ent and previous beats),
- The ratio r = RR1-to-RRo (RRn is calculated as the difference between the QRS peaks of the pres- ent and following beats; see Figure 4.3) and
- The QRS width w (calculated according to the Tompkins algorithm [59]).

In this way, each beat is stored as a three-element vector. Table 4.4 provides the most relevant parameters used in this type of analysis.

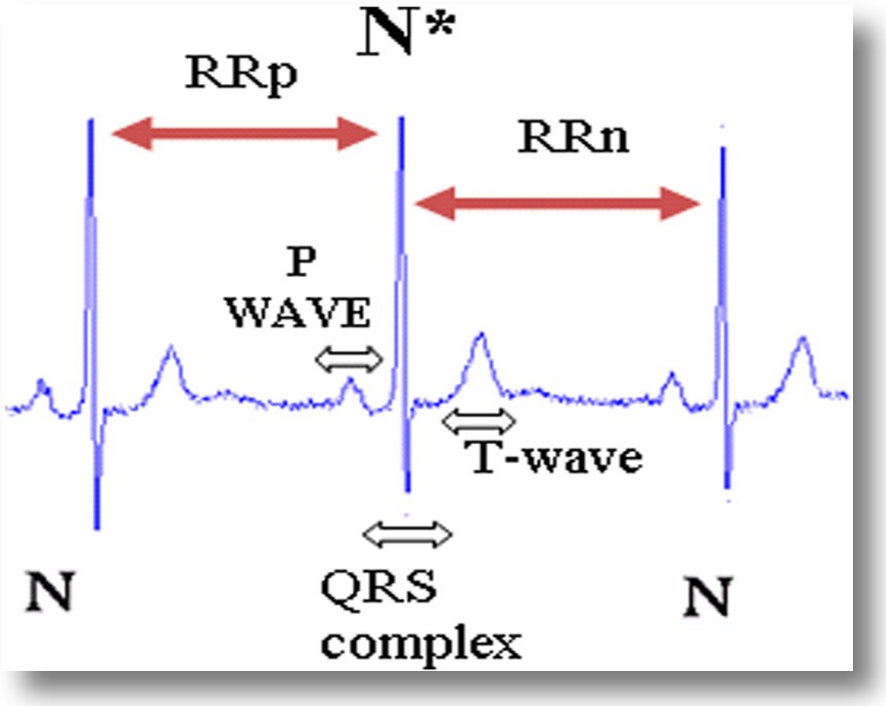


Figure 4.3 Standard ECG beat.

Record	N	A	J	V
101	1860	3	—	—
103	2082	2	—	—
104	163	—	—	2
105	2526	—	—	41
106	1507	—	—	520
107	—	—	—	59
108	1739	4	—	17
109	—	—	—	38
111	—	—	—	1
112	2537	2	—	—
113	1789	—	—	—
114	1820	10	2	43
115	1953	—	—	—
116	2302	1	—	109
117	1534	1	—	—
118	—	96	—	16
119	1543	—	—	444
121	1861	1	—	1
122	2476	—	—	—
123	1515	—	—	3
124	—	2	29	47
200	1743	30	—	826
201	1625	30	1	198
202	2061	36	—	19
203	2529	—	—	444
205	2571	3	—	71
207	—	107	—	105
208	1586	—	—	992
209	2621	383	—	1

Table 4.4 Evaluation data taken from the MIT-BIH database

4.4.4 Data visualisation

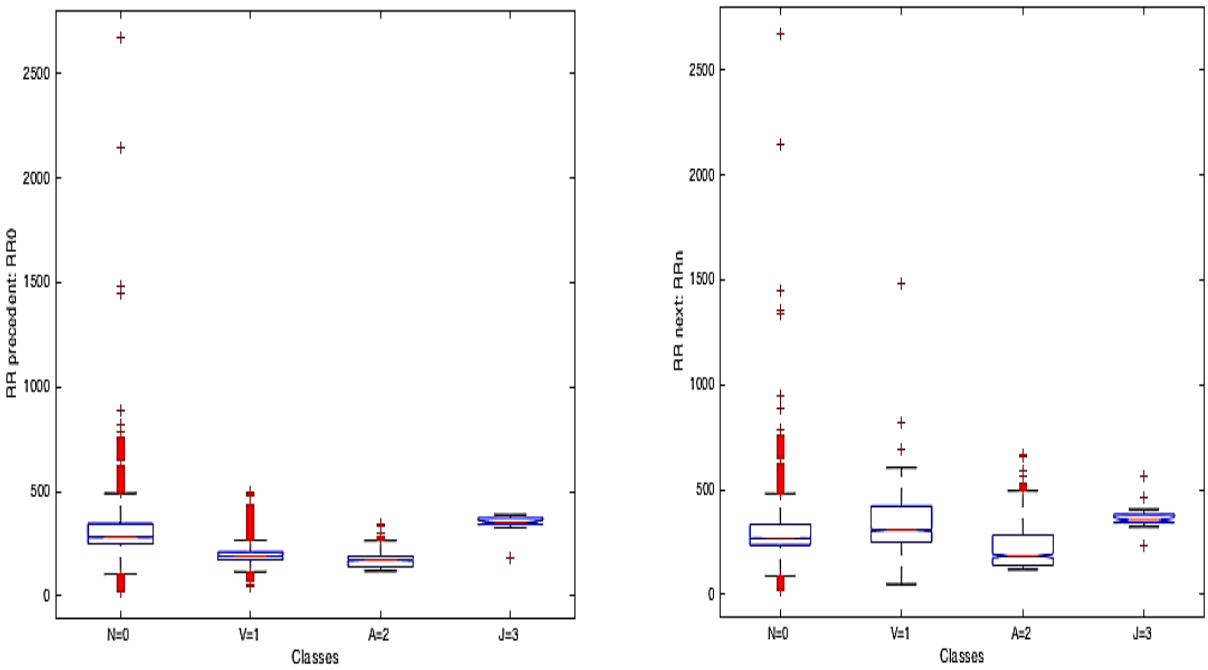


Figure 4.4 RR0&RRn features with classes

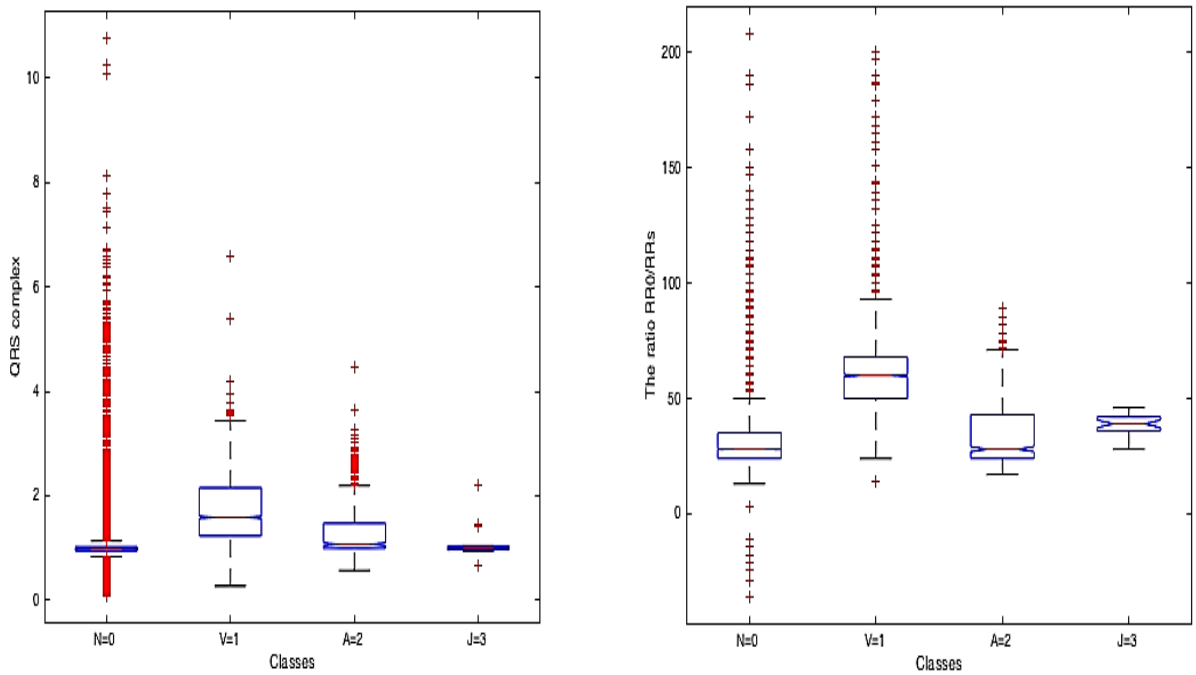


Figure 4.5 QRS&COMP features with classes

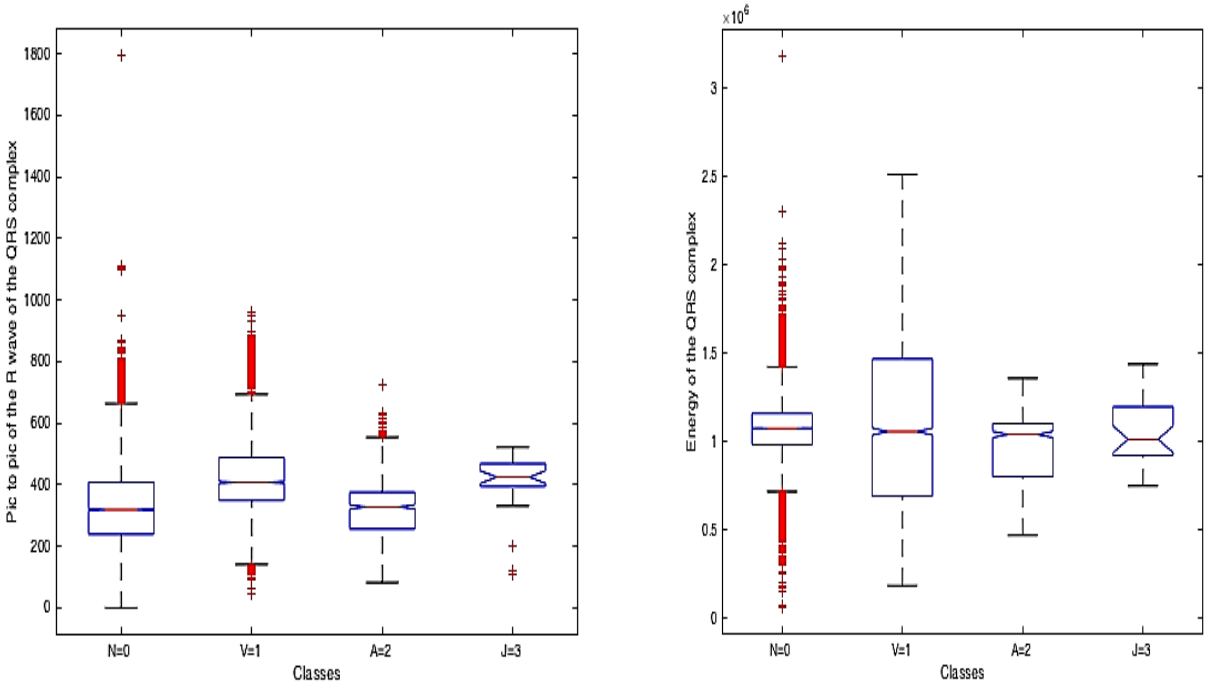


Figure 4.6 PP&ENERGY features with classes

Discussion:

Two-dimensional representation is very significant if you want to treat data and extract knowledge, especially if you want to know the behavior of the data in different features and classes.

clearly noticed was with only one features we can't not show the differences between classes. However, it is essential to know the degree of relevant of each feature.

We notice in figure 4.5 that the QRS represents a good discriminating features between PVC and normal case, the same thing for the COMP see Figure 4.5.

As conclusion, we can say that each one of the parameters is important in the diagnosis of his classes but with a different degree of importance.

4.4.5 Correlation study

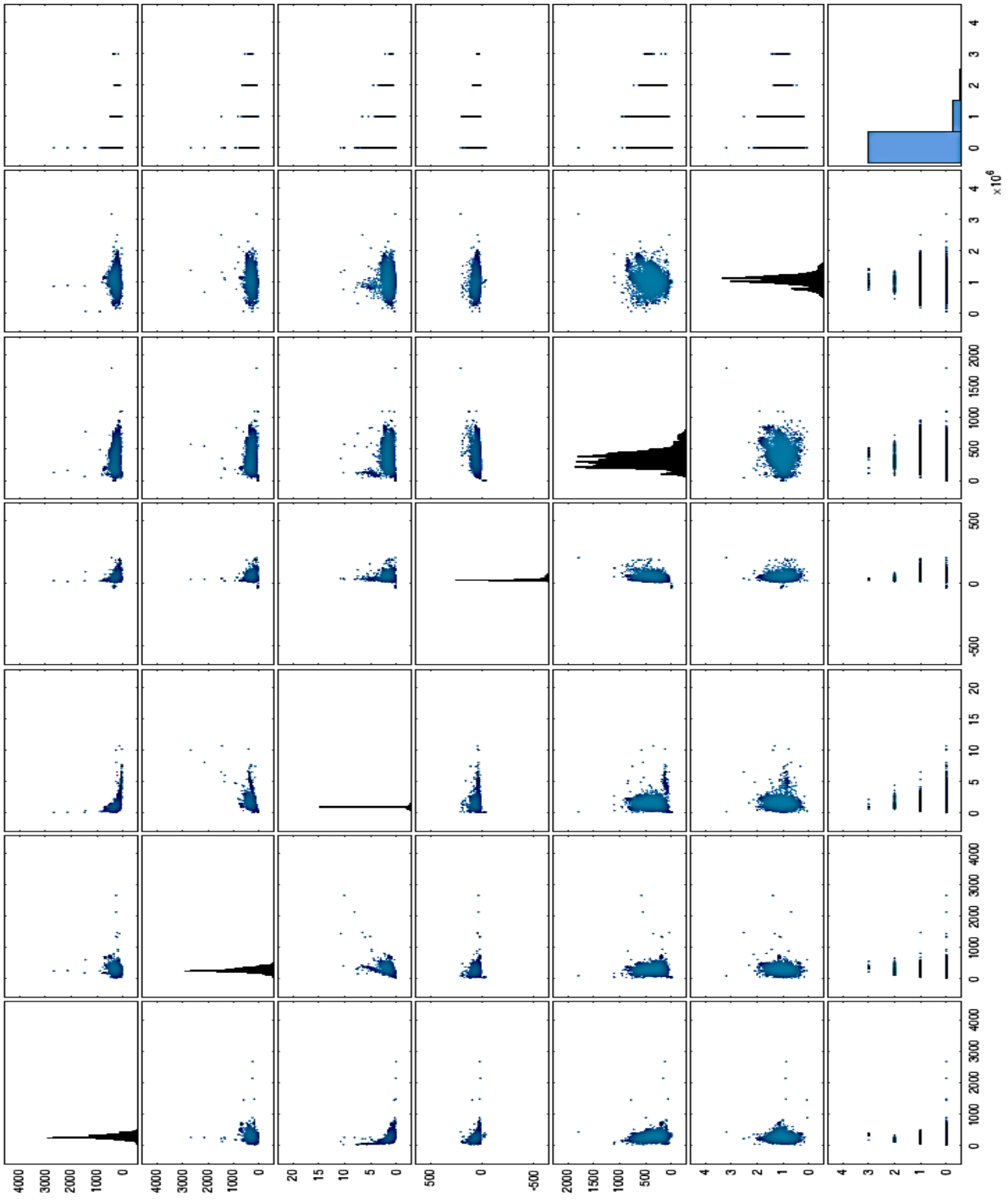


Figure 4.7 matrix of correlation

	RR0	RRn	QRS	COMP	PP	ENERGY	CLASS
RR0	1	0.49728	-0.43924	-0.28433	-0.03675	-0.09099	-0.36555
RRN	0.49728	1	0.41501	0.01896	0.05735	-0.03505	0.04703
QRS	-0.43924	0.41501	1	0.32951	0.07565	0.06068	0.41057
COMP	-0.28433	0.01896	0.32951	1	0.17278	0.05531	0.49441
PP	-0.03675	0.05735	0.07565	0.17278	1	0.18565	0.13435
ENERGY	-0.09099	-0.03505	0.06068	0.05531	0.18565	1	0.00306
CLASS	-0.36555	0.04703	0.41057	0.49441	0.13435	0.00306	1

Table 4.5 Correlation coefficients

Discussion:

It's important to study the correlation between all the features and the classes. To see if the variation is significant in database we make sure that we choose the main relevant features.

We can see clearly that we made good choice of features because, like RR and QRS with correlation of: and we can say that all features have one by one good correlation with classes and there is also correlation between some features like RRs & RR0 and COMP who is normal because COMP =RRS/RR0.

4.5 Fuzzy Partition Design

This section covers the left most parts, that is, those related to membership functions extraction from both the expert's input and experimental data. Note that the initial step considers the extraction from the former and subsequently some approaches for membership function design from data are introduced.

Attributes	**Meaning**
RR precedent: RR0	The distance between the peak of this beat R and R of the peak beat precedent
RR next: RRn	RRn between the peak of the present R beat and the peak of R beat following
QRS complex	Beginning of the Q wave and the end of the S wave
Comp	The ratio RR0/RRs
PP	Pic to pic of the R wave of the QRS complex
Energy	Energy of the QRS complex

Table 4.6 The various descriptors.

When defining expert knowledge, we assume that the linguistic variables of the system are sufficiently known. More specifically, experts may identify specific properties such as, the domain of interest within a physical range, who would be subsequently facilitated in the decision process by a given and possibly small, number of linguistic terms. In this study, we assume that a minimum information on membership function definition, which includes the definition of universes, number of terms and sometimes, prototypes of linguistic labels (modal points), is available [65]. Note that this is a reduced version of the interval estimation method [67], as the interval is reduced to a single point. Furthermore, if additional information is provided by the expert, this can be integrated into the system.

The knowledge base is split into two main parts: the data base (DB) and the RB. The former is defined by the description of the linguistic variables such as number, range, granularity, membership functions and their normalisation functions.

As discussed in [65], in most of the existing approaches, which focus on the generation from data of the fuzzy partitions, the automatic design of the DB is one of the most important steps in the definition of the overall knowledge base. However, the method we are proposing in our study consists of rules that integrate expert as well as data-based knowledge.

The automatic generation of fuzzy partitions is based on the definition of the best shapes of the membership functions, in terms of the optimal number of linguistic terms in the fuzzy partitions and the location of the fuzzy sets within the universe of the variable.

As discussed in [67], in this thesis, we follow an approach, which includes the following:

1. A non-supervised clustering process, which is performed to address the extraction of the DB from the available dataset as part of the preliminary design;
2. An embedded basic learning method, which derives the DB.

Simultaneous design [67] is another method, which cannot be successfully applied in this context. In fact, it usually generates a far more complex process in which the computational effort involved is not fully exploited because only the fuzzy partitions are considered. In addition, our interpretability requirements need some specific properties for the partitions. In fact, techniques generating multidimensional clusters cannot be successfully applied, because only one-dimensional membership functions are required. On the other hand, it is feasible to apply a one-dimensional optimisation technique if it includes some semantic constraints.

In the case of embedded design, when search techniques are used in the design of the DB, it is essential to include appropriate interpretability measures in the objective function, to provide suitable and optimal solutions. These include measures of completeness, consistency, compactness or similarity. Finally, at the end of the embedded design process, only fuzzy partitions are considered, which will subsequently lead to the creation of fuzzy rules.

4.5.1 Criteria for the evaluation of fuzzy partitions

The evaluation of fuzzy partitions is based on both *linguistic properties* and *partitioning properties* [65]. The former influence the shape of the fuzzy sets, as well as the relations between fuzzy sets corresponding to the same variable. The latter, on the other hand, focus on the data from the partitions that have been generated and more importantly, on their level of matching with the partitions derived from data. This is not the case for linguistic properties because their assessment does not involve data.

In the remaining section, we will give a brief overview of some partitioning features (Figures 4.8 – 4.13 and Tables 4.7 - 4.12), as introduced in [65]. We will focus on the following:

1. Methods based on the data distribution, which excludes methods based on an input–output relation;

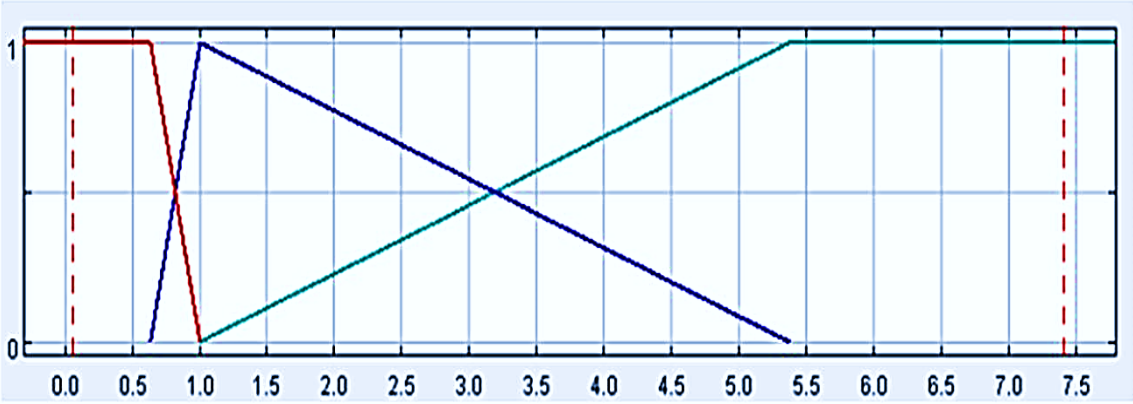


Figure 4.8 Fuzzy partition RR0 from K-means algorithm.

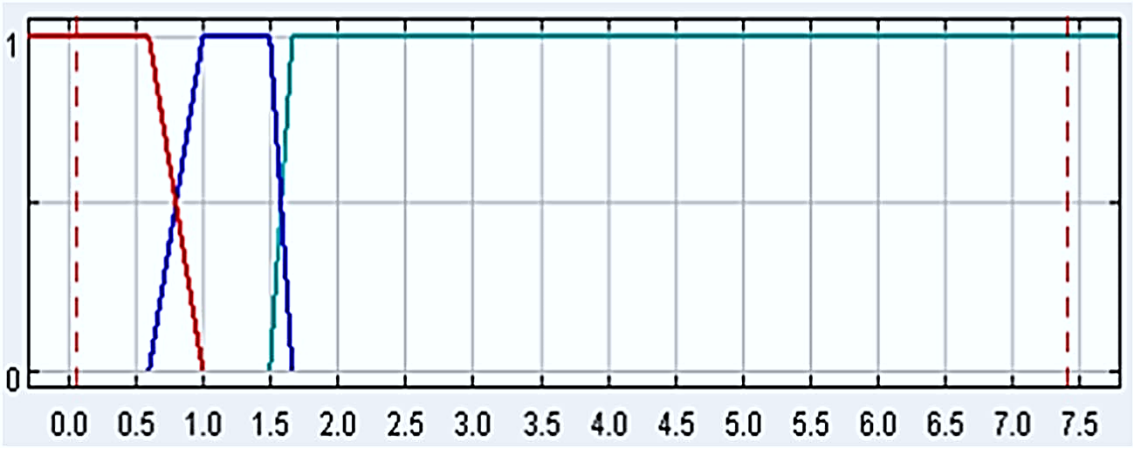


Figure 4.9 Fuzzy partition RRs from expert and TM

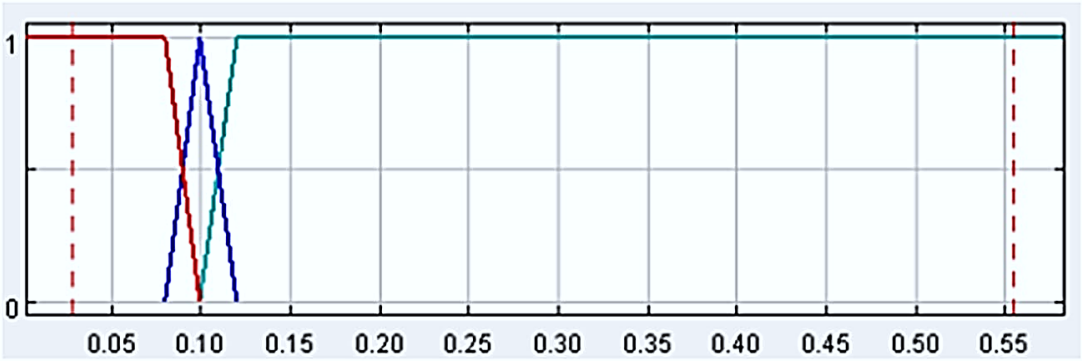


Figure 4.10 Fuzzy partition QRS from expert and TM.

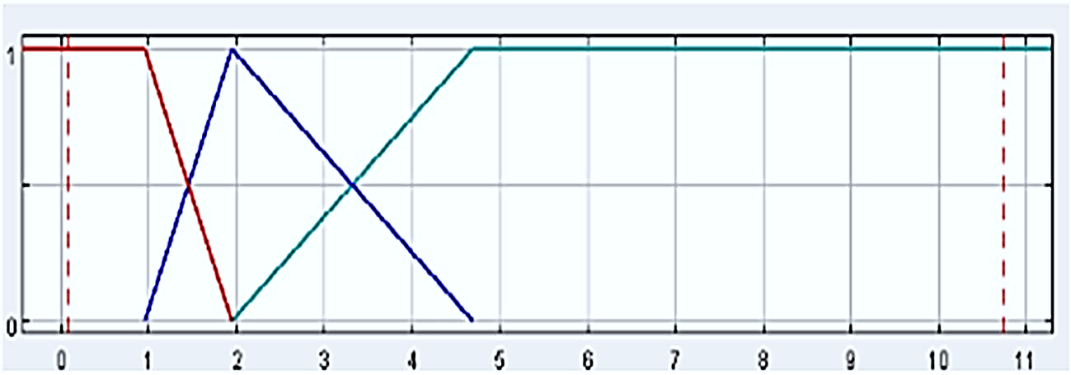


Figure 4.11 Fuzzy partition COMP from K-means algorithm.

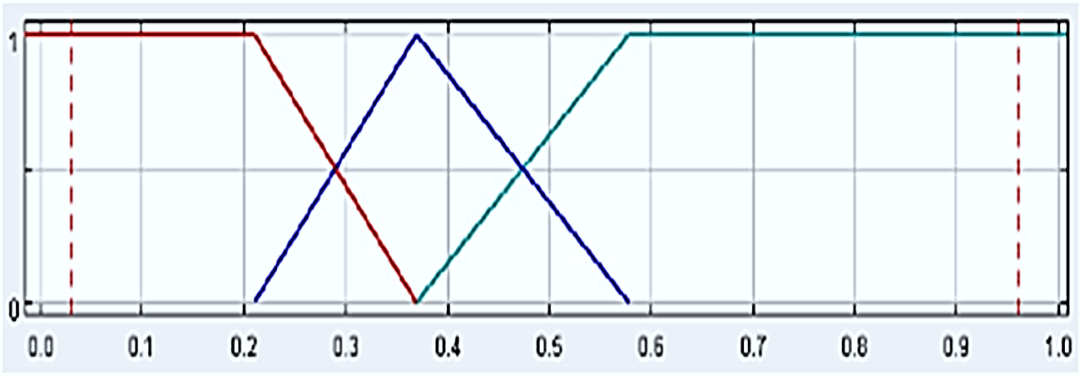


Figure 4.12 Fuzzy partition PP from K-means algorithm ($x10^3$).

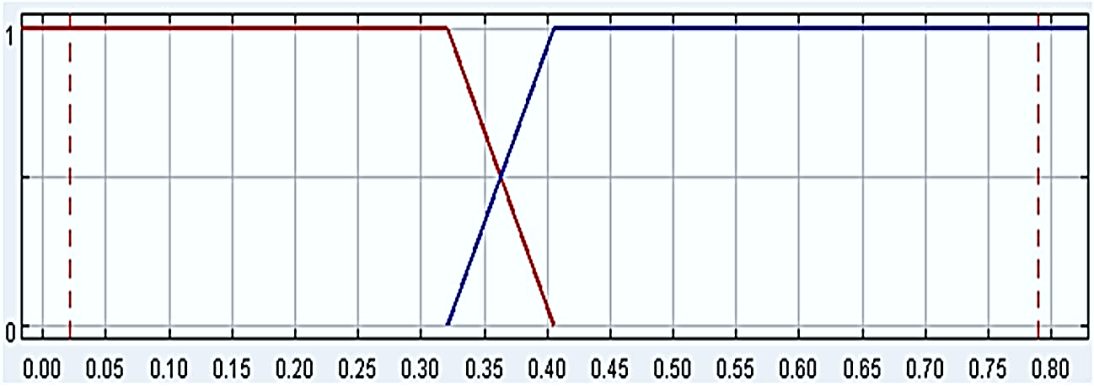


Figure 4.13 Fuzzy partition energy from HFP algorithm.

2. Methods typically applied in unsupervised clustering and not in supervised clustering;
3. The assignment of data elements to each of the items of the partition.

Partition	**Partition coefficient (max)**	**Partition entropy (min)**	**Chen index (max)**
HFP	0.77513	0.33406	0.77094
Regular	0.69936	0.47483	0.74665
K-means	**0.79948**	**0.30842**	**0.80794**
Expert and TM	0.78262	0.32504	0.78647

Table 4.7 RRo fuzzy partition quality (three labels).

Partition	**Partition coefficient (max)**	**Partition entropy (min)**	**Chen index**
HFP	0.77378	0.33614	0.76980
Regular	0.69705	0.47751	0.74360
K-means	0.77121	0.34723	0.77941
Expert and TM	**0.78300**	**0.32441**	**0.78668**

Table 4.8 RRs fuzzy partition quality (three labels).

Partition	**Partition coefficient (max)**	**Partition entropy (min)**	**Chen index (max)**
HFP	0.76812	0.34495	0.76312
Regular	0.66966	0.50649	0.70024
K-means	0.82214	0.26975	**0.82707**
Expert and TM	**0.84540**	**0.22046**	0.82512

Table 4.9 QRS fuzzy partition quality (three labels).

Partition	**Partition coefficient(max)**	**Partition entropy (min)**	**Chen index (max)**
HFP	0.68847	0.45294	0.67092
Regular	0.71224	0.45990	0.75945
K-means	**0.89473**	**0.18398**	**0.90993**
Expert and TM	0.81362	0.29566	0.82971

Table 4.10 COMP fuzzy partition quality (three labels).

Partition	Partition coefficient (max)	Partition entropy (min)	Chen index (max)
HFP	0.68847	0.45294	0.67092
Regular	0.71224	0.45990	0.75945
K-means	**0.89473**	**0.18398**	**0.90993**
Expert and TM	0.81362	0.29566	0.82971

Table 4.11 PP fuzzy partition quality (three labels).

Partition	Partition coefficient (max)	Partition entropy (min)	Chen index (max)
HFP	**0.78374**	**0.31852**	0.66240
Regular	0.52352	0.66896	0.17696
K-means	0.77454	0.33891	**0.67045**

Table 4.12 Energy fuzzy partition quality (two labels).

Let u_{ik} be the degree of membership of the k-th element of the dataset to the i-th element of the fuzzy partition. The partition coefficient is:

(we used the same notation as in [65])

$$\boldsymbol{PC = \frac{\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^{2}}{n}}$$

And the partition entropy is:

$$\boldsymbol{PE = -\frac{1}{n}\left\{\sum_{k=1}^{n}\sum_{i=1}^{c}[u_{ik}log(u_{ik})]\right\}}$$

Where c is the number of elements of the fuzzy partition and n is the cardinality of the set of data.

Furthermore, Chen [68] recently introduced the following index measure:

$$ch = \frac{1}{n}\sum_{k=1}^{n} max(u_{ik}) - \frac{2}{c(c-1)}\sum_{i=1}^{c-1}\sum_{j=i+1}^{c}\frac{1}{n}\sum_{k=1}^{n} min(u_{ik}, u_{jk})$$

An efficient partition should minimise the entropy and maximise the coefficient partition and the Chen index [65].

The most important part of the knowledge-based system is the reasoning mechanism that induces decision rules.

Because the fuzzy partition is at the core of induction methods in fuzzy logic, we propose to initialize the fuzzy partitions by two approaches: first, a purely automatic method of induction (K-means, HFP and regular) and, secondly, the approach resulting from information extraction from textual sources, as discussed in Section, to compare between the different methods. Subsequently, we have established linguistic terms to construct the rules and modal point.

4.6 Rule Base Generation

The process of generating rules from data is called *induction*, aiming to produce general statements, expressed as fuzzy rules in our case, valid for the whole set, from partial observations.

The observed output, for each sample item is part of the training set allowing supervised training procedures. Many methods are available in the fuzzy logic literature [69], but we are only interested in those one, which generate rules sharing the same fuzzy sets. Thus, we have chosen the methods, which are implemented in Fispro [69], and they are used by Knowledge Base Configuration Tool (KBCT) [45], as they can be run with previously defined partitioning.

4.6.1 Knowledge base accuracy

To obtain an accuracy measure, we need to compare the inferred output with the observed one in a real system. In classification systems, the most common index is defined as the number of misclassified cases. We will consider the three following indices:

- Unclassified cases (UC): number of cases from dataset that do not fire at least one rule with a certain degree.
- Ambiguity cases (AC): number of remaining cases for which the difference between the two highest output confidence levels is smaller than an established threshold (AmbThres). More specifically, we also have the following:
- AC (Total): all detected ambiguity cases.
- AC (Error): only those ambiguity cases related to error cases (EC) (observed and inferred outputs are different
- EC: number of remaining cases for which the observed and inferred output classes are different.

A good Knowledge Base (KB) should minimise all of them by offering an accurate (reducing EC), consistent (reducing AC) and complete (reducing UC) set of rules. They can be combined to define the next accuracy index:

$$Accuracy = 1 - \frac{EC + AC(Error) + UC}{Data(Total)}$$

$$Accuracy(CONS) = 1 - \frac{EC + AC(Total) + UC}{Data(Total)}$$

$$Accuracy(BT = 0) = 1 - \frac{EC + AC(Error) + UC(Error)}{Data(Total)}$$

4.7 Evaluation

The evaluation was carried out over a variety of experimentations. More specifically, the suitable fuzzy partition was investigated and subsequently, we induced the corresponding decision rules, calculated and assessed the quality criteria to measure the accuracy of each approach. Table 4.13 clearly shows that the best results correspond to FDT1, FDT2, FDT3, FDT5, FDT6, FDT9, FDT4, FDT7, FDT14 and FDT15, which are based on different algorithms for induction and partitions, with the following parameters:

- Coverage (cv%): percentage of data samples from the selected dataset that fire at least one rule in the RB with an activation degree higher than the pre-defined blank threshold (BT).
- Accuracy (ac): percentage of data samples properly classified.
- Accuracy (acons): percentage of data samples properly classified.
- Average confidence firing degree (acfd): mean value of the firing degree related to the winner rule for the whole dataset.
- Minimum confidence firing degree (micfd): minimum value of the firing degree related to the winner rule for the whole dataset.
- Maximum confidence firing degree (macfd): maximum value of the firing degree related to the winner rule for the whole dataset.
- Max error (me): maximum difference between the observed class and the inferred one.

- Mean square classification error (msce).

KB	cv%	ac	acons	abt = 0	acfd	micfd	macfd	me	msce
FDT1	100	0.914	0.908	0.914	0.865	0	1	3	0.007
FDT4	100	0.424	0.303	0.424	0.516	0	1	3	0.125
FDT7	100	0.873	0.87	0.873	0.448	0	0.842	3	0.054
FDT15	100	0.717	0.704	0.717	0.704	0.013	1	3	0.013
FDT14	99.794	0.939	0.934	0.939	0.925	0	1	3	0.007
FDT2	100	0.935	0.928	0.935	0.701	0	1	3	0.024
FDT3	100	0.318	0.277	0.318	0.373	0.005	0.742	3	0.068
FDT5	100	0.424	0.303	0.424	0.468	0	1	3	0.147
FDT6	100	0.930	0.924	0.930	0.645	0.002	1	3	0.035
FDT9	100	0.618	0.549	0.618	0.361	0	1	3	0.075

Table 4.13 Quality measurements

FDT14 was generated by attributes resulting from the TM extraction (RRs and QRS) integrated with K-means and HFP algorithms and it gave the best result with a classification rate of 93.9% and a 0.646 interpretability value. On the other hand, FDT15 is created by expert fuzzy partition with fuzzy decision tree induction algorithm and it gave a lower classification rate and interpretability value of 71.7% and 0.025, respectively.

Considering FDT1 (regular fuzzy partition with fuzzy decision tree induction algorithm), FDT4 (K-means fuzzy partition and fuzzy decision tree induction algorithm) and FDT7 (regular fuzzy partition and fuzzy decision with Wang and Mendel induction algorithm), we noticed that the interpretability value was zero; this is clearly explained by the very large number of rules.

4.7.1 Analysis of rules

In this section, we further discuss the FDT14 results, which have been shown to be more accurate, with a knowledge base consisting of the *'data and expert'* parts as a system partition. More specifically, we successfully built a simplified and optimised based knowledge, with 11 decision rules and nine induced rules from the database and two of the expert. Figure 4.16 depicts an example based on one sample from the dataset, which activates both rule 8 and rule 9 with 0.553 and 0.447 distributed as class 2. We also had an improvement of the interpretability as shown by the corresponding index of 0.646, providing a very good compromise between the accuracy, with value of 0.939, and the interpretability. Furthermore, this also yields Nauck's index [70] of 99.796, defined by the product:

$$\boldsymbol{Nauck's = Comp}\text{*Part*cov}$$

Where Comp represents the complexity of a classifier measured as the number of classes divided by the total number of premises. Part stands for the average normalised partition index overall input variables.

Rules

Rule	Type	Active	If RR0	AND RRs	AND COMP	AND QRS	AND PP	AND ENERGIE	THEN Class
1	I	yes			Late L				1.0
2	I	yes		NOT(Irregular R)	NOT(Late L)	Small			1.0
3	I	yes	Irregular L		NOT(Late L)	Average			1.0
4	I	yes	NOT(Irregular L)			Average			1.0
5	I	yes			Late R	Large	Small		1.0
6	I	yes		Irregular R	NOT(Late L)	Small			3.0
7	I	yes			Regular	Large	Small		2.0
8	I	yes			NOT(Late L)	Large	Average		2.0
9	I	yes			NOT(Late L)	Large	Tall		2.0
10	E	yes	Irregular L	Irregular R	Late R	Large	Tall	high	2.0
11	E	yes	Irregular R	Irregular L	Late L	Average	Tall	high	2.0

Figure 4.14 Rules of FDT14 (expert and text mining) and fuzzy decision tree algorithm. (taken from gauge software)

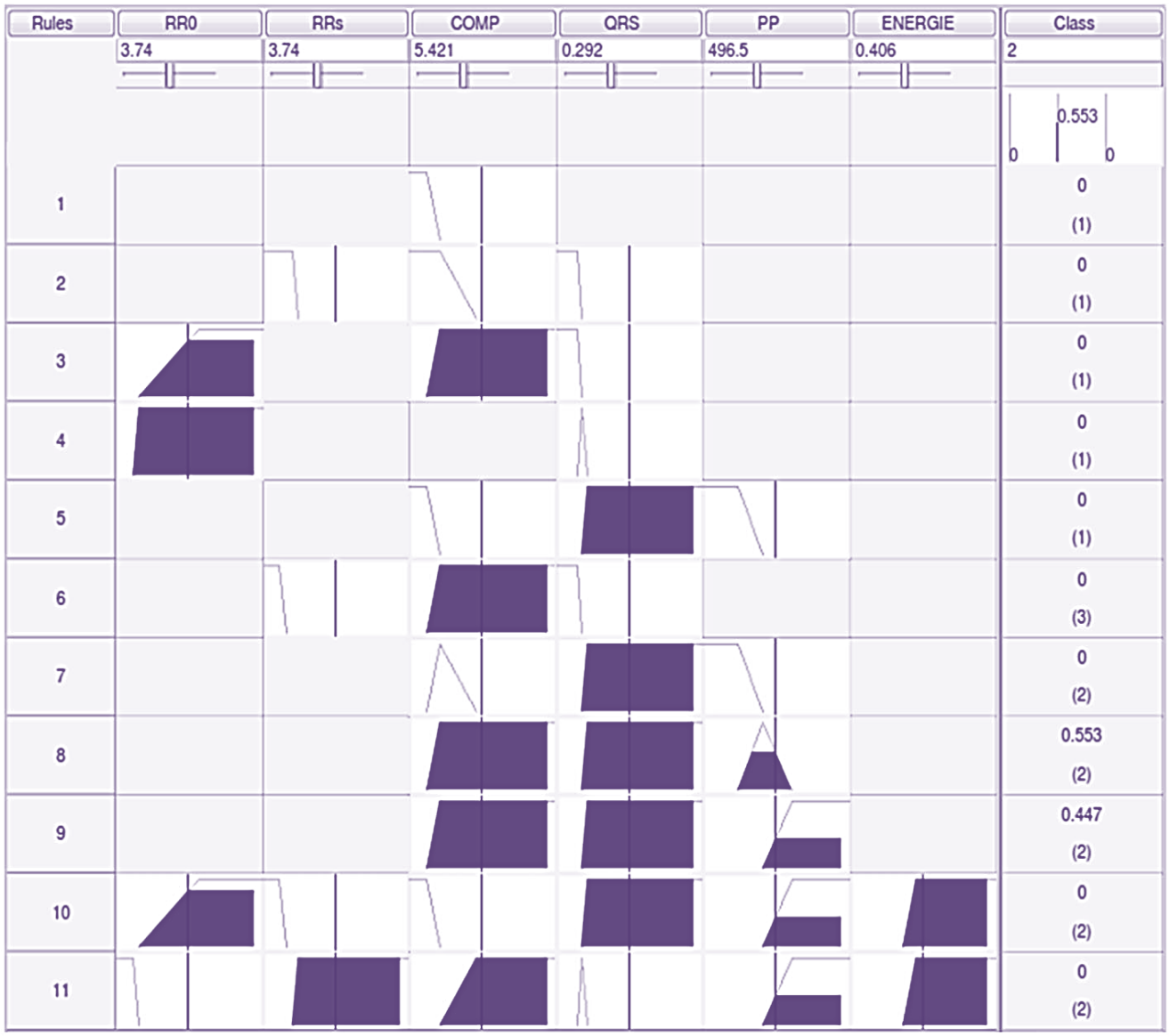


Figure 4.15 Inference rules.

Indices	Values
Converge (%)	99.796
Accuracy	0.939
Average confidence firing degree	0.925
Total rule length	36
Inferential fired rules (training) (max)	5
Inferential fired rules (training) (average)	2.342
Inferential fired rules (training) (min)	1
Accumulated rule complexity	11.213
Interpretability index (Fingrams)	0.646
Interpretability index (HILK)	0.123
Nauck's index	99.796

**HILK, Highly Interpretable Linguistic Knowledge*

Table 4.14 Fingrams measurement.

It is computed as the inverse of the number of labels minus one (two is the minimum number of linguistic terms in a partition) for each input variable.

Cov is the average normalised coverage degree of the fuzzy partition (Figure 4.15 and Table 4.14).

4.8 Comparative studies with other classification techniques

Techniques	Classification rate (%)
HMM	71
RNN	90
FDT	93.9
SOM	95
FCL	87.93

Table 4.15 Classification rate comparative study.

This section is intended to provide a comparative study between our work and different methods that are used in the literature, the database used is MIT-BIH.

We see clearly that different techniques are used to detect the best cardiac arrhythmias. In our study, we set two objectives:

A reliable classification which the rate is 93.9% and the rule of interpretable and understandable decision by doctors. Compared to neural networks RNA [71] and kohonen map [72], which have good classification rate (95%, 90% respectively) but they have one major drawback; they are black boxes and do not allow a good interpretability of the results, which does not encourage cardiologists to use.

Whereas other techniques such as hidden Markov HMM [73] chains and fuzzy inference systems FCL with respective rates (71% and 87%), comparable rate to our study, but the form of rules induced and their degree o interpretability is less good than our study.

4.9 Conclusion:

In this chapter, we have discussed a method to build a system based on rules, using three sources of knowledge. The use of fuzzy logic, as a platform for communication between the different sources of knowledge, proves to be a successful solution to manage the knowledge fusion in a database of common rules. The application of specific TM methods in the extraction of knowledge from the large textual datasets, provided by PubMed, has enabled an accuracy of 93.9% and interpretability index of 0.646. This is clearly a marked improvement compared with the existing algorithms, which may obtain high accuracy but lacking in interpretability. Furthermore, our method offers more flexibility and transparency in the system of detection, allowing expert's contribution to facilitate and guide the process of medical decision-making.

Chapter 5

Conclusion

This thesis presents a new method to build a system based on rules, using *three sources of knowledge*, with a good balance between accuracy and interpretability, and the use of *fuzzy logic*, as a platform for communication between the different sources of knowledge.
Three kinds of knowledge, *expert knowledge*, *knowledge extracted from data* and *knowledge from the large textual datasets*, provided by PubMed, are considered. They convey complementary information, and their *fusion* can lead to compact and robust knowledge bases. It has enabled an *accuracy* of **93.9%** and *interpretability index* of **0.646** this is clearly a marked improvement compared with the existing algorithms which may obtain high accuracy but lacking in interpretability. Furthermore, our method offers more *flexibility* and *transparency* in the system of detection, allowing expert's contribution to facilitate and guide the process of medical decision-making.

The ECG signal quality is a major constraint for the recognition of various diseases. As the mode of acquisition has a major role, to differentiate between premature ventricular, premature junctional, arterial premature conductions. Our data extracted from the MIT-BIH is mainly composed of beats taken from derivation DII which is a major handicap in the classification.

Currently fuzzy decision trees are, a major advantage in the classification because of their simplicity and interpretability. Note that in the medical field, every expert requires any automatic method of diagnostic support to justify his decisions, a characteristic absent in several techniques cited in the literature, in particular neural networks. *The system, we present in this thesis offers physicians an explicit knowledge base (as a rule) acquired in medical database*. The expert will be *able to accept the rules, modify, delete* or *add others*.

This work, however, leaves a number of open questions which could be further developed both mathematically and methodologically. In the following, we give some points that we consider relevant for future developments:
It would be interesting to extend our application to other diseases such as ventricular tachycardia and supraventricular tachycardia, biological data. it's more interesting also to analyze problem of classification and learning system, to improve the accuracy, use other algorithm of rules induction, raise interpretability and reach high performance of our system.

The work led to the development of a decision support system (DSS) in order not only to help and improve diagnosis and detection of arrhythmias but also to improve health of people, using many sources of information such as (bio bank, environment data, clinical trial, web data, hospital information system, clinical & molecular & DICOM data) and this system can be used by:

- Healthcare providers
- Patients
- Researchers including lawyers, ethicists, IT people
- Clinical Research Organisations (CRO)
- Public health authorities

For now, the continuation of this work is done in cooperation with the research Group from universities of Tlemcen and Oran (Algeria) with Derby, Edge Hill universities (United Kingdom) and we hope to achieve our objective in the following years.

Articles Published:

1- "Big data-based extraction of fuzzy partition rules for heart arrhythmia detection: a semi-automated approach",

Omar Behadada, Marcello Trovati, MA Chikh and Nik Bessis

Concurrency and Computation: Practice and Experience

Article first published online: 21 JAN 2015, DOI: 10.1002/cpe.3428

2- "An interpretable classifier for detection of cardiac arrhythmias by using the fuzzy decision tree"

Omar Behadada, M. A Chikh

Artificial Intelligence Research, 2013, Vol.2, No. 3,

DOI: 10.5430/air.v2n3p45 URL: http://dx.doi.org/10.5430/air.v2n3p45, www.sciedu.ca/air.

3- "A PVC BEATS RECOGNITION USING FUZZY CLASSIFIER"

M. A. CHIKH and Omar BEHADADA,

J.Mech. Med. Biol. 10, 327 (2010)

DOI:0.1142/S021951941000337X

Book Chapter:

"Information Extraction from Unstructured Data Sets: An Application to Cardiac Arrhythmia Detection"

Omar BEHADADA,

BOOK Title « Big-Data Analytics and Cloud Computing »,

SPRINGER INTERNATIONAL PUBLISHING, P127-145, DOI 10.1007/978-3-319-25313-8_9, Print ISBN 978-3-319-25311-4, Copyright 2015.

Proceeding Indexed:

"Fuzzy Partition Rules for Heart Arrhythmia Detection",

Omar BEHADADA, MA Chikh and Marcello Trovati,

2015 International Conference on Intelligent Networking and Collaborative Systems,

2-4 Sept. 2015, Taipei, p463 – 466, Print ISBN: 978-1-4673-7694-5 IEEE Computer Society © 2015

International Conferences:

1- ''Logistic Regression Multinomial for Arrhythmia Detection''

Omar Behadada, Marcello Trovati, M.A. Chikh, Nik Bessis, and Yannis Korkontzelos

The 2nd International Workshop on Data-driven Self-regulating Systems (DSS 2016),12 September 2016, Augsburg, Germany in conjuction with 10th IEEE International Conference on Self-Adaptive and Self-Organizing Systems (SASO).

2- ''Application flou dans la reconnaissance des arythmies cardiaques''

Omar Behadada, M.A CHIKH,

10ème Conférence maghrébine sur les technologies de l'informatiques.
Univ sciences et technologie d'Oran –Mohammed Boudiaf–, le 28-3 avril 2008.

3- ''Construction d'un système d'aide à la décision médicale pour la détection des arythmies cardiaques à l'aide d'arbres de décision flous'',

Omar Behadada, M.A CHIKH,

Conference internationale sur l'informatique et ses applications.
Univ tahar moulay saida. 03-04 Mai 2009

4- ''Classification explicite des données cardiologiques par l'approche,

Omar Behadada, M.A CHIKH,

Conférence Internationale des Technologies de l'Information et de la Communication.
univ ferhat abbas setif. 04-05 Mai 2009.

5- « La classification des arythmies cardiaques par les chaînes de Markov cachées »

TRIQUI Bouchra, BENYETTOU Abdelkader, BEHADADA Omar,

BIOMEIC'12 Biomedical Engineering International Conference Tlemcen-Algeria, October 15-16, 2014.

6- « Diabete recognition using Restricted Coulomb Energy Networks (RCE) »

Ammar Mohammed, Chikh Mohammed Amine, Behadada Omar

Second International Conference on Image and Signal Processing and their Applications ISPA 2010

Edited Book:

1- « Détection des arythmies cardiaques par les arbres de décision floue »

Omar Behadada, Chikh MA

Publisher : Editions universitaires europeennes (October 18, 2011) (French Edition)

ISBN-13: 978-6131592485

2- « Détection des arythmies cardiaques par les chaînes de Markov cachées »

Bouchra Triqui, Omar Behadada, Abdelkader Benyettou

Publisher : Editions universitaires europeennes (April 25, 2012) (French Edition)

ISBN-13 : 978-3841795403

Bibliography

[1] Guiraud-Carrier, C. (1996, December). FLARE: Induction with prior knowledge. In Research and Development in expert Systems XIII. Expert Systems'96, pp. 11 24. SGES Publications.

[2] Hu, J. and J. W. Rozenblit (1991). Knowledge acquisition based on explicit representation. Expert Systems with Applications 3(3), 303–315.

[3] Jeng, B., T.-P. Liang, and M. Hong (1996). Interactive induction of expert knowledge. Expert Systems with Applications 10(3-4), 393–401.

[4] Julien, B. and S. J. Fenves (1996, December). An environmental evaluation learning apprentice system. Engineering Applications of Artificial Intelligence 9(6), 589–599.

[5] Gamberger et al., 2003Gamberger, D., N. Lavrac, and G. Krstacic (2003, May). Active subgroup mining: A case study in coronary heart disease risk group detection. Artificial Intelligence in Medicine 28(1), 27–57.

[6] World Health Report 2017. (Retrieved 11, 02, 2017, from http://www.who.int/whr/en/)

[7] Alonso JM, Luis ML. *An experimental study on the interpretability of fuzzy systems. Proceedings of the Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference, Lisbon, Portugal, 2009; 125–130.*

[8] Gacto MJ, Alcala R, Herrera F. *Interpretability of linguistic fuzzy rule-based systems: an overview of interpretability measures.* Information Sciences 2011 ; 181(20):4340–4360. doi:10.1016/ j.ins.2011.02.021.

[9] Alonso JM. *Interpretable fuzzy systems modeling with cooperation between expert and induced knowledge.* PhD Thesis, 2007.

[10] Rumelhart D, Hinton G, Willams R. *Learning internal representations by error propagation. In Parallel Distribution Proceeding: Exploration in the Microstructure of Cognition, Foundations, Vol. 1, Rumelhart D, McClemlland J (eds).* MIT Press: Cambridge, M.A, 1986; 318–362.

[11] Meau YP, Ibrahim F, Naroinasamy SAL, Omar R. *Intelligent classification of electrocardiogram (ECG) signal using extended Kalman filter 441 (EKF) based neuro fuzzy system. Computer Methods and Programs in Biomedicine.* 2006; 82(2):157–168. Epub 2006 Apr 25.

[12] Yu S, Chou T. *Integration of independent component analysis and neural networks for ECG beat classification. Expert Systems with Applications* 2008; 34(4):2841–2846.

[13] Hosseini HG, Luo D, Reynolds KJ. *The comparison of different feed forward neural network architectures for ECG signal diagnosis. Medical Engineering & Physics* 2006; 28(4):372–378.

[14] Raghupathi W. *Data mining in health care. In Healthcare Informatics: Improving Efficiency and Productivity, Kudyba S (ed).* Taylor & Francis Group: Boca Raton, 2010; 211–223.

[15] Raghupathi W, Raghupathi V. *Big data analytics in healthcare: promise and potential.* Health Information Science and Systems 2014; 2:3.

[16] Andrew T and all, *« The Physiological Basis of theElectrocardiogram" chapter 1 from book* Advanced Methods and Tools for ECG Data Analysis Artech House, Inc. 2006

[17] Mark, R. G., *"Biological Measurement: Electrical Characteristics of the Heart," in Systems & Control Encyclopedia, Singh, M. G., (ed.),* Oxford, U.K.: Permagon Press, 1990, pp. 450-456.

[18] Mark, R. G., *HST.542J/2.792J/BE.371J/6.022J Quantitative Physiology: Organ Trans- Port Systems, Lecture notes from HST/MIT* Open Courseware 2004.

[19] Malmivuo, J., and R. Plonsey, *Bioelectromagnetism: Principles and Applications of Bioelectric and Biomagnetic Fields,* Oxford, U.K.: Oxford University Press, 1995,

[20] Barkley, D., M. Kress, and S. Tucherman, *"Spiral-Wave Dynamics in Simple Model of Excitable Media: Transition from Simple to Compound Rotation,"* Phys. Rev. A., Vol. 42, 1990, pp. 2489–2491.

[21] Ito, H., and L. Glass, *"Spiral Breakup in a New Model of Discrete Excitable Media,"* Phys. Rev. Lett., Vol. 66, No. 5, 1991, pp. 671–674.

[22] Katz, A. M., *Physiology of the Heart, 4th ed., Philadelphia, PA:* Lippincott Williams & Wilkins, 2006.

[23] Fletcher, G. F., et al., *"Exercise Standards; A Statement for Healthcare Professional from the American Heart Association"* Circulation, Vol. 91, 2001, p. 580.

[24] Marriott, H. J. L., *Emergency Electrocardiography,* Naples: Trinity Press, 1997.

[25] Nathanson, L. A., et al., *"ECG Wave-Maven: Self-Assessment Program for Student and Clinicians," http://ecg.bidmc.harvard.edu.*

[26] Pearl J, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference,* Morgan Kaufmann Press, 1998

[27] Sanchez-Graillet O and Poesio M, *Acquiring Bayesian Networks from Text.* Proceedings of LREC, 2004

[28] Girju R and Moldovan D I, *Text Mining for Causal Relations.* Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference, 2002

[29] Blanco E, Castell N and Moldovan D, *Causal Relation Extraction.* Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), 2008

[30] Inui T, Inui K and Matsumoto Y, *Acquiring Causal Knowledge from Text Using the Connective Marker Tame.* ACM Transactions on Asian Language Information Processing (TALIP), Vol. 4, 2005

[31] Cole S, Royal M D, Valtorta M G, Huhns M N and Bowles J B A, *Lightweight Tool for Automatically Extracting Causal Relationships from Text.* Proceedings of the IEEE, pages 125–126, 2006

[32] Karin V, William A. Baumgartner Jr, Unstructured *Information Management Architecture (UIMA)* Encyclopedia of Systems Biology ISBN: 978-1-4419-9862-0 (Print) 978-1-4419-9863-7 (Online)

[33] Toutanova C, Klein D, Manning C and Singer Y*. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network.* In Proceedings of HLT-NAACL, pages 252-259, 2003.

[34] Boguraev B and Neff M, *Navigating through Dense Annotation Spaces.* Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), 2008

[35] Danks D, Griffiths T L and Tenenbaum J B*. Dynamical Causal Learning.* NIPSMIT Press, pages 67–74, 2002

[36] Pearl J. *Causality: Models, Reasoning and Inference Econometric Theory,* Vol 19, pages 675–685, 2003

[37] Raghuram S, Xia Y, Palakal M, Jones J, Pecenka D, Tinsley E, Bandos J and Geesaman J, *Bridging Text Mining and Bayesian Networks.* IEEE Computer Society, pages 298–303, 2009

[38] Navigli R, *Word Sense Disambiguation:* A Survey ACM Computing Surveys, Vol. 41, N.2 2009

[39] Pedrycz,W. (1993). *Fuzzy control and fuzzy systems (2nd ed.). Studies in Fuzziness.* Research Studies Press Ltd, Tauton, Somerset, England.

[40] Pedrycz, W. (1994). *Why triangular membership functions?* Fuzzy Sets and Systems 64(1), 21–30.

[41] Miller, G. A. (1956). *The magical number seven, plus or minus two: Some limits on our capacity for processing information.* The Psychological Review 63(2), 81–97.

[42] J. A. Hartigan et M. A. Wong. Algorithm AS 136: A K-Means Clustering Algorithm, Journal of the Royal Statistical Society, Series C, vol. 28, no 1,) 108–100 p. ,1979 JSTOR 2346830)

[43] Guillaume, S. and B. Charnomordic (2003). *A new method for inducing a set of interpretable fuzzy partitions and fuzzy inference systems from data. In Casillas et al. (2003c), 148–175.*

[44] Guillaume, S., B. Charnomordic, and J.-L. Labl´ee (2002). *FisPro (Fuzzy Inference System Professional): An open source portable software for designing fuzzy inference systems.*

[45] Alonso, J. M., S. Guillaume, and L. Magdalena (2003). *KBCT: A knowledge management tool for fuzzy inference systems. Free software under GPL license, available in* http://www.mat.upm.es/projects/advocate/kbct.htm.

[46] Mamdani, E. H. (1977). *Application of fuzzy logic to approximate reasoning using linguistic synthesis,* IEEE Transactions on Computers 26 (12): 1182–1191.

[47] Serge Guillaume and Brigitte Charnomordic (2001). *Induire un partitionnement flou interprétable.* In LFA'01, pages 113-120, Toulouse, France, Cépaduès Editions.

[48] Ḟurnkranz, J. and E. Ḧullermeier. 2005. *Preference learning*. Ḳunstliche Intelligenz.

[49] Wang, L.X., and J.M. Mendel (1992). *Back-propagation fuzzy system as non-linear dynamic system identifiers.* In IEEE Conf. on Fuzzy Systems. pp. 1409–1416.

[50] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.I. (1984). *Classification and regression trees.* Belmont, Calif.: Wadsworth.

[51] Quinlan, J.R. (1985). *The effect of noise on concept learning. In R.S. Michalski, J.G. Carbonell & T.M. Mitchell (Eds.),* Machine learning. Los Altos: Morgan Kaufmann.

[52] Quinlan, J.R. (1985). *Decision trees and multi-valued attributes. In J.E. Hayes & D. Michie (Eds.),* Machine intelligence 11. Oxford University Press.

[53] Guillaume, S. and L. Magdalena (2006). *Expert guided integration of induced knowledge into a fuzzy knowledge base. Soft Computing* - A Fusion of Foundations, Methodologies and Applications 10(9), 773–784.

[54] Glorennec, P.-Y. (1999). *Algorithmes d'apprentissage pour syst`emes d0inf´erence floue.* Editions Herm`es, Paris.

[55] J.M.Alonso,L.Magdalena *"Generating Understandable and Accurate Fuzzy Rule-Based Systems in a JavaEnvironment", Lecture Notes in Artificial Intelligence* - 9th International Workshop on Fuzzy Logic and Applications, Springer-Verlag, LNAI6857:212-219, 2011 http://dx.doi.org/10.1007/978-3-642-23713-3_27

[56] J. M. Alonso, L. Magdalena *"HILK++: an interpretability-guided fuzzy modeling methodology for learning readable and comprehensible fuzzy rule-based classifiers",* Soft Computing, 15:1959-1980, 2011 http://dx.doi.org/10.1007/s00500-010-0628-5

[57]11. Manning CD. *Foundations of Statistical Natural Language Processing. MIT Press:* Cambridge, M.A, 1999.

[58] De Marneffe MF, MacCartney B, Manning CD. *Generating typed dependency parses from phrase structure parses,* LREC, 2006.

[59] Liu B. *Synthesis Lectures on Human Language Technologies. Sentiment Analysis and Opinion Mining.* Morgan & Claypool Publishers, 2012; 5(1):1–167. DOI:10.2200/S00416ED1V01Y201204HLT016

[60] PubMed*. (Retrieved 06 10, 2014, from http://www.ncbi.nlm.nih.gov/pubmed/)*

[61] Schönbauer R, Sommers P, Misfeld M, Dinov B, Fiedler F, Huo Y, Arya A. *Relevant ventricular septal defect caused by steam pop during ablation of premature ventricular contraction.* Circulation 2013; 127(24):e843–e844. DOI: 10.1161/CIRCULATIONAHA.112.130195.

[62] Soheilykhah S, Sheikhani A, Sharif AG, Daevaeiha MM. *Localization of premature ventricular contraction foci in normal individuals based on multichannel electrocardiogram signals processing.* Springerplus 2013; 2:486. DOI: 10.1186/2193-1801-2-486

[63] Moody GB, Mark RG. *The impact of the MIT-BIH arrhythmia database.* IEEE Eng in Med and Biol 2001; 20(3):45–50.

[64] Pan J, Tompkins WJ. *A Real-Time QRS Detection Algorithm.* IEEE Transactions ON Biomedical Engineering 1985; 32(3):230–236.

[65] Guillaume S, Magdalena L. *Expert guided integration of induced knowledge into a fuzzy knowledge base.* Soft Computing 2006; 10:773–784.

[66] Piotrkiewicz M, Kudina L, Mierzejewska J, Jakubiec M, Hausmanowa-Petrusewicz I. *Age-related change in duration of afterhyperpolarization of human motoneurones.* The Journal of Physiology 2007; 585:483–490.

[67] Casillas J. *Accuracy Improvements in Linguistic Fuzzy Modelling.* Springer Science & Business Media: Berlin Heidelberg, 2003.

[68] Chen T. *An effective fuzzy collaborative forecasting approach for predicting the job cycle time in wafer fabrication.* Computers & Industrial Engineering 2013; 66(4):834–848.

[69] Guillaume S, Charnomordic B. *Fuzzy inference systems: an integrated modeling environment for collaboration between expert knowledge and data using FisPro.* Expert Systems with Applications 2012; 39(10):8744–8755.

[70] NauckDD. *Measuring interpretability in rule-based classification systems. In Proceedings of the* FUZZ-IEEE, St. Louis, Missouri, USA, 2003; 196–201. DOI: 10.1109/FUZZ.2003.1209361.

[71] CHRISTOPHER M. BISHOP, *Neural Networks for Pattern Recognition,* CLARENDON PRESS, OXFORD, 1995.

[72] Kohonen, T. "Generalizations of the self organizing map", Neural *Networks, 1993. IJCNN '93-Nagoya. Proceedings of 1993 International Joint Conference on,* On *page(s): 457 - 462 vol.1 Volume: 1, 25-29 Oct. 1993*

[73] L. Rabiner, B. Juang. *An introduction to hidden Markov models, IEEE ASSP Magazine (Volume:3 , Issue: 1) Page(s): 4 – 16, ISSN :0740-7467, Jan 1986*

[74] L. A. Zadeh. *The concept of a linguistic variable and its application to approximate reasoning.* Parts I, II, and III. Information Sciences, 8, 8, 9:199–249, 301–357, 43–80, 1975.

[75] J. M. Alonso, L. Magdalena, and S. Guillaume. *HILK: A new methodology for designing highly interpretable linguistic knowledge bases using the fuzzy logic formalism.* International Journal of Intelligent Systems, 23(7):761–794, 2008.

[76] A. Mun˜oz, A. Vera, J. A. Bot´ıa, and A. F. G´omez Skarmeta. *Defining basic behaviours in ambient intelligence environments by means of rule-based programming with visual tools.* In 1st Workshop of Artificial Intelligence Techniques for Ambient Intelligence. ECAI, 2006.

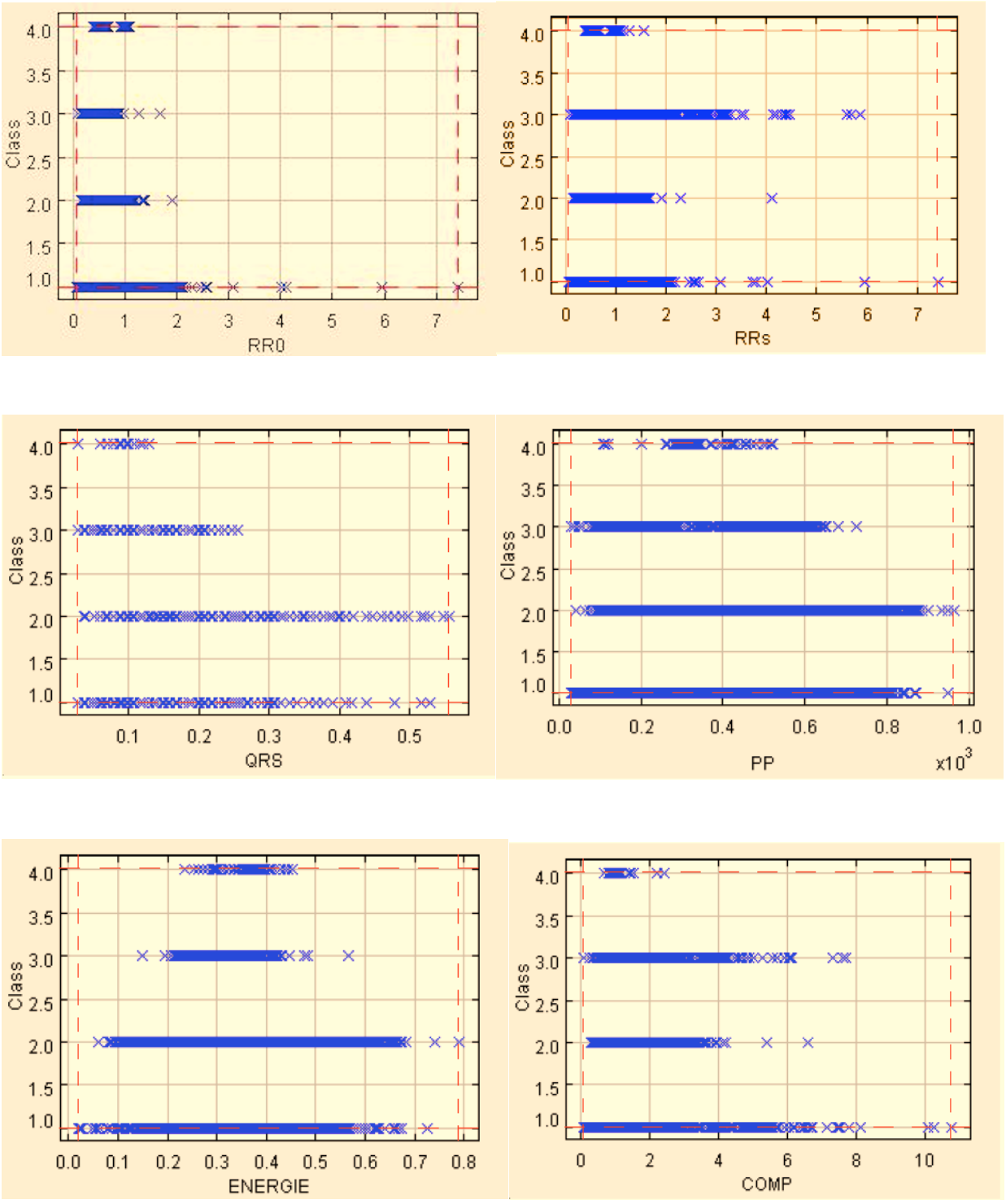
[77] F. Hayes-Roth, D. A. Waterman, and D. B. Lenat. *Building expert systems.* Addison-Wesley,1983.

[78] A. L. Kidd. *Knowledge elicitation for expert systems: A practical handbook.* Plenum Press, 1987.

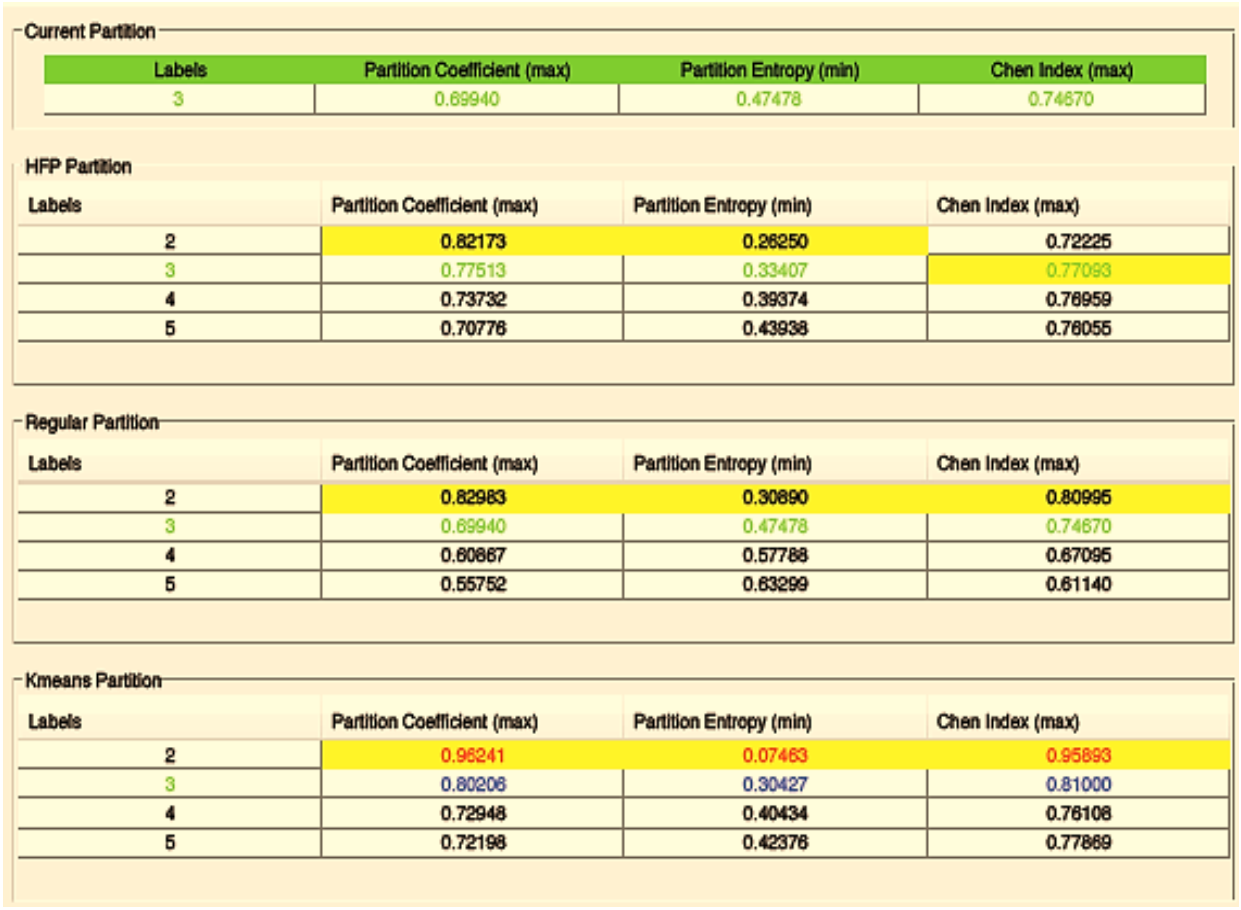
[79] E. H. Ruspini. *A new approach to clustering. In- formation and Control,* 15(1):22–32, 1969.

Annex A

A Experimentation results



Figures 7. (1,2,4,5,6) RR0,RRS, QRS, PP, COMP and ENERGY features with classes



Current Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
3	0.69940	0.47478	0.74670

HFP Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.82173	0.26250	0.72225
3	0.77513	0.33407	0.77093
4	0.73732	0.39374	0.76959
5	0.70776	0.43938	0.76055

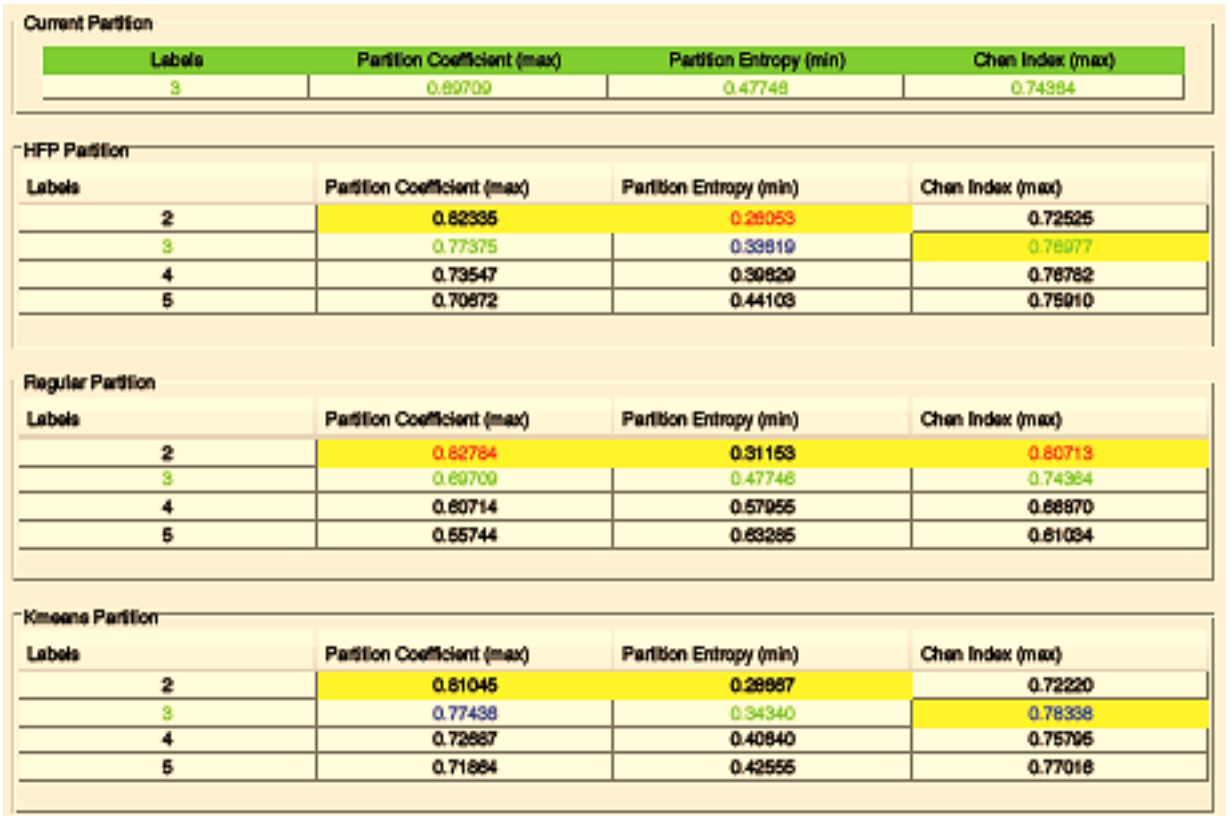
Regular Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.82983	0.30890	0.80995
3	0.69940	0.47478	0.74670
4	0.60867	0.57788	0.67095
5	0.55752	0.63299	0.61140

Kmeans Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.96241	0.07463	0.95893
3	0.80206	0.30427	0.81000
4	0.72948	0.40434	0.76108
5	0.72198	0.42376	0.77869

Figure 7.7 RR0 fuzzy partition quality.



Current Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
3	0.69709	0.47748	0.74384

HFP Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.82335	0.26053	0.72525
3	0.77375	0.33819	0.76977
4	0.73547	0.39829	0.76782
5	0.70672	0.44103	0.75910

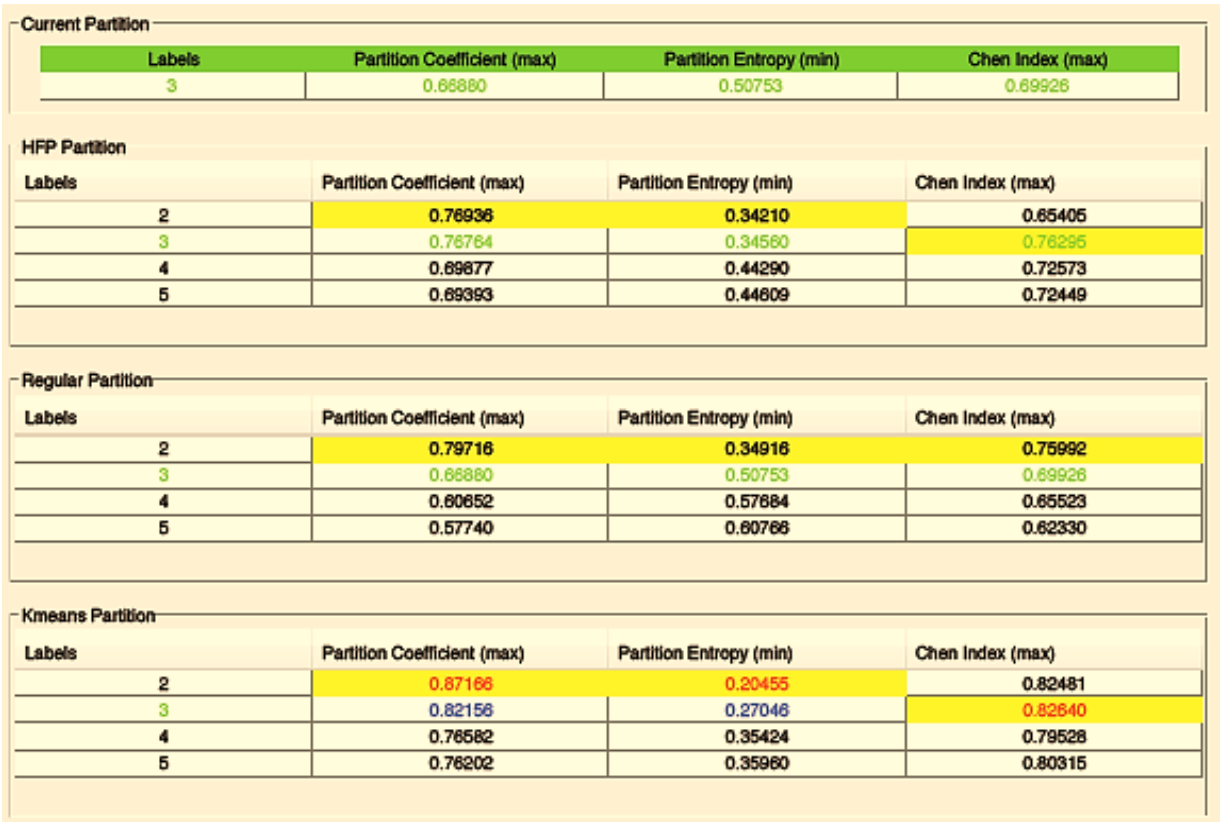
Regular Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.82784	0.31153	0.80713
3	0.69709	0.47748	0.74384
4	0.60714	0.57955	0.66970
5	0.55744	0.63285	0.61034

Kmeans Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.81045	0.28867	0.72220
3	0.77438	0.34340	0.78338
4	0.72887	0.40840	0.75795
5	0.71884	0.42555	0.77016

Figure 7.8 RRS fuzzy partition quality.



Current Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
3	0.66880	0.50753	0.69926

HFP Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.76936	0.34210	0.65405
3	0.76764	0.34560	0.76295
4	0.69677	0.44290	0.72573
5	0.69393	0.44609	0.72449

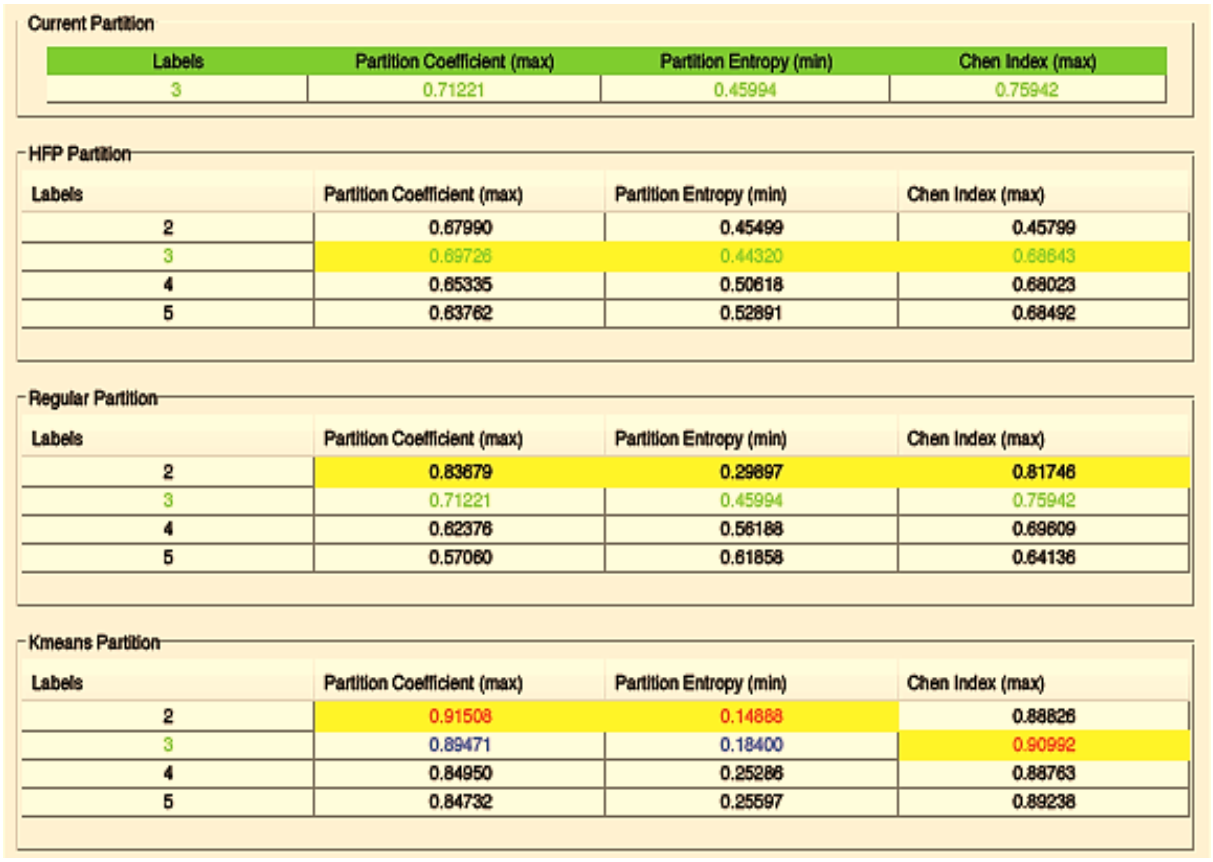
Regular Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.79716	0.34916	0.75992
3	0.66880	0.50753	0.69926
4	0.60652	0.57684	0.65523
5	0.57740	0.60766	0.62330

Kmeans Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.87166	0.20455	0.82481
3	0.82156	0.27046	0.82640
4	0.76582	0.35424	0.79528
5	0.76202	0.35960	0.80315

Figure 7.9 QRS fuzzy partition quality.



Current Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
3	0.71221	0.45994	0.75942

HFP Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.67990	0.45499	0.45799
3	0.69726	0.44320	0.68643
4	0.65335	0.50618	0.68023
5	0.63762	0.52891	0.68492

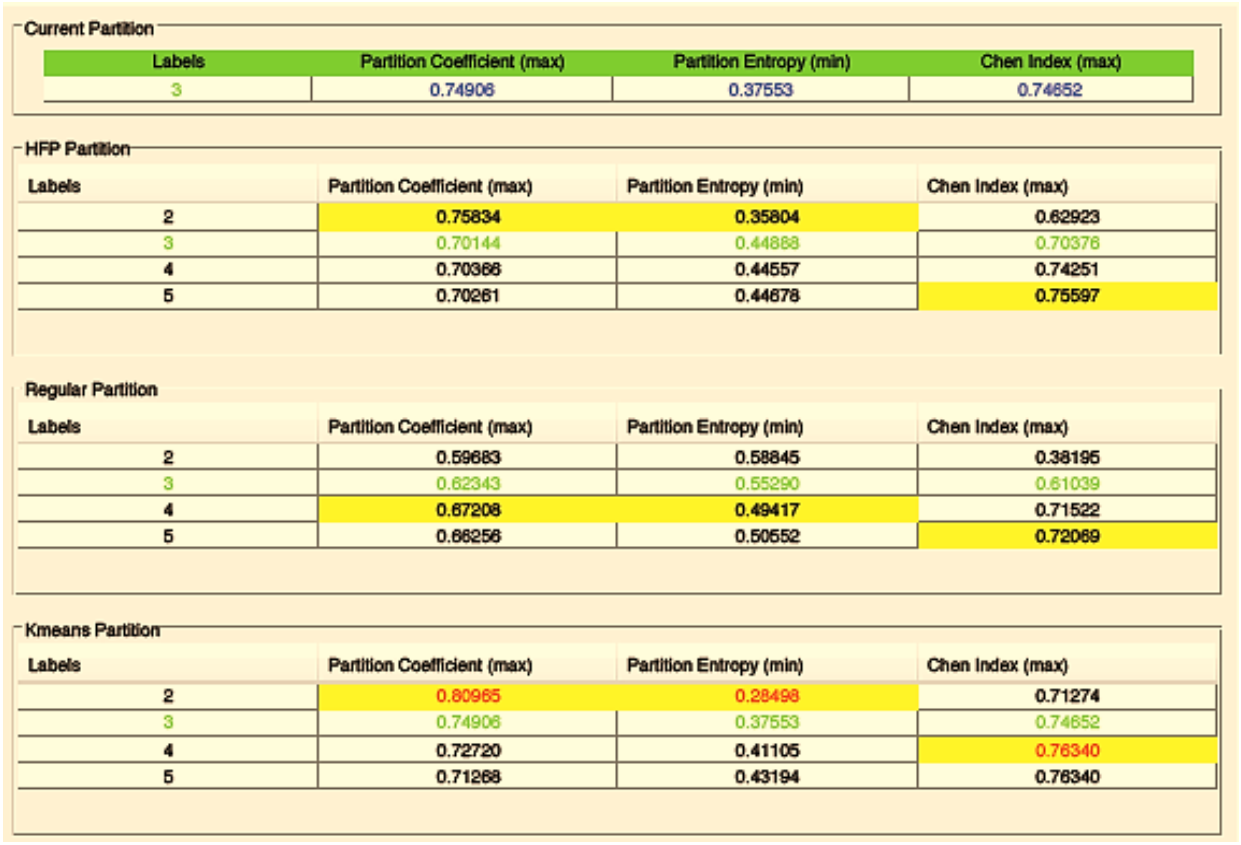
Regular Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.83679	0.29897	0.81746
3	0.71221	0.45994	0.75942
4	0.62376	0.56188	0.69609
5	0.57060	0.61858	0.64136

Kmeans Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.91508	0.14888	0.88826
3	0.89471	0.18400	0.90992
4	0.84950	0.25286	0.88763
5	0.84732	0.25597	0.89238

Figure 7.10 COMP fuzzy partition quality.



Current Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
3	0.74906	0.37553	0.74652

HFP Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.75834	0.35804	0.62923
3	0.70144	0.44888	0.70376
4	0.70366	0.44557	0.74251
5	0.70261	0.44678	0.75597

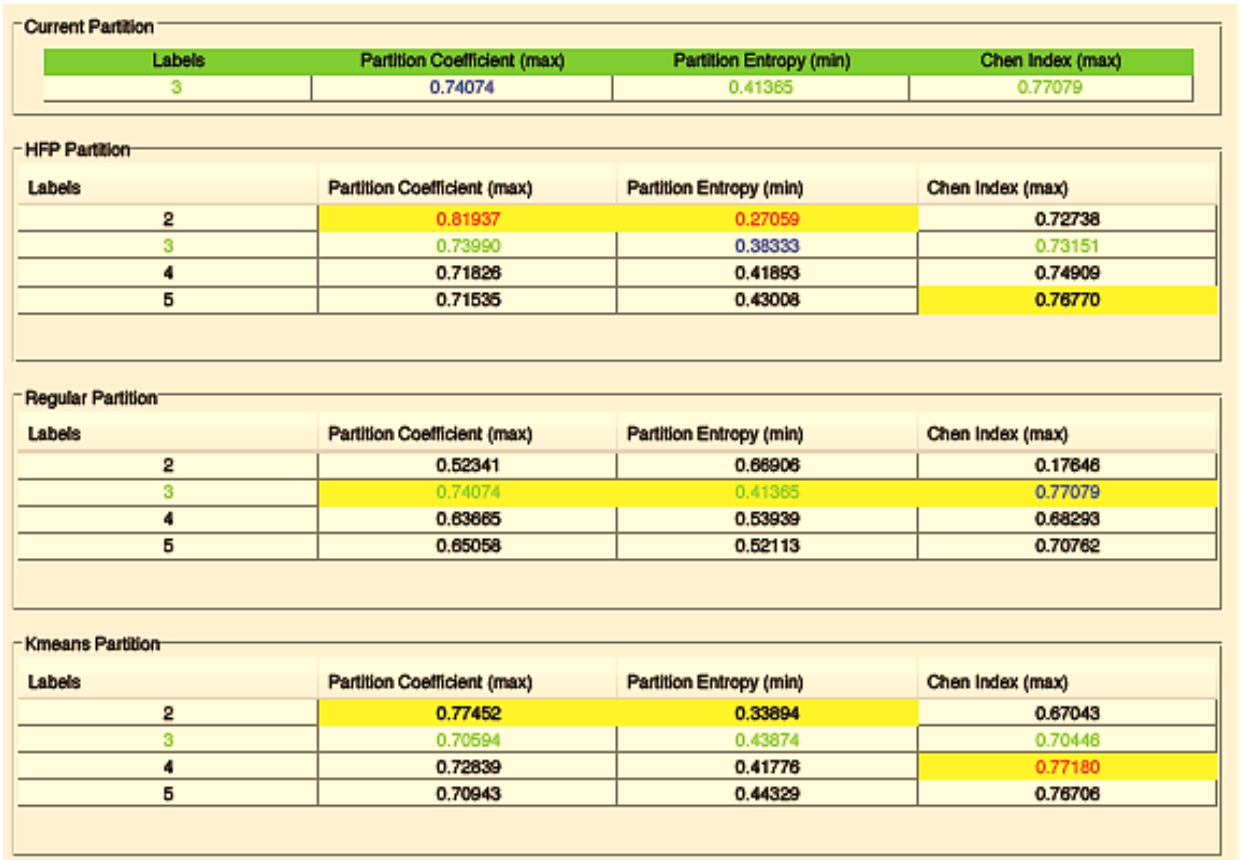
Regular Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.59683	0.58845	0.38195
3	0.62343	0.55290	0.61039
4	0.67208	0.49417	0.71522
5	0.66256	0.50552	0.72069

Kmeans Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.80965	0.28498	0.71274
3	0.74906	0.37553	0.74652
4	0.72720	0.41105	0.76340
5	0.71268	0.43194	0.76340

Figure 7.11 PP fuzzy partition quality.



Current Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
3	0.74074	0.41365	0.77079

HFP Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.81937	0.27059	0.72738
3	0.73990	0.38333	0.73151
4	0.71826	0.41893	0.74909
5	0.71535	0.43008	0.76770

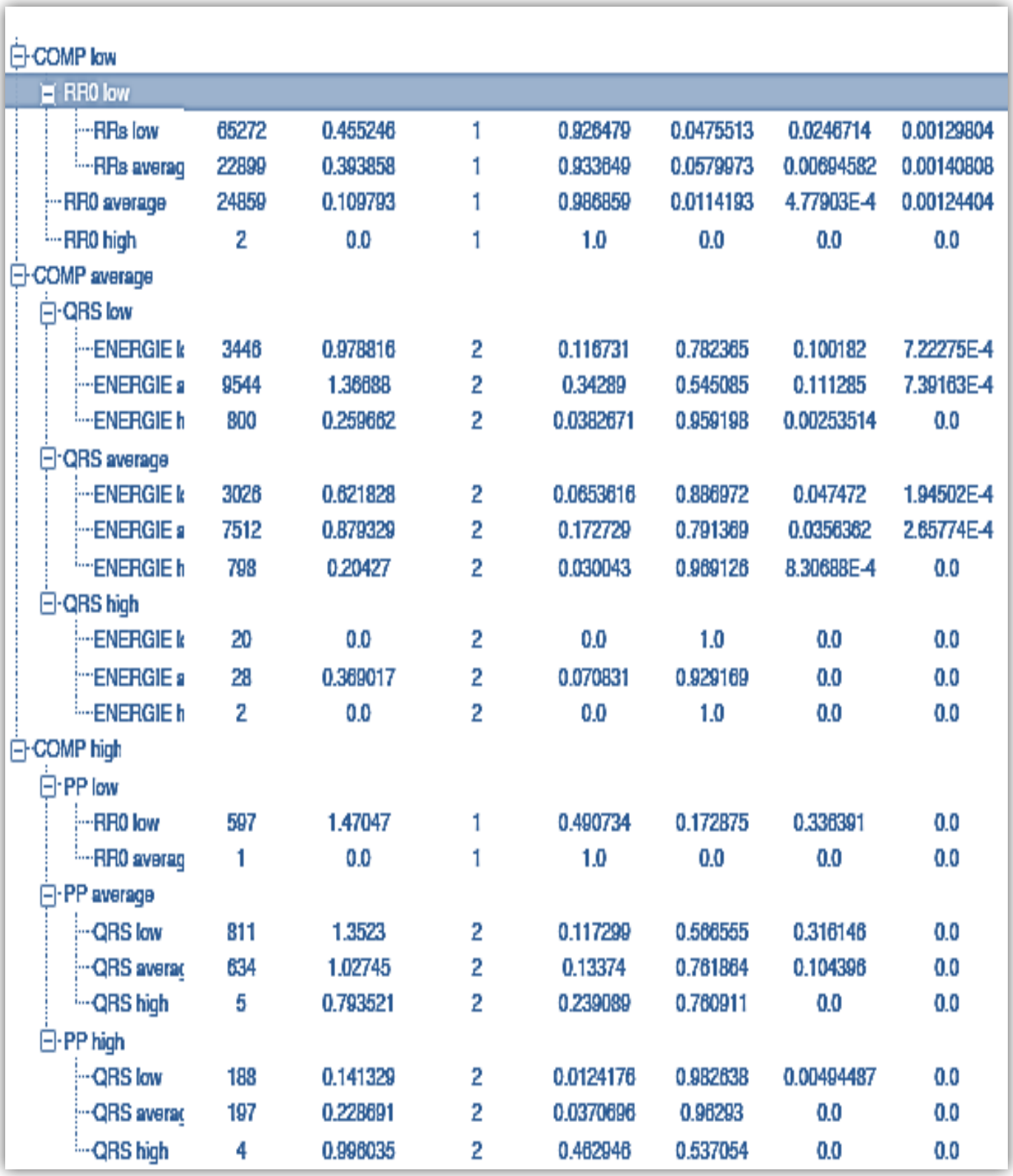
Regular Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.52341	0.66906	0.17646
3	0.74074	0.41365	0.77079
4	0.63665	0.53939	0.68293
5	0.65058	0.52113	0.70762

Kmeans Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.77452	0.33894	0.67043
3	0.70594	0.43874	0.70446
4	0.72839	0.41776	0.77180
5	0.70943	0.44329	0.76706

Figure 7.12 ENERGY fuzzy partition quality.



COMP low							
RR0 low							
RRs low	65272	0.455246	1	0.926479	0.0475513	0.0246714	0.00129804
RRs averag	22899	0.393858	1	0.933649	0.0579973	0.00694582	0.00140808
RR0 average	24859	0.109793	1	0.986859	0.0114193	4.77903E-4	0.00124404
RR0 high	2	0.0	1	1.0	0.0	0.0	0.0
COMP average							
QRS low							
ENERGIE k	3446	0.978816	2	0.116731	0.782365	0.100182	7.22275E-4
ENERGIE a	9544	1.36888	2	0.34289	0.545085	0.111285	7.39163E-4
ENERGIE h	800	0.259662	2	0.0382671	0.959198	0.00253514	0.0
QRS average							
ENERGIE k	3026	0.621828	2	0.0653616	0.886972	0.047472	1.94502E-4
ENERGIE a	7512	0.879329	2	0.172729	0.791369	0.0356362	2.65774E-4
ENERGIE h	798	0.20427	2	0.030043	0.969126	8.30688E-4	0.0
QRS high							
ENERGIE k	20	0.0	2	0.0	1.0	0.0	0.0
ENERGIE a	28	0.369017	2	0.070831	0.929169	0.0	0.0
ENERGIE h	2	0.0	2	0.0	1.0	0.0	0.0
COMP high							
PP low							
RR0 low	597	1.47047	1	0.490734	0.172875	0.336391	0.0
RR0 averag	1	0.0	1	1.0	0.0	0.0	0.0
PP average							
QRS low	811	1.3523	2	0.117299	0.586555	0.316146	0.0
QRS averag	634	1.02745	2	0.13374	0.761864	0.104396	0.0
QRS high	5	0.793521	2	0.239089	0.760911	0.0	0.0
PP high							
QRS low	188	0.141329	2	0.0124176	0.982638	0.00494487	0.0
QRS averag	197	0.228691	2	0.0370696	0.96293	0.0	0.0
QRS high	4	0.996035	2	0.462946	0.537054	0.0	0.0

Figure 7.13 decision tree of FDT1 (taken from gauge software)

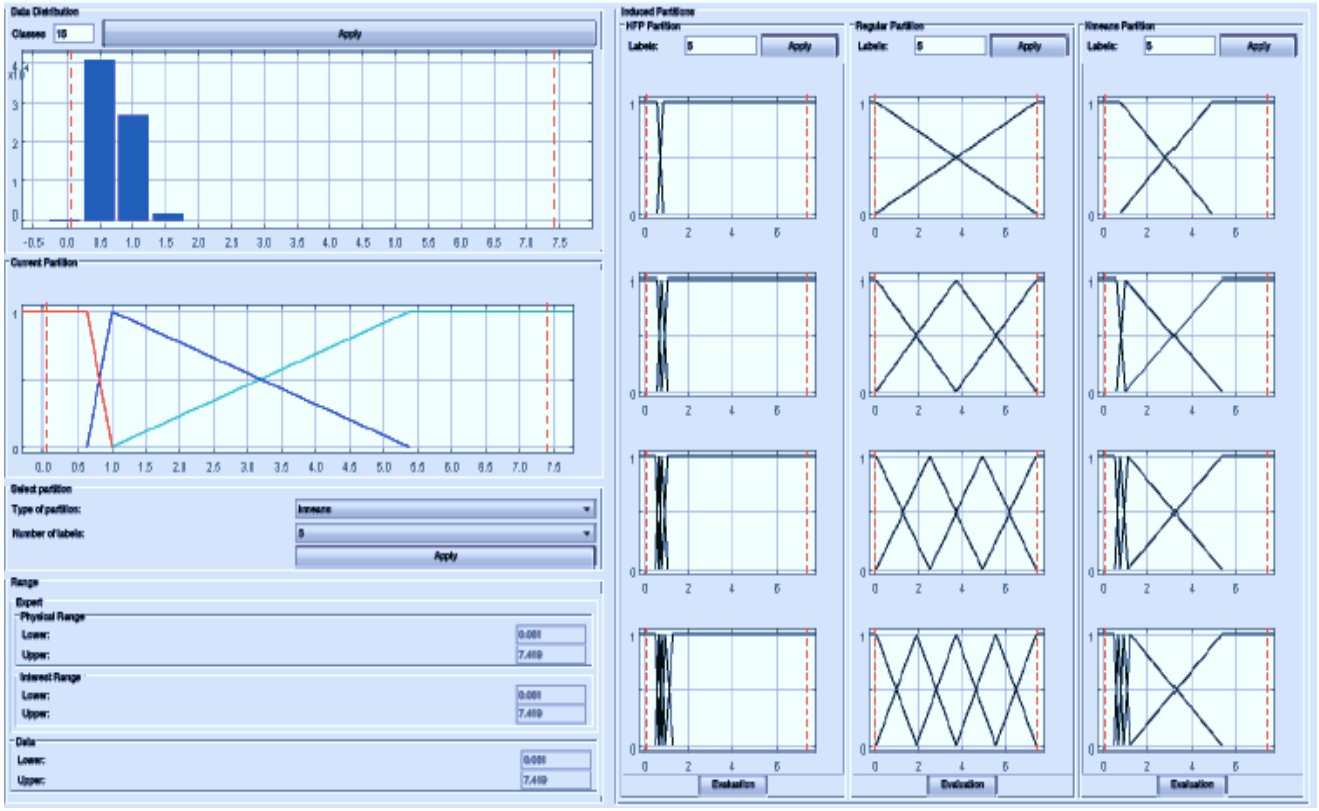


Figure 7.14 RR0 fuzzy partition

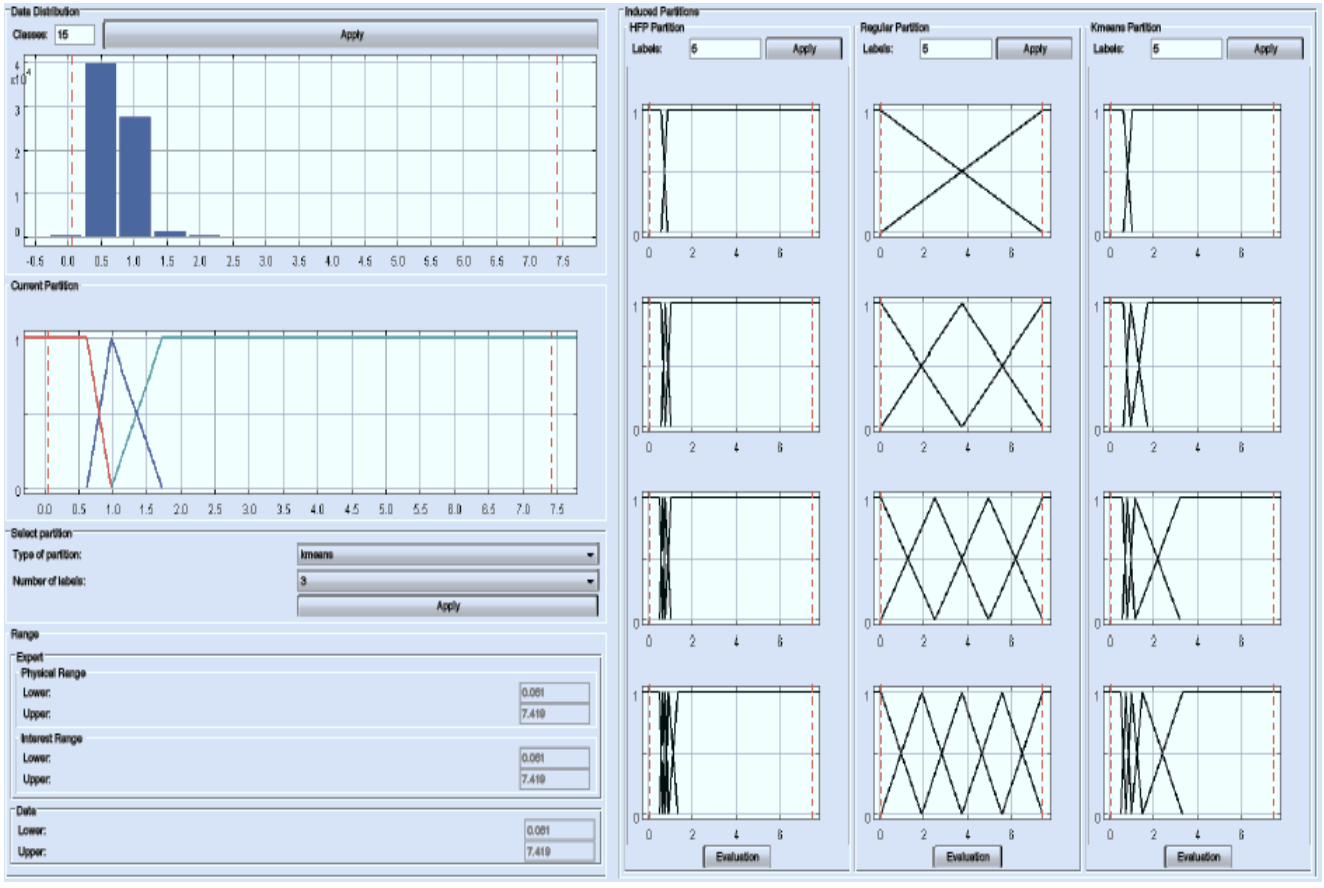


Figure 7.15 RRS fuzzy partition.

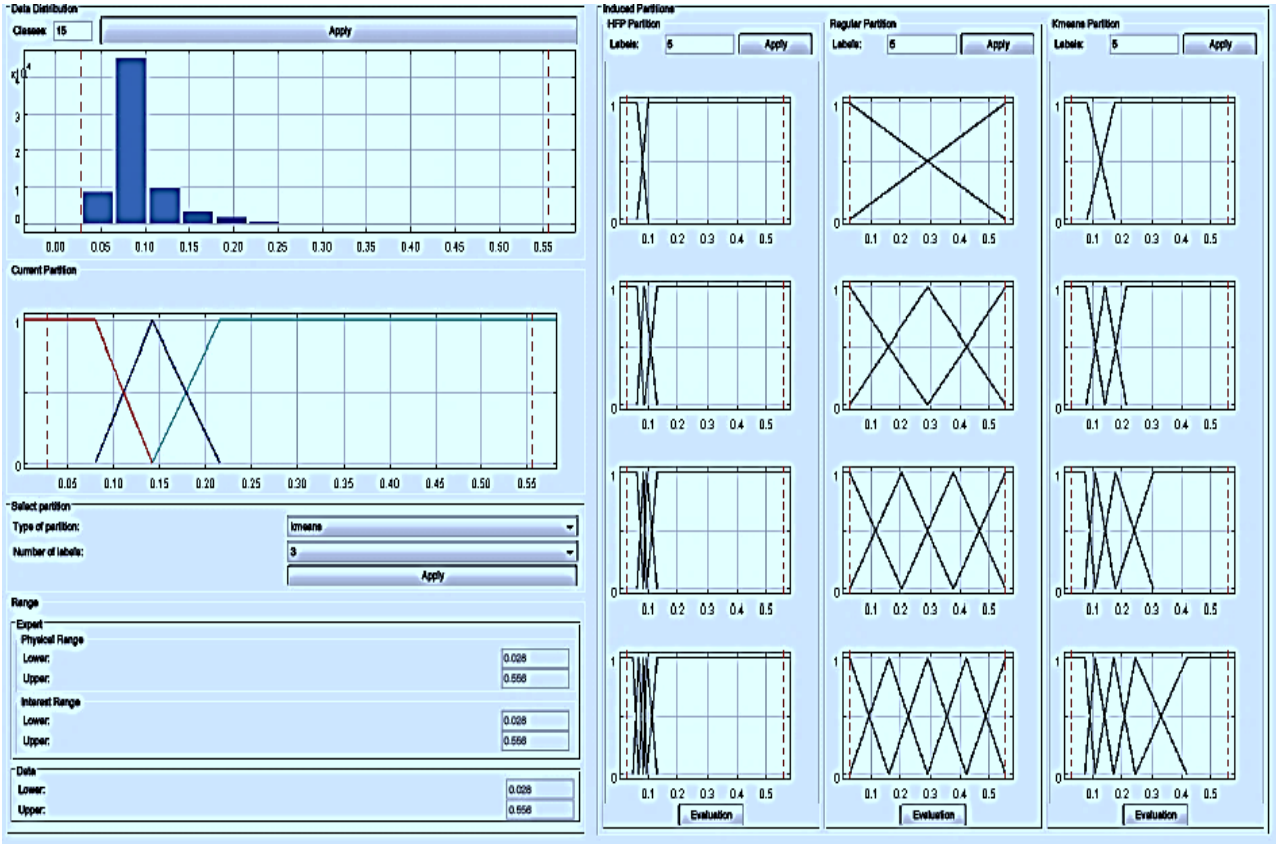


Figure 7.16 QRS fuzzy partition

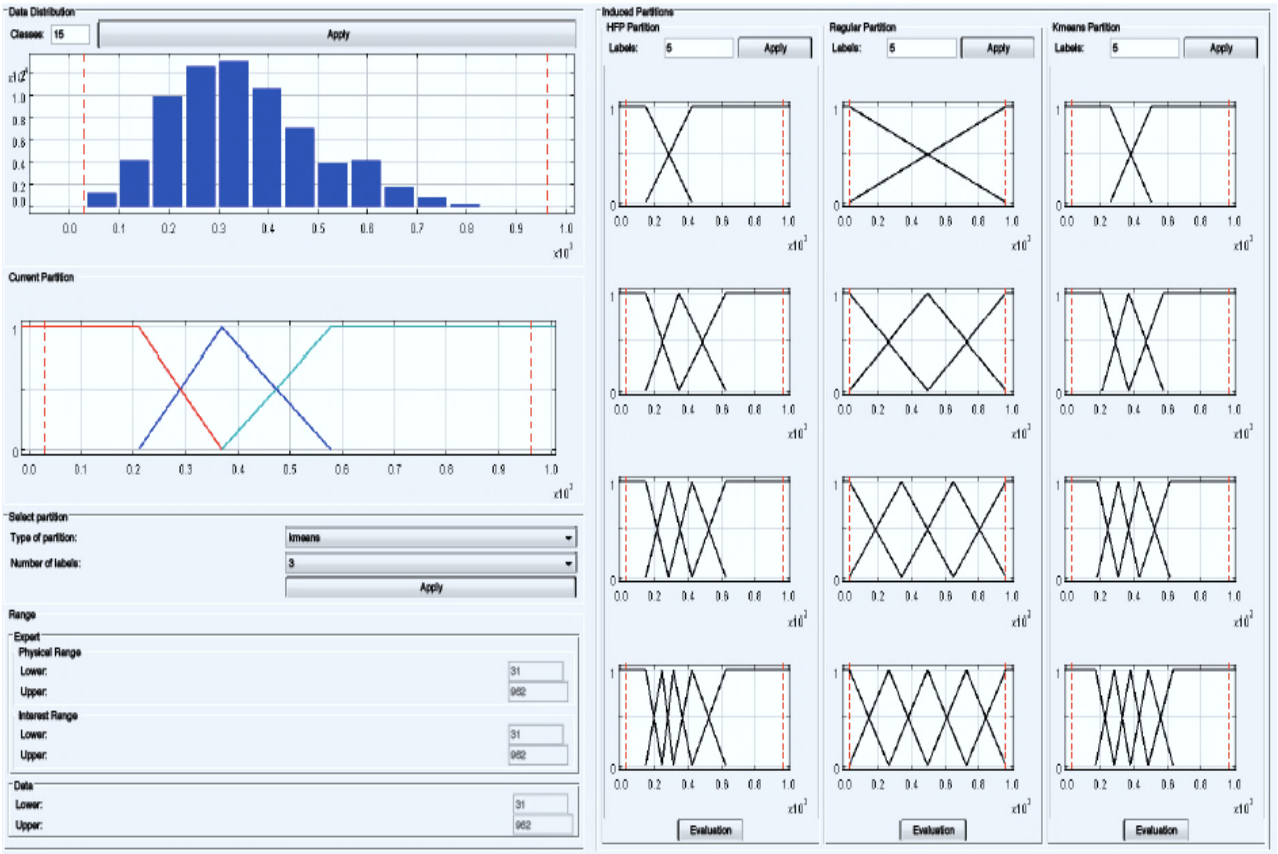


Figure 7.17 PP fuzzy partition

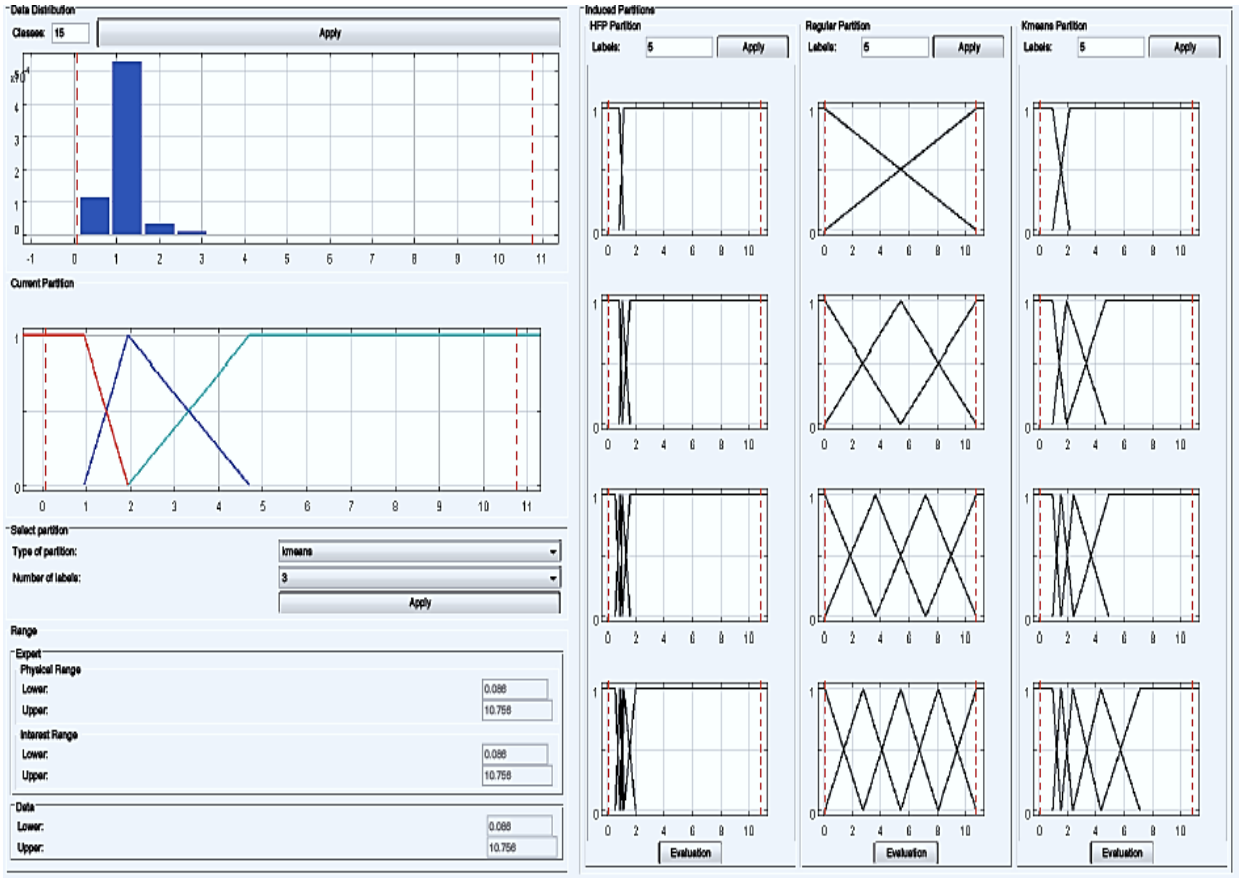


Figure 7.18 COMP fuzzy partition

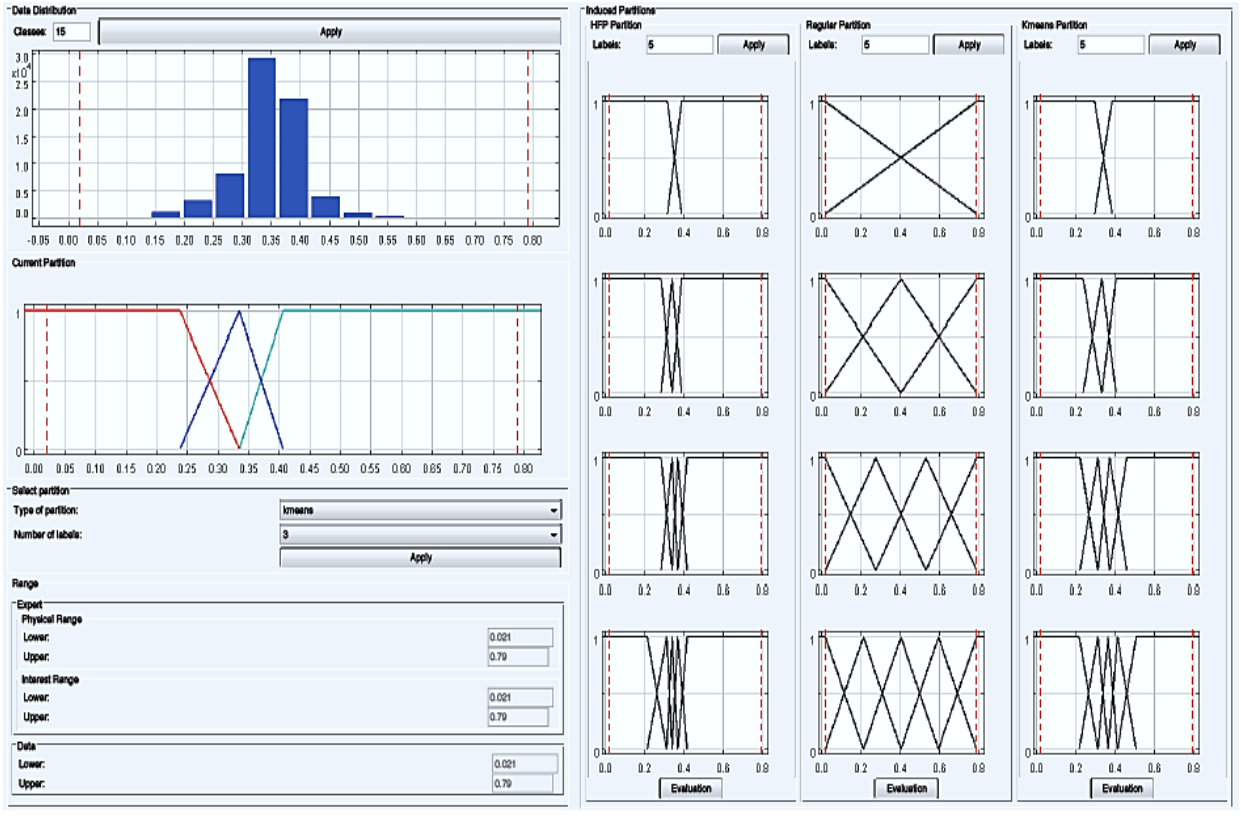


Figure 7.19 ENERGY fuzzy partition

	69115	0.666763	1	0.870882	0.0970846	0.0308327	0.0012009
COMP low	65277	0.423449	1	0.933412	0.0420289	0.0233043	0.00125474
QRS low	60633	0.208168	1	0.972095	0.00475455	0.0221337	0.00101649
QRS average	18041	0.978917	1	0.755452	0.203127	0.0383175	0.00310286
QRS high	3018	0.831532	2	0.211246	0.77708	0.0116741	0.0
COMP average	9715	1.24162	2	0.271898	0.638314	0.0891963	5.90889E-4
QRS low	4906	1.32808	1	0.631011	0.165424	0.202169	0.00139624
QRS average	6211	0.844492	2	0.154958	0.807842	0.0369295	2.69729E-4
QRS high	3355	0.25759	2	0.0302069	0.962468	0.00732537	0.0
COMP high	949	1.5602	1	0.390374	0.359953	0.249673	0.0

Figure 7.20 decision tree of FDT2. (taken from gauge software)

Current Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
3	0.73400	0.43148	0.78054

HFP Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.74456	0.37536	0.60696
3	0.71742	0.42603	0.72151
4	0.71276	0.43173	0.75003
5	0.67994	0.48000	0.73434

Regular Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.85120	0.27724	0.83541
3	0.73400	0.43148	0.78054
4	0.64839	0.53244	0.71196
5	0.59437	0.59332	0.65988

Kmeans Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.97037	0.06205	0.96865
3	0.93942	0.10802	0.95409
4	0.90835	0.14795	0.93315
5	0.86098	0.22370	0.89483

Figure 7.21 RR0 with 80 sample of each classes fuzzy partition quality.

Current Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
3	0.68424	0.49140	0.72750

HFP Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.82294	0.26535	0.73359
3	0.74131	0.38981	0.74409
4	0.69544	0.45588	0.73262
5	0.68837	0.46432	0.74001

Regular Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.81887	0.32319	0.79563
3	0.68424	0.49140	0.72750
4	0.59610	0.59054	0.64465
5	0.55446	0.63595	0.59493

Kmeans Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.96910	0.06832	0.96765
3	0.93435	0.12338	0.95133
4	0.81335	0.28835	0.83890
5	0.80803	0.29716	0.84472

Figure 7.22 RRS with 80 samples of each classes fuzzy partition quality.

Current Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
3	0.61441	0.57003	0.61912

HFP Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.75899	0.35838	0.63764
3	0.71403	0.42649	0.71641
4	0.74434	0.38469	0.76158
5	0.72644	0.41040	0.77737

Regular Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.75108	0.40909	0.69496
3	0.61441	0.57003	0.61912
4	0.58886	0.59602	0.62596
5	0.60026	0.57912	0.64829

Kmeans Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.94247	0.10262	0.93389
3	0.84830	0.23268	0.85153
4	0.81181	0.29240	0.83860
5	0.79871	0.30924	0.83629

Figure 7.23 QRS with 80 sample of each classes fuzzy partition quality.

Current Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
3	0.66904	0.49628	0.66243

HFP Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.82524	0.26199	0.73320
3	0.73654	0.38917	0.72890
4	0.70912	0.43202	0.73916
5	0.70443	0.44633	0.75757

Regular Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.57583	0.61138	0.31714
3	0.66904	0.49628	0.66243
4	0.64933	0.52298	0.69531
5	0.68885	0.47498	0.75173

Kmeans Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.81679	0.27878	0.72981
3	0.77262	0.35368	0.79125
4	0.71397	0.43459	0.75689
5	0.72588	0.42206	0.78138

Figure 7.24 PP with 80 sample of each classes fuzzy partition quality.

Current Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
3	0.66361	0.51607	0.69581

HFP Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.85402	0.23812	0.80909
3	0.72206	0.41228	0.71475
4	0.69579	0.45331	0.72998
5	0.69324	0.46069	0.74529

Regular Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.80181	0.34606	0.77181
3	0.66361	0.51607	0.69581
4	0.58541	0.60396	0.63873
5	0.56700	0.62097	0.62494

Kmeans Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.96272	0.07223	0.95911
3	0.87645	0.18836	0.87507
4	0.87187	0.19569	0.88762
5	0.86390	0.20870	0.88830

Figure 7.25 COMP with 80 sample of each classes fuzzy partition quality.

Current Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
3	0.72895	0.42148	0.73645

HFP Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.81052	0.28581	0.72103
3	0.72297	0.40985	0.71418
4	0.70750	0.43759	0.74521
5	0.71203	0.42986	0.75308

Regular Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.53804	0.65350	0.21161
3	0.72895	0.42148	0.73645
4	0.62322	0.55458	0.65993
5	0.70645	0.45444	0.77282

Kmeans Partition

Labels	Partition Coefficient (max)	Partition Entropy (min)	Chen Index (max)
2	0.84443	0.24000	0.77586
3	0.80389	0.31239	0.81906
4	0.72512	0.41977	0.77056
5	0.68883	0.47135	0.74630

Figure 7.26 ENERGY with 80 sample of each classes fuzzy partition quality.

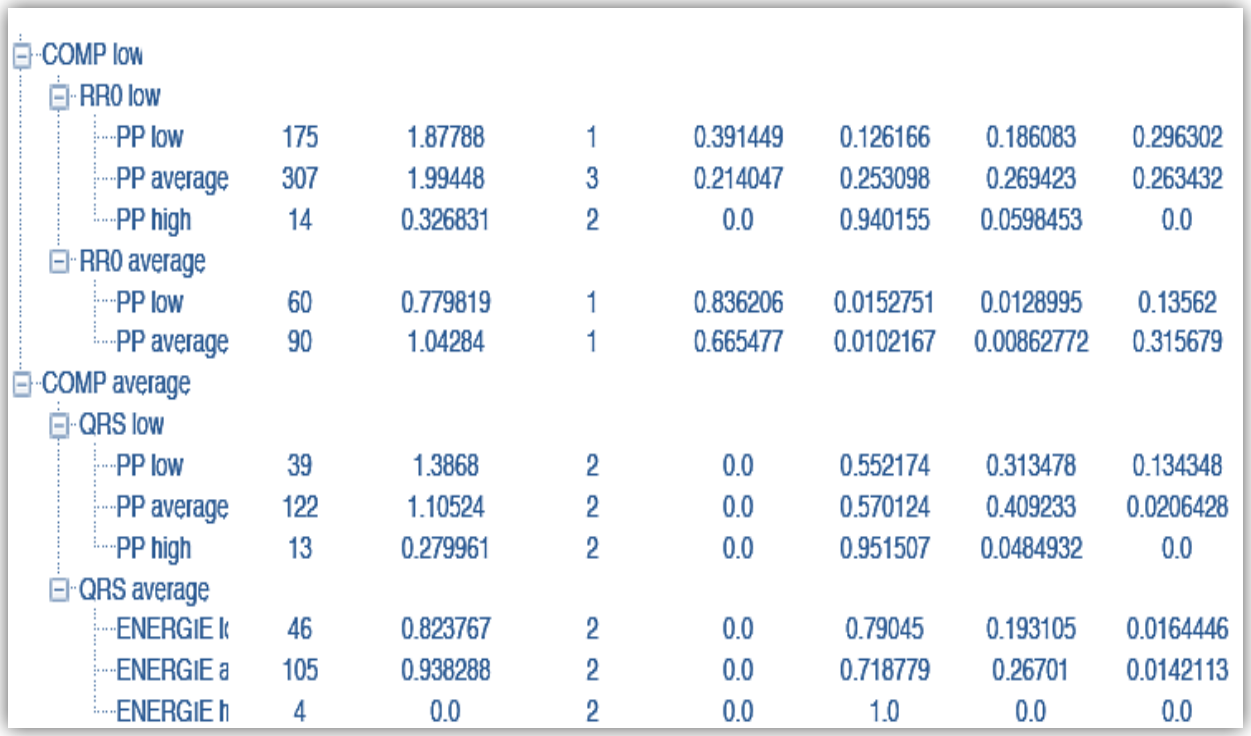


Figure 7.27 decision tree FDT3 (taken from gauge software)

COMP low							
QRS low							
ENERGIE l	65	1.66093	4	0.100242	0.0726554	0.406478	0.420624
ENERGIE a	134	1.53038	4	0.193119	0.0681552	0.125368	0.613358
QRS average							
ENERGIE l	42	1.49929	2	0.0	0.491937	0.292847	0.215216
ENERGIE a	60	1.13658	2	0.0	0.519135	0.0240245	0.45684
ENERGIE h	2	0.0	2	0.0	1.0	0.0	0.0
COMP average							
QRS low							
PP low	22	1.28446	3	0.0	0.164571	0.649285	0.186144
PP average	28	0.501535	3	0.0	0.110538	0.889462	0.0
PP high	7	0.0	3	0.0	0.0	1.0	0.0
QRS average							
ENERGIE l	17	0.998471	2	0.0	0.523013	0.476987	0.0
ENERGIE a	14	0.692598	2	0.0	0.814202	0.185798	0.0
ENERGIE h	1	0.0	2	0.0	1.0	0.0	0.0

Figure 7.28 decision tree of FDT4 . (taken from gauge software)

Rule	Type	If RR0	AND RRs	AND COMP	AND QRS	AND PP	AND ENERGI	THEN Class
1	I	low	low	low	low	low	average	1.0
2	I	low	low	low	low	average	average	1.0
3	I	low	low	low	low	low	low	1.0
4	I	low	low	average	low	average	average	2.0
5	I	low	low	average	average	high	average	2.0
6	I	low	low	low	average	average	low	2.0
7	I	low	low	average	average	low	average	2.0
8	I	low	low	average	average	low	low	2.0
9	I	low	low	average	low	low	low	2.0
10	I	low	low	low	average	low	average	2.0
11	I	low	low	average	average	average	average	2.0
12	I	low	low	low	average	average	average	2.0
13	I	low	low	average	average	average	low	2.0
14	I	low	low	average	low	average	low	2.0
15	I	low	low	average	low	low	average	3.0
16	I	low	low	low	average	low	low	3.0
17	I	low	low	low	low	average	low	4.0

Figure 7.29 Rules of FDT5. (taken from gauge software)

QRS low							
COMP low							
RR0 low	9883	0.167874	1	0.97664	4.75005E-4	0.0221655	7.19707E-4
RR0 avera	16269	0.215141	1	0.96605	5.56573E-5	0.0338257	6.88937E-5
RR0 high	11649	0.00464003	1	0.999663	7.83316E-5	2.58922E-4	0.0
COMP averag							
RRs low	11240	0.157122	1	0.979417	0.00106926	0.0183512	0.00116251
RRs avera	17270	0.378373	1	0.92733	8.74024E-5	0.0724282	1.54096E-4
RRs high	11056	0.0199073	1	0.998217	1.29015E-4	0.00159151	6.27738E-5
COMP high							
ENERGIE	549	1.18479	3	0.420762	0.0286153	0.545993	0.00462992
ENERGIE	1579	0.968199	1	0.653991	0.0036251	0.34152	8.64039E-4
ENERGIE	1041	0.521661	1	0.900451	0.0103188	0.0884779	7.52464E-4
QRS average							
COMP low							
RR0 low	14594	0.157786	1	0.981666	0.00832803	0.00781841	0.00218764
RR0 avera	20487	0.0832331	1	0.990792	0.00109608	0.00756492	5.46869E-4
RR0 high	13362	0.0242011	1	0.997867	6.033E-4	6.00582E-5	0.00146995
COMP averag							
RR0 low	16843	0.194793	1	0.976147	0.00872934	0.0126856	0.00243763
RR0 avera	20649	0.114127	1	0.986242	0.00132674	0.0119269	5.04532E-4
RR0 high	12915	0.0420815	1	0.996015	0.0017997	1.07422E-4	0.00207816
COMP high							
RR0 low	2314	1.23693	1	0.667026	0.215042	0.117154	7.77892E-4
RR0 avera	855	0.760179	1	0.835275	0.0258197	0.137476	0.0014295
RR0 high	238	0.593244	1	0.887434	0.094526	0.00422219	0.0138182
QRS high							
COMP low							
RR0 low	6001	1.0001	1	0.723926	0.246229	0.0269763	0.00286902
RR0 avera	7317	0.326435	1	0.953076	0.0210185	0.0247771	0.00112857
RR0 high	3628	0.299496	1	0.955075	0.0372274	0.00145168	0.00624594
COMP averag							
ENERGIE	5759	1.19867	1	0.658609	0.268048	0.0707318	0.00261157
ENERGIE	7203	0.469104	1	0.925246	0.025297	0.0463707	0.00308668
ENERGIE	6295	0.895093	1	0.744479	0.244748	0.00521083	0.00556152
COMP high							
RR0 low	6427	0.667349	2	0.114035	0.862927	0.0228911	1.46746E-4
RR0 avera	680	1.51344	2	0.323387	0.465922	0.210691	0.0
RR0 high	221	0.900908	2	0.26287	0.728169	0.00896117	0.0

Figure 7.30 decision tree of FDT6 (taken from gauge software)

Rule	Type	If RR0	AND RRs	AND COMP	AND QRS	AND PP	AND ENERGI	THEN Class
1	I	low	low	low	low	low	high	1.0
2	I	low	low	average	average	high	low	1.0
3	I	low	low	low	average	average	low	1.0
4	I	low	low	average	average	average	average	1.0
5	I	low	low	average	high	high	average	1.0
6	I	low	low	low	average	average	average	1.0
7	I	low	low	low	high	high	average	1.0
8	I	low	low	average	average	average	low	1.0
9	I	low	low	average	average	average	high	1.0
10	I	low	low	high	low	low	average	1.0
11	I	low	average	low	average	low	high	1.0
12	I	low	low	average	low	low	low	1.0
13	I	low	low	average	low	low	average	1.0
14	I	low	low	average	average	low	average	2.0
15	I	low	low	low	low	high	average	1.0
16	I	low	average	average	low	high	average	1.0
17	I	average	low	low	low	average	low	1.0
18	I	low	high	high	low	average	average	1.0
19	I	high	low	low	low	low	average	1.0
20	I	low	low	low	low	average	high	1.0
21	I	low	low	low	low	average	low	1.0
22	I	low	low	average	low	high	average	2.0
23	I	low	low	low	average	low	low	1.0
24	I	low	low	low	low	low	low	1.0
25	I	average	average	low	low	low	average	1.0
26	I	low	average	low	low	low	average	2.0
27	I	low	low	average	low	average	average	1.0
28	I	average	low	low	average	low	average	1.0
29	I	low	average	high	low	average	average	1.0
30	I	average	low	low	average	high	low	1.0
31	I	low	low	low	high	average	average	2.0
32	I	low	low	low	average	high	low	2.0
33	I	low	low	average	low	average	low	2.0
34	I	low	low	low	average	high	average	2.0
35	I	low	low	low	high	average	low	2.0
36	I	low	low	low	high	low	average	2.0
37	I	low	low	average	low	average	high	1.0
38	I	low	average	average	low	low	average	1.0
39	I	average	low	low	low	low	average	1.0
40	I	low	average	average	low	low	high	2.0
41	I	low	low	low	low	average	average	1.0
42	I	low	low	low	average	average	high	2.0
43	I	low	low	low	average	high	high	2.0
44	I	low	low	low	low	high	high	2.0
45	I	low	low	average	low	high	low	2.0
46	I	low	low	low	high	high	low	2.0
47	I	low	low	low	low	high	low	2.0
48	I	low	average	average	low	average	average	1.0
49	I	average	low	low	low	average	average	1.0
50	I	low	low	low	low	low	average	1.0
51	I	low	low	low	average	low	average	1.0
52	I	low	high	average	low	low	average	3.0
53	I	low	average	low	low	average	average	3.0

Figure 7.31 Rules of FDT7 (taken from gauge software)

25	I	low	low	average	average	low	average	2.0
26	I	low	low	low	low	low	low	1.0
27	I	average	low	low	low	average	low	1.0
28	I	low	low	average	low	average	low	2.0
29	I	average	average	low	low	average	low	1.0
30	I	average	average	low	average	low	average	1.0
31	I	low	average	average	average	low	average	3.0
32	I	low	average	average	low	average	low	2.0
33	I	low	low	average	average	average	low	2.0
34	I	low	average	low	average	average	low	2.0
35	I	average	low	low	average	average	low	2.0
36	I	average	average	low	average	average	low	2.0
37	I	low	average	average	average	average	low	2.0
38	I	low	average	low	low	low	low	2.0
39	I	low	low	low	average	low	low	2.0
40	I	average	low	low	low	low	low	1.0
41	I	average	average	low	low	low	low	1.0
42	I	low	low	average	low	low	low	1.0
43	I	low	average	average	low	low	low	2.0
44	I	low	average	low	average	low	low	2.0
45	I	low	low	average	average	low	low	2.0
46	I	average	low	low	average	low	low	1.0
47	I	average	average	low	average	low	low	2.0
48	I	low	average	average	average	low	low	2.0
49	I	low	low	low	low	high	average	1.0
50	I	low	low	low	average	high	average	2.0
51	I	low	low	average	average	high	average	2.0
52	I	low	low	average	low	high	average	2.0
53	I	low	average	low	low	high	average	2.0
54	I	low	average	low	average	high	average	2.0
55	I	low	average	average	average	high	average	2.0
56	I	low	average	average	low	high	average	2.0
57	I	low	low	low	low	high	low	2.0
58	I	low	low	low	average	high	low	2.0
59	I	low	low	average	average	high	low	2.0
60	I	low	average	low	low	high	low	2.0
61	I	low	low	average	low	high	low	2.0
62	I	low	average	low	average	high	low	2.0
63	I	low	average	average	average	high	low	2.0
64	I	low	average	average	low	high	low	2.0
65	I	low	low	low	low	average	high	2.0
66	I	low	low	low	average	average	high	2.0
67	I	low	low	average	low	average	high	2.0
68	I	low	low	average	average	average	high	2.0
69	I	low	average	low	low	average	high	2.0
70	I	low	average	low	average	average	high	2.0
71	I	low	average	average	average	average	high	2.0
72	I	low	average	average	low	average	high	2.0
73	I	low	low	low	low	low	high	1.0
74	I	low	low	low	average	low	high	1.0
75	I	low	average	low	low	low	high	1.0
76	I	low	low	low	average	high	high	2.0
77	I	low	low	low	low	high	high	2.0
78	I	low	low	average	average	high	high	2.0
79	I	low	low	average	low	high	high	2.0
80	I	low	average	low	average	high	high	2.0
81	I	low	average	average	average	high	high	2.0
82	I	low	average	low	low	high	high	2.0
83	I	low	average	average	low	high	high	2.0
84	I	low	low	low	high	average	average	2.0

Figure 7.32 Rules of FDT8 (taken from gauge software)

B Platform GAUJE

GUAJE ENVIRONMENT

The main novelty of the GUAJE approach is that, it is the first one combining several software tools (not only libraries) with the aim of building interpretable fuzzy models. Notice that, interpretability is the main requirement and it is taken into consideration along the whole modeling process. Of course, accuracy is not forgotten because all kind of system must achieve at least a minimum accuracy, being completely useless otherwise. What we want to highlight is the fact that our proposal is especially designed for humanistic systems (defined by Zadeh as those systems whose behavior is strongly influenced by human judgment, perception or emotions [69]) used in real-world applications (for instance decision-support systems in fields like education, robotics, medicine, etc.) where there is a huge human-system interaction and thus, the interpretability of the model is strongly appreciated. Moreover, some loss of accuracy may be tolerated in exchange for a more interpretable model.

The core of GUAJE is the last downloadable version of KBCT (version 3.0) which has been up-graded with new functionalities and implementing the HILK (Highly Interpretable Linguistic Knowledge) fuzzy modeling methodology [70]. GUAJE combines the following six preexisting tools (the first five ones are freely available as open source) making use of the tools and methods that they provide:

• KBCT. Open source software for knowledge extraction and representation which combines expert knowledge and induced knowledge (knowledge automatically extracted from data) [69]. The combination of both kind of knowledge is made carefully and it includes consistency analysis, simplification, and optimization tasks.

• FisPro. An open source tool for creating fuzzy inference systems (FIS) to be used for reasoning purposes, especially for simulating a physical or biological system. It includes many algorithms (most of them implemented as C programs) for generating fuzzy partitions and rules directly from experimental data.

• Xfuzzy. A free software development environment for generating FIS. It integrates a set of tools that ease the user to cover the several stages involved in the whole designing process, from their initial description to their final implementation, including simulation, edition and program synthesis. It is written in Java and all its tools are based on a common specification language named XFL3.

• ORE11 (Ontology Rule Editor) [71]. A java-based open source platform-independent application for defining, managing and testing inference rules on a model represented by a specific ontology.

• Weka. An open source tool providing lots of algorithms for data mining. It includes the implementation of many classical algorithms like for example J48 which corresponds to the well-known C4.5 algorithms.

• Matlab Fuzzy Toolbox. It is the most widely used commercial tool for fuzzy systems. Its main advantage is that it takes profit from the fact that it is fully integrated with all functionalities provided by Matlab environment which is commonly used in engineering for both educational and business applications.
The Figure 7.33 illustrates how all these tools cooperate in the GUAJE environment. It shows the tasks made by each tool.

We consider two main sources of knowledge, *experimental data* and *expert knowledge*. An expert is a person who perfectly knows the problem under analysis. The expert can describe the system behaviour and his/her background (knowledge, experience, preferences, etc.) is essential to get a good model. For that reason, as it can be seen in the diagram expert, knowledge plays a key role being present all along the modelling process. Although GUAJE can work in a fully automatic way it also lets expert supervision and interaction at each step. Such integration only is possible if both expert and induced knowledge are formalized using the same language. In this case, we use FL along with a set of constraints to guarantee the interpretability of the generated model.

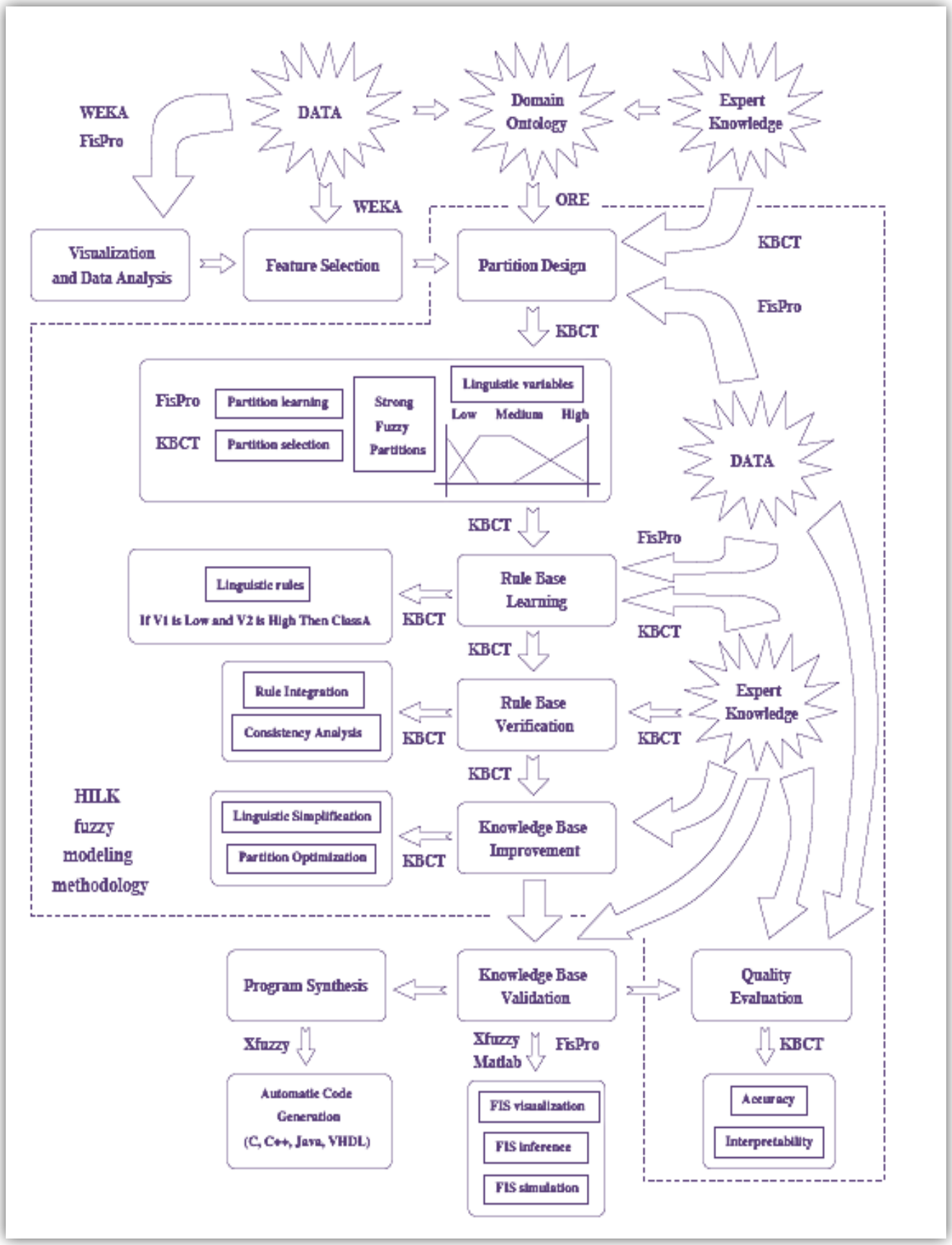


Figure 7.33 Scheme of the proposed GUAJE environment

Elicitation of expert knowledge is a complex task and it usually becomes a bottleneck in the whole modelling process [72, 73]. In order to make easier expert knowledge extraction and representation we can use an intermediate level, the *domain ontology*. There are so many web ontologies that it is easy to find one related to the problem under consideration. The hierarchy of concepts included in the selected ontology can be shown graphically to the expert who should identify the main influential input-output variables as well as a small set of basic expert rules [74]. All tasks to be done in relation with ontologies such as web searching, downloading, handling, representation, and so on are made calling to ORE functions directly from the KBCT graphical interface. As result of this preliminary stage, regardless of whether we use ontologies or not, we obtain a first simple expert knowledge base (KB) that must be complemented and upgraded later with knowledge induced from experimental data. The handling of data includes both visualization and analysis tasks that are carried out by KBCT making calls to Weka and FisPro algorithms. The first step, consists in feature selection which is crucial to keep interpretability when dealing with large data files made up of dozens of input variables. Weka provides many algorithms for feature selection. Then, it is possible to use algorithms provided by FisPro for partition and automatic rule generation from data. Such algorithms are especially designed for generating interpretable partitions and rules. As a result, we can build a whole KB directly from experimental data. It is important to remark that expert and induced KBs can be built in parallel or sequential steps. However, in order to get a unique KB the integration must be made carefully at both partition and rule levels. Therefore, for the sake of interpretability we recommend first closing the partition design stage (including both expert and data), i.e., defining fuzzy partitions with a global semantics before starting the rule base definition. Although it is not mandatory, we recommend the use of strong fuzzy partitions (SFP) [74] which satisfy most demanded semantic constraints (distinguishabilityy, coverage, normality, convexity, etc.) to design interpretable partitions. Notice that, linguistic comparison for consistency analysis is only feasible when all rules (expert and induced ones) are defined using the same linguistic terms defined by the same fuzzy sets. Once we have achieved a unique and consistent KB, it is time to think about the interpretability-accuracy trade-off. KBCT offers powerful algorithms for linguistic simplification with the aim of increasing even more the KB interpretability while preserving the accuracy. It starts looking for redundant elements (labels, inputs, rules, etc.) that can be removed without altering the system accuracy. Then, it tries to merge elements always used together. Lastly, it forces removing elements apparently needed but not contributing too much to the final accuracy. After getting a compact KB, accuracy can also be increased by applying optimization techniques for tuning the fuzzy partitions. Notice

that, KBCT provides some optimization algorithms which are strongly constrained with the aim of not penalizing too much interpretability. The partition tuning process must keep the matching in between fuzzy sets and linguistic terms which should be fully meaningful according to the problem context and the expert background. To sum up, the five main stages of the HILK methodology (partition design, rule base learning, rule base verification, knowledge base improvement, and quality evaluation) constitute the core of GUAJE environment. They are surrounded by a dash line in Figure7.33 and they are directly implemented in the new enhanced version of KBCT. Although, for the sake of clarity, the five steps are represented sequentially, in practice the tool is absolutely flexible and the whole process is iterative. Everything can be supervised by an expert who can decide going back to the previous steps at whatever moment. Finally, at the last stage, KBs built with KBCT can be exported to the format recognized by FisPro, Matlab, and Xfuzzy. In consequence, designed KBs can be used with the inference engines provided by such tools. In addition, we can take profit of the Matlab Simulink environment in order to include the modelled system as part of a more complex simulated system. Furthermore, the program synthesis made by Xfuzzy is useful for generating a final stand-alone module to be embedded in a real application.

SUMMARY

In this thesis, we introduce a novel method to define semi-automatically fuzzy partition rules to provide a powerful and accurate insight into cardiac arrhythmia. In particular, we define a text mining approach applied to a large dataset consisting of the freely available scientific papers provided by PubMed. The information extracted is then integrated with expert knowledge, as well as experimental data, to provide a robust, scalable and accurate system, which can successfully address the challenges posed by the management and assessment of big data in the medical sector. The evaluation we carried out shows an accuracy rate of 93% and interpretability of 0.646, which clearly shows that our method provides an excellent balance between accuracy and system transparency. Furthermore, this contributes substantially to the knowledge discovery and offers a powerful tool to facilitate the decision-making process.

Résume

Dans cette thèse, nous introduisons une nouvelle méthode pour définir des règles de partition semi-automatique flou pour fournir un aperçu puissant et précis de l'arythmie cardiaque. En particulier, nous définissons une approche d'extraction de texte appliquée à un ensemble de données important composé des documents scientifiques disponibles gratuitement par PubMed. L'information extraite est ensuite intégrée à des connaissances spécialisées, ainsi qu'à des données expérimentales, pour fournir un système robuste, évolutif et précis, qui peut répondre avec succès aux défis posés par la gestion et l'évaluation des grandes données dans le secteur médical. L'évaluation que nous avons effectuée montre un taux d'exactitude de 93% et une interprétation de 0.646, ce qui montre clairement que notre méthode offre un excellent équilibre entre précision et transparence du système. En outre, cela contribue de manière substantielle à la découverte de connaissances et offre un outil puissant pour faciliter le processus décisionnel.

خلاصة

في هذه الأطروحة، نقدم طريقة جديدة لتحديد قواعد التقسيم شبه غامض تلقائيا لتوفير نظرة قوية ودقيقة في عدم انتظام ضربات القلب. وعلى وجه الخصوص، نحدد نهجا للتعدين النص يطبق على مجموعة بيانات كبيرة تتألف من الورقات العلمية المتاحة مجانا التي تقدمهاPubMed . ثم يتم دمج المعلومات المستخرجة مع المعرفة الخبراء، فضلا عن البيانات التجريبية، لتوفير نظام قوي وقابل للتطوير ودقيقة، والتي يمكن أن تعالج بنجاح التحديات التي تفرضها إدارة وتقييم البيانات الكبيرة في القطاع الطبي. يظهر التقييم الذي قمنا به معدل دقة 93٪ وتفسير 0.646، مما يدل بوضوح على أن أسلوبنا يوفر توازن ممتاز بين الدقة وشفافية النظام. وعالوة على ذلك، يساهم ذلك بشكل كبير في اكتشاف المعرفة ويوفر أداة قوية لتسهيل عملية صنع القرار.