Design and Implementation of Advanced Bayesian Networks with Comparative Probability

Ali Hilal Ali

M.Sc. (University of Technology, 2004)

School of Computing and Communication Systems,

Lancaster University,

England

Submitted for the degree of Doctor of Philosophy

February 2012.

Abstract

Design and Implementation of Advanced Bayesian Networks with Comparative Probability

Ali Hilal Ali

Submitted for the degree of Doctor of Philosophy

February 2012

The main purpose of this research is to enhance the current procedures of designing decision support systems (DSSs) used by decision-makers to comprehend the current situation better in cases where the available amount of information required to make an informed decision is limited. It has been suggested that the highest level of situation awareness can be achieved by a thorough grasp of particular key elements that, if put together, will synthesize the current status of an environment. However, there are many cases where a decision-maker needs to make a decision when no information is available, the source of information is questionable, or the information has yet to arrive. On the other hand, in timely critical decision-making, the availability of information might become a curse rather than a blessing, as the more information is available the more time is required to process it. In time critical situations, time is an expensive commodity not always affordable. For instance, consider a surgeon performing cardiac surgery. With all the new advances in monitoring equipment and medical laboratory tests, there would be too much information to account for before the surgeon could decide on his

next “cut”. A DSS could help reduce the amount of information by converting it into the bigger picture through summarizing.

The research resulted in a new innovated theory that combines the philosophical comparative approach to probability, the frequency interpretation of probability, dynamic Bayesian networks and the expected utility theory. It enables engineers to write self-learning algorithms that use example of behaviours to model situations, evaluate and make decisions, diagnose problems, and/or find the most probable consequences in real-time. The new theory was particularly applied to the problems of validating equipment readings in an aircraft, flight data analysis, prediction of passengers behaviours, and real-time monitoring and prediction of patients’ states in intensive care units (ICU). The algorithm was able to pinpoint the faulty equipment from between a group of equipment giving false fault indications, an important improvement over the current fault detection procedures. In addition, the network was able to give to the aircraft pilot recommendations about the optimal speed and altitude that will result in reducing fuel consumptions and thereby saving costs and extending equipment lives. On the ICU application side, the algorithm was able to predict those patients with high mortality risk about 24 hours before they actually deceased. In addition, the network can guide nurses to best practices, and to summarize patients’ current state in terms of an overall index. Furthermore, it can use data collected by hospitals to improve its accuracy and to diagnose patients in real-time and predict their state well-ahead to the future.

Declaration

I hereby declare that the research contained in this thesis is my original ideas and work. However, the application of the comparative probability and dynamic Bayesian networks to ICUs was the result of thorough discussions between myself, my supervisor (professor Garik Markarian) and Stuart Grant of Manchester University. Nonetheless, I have taken extra care to properly reference any work conducted by others in order to distinguish my accomplishments from theirs.

The materials and work described in this thesis have not been previously submitted for the same degree in the current or any other form.

Acknowledgements

As my journey towards my PhD degree is concluding, I look back at my first day at Lancaster University just to realize how far I have come. The past four years were full of rich experiences, of achievements and disappointments, and of naïve thoughts and skilful approaches. I have learnt a lot so I would like to take this opportunity to thank everyone who lent me a hand during this amazing time as a PhD student at Lancaster University.

I would like to express my great gratitude and thanks to my supervisor professor Garik Markarian for believing in me since the start of this journey and for all his help, his patient throughout my falls, and for opening my eyes to the meaning of academic research. Secondly, I would like to thank Dr Plamen Angelov for supervising me during my work on the Svetlana project. It was his high standards of analysing and reporting results that put my feet on the right ground. I would also like to thank my friends and family for believing in me, for all the encouragement and for their dedication. For all of you who made the person I am thank you!

Finally, some of the research in this thesis has received funding from the European Union Seventh Framework Program (FP7/2007-2013) under grant agreement n° ACPO-GA-2010-265940 SVETLANA.

Related Publications

1- **Co-Author: (Patent)** A. H. Ali and others, "Monitoring System," United Kingdom Patent, Application reference number 1109215.2, July 2011.

2- **Author: (Journal)** A. H. Ali, "Utilizing BADA (base of aircraft data) as an on-board navigation decision support system in commercial aircrafts," Intelligent Transportation Systems Magazine, IEEE, vol. 3, pp. 20-25, 2011.

3- **Co-Author (Journal)** A. H. Ali, et al., "Feasibility demonstration of diagnostic decision tree for validating aircraft navigation system accuracy," Journal of Aircraft, vol. 47, 2010.

4- **Co-Author (Journal)** A. H. Ali, et al., “A survey of Mathematical Tools for Anomaly Detection and Isolation in Commercial Aircrafts” Submitted to the IEEE Transactions on Intelligent Transportation Systems.

5- **Co-Author (Journal)** A. H. Ali, et al., “Predicting the IRIS Score of ICU Patients using Comparative Probability Based DBN”, submitted to the IEEE Transaction on Computational Biology and Bioinformatics.

6- **Co-Author (Conference Paper)** H. A. Ali, et al., "Smart on-board diagnostic decision trees for quantitative aviation equipment and safety procedures validation," Proceedings of SPIE, 2010, p. 77090K.

7- **Co-Author (Conference Paper)** A. H. Ali and A. Tarter, "Developing neuro-fuzzy hybrid networks to aid predicting abnormal behaviours of

passengers and equipments inside an airplane," presented at the Proceedings of SPIE, Orlando, USA, 2009.

8- **Co-Author (Project Report)** A. H. Ali, et al., SVETLANA WP3.1 Mathematical tools identification, D3.1, V 5.0, 2011.

9- **Author (Project Report)** A. H. Ali, Final report to RNC Avionics, North West Development Agency Voucher Award, September, 2009.

Table of Contents

List of Figures

List of Tables

1. Introduction

We live in an ever-changing world where our convictions about the state of it update with time as we discover new information about our surroundings. As we acknowledge the imperfections of our knowledge repositories regarding the state of the world, we often need to make decisions despite all the missing details and the uncertainty of where our decisions might lead us to. A robot might use its sensory system, for instance, a sonar based sensor, to retrieve cues about its surroundings. Then it might use these cues to decide on which direction is best to turn to. Since the world behind the range of the robot's sensors is unknown, the robot may take a turn that leads to a dead end. Hence, the robot needs to make a decision in an environment where the only available information is that of its immediate surroundings. Even if the robot was in an exceptionally charted environment, its sensors might malfunction or degrade. In this case, the uncertainty arises not from the environment but rather from a lack of trustworthiness of the robot's sensors. In addition, the robot programming may contain bugs, the robot might trip and fall, or its battery may run out of power or be stolen. The list of events that the robot could possibly face in an environment grows infinitely as we consider more details. The problem of specifying all the exceptions a designer needs to consider is called the qualification problem [1, p. 268].

Uncertainty can arise due to external factors, such as noise. In statistics, noise refers to unexpected (or unexplained) variations in the observations of a process, as opposed to the explained variation where the mathematical model of the process can be estimated [2]. In digital communications, information may be sent as pulses with varying amplitudes that each represents a state. After random noise is added to the amplitude of the pulses throughout the transmission channel, the receiver has to estimate what state was sent given the random variations in the received signal due to the added noise [3].

In general, uncertainty might arise due to theoretical ignorance, as is the case when scientists have an incomplete understanding of phenomena; laziness because listing all the causes that orchestrate the observed behaviour of a phenomenon might be too much work; or practical ignorance when we are required to decide based on partial evidence, for instance, a physician trying to diagnose a patient without performing all the necessary laboratory tests [1, p 481].

Finally, in quantum physics, uncertainty is an objective property of reality. Certain pairs of particles' properties are constrained together in a precise inequality, such that the more that is known about the first of the pair, the less that is knowable about the second [4]. Consequently, a part of our world is always going to be fuzzier even as we gain more knowledge about the other part.

Probability theory is the main tool used to represent uncertainty arising from laziness and ignorance [1, p 482]. If we consider probability as a measure of how likely an event would be observed in an experiment repeated

a certain amount of times, then it could be used as a quantitative representation of our certainty of how likely that event might occur from among all other possible events. In this context, probability is interpreted as a degree of belief rather than a frequency of occurrence. It provides a quantifiable interface to an agent epistemological state regarding the world. For example, if 1 robot out of 100 suffered power problems then we could say that our belief that this robot would suffer power problems is 0.01.

Probability can also be used in decision-making where it is treated as the expression of an agent's judgement of how possible an event is. Probability in this context represents a decision not an estimate of errors [5]. Combined with utility, probability can be used to construct decision networks where various decision paths are plotted and assigned preferences that describe their usefulness to the decision-maker, and where the likelihood of each path is expressed in terms of probability. Thereby, the decision-maker can find the path that results in the maximum utility [6]. In addition, probability is used to model noise and random processes in digital communication systems to minimize the rate at which the receiver wrongly guesses which state the transmitter has actually sent. The likelihood of an outcome with respect to the sample mapped into a function of time represents the random nature of a process [3, p. 303].

An extensive amount of research and literature is available on the statistical modelling of noise and the probabilistic representation of uncertainty. Moreover, researchers have suggested various approaches on how to quantify uncertainty. The purpose of this thesis is to find an optimal

approach of dealing with decision-making under uncertainty when little information is available to the decision-maker at the time of making the decision. We will look into the objective of this thesis in the next section.

1.1 Motivations and aims

The main purpose of this research is to enhance the current procedures of designing decision support systems (DSSs) used by decision-makers to comprehend the current situation better in cases where the available amount of information required to make an informed decision is limited. It has been suggested that the highest level of situation awareness can be achieved by a thorough grasp of particular key elements that, if put together, will synthesize the current status of an environment [7]. However, there are many cases where a decision-maker needs to make a decision when no information is available, the source of information is questionable, or the information has yet to arrive. For example, consider a nurse in a public health centre who is responsible for admitting and assigning patients to be seen either by a doctor or a nurse. The assignment to a doctor should be based on a higher severity condition of the patient's symptoms relative to that of an assignment to a nurse. Since some patients might overstate their symptoms to be admitted to a doctor and thereby a better service, or conversely, they may understate their symptoms out of fear. Therefore, the nurse cannot be certain about the severity of those patients' illnesses.

In timely critical decision-making, the availability of information might become a curse rather than a blessing, as the more information is available the more time is required to process it. In time critical situations, time is an expensive commodity not always affordable. For instance, consider a surgeon performing cardiac surgery. With all the new advances in monitoring equipment and medical laboratory tests, there would too much information to account for before the surgeon could decide on his next “cut”. A DSS could help reduce the amount of information by converting it into the bigger picture through summarizing.

In the aviation industry, large aircraft often contain redundant measuring equipment. The accuracy of the navigation system can be verified by comparing the readings from two different equipment groups. For instance, an accurate altitude can be assumed when the altimeter reading of the pilot’s panel is identical to that on the flight officer’s panel. Otherwise, a search for a defective component is initialized, which, in turn, might involve manual procedures, such as switching to alternative air data, or observing the status of the altimeter for visual defection cues, such as a fluctuating pointer [8]. However, manual observations require the pilots to be in a high state of situational awareness where they would be able to comprehend the states of the aircraft, and in turn, make reasonable decisions. This would defeat the purpose of a DSS (or redundant measuring equipment), as they are supposed to raise pilot’s situational awareness instead of the other way around.

The work in this thesis was particularly applied to the problem of validating equipment readings in an aircraft, flight data analysis, and real-time monitoring

and prediction of patients' states in intensive care units (ICU). Each application will be discussed further in the upcoming chapters. However, the author feels it is necessary to introduce some basic notations and background topics before the main theory is introduced.

1.2 Decision Support Systems

DSSs is an umbrella term applied to any computerized system used in aiding making decisions in an organization [9, p 14]. One of the earliest definitions of DSSs comes from Keen and Morton in 1978, where [9, p. 12]:

> *Decision support systems couple the intellectual resources of individuals with the capabilities of the computer to improve the quality of decisions. It is a computer-based support system for management decision makers who deal with semi-structured problems.*

Classically, the process of designing a DSS was classified into three categories: structured, unstructured, and semi-structured [9, p. 11]. Structured DSSs are those that involve a straightforward decision-making process where standard procedures exist to make the required decision; for example, processing a new order in an online store. Unstructured DSS is where the problem of coming up with a decision is often complex, fuzzy, or has no standard solutions, for example, buying new software for processing documents in a firm. Finally, semi-structured DSSs are in-between cases,

where part of the decision-making process can be structured but others cannot. An example is selecting the best car insurance.

With respect to the application that has driven the design of a DSS, DSSs are classified into model-driven, data-driven, communication-driven, document-driven, and knowledge-driven [10]. Model-driven DSSs are those that simulate, optimize, and/or manipulate a process. They use parameters and/or rules provided by experts to aid decision-makers in the process of analyzing a situation and thereby come up with a more optimized decision/s. As the capabilities of computers dramatically grew, model-based DSSs grew in complexity and started to provide wider ranges of options, optimisability, and decision routes. Conversely, data-driven DSSs are designed to support better access and manipulation of a company's internal (or even external) data. They could be as elementary as a web-based query tool or as complex as real-time access and analysis of a huge data warehouse. Communication-driven DSSs use state-of-the-art communication technologies as a media to facilitate better collaboration and communicational-based decisions. Some of the commonly used communication technologies are video conferences, internet newsletters, and computer based bulletin boards. Document-based DSSs emphasize the accessibility and/or manipulation of documents from normally huge databases. As the World Wide Web grew in size and more documents became available, document-based DSSs became the main platform for usage in document searching and retrieval. Knowledge-based DSSs (Kb-DSSs) have the capability of recommending an action to a decision-maker rather than a passive analysis and/or accessibility, as with

previous types of DSSs. They usually use expert knowledge or artificial intelligence optimized to solve problems within a specific domain. One example is computer-based medical diagnosis tools. The overall aim of this thesis falls within the domain of Kb-DSSs.

To approximate the ways experts make a decision, several frameworks have been suggested to model human information processing. Simon's three-phase paradigm of intelligence [11] is one of the earliest. Simon's model is a

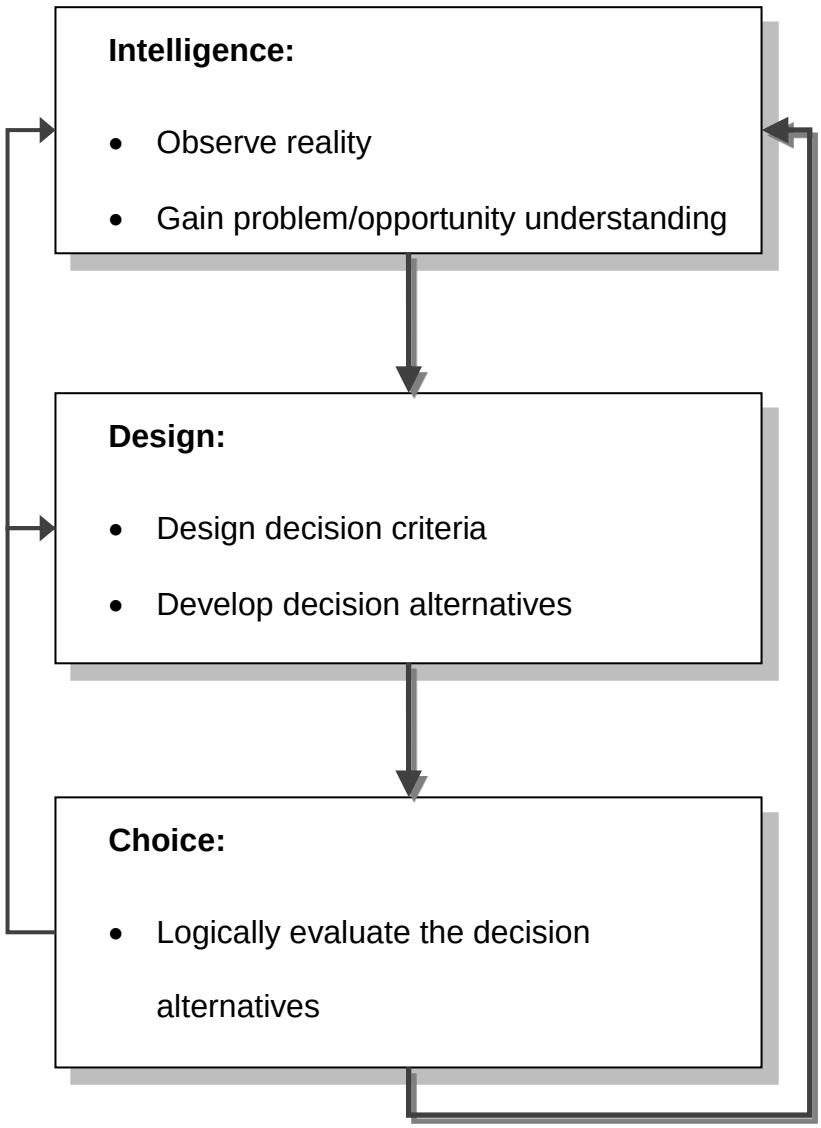


FIGURE 1. The three-phase paradigm of intelligence

conceptual model that, software-wise, can easily be implemented [12]. The model consists of three phases: intelligence, design, and choice (Figure 1) [11]. The first phase is a reconnaissance phase, where a decision-maker starts by collecting various cues from a situation, and then collects

information, detects opportunity, and comprehends the main drives behind them. The second phase is where the intelligence collected previously is used to model the problem/opportunity. The decision-maker would develop relationships between events motives and/or drives behind the situation at hand and in turn set up criteria that links his systematic model to expected results and their desired utilities and possible alternatives to an action. Finally, the decision-maker would apply his model along with the collected intelligence to produce an action or a list of actions summarizing the next course of action/s. An extra step would be a reflection phase, where the decision-makers evaluate the effectiveness of their model and come up with suggestions for the next cycle of decision-making, where they develop confidence and expertise in the process of decision-making and start the actual implementation plan [11]. In the next section, we briefly present the process of making choices under uncertainty, which characterizes the second step of the Simon's three phases of intelligence.

1.3 Choice under Uncertainty

When a decision-maker decides on which type of computer to buy for an office, the output of his choice is always certain and determined in the sense that if computer type A is bought, then computer type A is what the decision-maker will get. This is because the choice of the decision-maker mainly influences the outcome of the decision. However, there are many cases where unforeseen events that the decision-maker cannot be sure of influence the outcome of a decision. For example, imagine that a gambler would gain £100

if the outcome of a dice roll is 6 and £75 if the outcome is 5 or 4, but would lose £100 if the outcome is 3, 2 or 1. The gambler cannot be certain of the output of the dice roll because many factors affect, and thereby determine which face of the dice is going to face up, and these factors are out of his hands. In such a situation, the gambler needs to make his bet while remaining uncertain of the output of his dice roll. It is evident to assume that the gambler would have different preferences to each possible outcome of the dice roll. For instance, he would not want to roll 3, 2, or 1 since he would lose £100 but would prefer to roll 4, 5, or 6. As mentioned in the Introduction, combining preferences (or utility) with probability is the basis of our modern understanding of decision theory.

The earliest recorded attempt to combine probability with preferential value to make a choice was that of Blaise Pascal in the seventieth century in his famous Pascal wager [13]. Pascal argued that the expected value of making a choice giving *n* possible choices with values $\{v_1, v_2,\ldots,v_n\}$ and probabilities $\{p_1,p_2,\ldots p_n\}$ is given by:

$$EV = \sum_{1}^{n} p_n v_n \tag{1}$$

In 1728, Nicholas Bernoulli challenged this notation that a decision-maker needs only to consider expected value in what is now known as the *St. Petersburg paradox* [14]*:*

> *Suppose someone offers to toss a fair coin repeatedly until it comes up heads, and to pay you $1 if this happens on the first toss, $2 if it takes two tosses to land a head, $4 if it takes three tosses, $8 if it takes four tosses, etc. What is the largest sure gain you would be willing to forgo in order to undertake a single play of this game?*

Since the probability of getting heads on the first toss is ½, the probability of getting heads on the second toss is ¼, and the probability of getting heads on the *n*th toss is $1/2^n$, the expected value can be estimated using Equation 1 as:

$$EV = \frac{1}{2} \times \$1 + \frac{1}{4} \times \$2 + \cdots + \frac{1}{2^n}\ \$2^{n-1} = \frac{1}{2} + \frac{1}{2} + \cdots + \frac{1}{2} = \infty \quad (2)$$

The results of Equation 2 suggest that a gambler should accept the bet no matter what entry price is set for that game as the expected payoff is always higher, in fact, it is infinite. However, it is obvious that only few, if any, rational decision-makers would consider paying any amount of money to enter such a game. Gabriel Cramer and Daniel Bernoulli proposed the solution to this paradox by noting that a gain of $2 is not necessarily twice as useful as a gain of $1 [14]. They introduced the notion of expected utility function $U(.)$ and used it to access a gambling situation rather than the expected value. In this context, the utility of a choice becomes the multiplication of its odds by its utility. The utility of a choice considers many factors other than the financial outcome of it. For example, the amount of wealth and resources that the decision-maker currently possesses and is willing to risk, the concept of the

diminishing marginal utility of money, i.e., U($2n) < 2U($n), and whether there a casino willing to offer such a gamble exists. With the expected utility principle in mind, we can rewrite Equation 1 as:

$$EU = \sum_{1}^{n} p_n U(n) \tag{3}$$

where $U(n)$ is the utility of choice n. Assuming that the current wealth of the gambler is *W*, the sure gain ζ of the gamble of the previous example is [14]:

$$EU(W + \zeta) = \frac{1}{2} \times \mathrm{U}(\mathrm{W} + \$1) + \frac{1}{4} \times \mathrm{U}(\mathrm{W} + \$2) + \cdots + \frac{1}{2^{\mathrm{n}}} \times \mathrm{U}(\mathrm{W} + \$2^{\mathrm{n}-1}) \tag{4}$$

For example, if we assume a natural logarithmic utility function and that the gambler's wealth is about $1,000, then the sure gain will only be about $5.94. Despite the fact that the utility function has solved one of the classical paradoxes in decision theory, it does not tell us much about how to model preferences of a decision-maker. In economics, the utility function of consumers is modelled under the assumption that their preferences are consequentialist, that is, that consumers are indifferent to two compound gambles if they can be reduced to the same simple gamble; and continuity, that is, the utility of gamble A is higher than gamble B even when the probability of a new gamble C is added to gamble A [15]. However, research into the expected utility modelling is much more involved than the scope of

this thesis and is sometimes controversial [16]. In addition, the expected utility principle would only work if the probability distribution of choices is known. This is also one of the main criticisms of Bayesian probability [17].

1.4 Overview of thesis structure

This thesis is organized into six chapters. It started with a brief introduction to the aims, motivations of the thesis and DSS outlined in chapter 1. Chapter 2 is an introduction to the theory of probability which overviews the combinatorial calculus, probability theory and its results, Bayesian networks and decision-making within the framework of Bayesian Networks.

Chapter 3 details the analysis of various interpretations of probability. It sets the objectives for the wining interpretation and finally presents the proposed approach to comparative probability which will be used in the following chapters.

Chapter 4 is the first application of the developed algorithms. It starts with brief introduction to aviation safety. It gives two applications of the proposed algorithms to aviation safety. Chapter 5 is the second application of comparative probability. The application will be to ICU patients. Once more, we will show two applications of comparative probability to monitoring and analyzing patients states in ICU.

Finally, chapter 6 concludes the thesis with reminder of the objectives of the thesis and how they have been met. In addition, it outlines potential opportunities and future work which made possible following the results of the research in this thesis.

2. Bayesian Artificial Intelligence

The main objective of this thesis is to establish a framework for making decisions when little information is available to the decision-maker without resorting to the common mistake of extracting knowledge from ignorance. We have already seen in Chapter 1 that probability is the basic foundation of representing and quantifying uncertainty. In this context, we could think of probability as an intermediate domain between events and actions. In addition, the importance of probability to scientists and engineers is so obvious that it requires no further explanation or listing of examples. Finally, we saw that probability is an aspect of reality in the realm of quantum physics. However, many references, be it books, journal papers, or lecture notes, devise their own abbreviations, symbols, and nomenclature to represent various quantities and terms in probability theory. Therefore, it would only be reasonable to introduce a common notation that we will consistently refer to throughout the course of thesis. However, as probability theory is far more detailed than being summed up in one chapter of a thesis, referring to the references mentioned throughout the context of this chapter is recommended.

This chapter will walk through the basic concept of counting to advanced concepts in probability to Bayesian networks and their applications. It starts by

discussing the principle of counting and the basic notations of combinations and possible outcomes of experiments. Then it moves to probability theory from unconditional to conditional and joint probability distributions for both discrete and continuous variables. Having introduced probability theory, Bayesian network is discussed along with their importance as probabilistic graphical model of joint probability distribution and their role in decision-making. Finally, chapter two concludes by brief discussion of learning Bayesian networks structures from examples.

2.1 The principle of counting

In combinatorial analysis, counting refers to the way of finding the number of possible outcomes of an experiment or a series of experiments that somehow are related together. One formulation of the principle of counting is:

> *"Suppose that two experiments are to be performed. Then if experiment 1 can result in any one of* m *possible outcomes and if for each outcome of experiment 1 there are* n *possible outcomes of experiment 2, then together there are* m×n *possible outcomes of the two experiments [18, p. 2]"*

For instance, suppose that ice cream either comes in a cup or a cone and the available flavours are chocolate, vanilla, and strawberry. Since the shape of the ice cream can be regarded as experiment 1 with 2 possible outcomes and the flavour of it can be noted as experiment 2 with 3 possible outcomes, the overall number of outcomes of both experiment 1 and 2 is: 2×3=6. One

could express the relationship between experiment 1 and 2 in terms of a tree diagram. The tree diagram helps understand the relationship between the two experiments. See figure 2

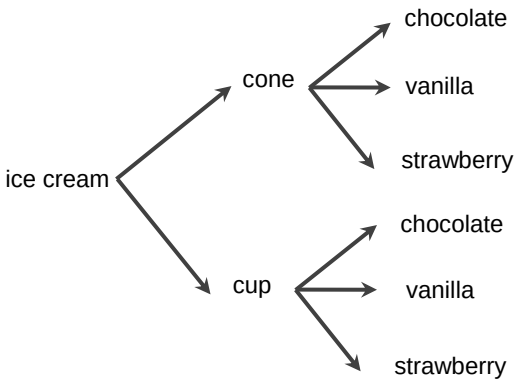


FIGURE 2. A tree diagram illustrations the principle of counting

The principle of counting can be generalized to more than two experiments. If an amount of r experiments are performed and the possible outcomes of experiment 1 were n_1, the possible outcomes of experiment 2 were n_2….and the possible outcomes of experiment r were n_r, then the overall number of possible outcomes is: $n_1 \times n_2 \times \ldots \times n_r$ [18, p. 3]. Each possible outcome in counting is referred to as a permutation. Although the principle of counting is very powerful, every so often we require a quick way of calculating the number of possible groups of *r* objects that can be arranged from a total of *n* objects. For example, a player in a word game may be interested in knowing how many permutations of 3 letters are possible out of the 10 letters he is holding. Since the first letter holder can contain any of the available 10 letters and the second letter holder can have any of the remaining 9 letters while the third one can hold any of the lasting 8 letters, it follows that the overall number

of permutation is: 10×9×8=720. However, this result assumes that the order of arrangement is relevant, that is permutations like ABC, BCA, BAC are accounted for. When the order of arrangement is irrelevant, then the overall number of permutation should be divided by the number of times the same letters are repeatedly re-arranged. In this case, it amounts to 3×2×1. In general, the number of possible combinations of r objects out of n objects where the order of permutations is not relevant can be expressed as [18, p. 6]:

$$\binom{n}{r} = \frac{n!}{(n-r)!\,r!} \tag{5}$$

Equation 5 is also referred to as the binomial coefficient because it plays an important role in binomial theorem [18,p. 15]. However, what if we are to divide the *n* objects into *r* distinct and non-overlapping groups? Since the groups are distinct and non-overlapping, and using the principle of counting, we can find [18,p. 11]:

$$\binom{n}{n_1\, n_2 \ldots n_r} = \frac{n!}{n_1!\, n_2! \ldots n_r!} \ldots (n_1 + n_2 + \cdots + n_r = n) \tag{6}$$

2.2 Basic concepts in probability

In this section, we will explore probability theory from its basic concepts to its greatest results such as the central limit theorem and the strong law of large numbers. However, probability theory is far more detailed and complex subject to be fit in a section of a thesis. Hence, most of the concepts

introduced here are as brief and abstracted as they could be. The main purpose of this section is not to introduce concepts that can be found in every first course book about probability but to establish consistent notation and reference basis upon which the main theory of this thesis can be build.

2.2.1 Events, sample space and their relationships

The word probability comes from Latin *probabilis* which means to that may be proved. It was also used in Shakespeare's Histories to mean worthy of acceptance or belief and having an appearance of truth [19]. However in modern everyday usage, it is used to refer to the degree of certainty that an event will occur [20,p. 15]. For example, the weather cast may indicate that there is a low probability the weather will be sunny during the next week in the North West of England. On the other hand, the theory of probability deals with quantifying and weighing of evidences and the likelihood of events. The probability calculus was proposed in the 17th century by Fermat and Pascal to tackle the problem of uncertainty in the outcome of gambling games [21,p. 6]. Later on, it was realized that probability calculus can also be applied to characterize ignorance. Probability became the very corner stone of science and weighing scientific observation that Bishop Butler considered it "*the very guide to life*" [21,p. 6].

If the output of an experiment cannot be deterministically estimated beforehand, then we might overcome that by deterministically estimating all the possible outcomes of the experiment. This is often referred to as the

sample space and denoted by the Greek uppercase letter Omega (Ω) whereas an outcome or subset of outcomes of the experiment is called an event and usually denoted by the Greek lowercase letter Omega (ω) [1,p. 484]. For example consider the case of dice toss. Since an ordinary dice has six faces labelled 1 to 6, the possible outcomes, or sample space, of the experiment will be:

$$\Omega = \{1,2,3,4,5,6\} \tag{7}$$

If the dice landed with side labelled 6 facing up then the event is represented as ω = 6. As previously discussed in the principle of counting section, we are sometimes interested in calculating the likelihood of an event when more than one experiment is performed. For instance, consider if we have two dices rather than one and they were tossed simultaneously. In this case the sample space of events is [18,p. 25]:

$$\Omega = \{(i,j): i,j = 1,2,3,4,5,6\} \tag{8}$$

Where *i* denotes the side label of the first dice and *j* denotes the side label of the second dice. Hence (*i,j*) denotes one event from the sample space Ω. Let the experiments of tossing two dices separately be regarded as E_1 and E_2 and event in experiment E_1 and E_2 is denoted as ω_1 and ω_2, then we define the new event $\omega_1 \cup \omega_2$ is the event that either ω_1 or ω_2 has occurred. This new event is referred to as the union of ω_1 and ω_2. Furthermore, the event that both ω_1 and ω_2 has occurred is denoted as $\omega_1 \cap \omega_2$ and referred to as the

intersection of events ω_1 and ω_2. The union and intersection of two events can be generalized to any number of events such as n to:

$$\omega_{1\cup n} = \bigcup_{i=1}^{n} \omega_i \tag{9}$$

for the union of events ω_1 to ω_n and to:

$$\omega_{1\cap n} = \bigcap_{i=1}^{n} \omega_i \tag{10}$$

for the intersection of events ω_1 to ω_n. The compliment of an event ω is defined as all the events over the sample space Ω where ω will not occur and is denoted by ω^c. If the subset of events described by ω_1 is also included in ω_2 then we say that ω_1 is contained in ω_2 which is usually denoted as $\omega_1 \subset \omega_2$ [18,p. 26]. When a subset of events such as ω_1 is contained within another ω_2 then the occurrence of ω_2 implies the occurrence of ω_1. Such consequential relationship plays an important role in reasoning and thereby in decision-making. On the other hand, if the subset of events in ω_1 is exactly that of ω_2, then the two events are equal and denoted as $\omega_1 = \omega_2$. The various relationships between events are usually expressed graphically by the so called Venn diagrams [21,p. 6]. In Venn diagram, a subset of events is represented in terms of closed shapes and the logical relationships between them are represented by symbolic intersections among these shapes. Figure 3 shows some of the previous relationships represented in Venn diagrams.

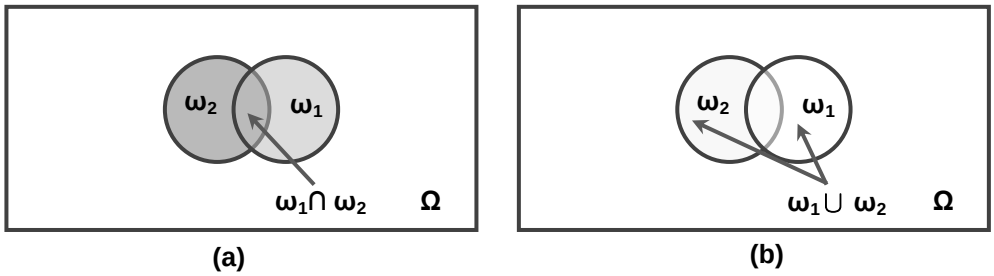


Figure 3. Venn Diagrams showing (a) intersection relationship (b) union relationship

2.2.2 Unconditional Probability

Consider an experiment in which a fair 6 faced dice is tossed. Since the dice hasn't been tampered with as to land on one of its edge, the dice should land on any one of its faces. We can express that in more abstract way by saying the outcome of a fair dice toss experiment should be any event from within the sample space defined as {1,2,3,4,5,6}. No matter how many times the same experiment is repeated, it's only intuitive that the result is always some value from within that sample space and that it is impossible to have an outcome that is 7, 9 or any other value that is not part of the sample space. Since we often express such intuitive in terms of probability, we might say that we are 100% sure that the experiment will result in any value of the sample space and 0% sure that it will result in any value outside that. If we normalize the percentage of our confidence and express the two mentioned intuitive expectations, we will get:

$$P(\Omega) = 1 \tag{11}$$

and

$$P(\Omega^c) = 0 \tag{12}$$

If the dice were biased in a way as to land with its side labelled 6 facing up, then, on average, we expect the event ω = 6 to take place more than the others. But as the dice is assumed fair, it is again intuitive to assume that each event within the sample space is as likely as the others. If we label the probability of occurrence of event ω as P(ω), then:

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) \tag{13}$$

Let us use the mathematical + sign to denote the probability of a union of two events such as ω_1 and ω_2, and using equation (11), we can write:

$$P(1) + P(2) + P(3) + P(4) + P(5) + P(6) = 1 \tag{14}$$

Since every event in (14) has the same probability, then:

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6 \tag{15}$$

Although equations (1) to (6) were derived intuitively, they are part of our modern understanding of probability which is build upon the basic three axioms of probability hence called axiomatic probability [18,p. 26]. The three axioms of probability, also known as Kolmogorov axioms state that [21,p. 6]:

Axioms of probability (16)

Axiom 1:

$P(\Omega) = 1 \ldots (16.1)$

Axiom 2:

$for\ all\ \omega \subset \Omega, P(\omega) \geq 1 \ldots (16.2)$

Axiom 3:

$for\ all\ \omega_1, \omega_2 \subset \Omega,\ \text{if}\ \omega_1 \cap \omega_2 = \emptyset, then\ P(\omega_1 \cup \omega_2) = P(\omega_1) + P(\omega_2) \ldots (16.3)$

Usually, the probability of an event is defined from a relative frequency of occurrence [18,p. 29]. In an experiment with a sample space of Ω which is repeated for *n* number of times under the same conditions, if an event like ω occurred $n(\omega)$ times during the course of performing the previously mentioned experiments, then we define $P(\omega)$ as:

$$P(\omega) = \lim_{n\to\infty} \frac{n(\omega)}{n} \quad (17)$$

Therefore, the probability of an event is the converging limit of occurrence of the event as the reptilian of the experiment approaches infinity. The assumption that the probability of an event should converge to some value can be considered as another axiom of probability or as a result of the previously mentioned axioms [18,p. 29]. Nonetheless, the axioms of probability can be used to derive other relationships such as the following:

Consequences of Kolmogorov axioms (17)

Probability of the empty set:

$P(\emptyset) = 0$

Probability of occurrence is 1 minus the probability of not occurring

$P(\omega_1) + P(\omega^c) = 1$

The addition law of probability

$P(\omega_1 \cup \omega_2) = P(\omega_1) + P(\omega_2) - P(\omega_1 \cap \omega_2)$

Probability of subset of event

$if\ \omega_1 \subset \omega_2, then\ P(\omega_1) \leq P(\omega_2)$

However, if our everyday world is deterministic, that is similar causes will result in similar effects, then shouldn't an experiment performed with the same conditions always lead to the same results? Where would the uncertainty in estimating the outcomes come from? Unquestionably, we would be uncertain about the output of an experiment if its initial condition cannot be guaranteed to be the same or if the slightest change in the initial condition will result in a butterfly chain of effects. On the contrary, this is not the assumption of the relative frequency definition of probability. One way to answer this paradox is to note that the previous definition of probability doesn't convey a proposition about reality but rather about logical possibilities. An experiment assumed to be carried out under the same condition is to assume that it favours no one

outcome over the others. Hence, a probability proposition asserts how logically possible an event would be if no other prior information is known. This type of probabilistic assertion is called unconditional, or prior, probability. The estimation of the conditional probability of an event require no more than knowledge of the sample space and no knowledge about the outer world is necessary. As soon as information about the actual world has arrived, conditional probability becomes invalid. Therefore, the likelihood of an event needs to be reassessed in light of the new information. The likelihood of an event in the presence of prior knowledge of the experiment is called conditional probability.

2.2.3 Conditional Probability

In philosophy, Kant distinguished between two types of judgements: analytical and synthetic judgement [22]. Analytical judgement deals with the way concepts and ideas are connected but it tells us nothing about the state of affairs in the actual world. Its truth requires nothing more than knowing the actual meaning of a concept or an idea, whereas synthetic judgements are those that their truths cannot be inferred without information about the actual world [23]. Hence, unconditional probability doesn't tell us anything about the actual world for it requires no knowledge about it other than the breadth of the sample space. If the unconditional probability of having a head in a coin flip is 0.5 then that shouldn't be considered what will happen in real coin flip experiment. Unconditional probability is an analytical judgement about possibilities not actualities.

Therefore, if the likelihood of an event in an experiment is to be estimated, information about the state of affairs surrounding that event should be gathered. When a condition of an experiment is known, then unconditional probability becomes void and a way to incorporate the new condition into the calculation of the event probability needs to be implemented.

Suppose that two ordinarily dices are to be tossed sequentially. If we know that the output of the first toss is 6, then how we are to incorporate this information into the estimation of how likely it will be to get an outcome that both adds up to 8 when the second die is tossed. We reason as follows: since the first dice roll is known, then there are only six possible outcomes out of the second experiments (6,1), (6,2), (6,3), (6,4), (6,5), and (6,6). In addition, there is only one way of getting an outcome of 8 namely (6,2), therefore, the conditional probability of the outcome 8 giving that 6 have occurred from the first dice roll is 1/6. In general, we define the conditional probability of event ω giving that evidence (or condition) *e* has occurred as [21,p. 7]:

$$P(\omega|e) = \frac{P(\omega \cap e)}{P(e)} \tag{16}$$

where $P(\omega|e)$ is called the conditional probability of ω giving that *e* has occurred. Equation 16 is also known as the product rule of probability written usually as [1,p. 486]:

$$P(\omega \cap e) = P(\omega|e)P(e) \tag{17}$$

The union between two events can also be expressed as $P(\omega, e)$ and in general written as:

$$P(\omega, e) = P(\omega|e)P(e) \quad (18)$$

The product rule of equation 18 can be generalized to any number of events or evidences as [18,p. 71]:

$$P(\omega_1\omega_2\omega_3 \dots \omega_n) = P(\omega_1)P(\omega_2|\omega_1)P(\omega_3|\omega_1\omega_2)P(\omega_n|\omega_1 \dots \omega_{n-1}) \quad (19)$$

It is of value to note that conditional probability satisfies all the three axioms of probability given in equations 16 [18,p. 102]. Some important theorems of conditional probability are given below [21,p. 8]:

Theorems of Conditional Probability (20)

Total probability:

$$P(\omega_1) = \sum_i \left(\omega_1 \middle| \omega_2^i\right) \quad \dots \quad (20.1)$$

The chain rule

$$P(\omega_3|\omega_1) = P(\omega_3|\omega_2)P(\omega_2|\omega_1) + P(\omega_3|{\sim}\omega_2)P(\omega_2|{\sim}\omega_1) \quad \dots (20.2)$$

where

$${\sim}\,\omega = \omega^c$$

is the complement of an event

2.2.4 Independence and conditional independence

In the previous section, we saw how the introduction of new information could affect the likelihood of an even in an experiment. However, not every change in state of affairs will result in a consequential update of the probability of an outcome. For example, the likelihood of obtaining a head when a fair coin is flipped doesn't change if we knew that the previous flipped resulted in a head or tail because the output of the first experiment doesn't change the number of combinations which the second experiment can result in. When the outcome of an event like ω_1 has no affect on the estimation of the likelihood of another event like ω_2, we say that ω_1 and ω_2 are independent (also marginal independent or absolutely independent) [1,p. 494]. The independent of two variables can be expressed as:

$$P(\omega_1, \omega_2) = P(\omega_1)P(\omega_2) \quad (21)$$

and for any number of events such as ω_1 to ω_n:

$$P(\omega_1, \omega_2, \omega_3, \dots \omega_n) = P(\omega_1)P(\omega_2) \dots P(\omega_n) \quad (22)$$

On the other hand, two events may seem to be dependent on a third event but the conditional probability of them does not seem to change when the likelihood of the third event is altered. For example, the likelihood of cloudy sky will increase dramatically if the sky is raining. Similarly, the likelihood of low temperature would increase if the sky is raining as well. Both the events cloudy sky and low temperature depends on the presence of rain. If we have no information about the condition of the sky, then looking at the thermometer

will alter our belief about the possibilities of the current weather. On the other hand, if we already know that it is raining, then looking at the thermometer wouldn't make more certain about the presence of clouds. That means that the two events: cloudy sky and low temperature are independent giving the event rainy sky. If we represent the event cloudy sky as ω_{cloud}, low temperature as ω_{temp}, and rainy sky as ω_{rain}, then [1,p. 498]:

$$P(\omega_{\text{cloud}} \cap \omega_{\text{temp}} | \omega_{\text{rain}}) = P(\omega_{\text{cloud}} | \omega_{\text{rain}}) P(\omega_{\text{temp}} | \omega_{\text{rain}}) \quad (23)$$

and:

$$P(\omega_{\text{cloud}} | \omega_{\text{rain}} \cap \omega_{\text{temp}}) = P(\omega_{\text{cloud}} | \omega_{\text{rain}}) \quad (24)$$

2.2.5 Bayes Theorem

Bayes theorem is an extension of the product rule of probability giving in equation 18 [1, p. 495]. It connects together the conditional probability between two events with its inverse. Despite its intuitive and simple nature, it has massive consequences on the interpretation of probability, approach to epistemology, hypothesis testing, and inductive logic [24]. It also forms the cornerstone of modern probabilistic reasoning in artificial intelligence, it is given by [1, p. 495]:

$$P(\omega_2 | \omega_1) = \frac{P(\omega_1 | \omega_2) P(\omega_2)}{P(\omega_1)} \quad (25)$$

Bayes rule comes in handy in cases where we have information about the probability of an effect giving some cause and we would like to estimate the likelihood of the cause when the effect at presence. This is particularly useful in diagnosis-wise flow of inference where we have symptoms and the most likely causes are to be inferred. But the real value of Bayes rule is that it shows how the likelihood of an event is updated as new evidences becomes available which is useful in inferring the likelihood of a hypothesis over another. It tells us that the likelihood of hypothesis *y* giving evidence *x* is equal to its likelihood times its prior probability before evidence x became available conditioned by the likelihood of evidence *x* itself. This process is referred to as conditionaliztion [21,p. 12].

Another application of Bayes rule is the subjective process of learning. In this context, learning is viewed as the a continuous process of updating believes about the likelihood of a state of affairs as new information is acquired [24]. For example, experience can alter our certainty about the truthful of previously held proposition. Bayes rule can also help eliminate irrational favourism such as the case with the principle of the weak evidence. It states that if an evidence like *e* with probability of P(*e*) does not increase the likelihood of a hypothesis (like h) over another (like h^*) and h was more believable than h^* then any new information that serve to strengthen P(*e*) will maintain a higher likelihood of h over h^* [24].

Although Byes rule is used widely in different disciplines ranging from philosophy to statistics to artificial intelligence, the concept of probability as a subjective belief is a controversial one [21,p. 12] that gets many philosophical

and research framework going in past years and years to come. We will introduce the application of Bayes rule in artificial intelligence in the next section.

2.2.6 Random Variables

Often, a gambler is not interested in the mere outcome of the two dice roll but rather the numerical sum of the number rolled, or in case of coin flip, the number of times of obtaining a head out of three repeated experiment. In process quality control, we are more interested in quantifying the number of times the output is above or within a certain range [25,p. 115]. In all these cases, the interest is on a certain function defined over the sample space of an experiment. Such function is often referred to as a random variable or a stochastic variable [18,p. 132]. The value of a random variable can be evaluated using the combinatorial calculus discussed earlier. For instance, the probability that sum of two dice rolls will be 10 can be calculated by counting the number of combinations where the sum of the two dice numbers is 8, namely: (6, 4), (4, 6), and (5, 5). Since there are 36 possible outcomes from a two dice roll, the probability of obtaining a sum of 10 is:

$$P(sum = 10) = \frac{3}{36} = \frac{1}{12} \tag{26}$$

A variable described by function over the sample space can be classified as either discrete or continues. A discrete random variable is that in which the function that defines it results in a finite number of possibilities such as the sum of two dice rolls which can be any of the group {1,2,3….12}, or an infinite

series of separate values such as the group of integer numbers. On the other hand, a continues random variable is that where its function can assume any possible value within a certain range or multi ranges of values [26].

Random variables obey the three axioms of probability and there values should sums up to 1. Usually an uppercase letter is used to denote a random variable and lowercase to denote a generic value of a random variable such that for the random variable X which has k discrete values [20,p. 20]:

$$\sum_{i=1}^{k} P\left(X = x^{i}\right) = 1 \qquad (27)$$

where x^i is the *i*-th value of the random variable *X*. Usually the probability function of a random variable, also known as the Probability Mass Function (PMF) , is presented in terms of a two dimensional graph. The x-axis of the graph is usually used to denote the range of values of the random variable whereas the y-axis is preserved for the corresponding probability value of that variable [18,p. 138]. Figure 4 shows the probability mass function of the sum of a pair of dice [27,p. 61].

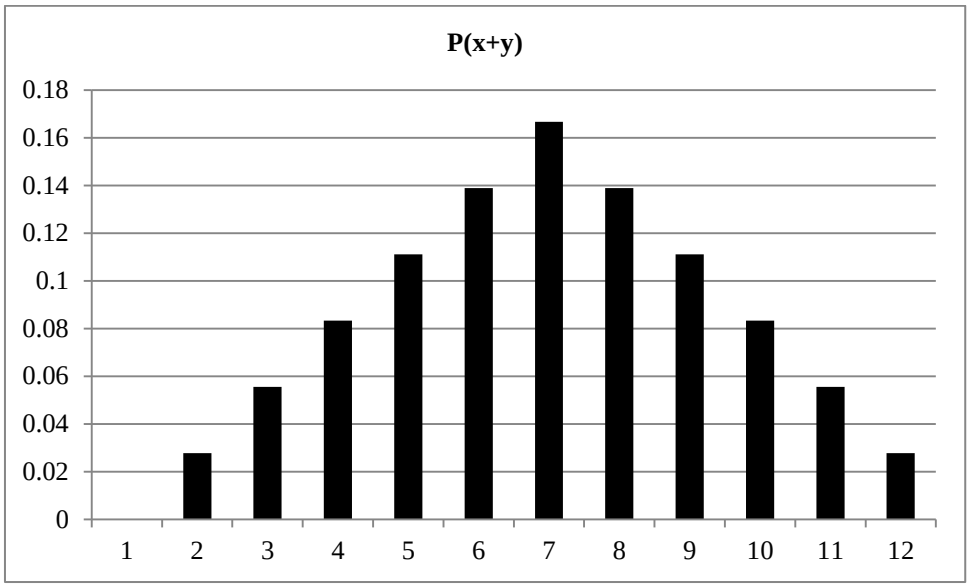


Figure 4. P(x+y) is the mass probability function of a pair of dice

Another way of representing a random variable function over a space sample is the Cumulative Distribution Function (CDF) which describes the probability that a random variable falls below a given value or simply the sum of all probabilities of the mass distribution function where it is less than or equal to some value like *x* [28]:

$$F_X(x_i) \equiv P(X \leq x_i) = \sum_{x \leq x_i} P(x) \tag{28}$$

where $F_X(x_i)$ is the cumulative distribution function of the random variable *X* when $x = x_i$. Since a CDF is essentially a sum of probabilities lying under a certain value, it is a cumulatively increasing function which starts always with a value of zero and ends with 1, and $F_X(a) > F_X(b)\ if\ a\ > b$ [29,p. 5]. Figure 5 shows the cumulative distribution function of probability mass function of a pair of dice rolls.

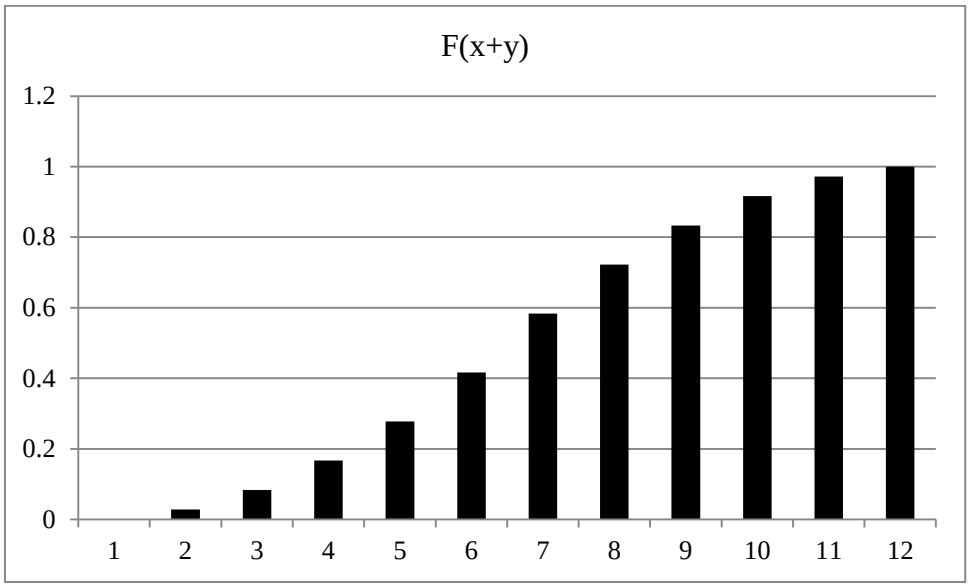


Figure 5. CDF of a pair of dice roll

One important and central concept in probability theory is the expected value of a random variable [30,p. 148]. The expected value of a random variable is defined as the weighted average of all the possible values in the sample space. Usually denoted by uppercase U, the expected value of random variable (X) is defined as [31,p. 127]:

$$E(X) \equiv \sum_{i=1}^{n} x_i P(x_i) \tag{29}$$

For example the expected value of fair dice roll is:

$$E(dice) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = \frac{7}{2} \tag{30}$$

The expected value represents another idealized concept in the frequency interpretation of probability just like the definition of probability itself [31,p. 127]. Therefore, the expected value of a random variable doesn't have to be a

directly measurable or even possible quantity that exists in the sample space. For example, the expected value of a dice rolls fond earlier as: 7/2 is impossible. From the point of view of the frequency interpretation of probability, it represents the ultimate average of the samples that the experiment should converge to when the observation is infinitely repeated. Since the ratio of observing an event such as ω from the sample space Ω would converge to P(ω) and that is true for all ω, then it follows that the average of observing ω is [18,p. 141]:

$$\sum_{i=1}^{n} \omega_i P(\omega_i) = E(\Omega) \tag{31}$$

The expected value of a function of random variable can also be calculated by noting that that function has a mass distribution function as the random variable has. If we designate that function as *g(X)* then the expected value of *g(X)* is [18,p. 145]:

$$E(g(X)) \equiv \sum_{i=1}^{n} g(x_i) P(x_i) \tag{32}$$

Although the importance of the expected value of a random variable, it does not some up all the properties of it. For example, we may be interested in knowing how wide the variable spread around the average. The spread of the probability distribution function, commonly denoted as the variance, is important in process control as it gives indication on whether the process is still under control or becoming uncontrolled [25]. The variance of a random

variable can also help measure the representation power of the average. If the variance is high then the average doesn't quite represent the data because it would imply that there are wide gaps between observed events [32]. If the variance is small then it means that the events are similar to each other. The variance of a random variable is given by [18,p. 149]:

$$\alpha(X) = E((X-\mu)^2) = \sum_{i}^{n} (x-\mu)^2\, p(x) = E(X^2) - E^2(X) \tag{33}$$

where $\alpha(X)$ is the variance of *X* and μ is the variable mean. For example the variance of a fair dice roll is:

$$\alpha(dice) = E(dice^2) - E^2(dice) = \frac{91}{6} - \frac{49}{4} = \frac{35}{12} \tag{34}$$

The square root of the variance is commonly known as the standard deviation (designated by the Greek letter σ). Not all random variables are discrete but there exists many examples that are continuous, for example, the measurement of a resistor value or the lifetime of a light bulb. Both are examples of measurements that result in uncertainty as to what the real value would be. In this case, we define the probability density function of the random continuous variable X over the sample space $x \in (-\infty, \infty)$ as f(x) and the probability that the random variable X will be within the set of real numbers B as [18,p. 205]:

$$P(x \in B) = \int_B f(x)dx \tag{35}$$

The continues probability counterpart to the discrete one should also abide the three axioms of probability. Therefore, the area under f(x) should always add up to 1, that is:

$$P(x \in (-\infty, \infty)) = \int_{-\infty}^{\infty} f(x)dx = 1 \tag{36}$$

Hence the cumulative distribution function of the continuous random variable X is:

$$F_X(x) = \int_{-\infty}^{x} f(x)dx \tag{37}$$

Using equation 29, the expected value of the continuous random variable X can written us:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx \tag{38}$$

In addition, if g(x) is a function defined over the continuous random variable whose probability distribution function is given by f(x), then the expected value of g(x) is given by:

$$E(g(x)) = \int_{-\infty}^{\infty} g(x)f(x)dx \tag{39}$$

Finally, the variance of the continuous random variable whose probability distribution function is given by f(x) is given by equation 33.

One of the most important probability distribution functions is the normal distribution function (also known as Gaussian distribution[20,p. 28]) pioneered by the French mathematician Abraham DeMoivre in 1733 to estimate the probability of binomial random variables and was later extended by Laplace and others [18,p. 218]. The normal distribution function is a one having a mean of μ and a standard deviation of σ is:

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/2\sigma^2} \quad (40)$$

The Gaussian distribution function is a bell shaped curve with a peak at μ. Figure 6 shows a typical Gaussian distribution function. The importance of the normal distribution function is that it gives a theoretical support to the practical notation of the behaviour of some continuous random variables such as the height of a person and it is considered one of the two greatest results of probability theory[1] [18,p. 218].

[1] The other is the strong law of large numbers

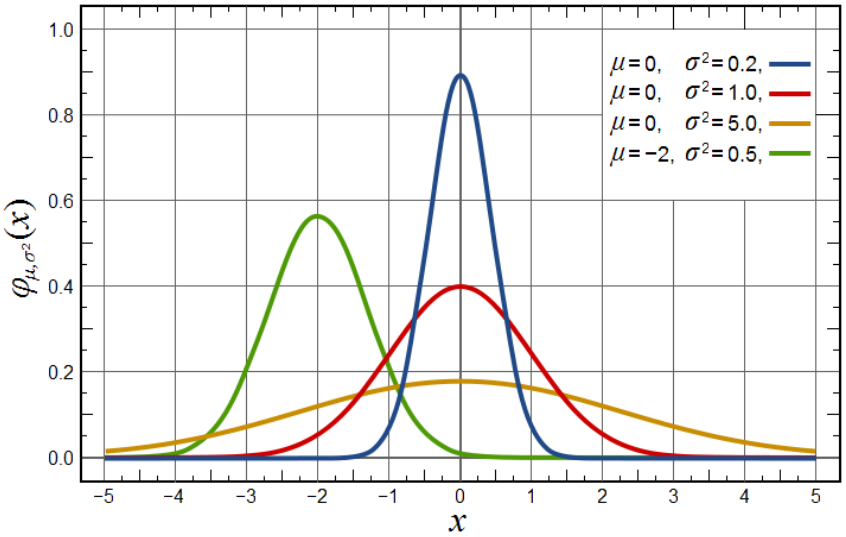


Figure 6. Normal distribution function from [33]

2.2.7 Joint probability distribution

In many scientific experiments which involves statistical measures, there are more than one random variable to be measured over the same sample space [34,p. 49] , for example, the pressure and volume of a gas, the resistance and temperature of a resistor, or the height and weight of a person. So far we have only introduces probabilistic concepts with regards to only one variable. In this section, we will briefly introduce basic concepts of *bivariate* distributions.

Let X and Y be two random variables from sample space Ω, we define the cumulate joint probability of X and Y by [18,p. 258]:

$$F(x,y) = P\{X \leq x, Y \leq y\} \quad and \quad -\infty < x,y < \infty \tag{41}$$

The distribution of both of X and Y can be derived from the joint probability of them which, in addition, could be used to answer all statistical enquiries about the joint probability of X and Y [18,p. 259]. Equation 42 gives the distribution of X and Y in terms of their joint probability [34, p. 50].

$$F_X(x) = P(X \leq x) = F(x, \infty) \quad \ldots(a)$$
$$F_Y(y) = P(Y \leq y) = F(\infty, y) \quad \ldots(b) \tag{42}$$

If X and Y are both continuous and their joint probability distribution (or density) function is f(x,y), then the joint probability of X and Y, written as P(X,Y) is [18,p. 261]:

$$P(X,Y) = \iint f(x,y)\, dx\, dy \tag{43}$$

hence the joint cumulative probability distribution function *f(x,y)* is [18,p. 262]:

$$f(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(x,y) dx\, dy \tag{44}$$

Therefore, the joint probability density function is the second derivate of equation 44 given by:

$$f(x,y) = \frac{\partial^2}{\partial x\, \partial y} F(x,y) \tag{45}$$

Previously, we have discussed the independence of two events and their corresponding conditional probability. It will be only natural to make the same inquiry regarding the independence of random variable. Let X and Y be two random variables. If x and y are any two sets of real valued numbers, then X and Y are independent when [35,p. 305]:

$$P(x,y) = P_X(x)P_Y(y) \tag{46}$$

The cumulative joint probability distribution function of X and Y follows by appealing to the three axioms of probability [18,p. 267] which will yield [35, p.307]:

$$F(x,y) = F_X(x)F_Y(y) \tag{47}$$

Using equation 46, the conditional distribution of X given Y and Y given X can be derived as [35, p.308]:

$$\begin{aligned} f(x/y) &= f_X(x) \quad \ldots(a) \\ f(y/x) &= f_Y(y) \quad \ldots(b) \end{aligned} \tag{48}$$

Equation 39 which gives the expectation of a single continuous random variable can be extended to the case of a function such g defined over the two joint random variables X and Y as [36,p. 141]:

$$E\big(g(X,Y)\big) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dx\,dy \tag{49}$$

2.2.8 Central limit theorem

The term central limit theorem traces back to a paper published by George Pólya back in the 1920s titled “Central Limit theorem in probability theory” and has been used since then [37, p. 1]. However, it was the result of the successive work of three of the most brilliant mathematicians of the eighteenth century: Abraham de Moivre, Simon Laplace, and Carl Gauss [38,p. 29]. Nowadays, the central limit theorem refers to an umbrella of statements that describe the convergence of some probability distribution functions of single or many random variables [37, p. 1]. The importance of the central limit theorem in probability theory comes from its diverse application and its ability to explain some of the widely used distributions such as the normal distribution [38,p. 29].

The first and the simplest limit theorem is the Markov inequality which tells us how likely a sample deviates from the mean. In addition, it applies to any random variable even those whose their distribution is unknown [39,p. 187]. If X is a random variable that can only take positive values then for any *x* larger than 0 [18,p. 430]:

$$P(X \geq x) \leq \frac{E(X)}{x} \tag{50}$$

However, Markov inequality is not always useful because if $x < E(X)$ then all it tells us is that $P(X \geq x)$ is less than a number larger than 1 which is an obvious statement as all probabilities are less than or equal to 1[39,p. 187]. Markov inequality is a generalization of Chebyshev’s inequality which works

for both positive and negative numbers. It is one of the most famous inequalities in probability theory and Chebyshev's best work [40,p. 75]. Equation 51 gives a mathematical formation of Chebyshev's inequality for random variable X which has a mean of μ and variance σ^2 for any value of k>1 [18,p. 431]:

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2} \tag{51}$$

Another importance of Markov's and Chebyshev's inequalities comes from the fact that when it is not always possible to know the distribution of the variable but rather its mean and variance, they can be used to set bounds on probabilities around the mean [18,p. 431].

The most important generalization drawn from Chebyshev's inequality is the weak law of large numbers [18,p. 433]:

The weak law of large numbers (52)

If X_1, X_2, … are random variables each with identical probability distribution function and a finite expectation value of μ_1, μ_2,…then for any ε>1:

$$\mathrm{P}\left(\left|\frac{\mathrm{X_1+X_2+\cdots+X_n}}{\mathrm{n}} - \mu\right| \geq \varepsilon\right) \to 0 \text{ as } \mathrm{n} \to \infty$$

The weak law of large numbers shows how the practically calculated probability through experiment is more likely to diverge from the theoretical one proposed by the frequency interpretation of probability [38,p. 19].

As previously mentioned, the most important result of probability theory is the central limit theorem [18,p. 434]. It simply tells us that averages (or sums) of n independent and identically distributed random variables each with mean of μ and variance of σ^2 tend to come close to a Gaussian distribution as n becomes boundlessly large [41,p. 47]. Hence, providing a theoretical framework to explain why many natural statistical phenomena have a bell shaped distribution. It also gives theoretical framework that deals with measurement errors by proposing that they should have normal distribution, in fact the central limit theorem was used to refer to as the law of frequency of errors in the seventieth and eightieth centuries [18,p. 442]. The central limit theorem in a very simplistic mathematical form, that is: for a single random variable only) is given by [18,p. 434]:

Central limit theorem (53)

If X_1, X_2, … are random variables each with identical probability distribution function and identical mean μ, and identical variance σ^2, then the distribution of:

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

will converge to normal distribution as $n \to \infty$ ***that is for a real number a:***

$$P\left(\frac{X_1+X_2+\cdots+X_n-n\mu}{\sigma\sqrt{n}} \leq a\right) \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} e^{-x^2/2}\,dx \quad \textbf{as} \quad a \to \infty$$

However, there are examples of superimposed independent effect that lead to non-normal processes [42,p. 28]. Although the existence of such process seem at first glance to invalidate the central limit theorem, careful

analysis of such processes shows that they posses infinite variance which places them outside the applicability of the central limit theorem [42,p. 28].

The strong law of large numbers states that, with perfect certainty, the averages of a sequence of random variables each with similar distribution will converge to the mean of the distribution [18,p. 443].

The strong law of large numbers (54)

If X_1, X_2, … are random variables each with identical probability distribution function and a finite mean of μ, then with probability =1 :

$$\frac{X_1+X_2+\cdots+X_n}{n} \rightarrow \mu \text{ as } n \rightarrow \infty$$

The strong law of large numbers shows that the averages of repeated experiments should converge to their expected value. For example, if a game of coin head or tail is repeated infinitely, then the proportion of heads or tails will be ½ with undutiful likelihood. Jacob Bernoulli was the earliest mathematician to prove the law of large numbers [43,p. 79]. Bernoulli was interested in developing mathematical tools to help make good decisions in civil, economic, and moral issues. He thought that by proving the strong law of large numbers, the relative frequency of observation can be a corner stone on which such decisions can be established [43,p. 79].

There are many other famous inequalities in the inventory of probability theory that deals with various situations or help simplify others such as the one-sided Chebyshev inequality [44,p. 70], Jensen's inequality, and Chernoff bounds. These are beyond the scope of this section which was mainly to

provide consistent mathematical background to establish the discussion of Bayesian networks in the next section and in chapter 3. Reference [44] gives quick introduction to them.

2.3 Bayesian Networks

We have seen in the previous section that all it takes to answer any statistical query about a random variable is knowing its probability distribution function or the joint probability for more than one random variable. However, it is not always possible, or practical, to obtain the full joint probability of some random variables. In addition, the size of the joint probability table in the case of discrete variables will increase dramatically as the number of random variables increases not to mention the required computation power for processing. For example, if a sample space has *n* variables each with only two possible outcomes, then the joint probability table will have 2^n entries [1,p. 493]. If a process requires 20 variables in order to be fully described and, for simplicity, each can have either of a binary state, then the joint probability table size is 2^{20}=1,048,576 entries. Processing such table size could be impractical assuming it was possible to calculate each entry in it. We have also seen that with independence and conditional independence, we can simplify some probabilistic queries by careful analysis of the relationship between the variables. It will be of great value if we could exploit this fact as to reduce the amount of calculation required to produce a joint probability distribution or to produce a more compact version of it.

Bayesian Networks are probabilistic graphical models that represent joint probabilities in way that emphasize statistical relationships between the random variables of a sample space [20,p. 51]. Bayesian Networks allow us to re-present probabilistic queries in a manner that flexibly reduce the amount of calculation and prior knowledge required to obtain the answer.

In this section, we will introduce key concepts and principles of Bayesian Networks and their role in statistical inference, making decision under uncertainty, learning and representation of knowledge, and dynamic update of belief with time.

2.3.1 Basic Bayesian Network Structure

A Bayesian network graph consists of nodes and arrows. Nodes are usually oval shapes that designate a variable whereas the arrows show the relationships between the variables. In order for the Bayesian Network representation to become superior to joint probability representation, the number of arrows has to be as minimum as possible. Otherwise, if the arrows were drawn such as to connect every variable to all the others, then the graph will become equivalent to a joint probability table. The usual way of drawing the arrows is to think of the casual relationship between the variables. In this context, an arrow is drawn from variable X to variable Y if the presences of X leads to Y. Casual relationships between the parameters are not to be strictly understood as real causal relationships but rather as implications or also known as statistical causality [45]. For example, suppose that a doctor has noticed that one of his patients is suffering symptoms of short breath (also known as Dyspnoea). The doctor knows that short breath can be the result of

lung cancer which, in turn, can be the result of pollution or smoking. She also knows that cancer can be identified with an X-ray. If the X-ray turns out positive result then Cancer is the cause of the patients symptoms but if turns out negative then it might be the result of some other causes such as bronchitis or tuberculosis [21,p. 30]. She also knows that the X-ray imagery result is not 100% trustworthy and that smoking is one contributor to lung cancer among others. Hence, she assigned prior probabilities to the trustworthiness of an X-ray machine, pollution, and smoking. In addition, she estimated the conditional probability of developing cancer giving that a patient is a smoker and having cancer when the patient is a polluted environment. With this information in mind, the Bayesian Network graph would look like the one in figure 7 [21,p. 31].

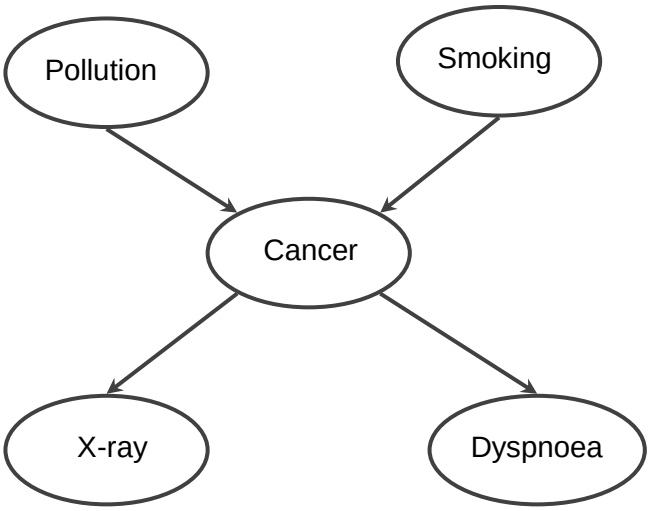


Figure 7. Bayesian network of the short breath patient

The advantage of Bayesian Network over joint probability table is clear from figure 7. The directions of the arrows tell us about the way evidences and observations flow throughout the graph and in turn how probabilities should be updated accordingly. Since there are no direct arrows between the node: Smoking and X-ray unless through the node: Cancer, then Smoking and X-ray are conditionally independent giving Cancer. This result makes sense when we recall the definition of conditional independence discussed in the last section. If the patient has been already diagnosed with cancer, then knowing that he is a smoker would not affect our confidence level with regards to the result of X-ray imagery. Similarly, Dyspnoea and X-ray are conditionally independent giving Cancer. In general, two variable are conditionally dependent if they are connected through a converging node and their probability would change if new evidences are added to the descendent node or either one of them [45,p. 8]. The requirement that no hidden connections between variables exist apart from those shown in the graph is called the Markov property [21,p. 33]. It is not necessary for Bayesian Network to adhere to Markov property but then the graph won't be optimal that is there will be redundant arrows that connect independent variables together. When the number of connections between the variables are so compact that no further reduction is possible, the graph is called an I-map (short for independent map) [21,p. 33], otherwise it is a D-map (short for dependent map). Graphs that are both an I-map and a D-map are called the perfect graphs [21,p. 33]. With all these little inferences regarding the statistical relationships between the variable, the estimation of any probabilistic query will be much simplified.

It is normal to describe Bayesian Network graphs with the aid of metaphors. For example, the node which results from another is called a child and the latter node is called a parent [21,p. 32]. A node is an ancestor of another if it appears before that other and the latter is called its descendent. The top node which is a child of none is called root whereas a node with no children is called a leaf. Markov Blanket is defined as the current node parents and children and the parents of its children [21,p. 32]. For example, Cancer is a parent of X-ray which is a leaf. Pollution and Smoking are roots and parents of Cancer. Markov blanket of Cancer is Pollution, Smoking, X-ray, and Dyspnoea.

2.3.2 Types of reasoning

Since a Bayesian network graph is essentially a simplified alternative way of developing the joint probability distribution of some random variables, it can also used to answer all statistical queries as the case with an ordinary joint probability table. Usually, we refer to that process as reasoning and classify them into four types [21,p. 34].

Figure 8 shows the four types of reasoning with application to the example shown in figure 7. For drawing clarification reasons, nodes names of figure 7 are reduced in figure to only their first letter so that the node Cancer is now only C. Figure 8(a) shows the first type of reasoning available through Bayesian network. The direction of reasoning for this mode is from results or effects to causes. Thus, it is referred to as diagnostic reasoning [21,p. 34]. For example, If the patients X_ray turns out to be positive, then the probability that the patient has lung cancer will increase because lung cancer has an effect on

the expected details of an X-ray image. Figure 8(b) shows the case where we have evidences that the patient is a smoker. Since there is colouration between smoking and lung cancer, we may infer that the patient will develop symptoms of lung cancer. The direction of inference in this case is predictive as it is project the state of thing into the future. In figure 8(c), the patient is assumed to have diagnosed with lung cancer. In such case, the likelihood that the patient is a smoker or lives in a polluted environment will increase. If we acquire evidence that the patient was a smoker, then that would explain his disease and in turn reduce the likelihood that he have developed it due to pollution. This form of reasoning is known as explaining away or intercasual [21,p. 35]. Not all queries can be fit in a diagnostic, predictive or intercasual fashion as the network can be queried at any node with any type of available evidences. Such type of reasoning is called combined reasoning [21,p. 35] and example of it is shown in figure 8(d).

Evidences are another term for the arrival of new information which could in turn be uncertain. For example, new evidences on the Polluted node can be an unconditional probability of how likely the patient has been exposed to pollution. Similarly, new information could arrive in terms of the conditional probability of detecting cancer in an X-ray image giving that the patient has developed a cancer. This kind of uncertain evidence is referred to as virtual evidence or likelihood evidence [21,p. 35].

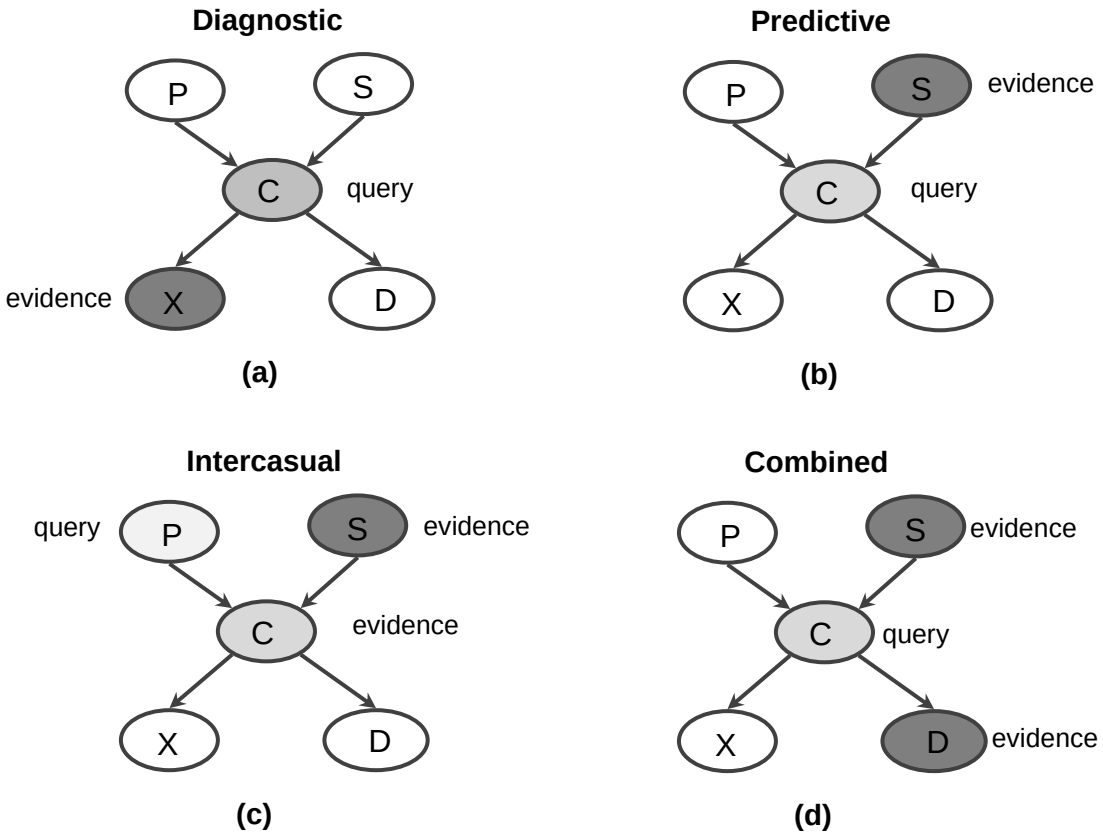


Figure 8. Four Types of reasoning in Bayesian Networks.

2.3.3 Inference in Bayesian Networks

The aim of any probabilistic network is to calculate the resultant, or posterior, probability on a given node(s) when evidences are available on other(s) [21,p. 53]. Usually, inference in Bayesian networks is either exact or approximate [21,p. 53]. The criteria for choosing which type of reasoning to adopt depends on the number of nodes in the network and the complexity of its interconnection [21,p. 53]. This section will briefly introduce key concepts of both reasoning types with application to the lung cancer example of figure 7. To unify the notation of queried nodes, evidences, and others, Uppercase

letters will be used to designate queried node such as X,P, or S, and lowercase letters to designate a specific evidence at a node such as x,p, or s.

Since a Bayesian network (BN) is a representation of joint probability table. Each node in a BN is represented by a conditional probability with regards to its parents and the multiplication of these together gives the joint probability table of the network as shown in equation 55 [1,p. 513]

$$P(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} P\left(x_i \middle| parents(X_i)\right) \quad (55)$$

If a node is a root, that is, it has no parent, then the unconditional probability is used instead of the conditional probability in equation 55. Hence, the joint probability of the lung cancer patient of figure 7 is given in equation 56.

$$P(P, S, C, X, D) = P(X|C)P(D|C)P(C|P, S)P(P)P(S) \quad (56)$$

Exact methods or algorithms use the joint probability representation of equation 55 and probabilistic relationships to compute posterior probability giving the availability of evidences on some nodes [1,p. 523]. For example, if the probability that the patient of figure 7 has cancer giving the availability of X-ray image and smoking status is to be calculated, then a procedure such as the following can be used:

$$P(C|x,s) = \frac{P(C,x,s)}{P(x,s)} \tag{57}$$

Equations 57 make use of the product rule given in equation 18. Let $\alpha = 1/P(x,s)$ be a normalization factor, equation 57 can be rewritten as:

$$P(C|x,s) = \alpha P(C,x,s) \tag{58}$$

The joint probability table of the nodes C,X, and S can be obtained by summing terms from the full probability table of all the nodes [1,p. 523]. Hence:

$$P(C,x,s) = \sum_{p}\sum_{d} P(C,x,s,p,d) \tag{59}$$

Where p,d is the sample space range of P and D respectively. Substituting equation 59 in 58, we get:

$$P(C|x,s) = \alpha \sum_{p}\sum_{d} P(C,x,s,p,d) \tag{60}$$

Finally using the full joint probability representation that we obtained through the Bayesian network graph in equation 56, we can rewrite equation 60 as:

$$P(C|x,s) = \alpha \sum_{p}\sum_{d} P(x|C)\, P(d|C) P(C|p,s) P(p) P(s) \tag{61}$$

Equation 61 can be enhanced further by noting that some terms are constant with regard to one or the two summations. The term $P(s)$ is constant with regard to both summations can be moved towards the far left end and the term $P(p)$ is constant in regards of the second summation over d. Hence, equation 61 becomes:

$$P(C|x,s) = \alpha\, P(s) \sum_{p} P(p) \sum_{d} P(x|C)P(d|C)P(C|p,s) \tag{62}$$

Such simple reducing observations can minimize the amount of computation power required to calculate the required query which can in turn be evaluated by looping through the two summations and multiplying the corresponding conditional probability tables entries on the go [1,p. 523]. Such tables need to be given prior to making any queries along with the unconditional probabilities of the root nodes. This approach is known as inference by enumeration. For a network with Boolean nodes, the complexity of calculation can be as high as $O(n2^n)$ [1,p. 523].

Researchers have proposed several methods to reduce the amount of calculations required to answer queries from a Bayesian network. For instance, Kim and Pearl's message passing algorithm works by performing a query in three steps: firstly the probability, or belief, of any node that its parents or children have received new evidence is updated, secondly the node calculates messages about the new evidence and send them to its parent in a bottom up propagation, thirdly the node computes a predictive factor and send it to its children [21,pp 57-59]. The attractive feature of the

Kim and Pearl's message passing algorithm is that everything is performed locally using the idea of estimating messages that reflects the availability of new evidences and their impact on the neighbouring nodes. As the number of parents connected to a node increases, the computation requirement of messages passages increases, in turn, as more cycles are required to update the status of the network [21,p. 61]. Alternatively, the variable elimination algorithm aims at reducing the amount of repeated calculations by saving them for later use [1,p. 524]. It starts by evaluating terms in a bottom up fashion, or in the case of equation 62, from right to left order, then as each term in equation 62 is a function of some variable and constant parameters, the portion of calculations that is variable independent is computed first and stored as cache complete further computations that are variable dependent [1,p. 524]. Once again, the amount of computation requirement of the variable elimination algorithm depends on the structure of the network and the amount of queries it is required to answer [1,p. 529]. It works best in networks where there is only one path connecting any two nodes which are known as singly-connected, polytree or forest networks [21,p. 56]. The upper half of the lung cancer patient shown in figure 7 is an example of a polytree network. If the network has more than one path between any two nodes then it is referred to as multiply-connected [1,p. 528].

The clustering algorithm belongs to the family of exact Bayesian network inference [1,p. 528]. It aims at reducing the complexity of multiply-connected networks by joining nodes together in order to transform a multiply-connected network to a singly-connected one. Figure 9 shows how that can

be achieved. Figure 9(a) is a multiply-connected network which can be simplified into singly-connected network if nodes B and C are merged together to form node B&C shown in figure 9(b). Then the belief updating or the variable elimination algorithm can be applied to the result [21,p. 67].

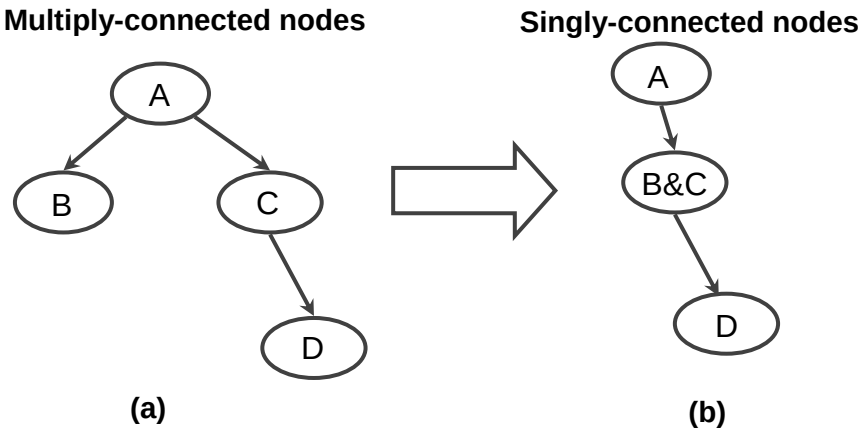


Figure 9. Grouping nodes together with the clustering algorithm

However, the transformation step might be greatly involved if the network is highly interconnected. Consequently, the amount of memory requirement for the transformation will also increase [21,p. 67].

An improved algorithm over the clustering approach is called the junction trees [21,p. 68]. It aims at increasing the efficiency of clustering through a methodological approach that starts by connecting all the parents together and removing their arrows pointers which produces the so called moral graph, then adding arcs to every groups of nodes larger than 3 which will result in a triangulated graph, then identifying the new merged nodes from the triangulated graph so as to produce a junction tree, and finally creating separators from the arcs that results from the intersection of adjacent nodes [21,p. 68]. As soon as the network is simplified, the algorithm proceeds to

calculate the new probability tables for the new combined nodes and then update the probabilities across the network by the message updating algorithm [21,p. 68]. Once more, the junction tree algorithm adds a substantial overhead to making queries through the transformation phase although it only needs to be done once. In addition, the new probability tables may have many entries that are simply zeros and thus takes occupy unnecessary memory [21,p. 69].

In general, the exact inference approach in BN reduces the task of evaluating an exponentially increasing joint probability tables by dynamic programming or transformation. However, exact inference is still bound by the worst case scenario of an exponential performance such as the case with some Bayesian networks build to model pixels in an image [20,p. 336]. In addition, new challenges will be introduced to the exact inference approach if the network variables were continuous rather than discrete [20,p. 337]. Particularly, when the new joint probability tables are calculated.

The exact approach to inference in Bayesian network is usually used for small to medium sized network in which the number of nodes is up to about three dozen [21,p. 72]. For networks with higher amount of nodes or multiply-connected network with high density of connection, an approximate inference should be followed [21,p. 72]. There are several approaches to approximate the inference process in Bayesian networks however the most common ones are those which depends on performing stochastic simulation such as the logic sampling methods [21,p. 72] also referred to as the direct sampling methods [1,p. 530]. In these methods, the network is used to generate cases

based on random initiation of evidences, then the posterior probabilities of the child nodes are estimated from the direction of roots to leaves, finally the procedure are repeated while the estimated probability of the queried conditional probability, like P(X/E) is updated. In order to estimate the value of P(X/E), the number of cases where both X and E are true are counted and divided by the number of cases where only E is true. According the law of large numbers, the updated probability estimate should converge to the exact value [21,p. 72]. Equation 63 gives the mathematical formulation of how the approximate posterior probability is calculated [21,p. 72]:

$$P(X = x_i|E = e) = \frac{Count(x_i, E = e)}{Count(E = e)} \quad (63)$$

Hence the sampling methods become mathematically inefficient when the chosen evidence is unlikely as some of the cases which do not contribute to the count of equation 63 will be discarded [21,p. 74]. In addition, the process of inference is directed, its power can mostly be observed in directed networks [20,p. 540].

The performance of the direct sampling methods can be enhanced by giving more attention to cases that are more consistent with the evidences [1,p. 532]. One example of such approach is the likelihood weighting algorithm which arises from the importance sampling technique in statistics but modified for Bayesian inference [1,p. 532]. If the posterior probability P(X/E) is queried, then the evidence nodes E values are set as constant while the other nodes

are samples as to generate cases, then each case is weighted by the likelihood of that evidence combinations [21,p. 74]. Although the weighting likelihood is more efficient than the direct sampling method, its performance will start to degrade as the number of evidences increase.

Another simulation based algorithms is the Markov chain Monte Carlo simulation methods (MCMC). MCMC works not by randomly initializing every case and manually working out the posterior probabilities but rather by making some random changes to the current case so as to obtain the next case [1,p. 535]. For example, the Gibbs sampling method works by an arbitrarily initialization of a case where the evidences are fixed at their observed values then the next case is obtained by applying random changes to one of the unobserved variables such as X_i which is then conditioned on the Markov blanket X_i [1,p. 536].

There are many other approximate inference methods that don't use the random case generation approach. For example, the search methods that instead of generated cases randomly, they try to pay more attention to cases with high likelihood. Therefore they don't generate an unbiased posterior probability estimate but a good upper and lower bound [20,p. 540]. Although, sampling methods are widely used to make good approximate to posterior probability in many Bayesian networks configurations, their performance is not easy to expect. In particular with complex probability distribution where the estimate obtained from generating cases is considerably inaccurate [20,p. 541].

In summary, although Bayesian networks provide simplified graphical representation of the joint probability table of some random variables, there are many instances where even the resulting joint probability is computationally expensive to query. Thus a need for better inference algorithms is justified and reflected by ever-active research efforts to reduce the time and computation power required to query a Bayesian network. In general, inference can be classified as exact and approximate. In the exact inference, the complex structure of the network is often reduced so as to obtain a new structure that is known to be more computationally efficient. Approximate inference relies on simulating cases so as to generate a large amount of samples that would converge to the value of the posterior probability resulted from exact inference. Exact and approximate methods can be combined together to obtain an algorithm that mosaic-wise combine the features of the two [20,p. 541]. Finally, the performance of the two approaches is bounded by factors like the complexity of network, the amount of connections between nodes, the complexity of distribution and the likelihood of evidences.

2.3.4 Dynamic Bayesian Networks

The previous section has shown the advantages of Bayesian networks as a simplified graphical representation of the joint probability of some random variables defined over a given process. The resulting network can then be used to answer any probabilistic query given the availability of some evidences. For example, the BN of the lung cancer patient shown in figure 7 can be used to estimate the likelihood of a patient having lung cancer giving

that he/she is a smoker. However, the network assumes that all patients' cases can be represented by the same variables connected together in a fixed structure fashion. This type of modelling is referred to as variable-based modelling [20,p. 199]. There are many applications where the process changes over time and we are more interested in capturing the dynamic behaviour of it. For example, in the case of inferring patients' states in an ICU, the states of the patients change over time in a way that the next state depends on the current one and some observed variables that are samples at some time intervals such as heart rate, blood pressure and urine output. While an ordinary BN can model the relationship of current patient state in terms of some observed variables, it fails to capture the dynamic nature of how that state evolves over time and, in turn, fails to represent the distribution of the patient's state over time. In addition, there are other classes of problem where the structure changes with every case. Consider, for example, the modelling of a genetic inheritance. In this case, each family has its own members which in turn have their own variables [20,p. 199]. Nonetheless, the way in which genes are inherited is the same for every family [20,p. 199]. This calls for a better way of representing dynamic processes then a mere variable-based fashion such as Dynamic Bayesian Networks (DBN). DBNs are models that work as templates to represents the temporal dynamics of an entire class of distribution in a compact way [20,p. 199]. The basic assumption of a DBN is that the world consists of successive temporal snapshots where each one has some random variables which are either observed or hidden [1,p. 567], and that the way the system evolves over time, called transition model, depends on a fixed number of previous states so as to prevent the transition probability

between the current and next state from becoming infinitely unbound [1,p. 568]. This assumption is known as the Markov assumption and the process that satisfies it is known as Markov chain [1,p. 568]. If the next state in a temporal transition model depends only on the previous state, then it is called a first order Markov chain whereas if it depends on the previous two states then it is called second order Markov chain [1,p. 568]. The set of system states at a given time instant like t is often designated as X_t and the set of evidences as E_t [1,p. 567]. Thus the transition model of a first order Markov chain can be expressed as [1,p. 568]:

$$P(X_t|X_{0:t-1}) = P(X_t|X_{t-1}) \quad (64)$$

Whereas the transition model of a second order Markov chain can be expressed as [1,p. 568]:

$$P(X_t|X_{0:t-1}) = P(X_t|X_{t-1}, X_{t-2}) \quad (65)$$

In addition, the transition model is assumed to be fixed over time. That is to say the temporal-based conditional probability is constant regardless of the current time. Using the chain rule of probability, the temporal joint probability distribution of a Markovian process can be expressed as [20,p. 201]:

$$P(X_{0:t}) = P(X_0)\prod_{i=0}^{t-1} P(X_{i+1}|X_{0:i}) \quad (66)$$

The Markov assumption can be further extended to the case of evidences. Evidences, or observed variables, may also depend on previous variables as well as the process states. However, careful modelling of the process states would make it suffice to generate the observed variables entirely so that the Markov assumption of the evidence, known as the sensory model, can be written as [1,p. 568]:

$$P(E_t|X_{0:t}, E_{0:t-1}) = P(E_t|X_t) \tag{67}$$

Combining equation 67 with 66 yields the general template temporal model of DBN that satisfies the Markov property given in equation 68 [1,p. 569]:

$$P(X_{0:t}, E_{1:t}) = P(X_0) \prod_{i=0}^{t} P(X_i|X_{i-1})\, P(E_i|X_i) \tag{68}$$

Equation 68 assumes that the evidences, or observations, start to arrive from time slice 1. Hence at time slice 0, there are no evidences to have a conditional probability and the only information available about the process is its unconditional probability which is an intuitive conclusion giving that the unconditional probability of a variable is its likelihood in the event of no available evidences. In addition, equation 68 shows that a DBN can be represented by three sub-models: the transition model $P(X_i|X_{i-1})$ which structures the evolution of the process variables between the current and next time slice, the sensor model $P(E_i|X_i)$ that connects the current process states with the observed sensors, and the unconditional probability distribution of the process variables $P(X_0)$ [1,p. 591]. Hence, it is more convenient to only plot

one slice of the DBN that shows the prior unconditional variables, the transition model, and the sensory model [1,p. 591]. Figure 10 shows a DBN representation of patient monitoring in ICU.

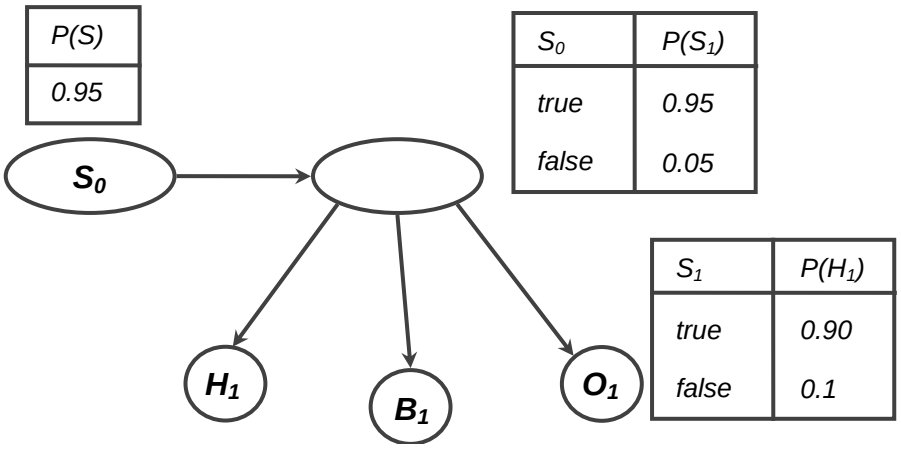


Figure 10. Simple DBN for monitoring patients at ICU

The state variables are shown as oval shapes whereas evidences are shown as circles. For simplicity, the state of the patient (denotes as S) can either be *true* or *false*. A *true* state indicates that the patient is a live whereas *false* indicates that the patient is deteriorating. The sensory model consists of three variables: heart rate (H), blood pressure (B) and oxygen saturation (O). The probability tables are filled with arbitrary values to serve as a demonstrating example of how it would look like. Although, figure 10 only specifies the sensor probability table of the heart rate sensor, the remaining tables follows the same structure of the heart rate probability table. As similar to the sate variable, the sensor conditional probability table can assume any of two states: *beating* or *non-beating*. The event of patient deteriorating while the heart rate sensor is still showing *beating* is assigned a probability of 0.1 to emphasize the likelihood of sensor failure. It represents the simplest modelling of a sensor failure commonly known as the transient failure model [1,p. 593].

It makes it possible to distinguish between nonsensical sensor reading due to sensor failure and reliable readings. If the predicted likelihood of *non-beating* heart rate sensor state giving all the past patient states is much less than the probability of transient sensor failure then the best explanation of the previous event is that the sensor has failed [1,p. 593]. Equation 69 gives a mathematical criterion of detecting the event of heart rate sensor failure at time slice *t*:

$$P(H_t = nonbeating|S_{1:t-1}) \ll P(H_t = nonbeating|S_t = true) \quad (69)$$

While the transient model helps smooth out the recorded history of sensor readings by removing the less probable ones according to equation 69, it still fails in cases where the failure is persistent [1,p. 593]. For example, if the heart rate sensor lead attached to the patient is disconnected. In order for the DBN to accommodate persistent failure, a persistent sensor model needs to

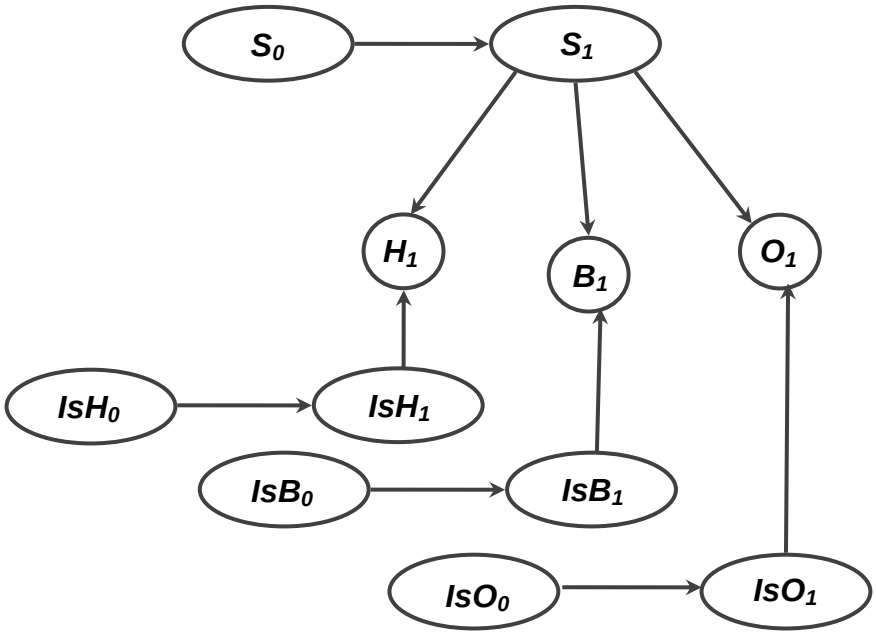


Figure 11. Modified DBN of figure 10

be developed where the sensor itself will have a hidden state that would be interpreted using the available evidences. If the state of a sensor is designated with a prefix of (*Is*), the DBN of figure 10 can be expanded to that of figure 11 which introduces three new states describing the conditions of the sensors.

Inference in DBN can also be classified as exact or approximate and the same techniques used to query an ordinary BN can also be used with a DBN [1,p. 595]. However, the basic models of figures 10 or 11 need to be replicated, or unrolled, until it fits the present amount of observations [1,p. 595]. Figure 12 shows the unrolling of figure 10 to time slice 3 where the three observations nodes are combined into one node for simplicity of drawing. There are many inference techniques proposed by researcher to reduce the amount of computations required to accomplish the task of probabilistic querying. Reference [20] discusses some of them in more details.

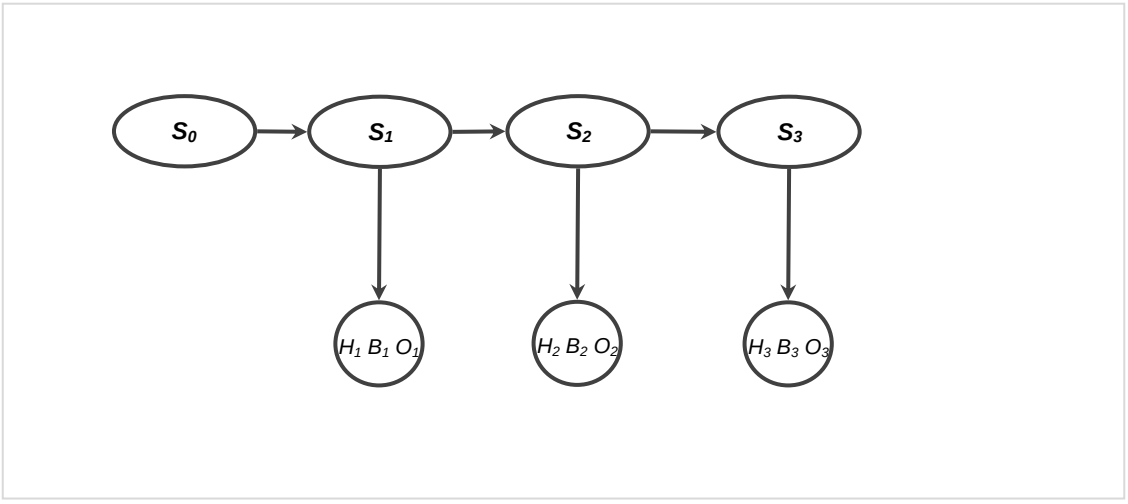


Figure 12. DBN of figure 10 rolled to time slice 3

2.3.5 Decision networks

In order to make good decision under uncertainty, two factors need to be known; i) the likelihood of every possible outcome, and ii) the preferences of the decision-maker with regard to each of the outcomes. Bayesian networks provide a sound methodology to obtain the probability of outcomes as discussed in the previous sections. Combining Bayesian networks with preferences will give a powerful foundation of making decisions under uncertainty [21,p. 89]. As discussed in section 1.3, preferences are better expressed in terms of a utility function the maps an outcome to a numerical value that conveys a useful aspect of the outcome to the decision-maker. Once the utility function is defined, the expected utility of each decision is calculated by [21,p. 89]:

$$EU(A|E) = \sum_i P(O_i|E,A)\, U(O_i|A) \tag{70}$$

where O_i is i-th the possible outcome, *A* is the actions for outcome O, $U(O_i|A)$ is the expected utility of each outcome when action *A* is made and $P(O_i|E,A)$ is the conditional probability of the i-th outcome giving the current evidences *E* and action *A* is made. The action with the highest utility is often selected if the principle of maximum expected utility is followed [21,p. 90]. The principle of maximum expected utility states that rational agents have a tendency to prefer the action that results in the maximum possible utility [21,p. 90]. Decision networks may be expressed graphically by extending BN graph with decisions and utility nodes [21,p. 91]. Decisions, or actions, are often

represented with a rectangular shape and the utility nodes with diamond shapes.

For example, the BN for monitoring an ICU patient given in figure 10 can be extended by considering what decisions a nurse would make for every possible state of the patient and the utility of each decision. Since there are only two possible states for the patient: alive and deteriorating then the nurse may only make one decision which is to contact the doctor in case the patient is deteriorating and to continue monitoring otherwise. The expected utility of contacting the doctor has an effect on the next state of the patient so a second node should be added to simulate temporal relationship between the current state of the patient, next state and the undertaken action. The utility function itself is used to map the decision of contacting the doctor to a numerical value that reflects whether the decision led to the recovery of the patient or further deterioration, see figure 13.

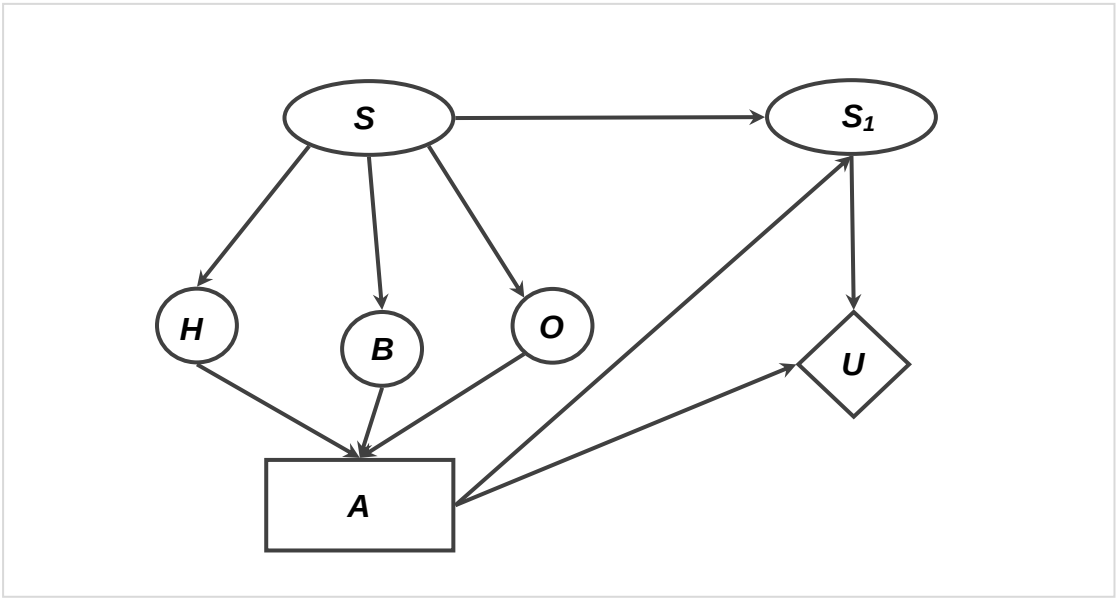


Figure 13. A simple decision network based on the DBN of figure 10

The decision block *A* belongs to the class of actions known as intervening actions [21,p. 97]. Intervening actions are those actions that have an effect on the probability of the outcome of the network. In the case of figure 13, making a decision to call the doctor would change the likelihood of the current state of the patient. Non-intervening actions are those which do not affect the probability of the system for example, betting in a gambling game [21,p. 97]. Although the decision network of figure 13 is very simple, it can be extended to include more than one decision in a sequential decision-making fashion such as to approximate what decision-maker would do in a course of actions. For example, the nurse may decide to make some test before deciding to contact the doctor to further confirm that the patient is really deteriorating. Such type of actions are referred to as test nodes [21,p. 98]. However, by the time the test is performed, it may be too late for the patient so a test node should be accompanied by a cost node. Similarly to the expected utility node, the cost node maps a cost-wise aspect of performing a test into a numerical value [21,p. 98]. A test has an effect on making further decision and can be regarded as evidences but it has no effect on the states of the process. Figure 14 shows a simple addition of a test node (T) with cost (C) that a nurse can undertake to confirm the readings of the ICU monitors.

The dashed arrow between the test node (T) and the action node (A) shows which one should be performed first and is known as the precedence link [21,p. 98]. In order to evaluate the utility of each decision, the decision network is transferred into a decision tree model [21,p. 101]. Each possible outcome of an action or test is represented by a branch starting with the

action/test that has the highest precedence and continue to divide each branch according to the possible actions/tests in the sequence of actions/tests, then the tree is further branched based on the possible outcomes of the states nodes and finally each ending leaf is weighted by its

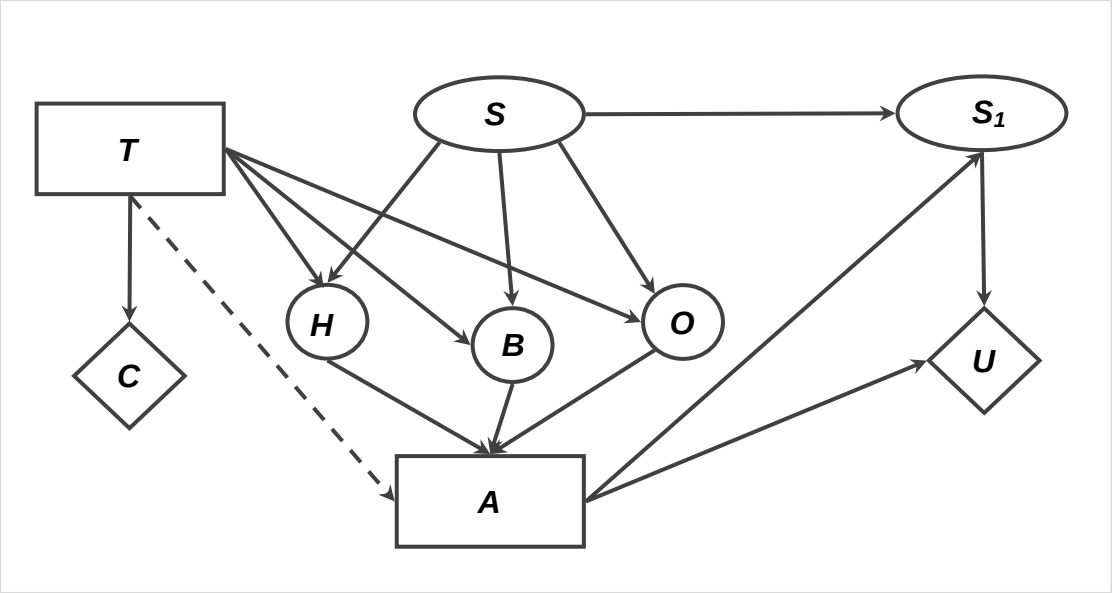


Figure 14. The addition of a test node to the network of figure 13

expected utility. Once the decision tree diagram is plotted, the calculation of the expected utility of actions follows from the bottom leaves where the utility nodes reside to the action/test node of the highest precedence by multiplying the value of the expected utility by its likelihood and then summing over the next branch until the root is reached [21,p. 103]. Once the utility of each decision is estimated, the one with the highest utility is selected if the principle of maximum expected utility is followed. Although analyzing the decision network through a decision tree seems appealing from simplicity point of view, it is computationally inefficient as it involves repetition of similar mathematical terms. It can be improved using the same techniques introduced in the

previous section for Bayesian network inference such as structure transformation and variable elimination [21,p. 104]. Decision nodes can also be added to a DBN so as to model the temporal evolution of actions through time and thereby creating a dynamic decision network (DDN). Figure 15 shows how the decision network for monitoring an ICU patient of figure 13 can be combined with the DBN of figure 12.

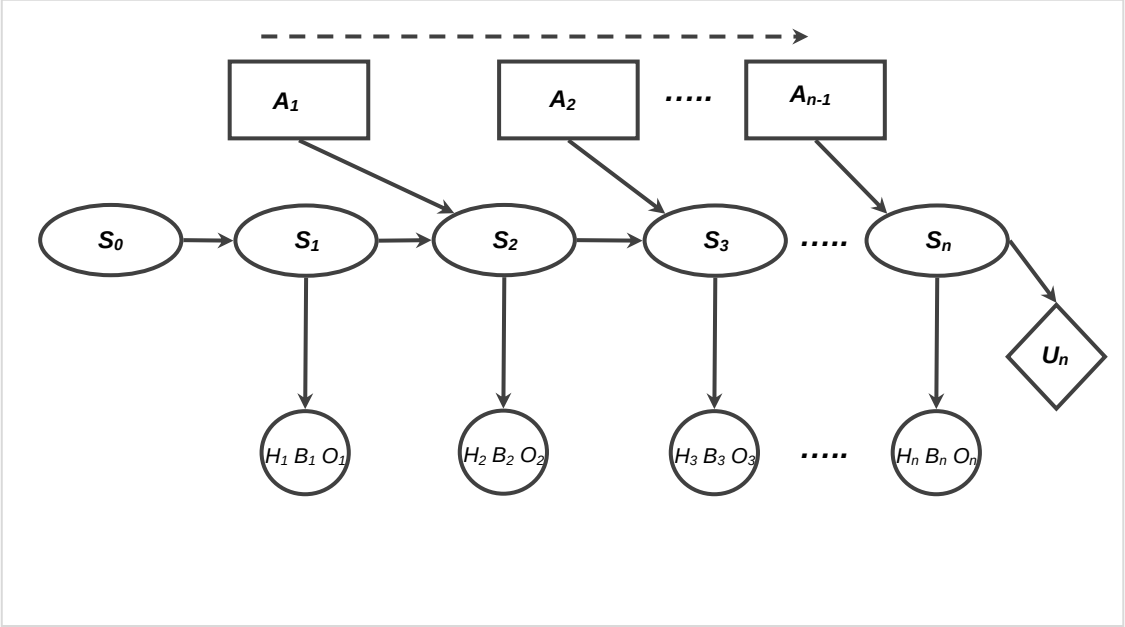


Figure 15. An example of DDN based on figure 12

The DDN of figure 15 assumes that the sequence of actions (shown as A_1, A_2 and A_n) starts after the arrival of the first evidence and that they have precedence from left to right as indicated by the dashed arrow. In addition, it assumes that the decision-maker is interested in the utility of making the first (n) sequence of evidences which is modelled by the inclusion of only one utility node after unrolling the DDN for (n) times. If a utility node is added to

each slice, then the measured utility will be the change in utility from between the previous and the current action [21,p. 110].

2.3.6 Learning Bayesian Network

Up until now, we have seen several approaches to answer probabilistic queries over some random variables and how, in turn, the results can be used to make decisions or even further queries. In addition, we have seen how to computationally reduce the requirements of performing a query through structure and variable transformations. However, the availability of a model that fully describes the variables is assumed a priori. Such a model is not always available in the domain where a process is to be modelled [20,p. 697]. Firstly, it may be too much work to construct a model that describes a very complex process with hundreds of underlying variables or there can be no expert who fully understands the process to come up with a causal model that describes the interconnection between the inputs, system states, and the outputs. Secondly, the more the model becomes detailed and involved the more inflexible it gets because the model would need to be modified in order to fit another process within the same domain. Hence, an expert is needed to update the model every time the process changes or upgraded in some sort [20,p. 697]. Finally, the resultant model needs to be thoroughly tested with examples for the results are known but the accuracy of the model under circumstances where it was not fully tested would be unknown. However, there are many cases where we have an extensive amount of data that shows input-output or situations-results pairing that could work as examples derived from the distribution to be modelled [20,p. 697]. For example, a flight data

analysis program normally has access to an extensive set of flight data recorded during the flight time of an aircraft, known as flight data recorder (FDR). If an anomaly occurred during the flight time of an aircraft, then an investigation will be initiated and the causes of the anomaly will be identified. In such case, the flight data analysis program has access to both situations and labels of the data that can be used to construct the probability distribution that describes the relationship between the FDR recorded variables at a given time instance and the normality/abnormality. Similarly, hospitals often have periodic tables that list the conditions of patients throughout their admittance to the hospital. A patient data may include several key variables that can be used to model the relationship between the current symptoms and the diagnostic suggested by a doctor.

In this section, we will briefly introduce a common technique to learn the distribution of a process from sets of examples. This approach is generally performed to learn either the model underlying the variables or the probability distribution over them. In addition, a goal or a set of goals are defined to describe the end result of the learning process so as to measure the convergence of the learnt model to the data from the actual process [20,p. 698].

Sewall Wright pioneered the work on using graphical models to make probabilistic inference in early twentieth century [21,p. 153]. His work was mainly focused on using linear models to represent casual representation to estimate the coefficients of such models in an approach that later became to be known as path modelling and gained extensive popularity in social science

[21,p. 153]. In essence, the relationship between three variables such as X, Y, and Z in a linear model can be expressed as [21,p. 153]:

$$Z = aX + bY + U \tag{71}$$

where a and b are coefficients that determine the contribution of X and Y to Z and U is the amount of Z that could not be approximated by the linear model. Hence, equation 71 shows that X and Y are statistical causes of Z. This fact can, in turn, be used to construct casual models. A typical procedure to accomplish that would be to search for correlations among the variables, inventing a casual model that would explain the correlations between them, then use Wright's decomposition rules to estimate the coefficients of the linear model and, finally, to test the model using some examples and measure the model accuracy [21,p. 160]. Wright's decomposition rules are guidelines that relate the observation of correlation to the way the weights of the linear model coefficients can be estimated [21,p. 155]. However, in order to use Wright's approach within the artificial intelligence realm, a computerized approach to inventing the casual models need to be developed. One realization of such approach is to observe conditional independence within the data. Conditional independence is the cornerstone that BN uses to reduce the amount of calculations needed to construct the joint probability tables of random variables. Thus, knowing that X is independent of Y giving Z would give us an idea of how to graphically represent the relationship between X, Y, and Z. The conditional independence learner algorithm is one example of such an approach [21,p. 161]. However, this leaves the question of how to know the

actual conditional independence between variables unanswered. The PC algorithm answers that question by the introduction of statistical significance test for conditional independence where independence is represented by the remaining correlation between two variables when a third is held constant, also known as vanishing partial correlation [21,p. 167]. Although these methods are developed for linear systems, they can be further extended to non-linear system by non-linear to linear transformation functions [21,p. 153]. Another approach to constructing casual network for discrete process is the Bayesian Metric (BM) [21,p. 197]. Rather than using statistical test and relationships, BM searches the casual model space using a metric conditional function like P(·/e) or an approximation to it and looks for the best model that maximizes that function [21,p. 197]. There are many algorithms within the family of BM and reference [21, ch 8] lists and discusses them in a chronological order.

2.4 Summary

This chapter serves as an introductory stage to the main theory of this thesis that will be discussed in chapter 3. Probability calculus is usually build upon the principle of counting and how to theoretically estimate the number of possible combinations of an experiment or a series of experiments. The probability of an event is the proportion of outcomes where the event occurs to the total number of outcomes. The most important conclusions of our modern theory of probability are the central limit theorem and the strong law of large numbers.

All is needed to answer any probabilistic query in a process is the joint probability table, or distribution, over the process variables. However, construction of such a table is often not practical with processes that have hundreds of variables. Bayesian networks are graphical representations of the joint probability table that greatly simplifies the computations required to make probabilistic queries. BN network can be integrated with preferences to construct decision network that helps simplify the process of probabilistic decision-making. DBN are template based representation of a process with emphasis on the dynamic evolution of variables over time. Finally, BN models can be achieved through data mining approach by using examples to construct the relationship underlying some variables or to estimate the probability distribution of the variables.

3. Theory of Comparative Probability

In chapter 2, we have seen how probability can be used to represent uncertainty and how, in turn, this can be combined with preferences to make decisions under uncertainty. In addition, chapter 2 briefly introduced the theoretical foundation of the frequency interpretation of probability and its relation to combinatorial analysis and well-known distributions such as the Gaussian or normal distribution. It serves as an entry point to emphasise the importance of the joint probability distribution/table in answering probabilistic queries. More importantly, it contains a discussion of the use of BN to simplify the creation of a joint probability table and the construction of these networks in the light of the information age, where an extensive set of examples is more likely to exist for use in constructing a BN or estimating the probability density.

However, the availability of examples may turn into a curse rather than a blessing because the amount of computational power and time required to process terabytes of information becomes impractical. In addition, the representation power of a probability density estimator for a process that has thousands of variables cannot be easily measured because there might not be enough examples to adequately profile the sample space. Moreover, many

processes have extensive sets of examples that show one mode or a few possible modes of operation. For example, a web server access log would have many examples of normal behaviour for users connected and browsing throughout the sites hosted by that server, but the behaviour of a hacker trying to exploit the server is rarely covered by a typical sample from it. As another example, consider typical data extracted from the flight data recorder of a commercial aircraft during a normal flight between two airports. Unquestionably, these examples do not fully represent the normal behaviour of the aircraft equipment during normal conditions because there are many variables that would drastically change if the weather or the flight route changes.

One way to tackle the problem of extensivity of information and save memory is to process the set of examples recursively on-the-fly. In such a scenario, the network would not have an offline period where it works out its internal structure while processing the given examples, but rather it will be available as soon as it is initialized and the first examples start rolling in. Some of these techniques can even evolve in the sense of being able to change their structure to better represent the data. However, this approach leaves the issue of the representation power of the network during the transient initialization period unanswered because the amount of information available to the network may be sparse.

The purpose of this thesis is to investigate similar conditions when an online system initialises and there is insufficient information to make a good decision. Furthermore, it examines the optimum initialisation values of the

unconditional probabilities values without resorting to ad hoc assumptions or ignorance. If the assumptions made during initialising an online system are to be analysed, it is only logical to study the various meanings of probability and the relation of each to the availability of information. Therefore, a few background discussion points the need this to be made before we address the main theme of the thesis.

Chapter 3 will introduce more focused probability topics and an interesting approach to representing knowledge. It first introduces various interpretations of probability and sets the requirement for the best interpretation to solve the main question of the thesis. Then we will walk through the proposed interpretation, from its axioms to its advanced results. The chapter will conclude by introducing an innovative way to represent probabilistic knowledge, which combines the strong points and results of various probability interpretations and theories.

3.1 Interpretation of probability

Consider a factory that produces cubes each having a side length of one unit. What will be the probability that a cube made by the factory has a side length less than or equal to ½? The answer to this question should be obvious from the principles of probability introduced in chapter 2. If the manufacturing process has a uniform probability distribution of cubes with side lengths between 0 and 1, then the answer is ½ [46]. If we change the description of the problem to finding the probability of a cube with a side area

of less than or equal to ¼, then the answer would be ¼, assuming the same uniform distribution again. Surprisingly, both problems describe the same event because a cube with a side length $\leqslant$ ½ has a side area of $\leqslant$ ¼, yet the probability of the former is twice that of the latter. It becomes even worse if the same problem is described in terms of the volume, as the probability of a cube with a volume $\leqslant$ 1/8 is 1/8! The inconsistency described in this example is known as the Bertrand paradox [47,p. 77]. It highlights issues with the classical interpretation of probability. If the description of the random variables in a process is so vague as to lead us to assign different probability values to each possible way of describing them, then a decision-maker trying to assess a situation where the available evidence, or data, is not enough to clearly define its variables may fall into the same situation, resulting in different probability values for each different definition of the same situation. Therefore, the theory of probability needs to be re-investigated in hopes of finding an interpretation that better suits the situation of online learning or decision-making based on sparse evidence.

Interpretation of probability is the task of providing analysis of the basic concepts of probability or the transformation of informal concepts used during everyday use to formal ones suitable for scientific theorizing [46]. There are many interpretations of probability in the literature of philosophy and science, but they generally fall into one of the three categories: the objective interpretation of probability, the subjective interpretation of probability and the logical interpretation of probability [48]. Before any candidate from within these categories is analysed, a framework of assessment needs to be

defined. The criteria help determine the adequacy of the interpretation to a set of standards. The Kolmogorov axioms of probability may provide criteria for assessing a candidate, but these axioms can satisfy not only a probability system but a non-probabilistic system, as well, such as the volume of a sphere. In addition, the Kolmogorov axioms of probability do not tell us much about how to assign a probability to an event outside the boundary cases, where the probability of all possible events should be 1 and the probability of a void event is 0. Salmon proposed three criteria to investigate various interpretations of probability [49,p. 63]. Although they are intuitively simple, Salmon noted that it is surprisingly difficult to find a candidate that satisfies them all. His criteria are:

a) *Admissibility*. This criterion ensures that a proposed interpretation adheres to the theory of probability and mathematics in general. It requires that the terms of an interpretation resolve the formal axioms of probability theory into true statements [49,p. 63].
b) *Ascertainability.* This criterion requires the availability of methods to assess the value of a probability, because the probability theory would be useless without the ability to calculate probabilities of events [49,p. 64].
c) *Applicability.* The interpretation of probability should be useful in the sense of being able to predict or to be used to assess some situation and have some bearing on which conclusions could be drawn. Its driving force is the famous Bishop Butler quote, “Probability is the very guide to life” [49,p. 64].

As stated earlier, there are non-probabilistic systems that satisfy the axioms of probability. For example, a cube of unit volume divided into several smaller cubes will satisfy the axioms of probability because the sum of all of these little cubes' volumes is 1, the volume of the void cube (no cube) is zero and all cubes' volumes are non-negative. In addition, the formal system of logic can satisfy the second criteria of ascertainability because the probability of a true statement can be regarded as 1 and the probability of a false statement as 0 [46], yet we do not consider formal logic as an interpretation of probability, although it is admissible, ascertainable and applicable.

A better approach, at least for scientific-oriented minds, is to focus on applicability, the power of the interpretation in explaining common probabilistic observation, [46] and its prediction power. After all, the purpose of choosing a specific interpretation is to make good predictions in terms of rational choices and modelling the behaviour of a process.

The classical interpretation of probability was championed by Laplace but can also be found in the work of Pascal, Bernoulli, Huygens, and Leibniz [46]. Estimating the probability of an event proceeds by breaking down events of the same kind into simpler events until they become equally possible; then the probability of the event is the ratio of the number of times it occurs to the number of all possibilities [49,p. 65]. Although chapter 2 referred to such probability as unconditional probability, it should be noted that the two are fundamentally different. Laplace does not use the frequency of appearance to estimate his probability calculation, but rather the sample space of events. That means the odds of heads in a coin toss and that of a yes answer in a

wedding proposal are fundamentally the same, which seems absurd. If probability is to be the guide for life, then it must be estimated from life a posterior; not a priori. Furthermore, Laplace's interpretation seems to draw its conclusion from ignorance, because in the absence of evidence to favour one event over the other, which he refers to as equally possible, then how we are to conclude that all events are equally possible and, in turn, what use is it as predictive tool? In addition, in a deterministic world, we would presume that a coin's side would be determined by the initial conditions under which the coin was tossed. Therefore, how would two events be equally possible? Even if we artificially set up initial conditions that would favour no side of the coin over the other, would it not be more preferable to presume that the coin would land on its edge? Finally, since the outcome: *edge* is a valid possibility, then there should be three possible outcomes, each with a probability of 1/3. It would turn out that we need to assign a highly absurd outcome the same probability as a normal or expected outcome. Clearly, we can conclude that the Laplace assumption—that during initial conditions and with no prior knowledge of the likelihood of outcomes, they should be assigned equal probability—will not be useful to answer the core question of this thesis.

3.1.1 Objective interpretations of probability

One important interpretation of probability that generalises the assumptions of the classical interpretation [46] is logical probability or probabilistic logic. The importance of the logical interpretation of probability comes from the fact that it can be used in science as a deductive tool to quantify the supporting power of evidence for a given hypothesis [50]. Hence,

it is applicable to the field of artificial intelligence through the automation of the process of updating the likelihood of the truthfulness of a hypothesis in the light of new evidence. As opposed to the classical interpretation of probability, the logical view can assign different weights to different probabilistic outcomes and can accommodate evidence [46]. In order for the logical interpretation to be useful, the number of hypotheses should be limited and known, which is not always possible. Moreover, the probability of each hypothesis is initially assigned equal weight [49,p. 73], which seems as if it is utilizing the principle of indifference, but certainly not every possible hypothesis should be assigned equal probability because ignorance as probability is a measure of possibility, not ignorance [50]. Finally, the logical interpretation lacks the ability to adapt to new changes in evidence and hypotheses. For example, if the number of hypotheses increase or new features or outcomes are discovered then the degree of confirmation becomes void and needs to be re-initialized [46]. But, if the learning process needs to reinitialize every time we discover something new, then we are not updating our confirmation degree regarding the truthfulness of a hypothesis but rather measuring its likelihood using a fancier term than the mere classical interpretation of probability offers.

As opposed to the analytical approach of the classical and logical interpretations of probability, there is a class of interpretation that regards probability as an objective property of things in reality. The frequency interpretation of probability introduced in chapter 2 is one example. From the frequency interpretation point of view, the probability of an event like A in sample space B should be estimated from its relative frequency of occurrence

within B. The frequency of occurrence is estimated from experience, where an experiment is set up so as to randomly output events from the sample space B and the number of times that A occurs is recorded as the experiment is repeated indefinitely. The frequency interpretation is de facto in science because it can be estimated from empirical observations of some experiments and because it can express less frequent events in terms of lower numerical probability value. However, the very definition of the frequency interpretation of probability raises more issues regarding its suitability for this thesis. A major issue is the so-called problem of the single case [46]: what would be the probability of an event in an experiment repeated once or a few times? Consider, for example, a coin flipped once: The probability of heads will be either 1 or 0, depending on which side the coin lands on by mere chance. Moreover, if the experiment of flipping a coin is repeated a thousand times, we can still consider the one thousand flips as a single case [46], albeit a synthetic one, and we will be back to the problem of the single case again. Even if we assume the availability of an infinite series of experiments where a coin is flipped and the outcome is recorded, we will still arrive at different probability values as the order of outcomes is rearranged. In fact, the probability of heads could be made to converge to any value from 0 to 1 [46]. Finally, how should we deal with imperfections in the collected data? The frequency approach seems to assume that the data are a perfect replication of reality, which is a very crude assumption. In many cases, the gathered data are noisy, incomplete, or even nonsensical. Hence the estimated probability of an observation is by itself probable. It will be highly absurd and less useful to express the likelihood of an A as the probability of the probability of A. If

probability is to be estimated in an objective fashion from a few examples, then we should not completely rely on an interpretation that does not work well under such circumstances.

The other “horn” of the objective approach to probability is the propensity interpretation proposed by Carl Popper in 1959 [51]. Rather than considering the frequency of occurrence as the objective realization of probability, the propensity interpretation considers probability as a sort of disposition [52], that is, the coin has a tendency to land heads half of the time. That tendency is a natural property of the coin, even if it has not been flipped yet. Likelihood and chance are viewed as real properties of things that cannot be reduced to other properties or systematic set-ups [52]. All philosophical issues with the propensity interpretation put aside, it is unclear how to estimate the probability of an event and why would it be this value and not any other. For example, why does a fair coin land heads half the time? Why does this not occur one third of the time? Without a well-defined math for the propensity interpretation, it would be of less application to the field of artificial intelligence and, in turn, this thesis, although it may serve well to explain some mind-boggling observations in quantum mechanics.

3.1.2 Subjective interpretations of probability

Subjective interpretations of probability refer to the class of probability analysis that aims at rationalizing everyday notations of probabilistic statements. For the subjectivists, probability is an epistemic statement about reality that expresses a belief in the trustworthiness of statements rather than about properties of reality itself, as is the case with the objective prospective.

Hence, probability can be applied to single events. In addition, subjective probability can be applied even to deterministic outcomes because its statement does not convey chances in real world, but rather how much an agent believes a particular outcome will result [53,p. 2]. The main difference between subjective and objective probability interpretations is the attitude toward knowledge extracted from events. For example, the objectivist will describe the likelihood outcome of a biased coin flip as P(outcome) $\neq$ ½, whereas the subjectivist would continue to consider the outcome as P(outcome) = ½ because the subjectivist does not know the direction of the bias so as to favour one outcome over the other [53,p. 2]. In addition, the subjectivist would revise the probability assigned to a belief as more evidence is discovered about it, whereas the objectivist will assign a constant probability that does not change as long as the probability of the evidence stays the same.

Bayesian probability is the well-established member of the subjective family of interpretations. Bayesian probability and its application were introduced in chapter 2. Despite being subjective, Bayesian probability often relies on the frequency approach to estimate the likelihood of events [45]. That is because the quantification of an agent's belief has to be rationalized in some way in order to be admissible and practical. Un-rational agents may believe in contradictory statements or assign high probability to impossible events or appeal to emotions or authority in their assessment of the truthiness of statements. Hence, the Bayesian probability calculus is not purely a subjective one, as the frequency counterpart is not a pure objective. An

intensively loyal Bayesian would defend his position by emphasizing that rational agents should base their beliefs about the truthiness of a statement on the relative frequency of times it turns out to be true. But the requirement of having enough data to make rational beliefs about their likelihood leads us back to square one, as the purpose of this thesis is decisions based on little evidence.

Terrance Fine in his magnum opus “Theories of Probability” gave another analysis of subjective probability based on the so-called comparative probability (CP) [54]. Comparative probability is a type of weak subjective interpretation of probability that rationalizes informal statements like A is at least as probable as B [54,p. 15]. This effort is by no means a new one because it was also proposed and defended by de Finetti [55], Savage [56] and Koopman [57]. However, for de Finetti, Savage, and Koopman, CP was the foundation from which the axioms of probability are derived [58]. Thereby, it serves as a more primitive entry point on which the modern theory of probability is built. On the other hand, Fine [54], Fishburn [59] and Keynes [60] proposed CP as an independent interpretation of probability, with its own axioms and calculus. Although CP received less interest from scientists, statisticians, and engineers, Fine highlights the following benefits of the CP framework [54,p. 15]:

1- CP results in a more pragmatic approach to random variables when the amount of information and data available are not enough to estimate the random variables quantatively.

2- There is more than one model to represent the probability in, as opposed to the strict one-model approach of the quantitative probability estimation.
3- As is the case with de Finetti approach, CP has the benefit of supporting the Kolmogorov axioms of probability.
4- CP naturally supports decision-making between other applications, as it describes probabilities in a qualitative way.

The justification for 1 comes from the fact that CP offers more relaxed claims than the strict numerical values that the other interpretations have to come up with. For example, suppose a coin is flipped 10 times, which results in 7 heads and 3 tails. A frequentist has to assign the value 0.7 to the probability of heads and deals with the critics who may find the drift in results from the ideal case of ½, as if the coin was tampered with or more “flips” should have been carried out. A comparativist, on the other hand, may describe the event as heads are more probable than tails [54,p. 16]. As a result, CP seems more suitable to the situations this thesis aims at solving than any method discussed so far. However, CP statements are qualitative in nature, which leaves the question of abstracting CP statements in a manner compatible with computers. A computer cannot directly deal with qualities like (is A > B?) unless A and B have numerical values. A computer should be equipped with an intermediate layer to transfer between the higher level qualitative description that CP offers and the quantative nature of computers. The ability of CP to describe the same event in more than one way gives us flexibility that many other methods can not. For example, we can describe the

result of the coin flip given previously as heads are at least as probable as tails, which would be closer to the frequency interpretation. Point 3 would enable us to define formulas to transfer between the space of comparative probability and frequency-based probability. The benefit of having such transformation is that an algorithm that deals with extracting knowledge online should be versatile. For instance, it should be able to present the knowledge collected so far in a fashion understandable by frequency-based algorithms. CP does not need to replace the current model of probability but rather to help represent knowledge when little data is known and to continue to do so until enough data is available. This is when CP gets its queue signal to pass the control to the frequency-based probability calculations.

This section discussed two approaches to the analysis of probability: objective and subjective. While the objective interpretation of probability is the de facto in science and even the subjective approach relies in its realisation on objective approach like the frequency interpretation, its performance is doubtful when little information is known about a situation. The comparative interpretation of probability seems appealing within the context of this thesis and will be discussed further in the next section. However, it should be noted that CP is a subjective probability approach and, therefore, it does not contradict the frequency interpretation of probability, because CP gives epistemological statements about reality that describe beliefs rather than a quantitative/qualitative representation of reality.

3.2 Axiomatic Comparative Probability

The task of analysing probability aims at formalizing everyday usage of probabilistic statements. In the context of comparative probability, there are usually binary statements that compare and contrast the likelihood of two events, like team A is more likely to win than team B, event C is at least as likely as D, or E is as likely as F. In everyday usage, a person may believe in statements that contradict one another or appear to be absurd in some situations. The task of axiomising probability is the task of finding the very fundamental statements of probability that look as intuitive as possible, so that they do not raise any objections or contradict each other. Thereafter, all other probabilistic assertions should be built upon these statements. For CP, the pursuit of axiomization should consider other important factors, such as the relationship to quantitative probability—in particular, the Kolmogorov axioms—the usefulness of the axioms to generate predictable results that can be observed empirically, and the applicability to decision making and DBN [54].

The notations for CP that will be used throughout this thesis follow those of both Fine and Fishburn [54, 59] but will also adhere to the notations of chapter 2. Let Ω be a sample space of finite outcomes labelled $\{\omega_1, \omega_2,\ldots, \omega_n\}$, where ω_n is the nth possible outcome, or subset of outcomes, from within the sample space Ω, we denote the comparative relationship ω_1 is at least as probable than ω_2 as[54]:

$$\omega_1 \succcurlyeq \omega_2 \tag{72}$$

the comparative relationship ω_1 is as probable as ω_2 as [54]:

$$\omega_1 \approx \omega_2 \tag{73}$$

and the comparative relationship ω_2 is not as probable as event ω_1 as [54]:

$$\omega_1 \succ \omega_2 \tag{74}$$

Fishburn [61] used a radically different notation to the intuitive notations presented in 72 to 74. He used real-valued function representations rather than inequalities, for example:

$$\omega_1 \succ \omega_2 \leftrightarrow \rho(\omega_1, \omega_2) > 0 \tag{75}$$

Since ρ is real-valued, it enables us to establish a bridge to the quantitative counterpart of probability, since under the assumption of simple linearity, 75 can be re-written as:

$$\omega_1 \succ \omega_2 \leftrightarrow P(\omega_1) - P(\omega_2) > 0 \tag{76}$$

However, what would be the assumptions of CP that can be set as axioms for their simplicity and intuitively? First, the axioms should not point out trivial facts or non-probabilistic assertions. In order to achieve that, possible events should have a probabilistic value such that the comparative probability of the

sample space—that is, all possible events—is more probable than the impossible events:

$$1)\ non-triviality: P(\Omega) \succ \phi \tag{77}$$

where ϕ is the empty set or the set of impossible to occur outcomes from the sample space Ω [54]. Consequently, an event like ω is not more certain or impossible than it is probable, or in CP terminology: ω is more probable than the impossible, which is the null or empty set [62]:

$$2)\ improbability\ of\ the\ impossibility: P(\omega) \succ \phi \tag{78}$$

Second, two comparative probability statements should not contradict each other. If ω_1 and ω_2 are both possible outcomes from the sample space Ω, then the two statements, ω_1 is at least as probable as ω_2 and ω_2 is at least as probable as ω_1, cannot be both true, that is [54],

$$3)\ Comparability: \omega_1 \succcurlyeq \omega_2 \vee \omega_2 \succcurlyeq \omega_1 \tag{79}$$

Third, another candidate for the axiom of CP comes from the property of transitivity in the mathematics of inequalities [63]:

$$4)\ transitivity: \omega_1 \succ \omega_2 \wedge \omega_2 \succ \omega_3 \supseteq \omega_1 \succ \omega_3 \tag{80}$$

However, the axiom of transitivity did not go unchallenged, as May [64], Tversky [65] and Fishburn [59] showed that in multidimensional events, cyclic patterns can arise, which would violate the implication of 80. Fishburn's [59]

example of a cyclic pattern assumes a hypothetical situation where an agent named Sue is supposed to meet a famous author named Mike. She has not met him before but has the following expectations about his attributes:

Height (*ht*): 6'.0" ≻ 6'.1" ≻ 6'.2",

Age (*ag*): 40 ≻ 50 ≻ 60,

Hair Colour (*hc*): brunette ≻ red ≻ blonde.

Based on these three attributes, Mike may be any of the following composites:

A = 6'.0" 60-year-old redhead;

B = 6'.1" 40-year-old blonde;

C = 6'.2" 50-year-old brunette

Sue would consider one of these composites more probable than the others if at least two of its attributes are more probable than the others. Hence A ≻ B, B ≻ C, but C ≻ A. As plausible as the objection to the axiom transitivity appears to be, it received less attention from other researchers, who continue to consider it as an axiom [58, 66, 67], albeit with caution [54]. It is worth noting that Sue's decision regarding the probability of how Mike will look is un-formalized and rather breaches the some of the basic rules of calculus. One central requirement of this thesis is for the interpretation and usage of probability to be admissible, that is, to comply with the theory of calculus.

More rationalised and more formalised analysis of the previous situation requires us to notice that the composite event of how Mike would

look is the joint probability of height, age and hair colour P(*ht, ag, hc*). It seems reasonable to assume that the three attributes are independent, hence:

$$P(ht, ag, hc) \approx P(ht)P(ag)P(hc) \tag{81}$$

If $P\left(ht = 6^{'}\ .0"\right) \succ P\left(ht = 6^{'}\ .1"\right)$, $P(ag = 40) \succ P(ht = 50)$ and $P(hc = \text{brunette}) \succ P(hc = red)$, then:

$$P\left(ht = 6^{'}\ .0", ag = 40, \text{hc} = \text{bruntte}\right) \succ P\left(ht = 6^{'}\ .1", ag = 50, \text{hc} = \text{red}\right) \tag{82}$$

But one cannot infer that B $\succ$ C because on one hand, we have:

$$P_B\left(ht = 6^{'}\ .1", ag = 40\right) \succ P_C\left(ht = 6^{'}\ .2", ag = 50\right) \tag{83}$$

and on the other:

$$P_B(hc = blonde) \prec P_C(hc = brunette) \tag{84}$$

So, a rational agent will be indifferent to the likelihood of B and C because 83 and 84 cannot mathematically be combined together to yield either B $\succ$ C or B $\prec$ C. Since the rational agent has no way to favour one over the other, he/she will assume they are equiprobable. A classical frequentist may approach the problem by applying the principles of counting: There are 3×3×3= 27 possible “looks” for Mike and there is only one way to count each

of A, B or C. Therefore, they are all equiprobable, which agrees with the results of personal subjective analysis. Consequently, we can intuitively set as an axiom the claim that if ω_1 is not more probable than ω_2 and ω_2 is not more probable than ω_1, then ω_1 is as probable as ω_2:

$$5)\ \omega_1 \not\succ \omega_2 \ \wedge\ \omega_2 \not\succ \omega_3 \rightarrow\ \omega_1 \approx \omega_3 \tag{85}$$

Based on the previous five axioms, one can easily conclude the following consequences [54]:

$$\omega\ \subseteq\ \Omega\ \rightarrow P(\Omega) \succ P(\omega) \tag{86}$$

Equation 86 states that if an event like ω is a subset from the sample space Ω, then it will make perfect sense to assume that the probability of every possible event or set of events is higher than the probability of a single event or set of events. Consequently, 86 can be generalized to:

$$\omega_1 \subseteq \omega_2\ \rightarrow\ \omega_2 \succcurlyeq \omega_1 \tag{87}$$

Since CP is a binary and linear relationship, if one event is more probable than another, the negation of that event is less probable than the negation of the other [68]:

$$\omega_1 \succcurlyeq \omega_2 \ \rightarrow \ \omega_1^c \preccurlyeq \omega_2^c \tag{88}$$

If we hold true the definitions of joint and union of events from chapter 2, then if the joint of two events is null, and [54]:

$$\omega_1 \succcurlyeq \omega_2 \wedge \ \omega_3 \succcurlyeq \omega_4 \wedge \omega_1 \cup \omega_3 = \ \phi \rightarrow \ \omega_1 \cup \omega_3 \succcurlyeq \omega_2 \cup \omega_4 \tag{89}$$

Further conclusions are possible and were discussed in [63]. Generally speaking, the axiomization of CP starts by selecting either $\omega_1 \succcurlyeq \omega_2$ as the very basic foundation of CP or $\omega_1 \succ \omega_2$ [59]. However, we have followed a mixed approach to that, where we started with the latter and then used the former to strengthen the weaker assumptions of CP that often generate controversy. As CP is still an infant concept as compared to other well-established probability interpretations, the axioms discussed so far are by no mean presented as complete or unchallengeable. They can be regarded as a guide to formalize everyday pseudo-rational statements regarding probabilities of events, but more importantly, they present how the quantitative probability can be deducted from the qualitative probability. The latter is the subject of the next section.

3.2.1 Compatibility with quantative probability

At first glance, CP seems to be compatible with quantative probability and its axioms, put forward by Kolmogorov. It seems plausible that every probability axiom of the Kolmogorov probability (KP) is compatible with the five

CP axioms developed earlier. Unfortunately, this is not the case because CP is defined on lexicographical order sample space, whereas KP is defined over numerical order sample space. A numerical set like [0,1] can have an infinite number of subsets but a lexicographical set like [a,b,c,d] can have only limited sets. Therefore, it is perfectly possible to think of a situation where the infinite becomes contradictorily finite [54,p. 18]. Consider, for the sake of the argument, that CP is defined over a topological space (R^n) like Ω = [0,1], having subsets that adhere to the Borel field of Ω. This choice is not coincidental but rather represents the best candidate for space that could become the bridge between the quantities and qualities of probabilistic statements. In addition, let λ(A) be the Lebesgue measure of subset A defined in Ω. Furthermore, let φ(A) be another measure of A dominated by λ(A), such as the triangular density. If the comparative relationship $A \succcurlyeq B$ is defined as:

$$A \succcurlyeq B \equiv \lambda(A) > \lambda(B) \vee \lambda(A) = \lambda(B) \wedge \varphi(A) \geq \varphi(B) \qquad (90)$$

then the definition satisfies all the axioms of CP presented in the previous section and thereby represents a compatible one-to-one relationship between CP and KP. This is evident because if A is the set (1-*x*,1) and B is the set (0,*x*) where $0 < x < 1$, then $\lambda(A) = \lambda(B) = x$ and $\varphi(A) = x(2-x) \geq \varphi(B) = x^2$ [54,p. 18] satisfies equations 77,78,79,80,85 and 90. However, this results in a contradiction, because *x* can be any value out of infinitely many values in the range [0, 1], whereas the lexicographical order space has only finite sets. Thereby it would contradict the existence of a one-to-one relationship. To

alleviate the possibility of contradiction, a new axiom should be added to the inventory of CP axioms, so that it guarantees the compatibility between the collation of lexicographical and numerical.

One way to define such an axiom is to imagine a topological space like R having a collection of subsets τ such that (R, τ) has a countable base or (R, τ) has an accountable order dense set [69, 70]. However, the axiom of countable base does not admit, or guarantee, a unique probability value for an event; rather, we can define as many functions as we like that satisfy the axioms of CP and are compatible with quantitative probability. Fortunately, if CP is to be compatible with KP, then we can simply choose the probability function that satisfies P(Ω) = 1 [54,p. 19].

In order for CP to be fully compatible with KP, it should be compatible with the finite additivity and in turn with the third axiom of KP, that is to say, if a comparative relationship satisfies the six axioms of CP, then there should exist a function like G of two variables, such that [54,p. 22]

$$A \cap B = \phi \;\Rightarrow P(A \cap B) = G(P(A), P(B)) \tag{91}$$

and it should also be symmetric, strictly increasing, and associative. However, Kraft [71] proved that such a function cannot exist, challenging the previous six axioms. An example of situation where 91 is not satisfied is evident when Ω={a,b,c,d,e} and τ is all the subsets of the following order [54,p. 22]:

$$\begin{aligned}\phi \prec a \prec b \prec c \prec ab \prec ac \prec d \prec ad \prec bc \prec e \prec abc \prec bd \prec cd \\ \prec ae \prec abd \prec be \prec acd \prec ce \prec bcd \prec abe \prec ace \\ \prec de \prec abcd \prec ade \prec bde \prec abde \prec bcde \prec abcde \\ = \Omega\end{aligned} \tag{92}$$

Equation 92 satisfies the six axioms of CP; however, there is no such G to satisfy 91. To see that, let P(a) = A, P(b) = B, P(c) = C, P(d) = D and P9e) = E; then from 92, it follows A + C < D, A + D < B + C and C + D < A + E, which can be simplified to: A + C + D < B + E, hence acd ≺ be, which contradicts 92 [54,p. 22]. One way around this contradiction is to introduce a condition that CP should satisfy in order to become fully compatible with finite additivity. Luce [72] introduced such sufficient a condition. Although others have proposed different approaches [56], Luce's seems more appealing [54,p. 25]. Luce used results from the theory of extensive measurement to prove his theorem by proposing the criterion [72]:

$$\begin{aligned}(A \cap B = \phi \wedge \quad A \succcurlyeq B \wedge \quad B \succcurlyeq D) \Rightarrow \left(\exists C', D', E\right)(C' \cap D' = \\ \phi \wedge E \approx A \cup B, C, D' \approx D, E \supseteq C' \cup D')\end{aligned} \tag{93}$$

Equation 93 is appealing because it does not require the sample space to be strictly infinite, as Savage and Kraft's proposals do [54,p. 25]. Since CP is now compatible to an acceptable extent with finite additivity, the next step is to look at countable additivity. If CP is compatible with countable additivity, then it will be compatible with KP.

In the measure and probability theory, countable additivity becomes equal to finite additivity if the following condition holds true [73]:

$$A_1 \supset A_2 \supset A_3 \supset \cdots \supset A_n \downarrow \phi \Rightarrow \lim_{i \to \infty} P(A_i) = 0 \quad (94)$$

The condition given in 94 is also known as the continuity condition, which was adopted by Kolmogorov as an axiom for the KP [74]. Hence, 93 can be relaxed to accommodate for continuity condition:

$$(\forall\{A_i\})(A_i \downarrow \phi \Rightarrow \bigcap_{i=1}^{\infty}\{B: \phi \prec A \preccurlyeq A_i\} = \phi) \quad (95)$$

However, having CP compatible with countable additivity is not always desirable. De Finetti [54, 74] argued against such an approach, as it would result in absurd situations, for example, the experiment of picking a positive integer number at random. In this experiment, the sample space Ω can be thought of as being Ω = {1,2,3,….}, which is clearly an infinite space. The power set of all positive integers is also infinite and the probability of each element within the power set is 0. Hence the probability of Ω, which is a member of the power set as well, is 0! But KP requires P(Ω) = 1. In addition, how are we to rationally justify that the probability of picking the number 1 is equal to the probability of picking a number from within the range $(1,10^9) = 0$? Shouldn't the latter weigh more than the former? Bertrand Russell viewed a set that has itself as a member as paradoxical [75] and required that no set be a member of itself. However, the strongest and most traditional argument against a subjective approach to probability comes from Ellsberg's analysis of

Savage's axioms [76], where he showed two examples of subjective judgements leading to absurd results. Ellsberg's first example is a traditional betting situation, where a gambler is asked to bet on the label of a randomly selected ball from an urn of 100 balls. The urn contains 25 balls labelled R1, 25 labelled B1, and the remaining 50 balls are labelled either R2 or B2, but their proportion is known to the gambler. The betting versus winning options are:

r1: wins $1000 if the chosen ball is R1, but nothing otherwise.

b1: wins $1000 if the chosen ball is B1, but nothing otherwise.

r2: wins $1000 if the chosen ball is R2, but nothing otherwise.

b2: wins $1000 if the chosen ball is B2, but nothing otherwise.

Presumably, the gambler would think the odds for r1 and b1 are the same and since r2 and b2 proportions are known, the gambler would be indifferent to any of them. In addition, the gambler would prefer r1 over r2 and b1 over b2 because, once again, the quantities of r1 and b1 are known, whereas those of r2 and b2 are unknown. Using CP terminology, we can specify the gambler's preferences as follows:

$$r1 \approx b1, r2 \approx b2, r1 \succcurlyeq r2, b1 \succcurlyeq b2 \tag{96}$$

Now, let us assume the game is updated with two further compound bets, as follows:

c1: wins $1000 if the chosen ball is R1 or B1, but nothing otherwise.

c2: wins $1000 if the chosen ball is R2 or B2, but nothing otherwise.

The new bets would seem less ambiguous in the eyes of the gambler because the R1+B1 ball count is known, as is R2 + B2. In fact, they both add up to 50. Once more, the gambler would give the same odds for each of them.

$$c1 \approx c2 \rightarrow \mathrm{r1} \cup b1 \approx r2 \cup b2 \tag{97}$$

But if the additive axioms are to hold true and since $\mathrm{r1} \approx \mathrm{b1}, \mathrm{r2} \approx \mathrm{b2}$, then $\mathrm{r1} \approx \mathrm{r2}, \mathrm{b1} \approx \mathrm{b2}$, which clearly contradicts 96 [61, 76, 77]. One way to resolve this contradiction is to involve rational subjective judgements. If a rational agent prefers $r1 \succcurlyeq r2$, then it implies that the number of balls labelled R1 is higher than or equal to R2; hence, the number of B2 balls is higher than or equal to B1, therefore $r1 \succcurlyeq r2 \text{ and } b1 \succcurlyeq b2$ cannot be both true at the same time because $r1 \succcurlyeq r2$ implies $b1 \preccurlyeq b2$. Ellsberg's analysis is just another example of why drawing knowledge from ignorance leads to contradictions. That is because Ellsberg presumed that because we do not know the proportion of R2 and B2, we should be indifferent to their probabilities. It is similar to the Laplacian approach to probability we discussed earlier.

A more rational standpoint to the Ellsberg example is:

$$r1 \approx b1 \wedge r2 \approx b2 \rightarrow \ r1 \succcurlyeq r2 \vee b1 \succcurlyeq b2 \tag{98}$$

Now our gambler can either prefer that $r1 \succcurlyeq r2$ or $b1 \succcurlyeq b2$ but not both. In turn, the contradictions that resulted from 97 cease to hold. Fishburn's approach is to add another axiom to CP that requires [61]:

$$A \cap B = \phi \rightarrow \rho(A \cup B, C) + \rho(\phi, C) = \rho(A, C) + \rho(B, C) \quad (99)$$

and if ρ is normalized against Ω:

$$\rho(\Omega, \phi) = 1 \quad (100)$$

then Fishburn proved [61] that if A_1, A_2…,A_n are disjoint pairwise events and similarly forB_1, B_2,…,B_n, then:

$$\rho\left(\bigcup_{i=1}^{n} A_i, \bigcup_{j-1}^{m} B_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \rho(A_i, B_j) - (m-1)\sum_{i=1}^{n} \rho(A_i, \phi) + (n - 1)\sum_{j=1}^{m} \rho(B_j, \phi) \quad (101)$$

The advantage of 101 is that it only needs values for ρ to be specified at the most elementary pairs of events, that is, A and B, but Fishburn did not provide a systematic way of estimating the values he chose for ρ apart from an ad hoc table with values already there. It is assumed that the values can be chosen arbitrarily but within the constraints of not violating his axioms of CP.

In conclusion, CP is still an infant approach to the analysis of probability. Its main playground is philosophical and logical formalisation of its axioms. This section has presented a brief bridge that transposes qualitative statements to quantitative statements. The debate about the axioms of CP

and its relation to KP is far from conclusive and there are many active researchers formalising and criticising the work already done for CP. However, the main objective of this thesis is not the minor dilemmas that always exist in any formation of axioms in existence. There are as many paradoxes for KP as there are for CP and the important lesson to learn, in the context of science, is the usefulness of a method in answering a scientific inquiry.

3.2.2 Conditional comparative probability

So far, we have only been concerned with elementary probability representation and relationship. This section expands on the axioms of CP developed in the previous sections and the analysis and limitations of CP as compared to KP.

The derivation of conditional comparative probability (CCP) follows either a ternary or a quaternary approach. In a ternary approach, CCP is assumed to be a ternary relationship over the space $\mathfrak{F} \times \mathfrak{F} \times \mathfrak{G}$, where $\mathfrak{F}$ is a field of events and $\mathfrak{G}$ is a set defined over $\mathfrak{F}$ [54,p. 28]. In other words, CCP defines a comparative relationship between two variables conditioned over a third, read as given by a third. If $A, C \in \mathfrak{F}$ and $B \in \mathfrak{F}$, then a ternary CCP relationship is defined as [54,p. 28]:

$$(\forall A, C \in \mathfrak{F})\ A|B \succcurlyeq C|B \tag{102}$$

As is the case with the theory of KP in chapter 2, CCP also satisfies the axioms of CP, as well as [54,p. 28]:

$$(\forall B, C \in \mathfrak{F})\ B \supseteq C \Rightarrow (A|C \succcurlyeq D|C \Leftrightarrow A \cap C|B \succcurlyeq D \cap C|B) \quad (103)$$

The importance of TCCP comes from the fact that it can be used to calculate posterior probability given some evidence, albeit in a weaker form. This is important in situations where a decision must be made in the light of some evidence or for estimating the probability of a variable recursively. The latter is of higher interest for this thesis, as its intention is to develop a theoretical background that deals with knowledge representation under lack of data and/or ambiguity.

For the quaternary CCP (QCCP) approach, the axioms are direct counterparts of the ones developed earlier. In his magnum opus paper, Luce introduced seven axioms for QCCP [78]. First, if Ω is the sample space of events where $\mathfrak{G}$ is a subset of it, then:

$$L1)\ X|X \succcurlyeq A|B \wedge X|X \approx A|A \quad (104)$$

$$L2)\ A|B \approx A \cap A|B \quad (105)$$

$$L3) A \cap B = A' \cap B' = \phi\colon A|C \succcurlyeq A'|B' \wedge B|C \succcurlyeq B'|C' \rightarrow A \cup B|C \succcurlyeq A' \cup B'|C \quad (106)$$

$$L4)\ A \subset B \subset C \wedge A' \subset B' \subset C' : (A|B \succcurlyeq A'|B' \wedge B|C \succcurlyeq B'|C') \vee (A|B \succcurlyeq B'|C' \wedge B|C \succcurlyeq A'|B') \rightarrow A|C \succcurlyeq A'|C \tag{107}$$

The beauty of QCCP comes from the fact that it can be used to derive a weak form of Bayes' theorem [54]. Assume that A,C is in field $\mathfrak{F}$ and $B, D: \mathfrak{G} \subseteq \mathfrak{F}$ but does not include the null set, and that $\succcurlyeq$ satisfies L1-L4. Then there exists P agreeing with CP axioms and two real-valued functions F and G that if $A|\Omega \succcurlyeq \phi|\Omega$, then [54]:

$$P(B|A) = F\left(G\left(P(A|B), P(B|\Omega)\right), P(A|\Omega)\right) \tag{108}$$

Equation 108 will enable us to infer the characteristics of $B|A$ from $A|B$,$B|\Omega$ and $A|\Omega$. Furthermore, Luce's axioms will enable us to make QCCP compatible with KP because axioms L1-L4 can be used to show that there exists a P agreeing with L1 to L4, such that [78]:

$$P(A \cap B|C) = P(A|B \cap C)P(B|C) \tag{109}$$

Equation 109 is the CP equivalent of the product rule conditioned on C. However, the theory of comparative probability would not be complete without the notion of independence. Since CP formally represents a relationship between events in a qualitative fashion, the notion of independence would only make sense if it was event-wise rather than experiment-wise, as is the

case with KP [54,p 33]. The difference between the two is important in the light of subjective probability. Independent experiments refer to experiment outcomes being statistically unrelated to each other, which is justified by the combinatorial calculus, whereas independent events are those where the occurrence of one does not change our expectations about the occurrence of the other, which is justified by axiomisation. If A and B are both events from the sample space Ω, then the simplest axiom that can be drawn from the independence of A and B is that if A is unrelated to B, then the same goes for B and A [54,p 33]:

$$F1)\ A \perp B \leftrightarrow B \perp A \tag{110}$$

where ⊥ exemplifies independence, or unrelatedness, between two events. Furthermore, an event is a subset of Ω, so the occurrence of one is unrelated to the other, as it does not change one's belief about the likelihood of the other. Using the same justification, we can intuitively see the validity of the following axioms [54,p 33]:

$$F2)\ A \perp \Omega \tag{111}$$

$$F3)\ A \perp \mathrm{B} \rightarrow A \perp \mathrm{B}^{\mathrm{c}} \tag{112}$$

$$F4)\ B \cap C = \phi, A \perp C \rightarrow A \perp (B \cup C) \tag{113}$$

Further intuitive axioms are also possible, and a weaker experiment-wise independence could also be driven, but they are not essential to the purpose of this thesis. For convenience, this section will conclude with a box of all the axioms of CP, which will be easier to refer to later on.

Axioms of CP (114)

$C1)\ P(\Omega) \succ \phi$ ... (114.1)

$C2)\ P(\omega) \succ \phi$ ... (114.2)

$C3) \omega_1 \succcurlyeq \omega_2 \vee \omega_2 \succcurlyeq \omega_1$ ... (114.3)

$C4)\ \omega_1 \succ \omega_2 \wedge \omega_2 \succ \omega_3 \supseteq \omega_1 \succ \omega_3$ ... (114.4)

$C5)\ \omega_1 \not\succ \omega_2 \wedge \omega_2 \not\succ \omega_3 \rightarrow \omega_1 \approx \omega_3$ ... (114.5)

$C6)\ A \succcurlyeq B \equiv \lambda(A) > \lambda(B) \vee \lambda(A) = \lambda(B) \wedge \varphi(A) \geq \varphi(B)$ ... (114.6)

$C7)\ (\forall\{A_i\})(A_i \downarrow \phi \Rightarrow \bigcap_{i=1}^{\infty}\{B: \phi \prec A \preccurlyeq A_i\} = \phi)$ ... (114.7)

$C8)\ A \cap B = \phi \rightarrow \rho(A \cup B, C) + \rho(\phi, C) = \rho(A, C) + \rho(B, C)$ ... (114.8)

$C9)\ X|X \succcurlyeq A|B \wedge X|X \approx A|A$ ... (114.9)

$C10)\ A|B \approx A \cap A|B$ ... (114.10)

$C11)\ A \cap B = A' \cap B' = \phi: A|C \succcurlyeq A'|B' \wedge B|C \succcurlyeq B'|C' \rightarrow A \cup B|C \succcurlyeq A' \cup B'|C$ ... (114.11)

$C12)\ A \subset B \subset C \wedge A' \subset B' \subset C' : (A|B \succcurlyeq A'|B' \wedge B|C \succcurlyeq B'|C') \vee (A|B \succcurlyeq B'|C' \wedge B|C \succcurlyeq A'|B') \rightarrow A|C \succcurlyeq A'|C$ ... (114.12)

$C13)\ A \perp B \leftrightarrow B \perp A$ ... (114.13)

$C14)\ A \perp \Omega$ ... (114.14)

$C15)\ A \perp \mathrm{B} \rightarrow A \perp \mathrm{B^c}$ ... (114.15)

$C16)\ B \cap C = \phi, A \perp C \rightarrow A \perp (B \cup C)$ ... (114.16)

3.2.3 Comparative probability: Decision-making prospective

So far, we have only been concerned with CP from mathematical, logical and philosophical standpoints. Philosophy is an arena of debate, whereas science is one of research, analysis of empirical observations and application of theories to yield useful products. Therefore, it will be of value to know whether CP has managed to escape the “ballrooms” of philosophical debates and mathematical theorising into the realm of practical implementation. In section 3.1, we looked at the criteria of a good interpretation of probability, and CP should not be an exception to them. The emphasis of a good probability theory should be on its applicability as a framework is simply useless to science if it cannot be utilized in any way.

There are many frameworks in the literature regarding the formalisation of CP axioms and sometimes for developing some guidelines as to how it would be used for decision-making and inference. As much as these frameworks are packed with long mathematical formalizations, they are short on comparison to their quantative probability counterpart. Without justification for preferring CP over the de facto interpretation of probability in science, which withstood the test of time and was there during all of our scientific endeavours, why would anyone choose CP? In this section, we will be looking at some of the interesting frameworks in CP, their applications, and limitations.

Peter Fishburn is one of the well-known names in decision-making and the axiomisation of CP as an independent interpretation of probability [59, 61, 70, 79, 80]. The previous section has already presented some of his contributions to CP. The main framework of Fishburn was answering de

Finnite's question on the existence of an order-wise relationship that is sufficient for the existence of order-preserving probability measure [80]. The answer to the question was to introduce more basic but constraining limits on CP in order to preserve the order-wise nature of CP while solving any paradoxical objection to it [61]. Fishburn showed an example of how his version of CP can solve Ellsberg paradoxical examples of subjective probability through the introduction of the skew-symmetric function ρ [61]. As mentioned earlier, there was no discussion on how the performance of CP compares to KP. Such a comparison, if it was in favour of CP, would prove the case of CP as an interesting alternative to KP that scientists should start to use, rather than shelve it along with the other mathematical constructs with internal inconsistencies. In addition, Fishburn did not provide us with a clear algorithm that explains in a step-by-step fashion how to use his framework to solve problems in decision-making beyond some isolated examples and more mathematical constraints.

Terrence Fine is another example of the independent interpretation approach to CP [54, 69, 81]. However, Fine's approach seems relaxed and less constrained than Fishburn's or Luce's, as examples [54]. Fine developed five axioms that characterize a rational decision-making process and expectations, all in terms of comparative-like inequalities [54]. However, all of these axioms were incomplete in showing a single example of how to use them to come up with a rational decision within any context, not even a game of chance. Fine admitted the existence of the problem of measuring subjective probability or preferences and even the psychological factors

leading to constraining the process of extracting those from decision-makers [54,p. 233]. Along with Walley, the framework seemed to be shifted toward establishing a unified framework of upper and lower probability, or imprecise probability [82, 83]. Imprecise probability refers to classes of mathematical models that deal with uncertainty and the availability of partial information [82]. Walley tried to unify many of the proposed models of imprecise probabilities, including CP itself, using the subjective framework of Bayesian networks [82]. When there is not enough data to infer a descriptive probability distribution, then upper and lower bounds are defined and the gap between the upper and lower limit is supposed to decrease as data is gathered, until the gap is closed and what was imprecise is now precise [84]. Walley's focus was on the mathematical level of generality that will be needed to achieve such unification [82]. His framework was further applied to graphical models' [85] belief functions [86], among others. Walley's framework seems interesting within the context of this thesis; however, it still has the drawback we mentioned earlier, namely, no step-by-step algorithm was specified that could aid a decision-maker in making the decision and no comparison with KP was attempted.

The third example of CP framework is Andrea Capotorti, who proposed some interesting CP axioms that can be described algorithmically and implemented on computers [66, 87]. For Capotorti, the reason a decision-maker would prefer CP over the other interpretations of probability is that they are not compatible with the psychology of human preferences and sometimes even violate the axioms of KP [66], not to mention the "where are all the numbers coming from?" argument [66]. However, the same argument goes

against CP because if it was valid for us to wonder where the frequentists take their quantities from, then by the same logic it is valid to wonder where the comparativists get their qualities from. The Capotorti algorithm works by constructing qualities and constraints that describe a situation [66]. For example, if heads was more probable than tails, then we can describe it as $\phi \preccurlyeq P(tails) \preccurlyeq P(heads) \preccurlyeq \Omega$. The algorithm continues to use constraints and new information to update the qualities until a decision, or inference, is possible [66], although it was not clearly specified how such an update is made. In addition, it seems like interfacing with such DSS, if ever implemented, will be extremely difficult because it does not provide a quantified output, nor does it have an objective procedure for converting sensor measures into qualities. Finally, it relies heavily on expert knowledge to come up with the qualities that represent a situation.

In conclusion, this section tried to summarize the most important frameworks of CP as a tool for inference and decision-making. We have seen how CP frameworks were shifted when faced with different challenges in regard to measuring personal preferences, restricting their flexibility by the addition of more axioms and pitching for unification with other frequency-based probabilistic theories. We have also seen that the closest framework to the objectives of this thesis was that of Peter Walley, which aimed at unifying CP with the upper and lower probability model of imprecise probability. What steps are required in order to improve Walley's axioms? Will it be possible to propose an algorithm that automates the process of inference in a way that proves more beneficial than the current conventional methods? If so, how do

they compare? We will explore the answers to these questions in the next section.

3.3 Proposing a new approach to CP

Up to this point, we have discussed various concepts in probability theory, from the basics to some of its advanced concepts and results. Moreover, we took a step back to understand where probability comes from and how we should interpret it. Our aim was to search for a better representation of knowledge in situations when little information is available. From within the discussion, a candidate emerged that seemed to be up to the task of answering the question, which we referred to as CP. But CP is still far from complete, both as a theory and as a practice. It is more a philosophical concept than a scientific method. However, this is not to say that CP literature is sparse but rather to emphasize the fact that, apart from making very simple decisions, it has not been used for much. Therefore, if we want to keep thinking that CP is the promising answer to our questions, then what modifications, if any, are necessary to make it work? The answer to the question requires us to specify the concepts of probability theory that we would like to keep and what sort of utility we would like CP to work with.

3.3.1 Requirements, assumptions and aims

First, CP should not replace KP or provide a standalone interpretation of probability that works completely parallel to KP or even contradicts KP. The reason behind this requirement is that KP is well established and has been used in almost every scientific discipline. It would not make sense to throw away a very successful theory such as KP, given how widely it is used and how successful it is as a scientific tool. In addition, we would like to utilize the greatest results of the modern theory of probability without worrying about them not being compatible with CP's axioms. Results such as the central limit theory and the strong law of large numbers are so much appealing to any researcher that it would be desirable to have them in scientific endeavours. Hence, the first requirement simply states that KP and its results are true a priori.

Second, we assume that probability exists as an objective property of things in reality and that it can be determined through experiments and empirical observations. This assumption is backed by both scientific and philosophical justifications. The scientific justification is based on the fact that empirical observations are the essence of the scientific method. Therefore, the quantification of the properties of an observable phenomenon in reality should reflect the objectivity of the property itself. The philosophical justification comes from our conceptualizing of probability as being the guide to life. If probability is to guide life, then it should not be mere analytical statements, for analytical statements do not describe reality nor provide a guide for it. Probability should be inferred from reality in order for it to become the guide

for life. In other words, probability should be estimated from data posterior, not from the space of possibilities, as is the case with the Laplacian probability. If no data is available, then probability still exists but it is unknown.

Third, when the number of experiments is high enough, then the probability of an outcome should be quantified using the relative frequency interpretation of probability. This requirement is important because without a way to objectively quantify a probability it will be useless, for an unknown quantity that can never be measured is worth nothing in the context of science. The third requirement is also essential because we want to use CP to write algorithms that are integrable with others. Hence it should "speak" the same language that the state-of-the-art algorithms speak. What is essential to our CP theory is to be able to give a hand to probability estimations in an online scenario, up to the point where the algorithms are ready enough to do it on their own.

Finally, we accept the relative frequency position that if an experiment is repeated often enough, then the probability of an outcome approaches its relative frequency of occurrence. Consequently, CP becomes a background tool, while KP is the foreground methodology for estimating probability. This assumption sets the new approach apart from that of Savage, Fine, Fishburn, and others because CP was either considered a standalone interpretation for some of them or an approach to the frequency interpretation for the others.

Since the process of scientific theorizing revolves around usefulness and utility, the new approach should prove useful in terms of its results when compared to other similar tools in inference and decision-making. But proving

the usefulness of a tool in making decisions or inference is not by itself enough to make a good case, because making good decisions under uncertainty only works on average. A process of decision-making may sometimes lead to a very bad decision but if it was compensated by other decisions that will overall produce positive utility, then we would still consider the process of making the decision a valid one. Instead, the benchmark should prove to be useful as a predictive tool; after all, scientific theories should be able to predict some measurable observations upon which the theory can be validated. Since the central aim of the thesis is to produce a theory to better represent knowledge in terms of probabilistic statements, then in order to prove the case of this thesis, the resulting theory should be better in estimating and representing probability than all the known competitors.

In conclusion, in order to propose a good approach to CP that proves useful as a scientific tool, some assumptions and compromises need to be made in order to ensure that the end results are on target. We require that CP should be fully compatible with KP, that CP is a way of representing probabilistic knowledge about reality where probability exists objectively, and that the probability of an event is its relative frequency of occurrence only when the number of experiments over the space where the event belongs is high enough. The proposed theory should prove useful in terms of probabilistic predictivity.

3.3.2 Axioms and theories of the proposed approach

The previous section sets the requirements of a good solution to have in this thesis. It is clear that the requirements favour the relative frequency

interpretation but use the comparative probability interpretation as a way of describing knowledge that may be subjective. How can we incorporate all of these ingredients to produce an online, real-time, dynamic probability estimator?

First of all, we presume the probability of an outcome or a set of outcomes defined over a sample space to exist and to have a single value like $P(\omega) \in [0,1]$. The exact value of $P(\omega)$ may be unknown, but we can subjectively suspect that it lies within a region of doubt like ε, such that the subjective upper and lower limits of $P(\omega)$ within which $P(\omega)$ should exist are:

$$\overline{P}(\omega) \approx P(\omega) + \frac{\varepsilon}{2} \quad \ldots \quad (115.1)$$

$$\underline{P}(\omega) \approx P(\omega) - \frac{\varepsilon}{2} \quad \ldots \quad (115.2) \qquad (115)$$

where $\overline{P}(\omega)$ is the subjective probabilistic upper bound of $P(\omega)$, $\underline{P}(\omega)$ is the corresponding lower bound and $\approx$ denotes an "as probable as" CP relationship. The reason for using comparative relationship here is to allow for unsymmetrical upper and lower bounds; otherwise, $P(\omega)$ will be simply the mathematical average of $\overline{P}(\omega)$ and $\underline{P}(\omega)$. In addition, we are using the term subjective in an epistemological fashion to convey the fact that the upper and lower bounds are not "real" probabilities but rather a belief about probability. From equation 114 we can infer that:

$$\underline{P}(\omega) \preccurlyeq P(\omega) \preccurlyeq \overline{P}(\omega) \qquad (116)$$

Since the estimation of $P(\omega)$ should be available in real-time while the data, or the experiment's outcomes, roll in, we will define $\tilde{P}_n(\omega)$ to be the approximate value of $P(\omega)$ at experiment number (n), which should eventually converge to $P(\omega)$ as the number of experiments increases indefinitely. As is the case with $P(\omega)$, $\tilde{P}_n(\omega)$ should also be within the upper and lower bound of the region of doubt:

$$\tilde{P}_n(\omega) \succcurlyeq \underline{P}(\omega) \cap \tilde{P}_n(\omega) \preccurlyeq \overline{P}(\omega) \tag{117}$$

and:

$$\lim_{n\to\infty} \left(\tilde{P}_n(\omega), \underline{P}(\omega), \overline{P}(\omega)\right) = P(\omega) \tag{118}$$

In chapter 2, we saw how an estimate of probability can be achieved using Markov and Chebyshev inequalities and how they can be used to prove that the average of a random variable converges to its expected value as the number of experiments increases in the well-known strong law of large numbers. But, these two inequalities, although very useful, are not enough to provide us with a powerful way to update our probability estimates as new experiments become available. For that end, Chernoff bounds provide a better and more restricted estimate [18, 88, 89]. For Bernoulli variables, the Chernoff bound is given by [90]:

$$P_n\left(\left|P_n(\omega) - \tilde{P}_n(\omega)\right| > \frac{\varepsilon}{2}\right) < 2e^{\frac{-n\varepsilon^2}{2}} \tag{119}$$

Equation 114 defines the upper bound, or tail, of $P(\omega)$. Since our goal is to better estimate $P(\omega)$ given $\tilde{P}_n(\omega)$, we want to decrease the size of the uncertainty region as more data starts rolling in. Therefore:

$$P_{n-1}\left(\left|P_{n-1}(\omega)-\tilde{P}_{n-1}(\omega)\right|>\frac{\varepsilon_{n-1}}{2}\right)\succcurlyeq P_n\left(\left|P_n(\omega)-\tilde{P}_n(\omega)\right|>\frac{\varepsilon_n}{2}\right) \quad (120)$$

Using equation 87, equation 120 implies that

$$\left(\left|P_{n-1}(\omega)-\tilde{P}_{n-1}(\omega)\right|>\frac{\varepsilon_{n-1}}{2}\right)\succcurlyeq\left(\left|P_n(\omega)-\tilde{P}_n(\omega)\right|>\frac{\varepsilon_n}{2}\right) \quad (121)$$

Since:

$$\frac{\varepsilon_{n-1}}{2}\geq\frac{\varepsilon_n}{2} \quad (122)$$

then:

$$\left|P_{n-1}(\omega)-\tilde{P}_{n-1}(\omega)\right|\succcurlyeq\left|P_n(\omega)-\tilde{P}_n(\omega)\right| \quad (123)$$

and:

$$2e^{\frac{-(n-1)\varepsilon_{n-1}^2}{2}}\geq 2e^{\frac{-n\varepsilon_n^2}{2}} \quad (124)$$

Since the lower limit of $\left|P_n(\omega)-\tilde{P}_n(\omega)\right|$ is $\frac{\varepsilon}{2}$, using equations 115 and 124, we can solve for $\underline{P_n}(\omega)$:

$$\underline{P_n}(\omega) \approx P_n(\omega) - \sqrt{\frac{n}{n+1}\left(P_{n-1}(\omega) - \underline{P_{n-1}}(\omega)\right)^2} \quad (125)$$

Similarly:

$$\overline{P_n}(\omega) \approx P_n(\omega) + \sqrt{\frac{n}{n+1}\left(P_{n-1}(\omega) - \overline{P_{n-1}}(\omega)\right)^2} \quad (126)$$

where

$$P_n(\omega) = \frac{1}{n}\sum_{1}^{n} \omega_n \quad (127)$$

is the mathematical average of ω. Since ω is a binary variable, its average is also its probability when the number of experiments is approaches infinity. Notice that 125 and 126 do not give the upper and lower tail of Chernoff bounds but rather a mirrored upper tail and a mathematically mirrored tail. In order to show how the above two equations can be used to estimate epistemologically stricter bounds for the upper and lower probability, we will use a coin flipping example. Let us assume a decision-maker is asked to predict the upper and lower probability bounds of heads in an experiment of coin flipping. The coin may be biased to heads or tails, but that is unknown to the decision-maker. Since the decision-maker is indifferent to whether the coin is biased or unbiased, he would assume an initial probability of ½ for the probability of heads, making the upper and lower probability as probable as

heads itself. The decision-maker is hoping to update the bounds as more experiments are performed. She/he decided to use equations 125 and 126.

Let us imagine the experiment of tossing a coin was repeated 1,000 times. The decision-maker would like to calculate the accuracy of equations 125 and 126. Using Matlab, we can generate pseudo-coin flips with uniform probability distribution. Figure 16 shows a plot of the upper bound (in green), the lower bound (in black) and the average probability calculated over the course of the 1,000 experiments.

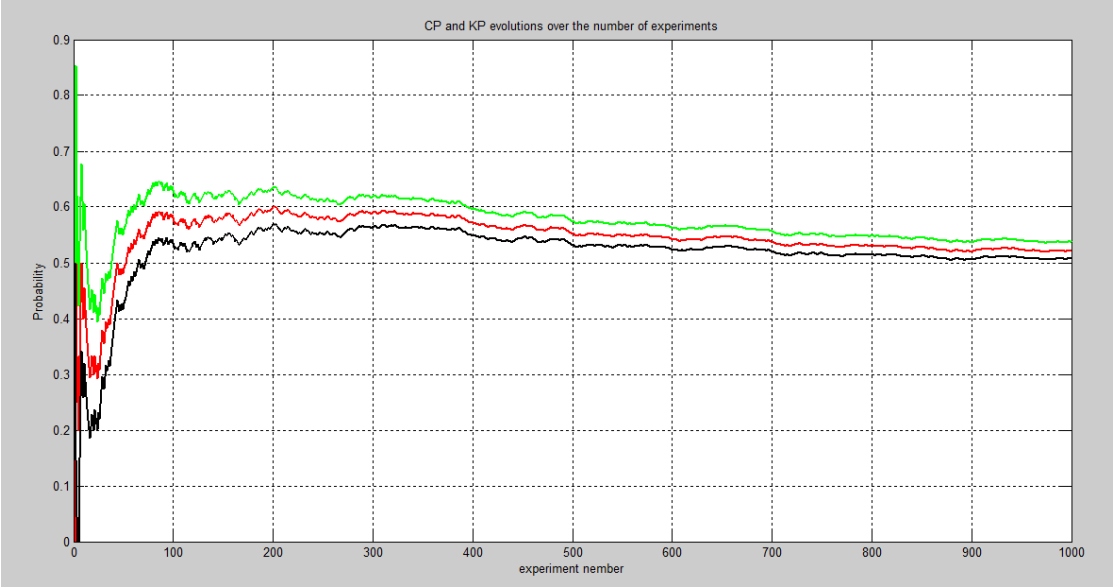


Figure 16. The upper (in green) and lower (in black) bounds of probability

Notice how the uncertainty gap between the upper and lower bounds starts to close with the increase in experiments. Let the error of representation be defined as the mathematical average of deviation of the current calculated probability from the ideal probability, which in this case is ½. The lower bound value will be in error if it was higher than ½, whereas the upper bound value will be in error if it was lower than ½. With that in mind and using figure 16, the average error in representing the lower probability is 0.03, the average error in

representing the upper probability is 0.0013 and the average error in using the relative frequency interpretation, equation 127, is 0.057. This clearly proves CP's upper and lower bounds as a better representation of probability. But it may seem ad hoc to assume the initial probability to be ½ and then prove that it is the case; after all, not every coin is a fair coin. In order to prove that the upper and lower bound method would still work in any other situation, let us imagine two situations where the decision-maker suspects the coin to be biased but it is not and where the decision-maker does not suspect any bias but the coin is biased. Figure 17 shows the first situation, where a decision-maker assigned P(heads) = 0.3 to the upper and lower probability values. Nonetheless, his/her belief starts to update towards the correct end pretty quickly.

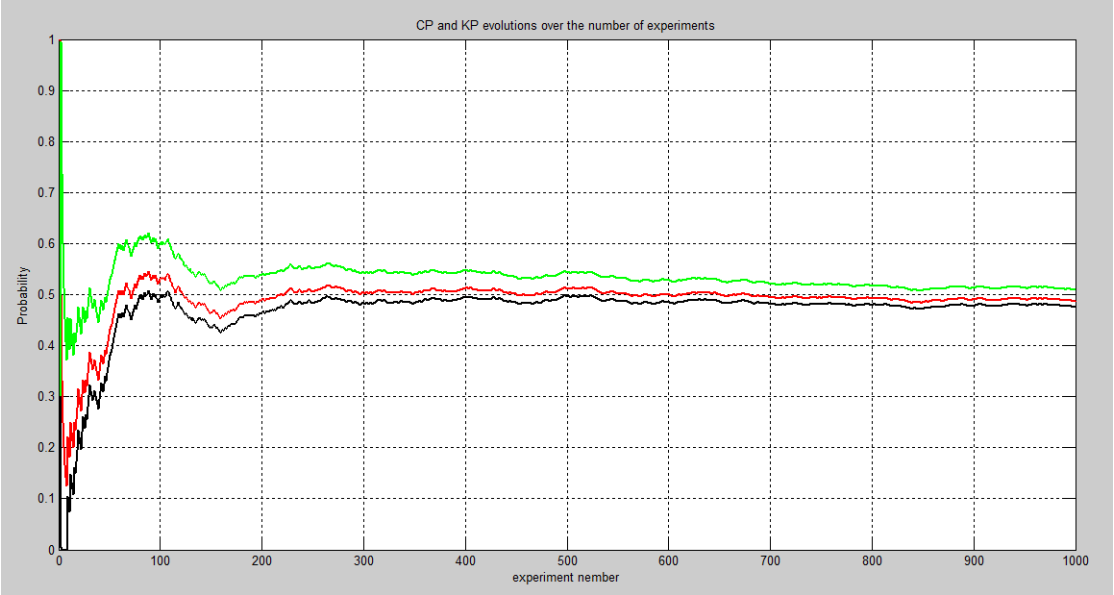


Figure 17. The upper (in green) and lower (in black) bounds when changing the initial probability to 0.3 rather than 0.5.

For the second case, the simulation of coin flipping was set so as to produce more heads than tails, with a ratio of 7:3, but as the decision-maker is

indifferent to that, he/she initially assigned P(heads) = ½ . Figure 18 shows the results. Notice how the upper and lower bounds almost always represent the ideal probability correctly, as it always lies within the region of uncertainty. This result is intuitive, given the power of Chernoff bounds and CP.

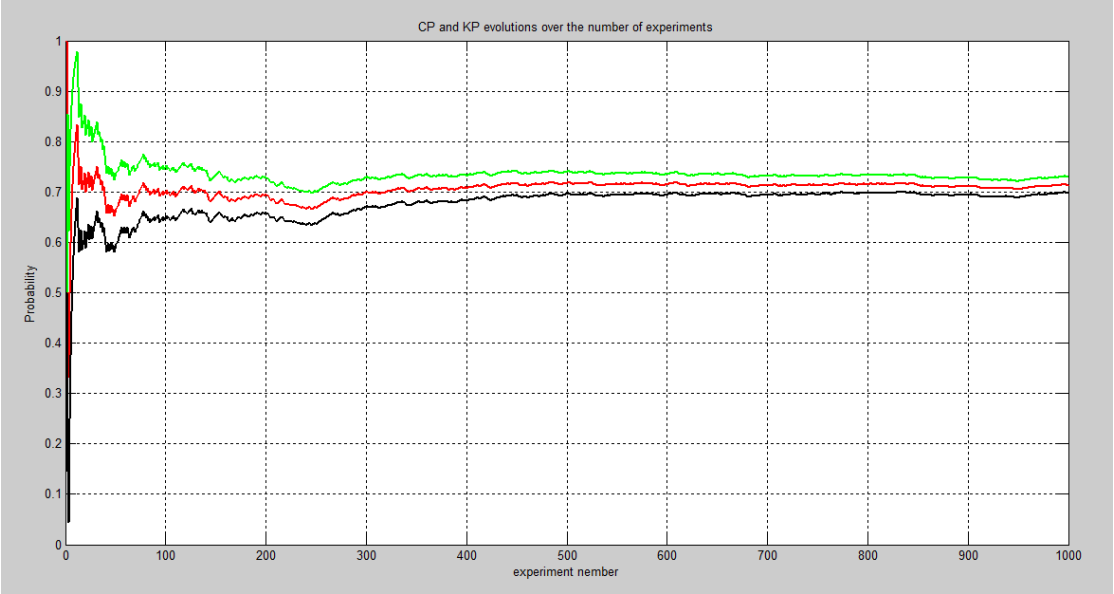


Figure 18.The upper (in green) and lower (in black) bounds for a biased coin with p(heads) = 0.7

Finally, because the thesis advocates better knowledge representation on less data, it will be of value to focus the lens on the performance of the upper and lower bounds on the first few experiments. Figure 19 shows 100 coin flipping experiments simulated with Matlab. Once more, the upper and lower bounds, although wider, came closer to the ideal probability of ½.

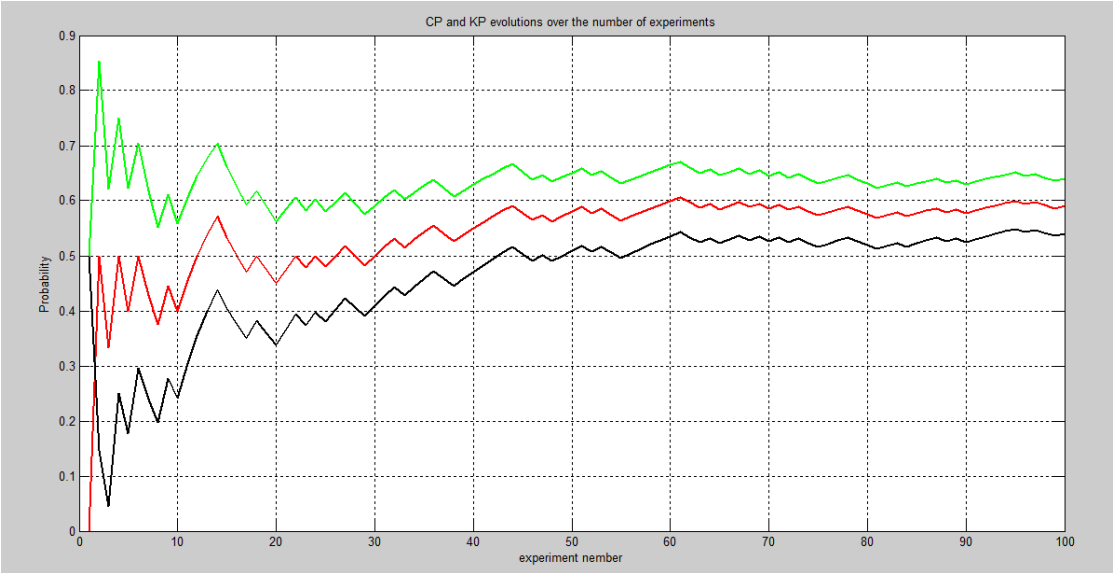


Figure 19. The upper (in green) and lower (in black) bounds for 100 coin flip experiments with p(heads) = 0.5

Table 1. Summary of determining the upper and lower probability bounds for a coin flip

No of experiment	Initial probability	Actual probability	Average error in $P_n(\omega)$	Average error in the lower bound	Average error in the upper bound
1000	0.5	0.5	0.057	0.03	0.0013
1000	0.3	0.5	0.018	0.00005	0.0021
1000	0.5	0.7	0.20	0.17	0
100	0.5	0.5	0.068	0.0138	0

Table 1 shows a summary of the previous four experiments. The results should not be surprising at all, as the upper and lower bounds provide a more relaxed way of describing what the probability is. What is important here is that although the relative frequency of heads moves above and below

the actual probability, the bounds stay on target, widening up when the number of experiments is low and narrowing down as the number of experiments increases. Equations 125 and 126 provided us with a way to change the representation of knowledge dynamically and online. To emphasize this point more, the four experiments show that we can use 125 and 126 to make decisions right from the beginning of the process, system, or observations. No period of offline time is needed to calculate the probability of an outcome; it will be computed dynamically online and in real-time. This very simple result in concept was not possible without CP to represent knowledge and KP to quantify the knowledge. Finally, we can summarize the algorithm used to estimate the probability of heads in the previous four experiments as follows:

Algorithm 1: Probability update of a Bernoulli variable

1- *Start with n=1 and let* $\underline{P}(\omega) = \overline{P}(\omega) = P$ *where P is a subjective belief or an unconditional probability of a similar variable.*
2- *Calculate the average occurrence of the variable using equation 127.*
3- *Calculate* $\underline{P}(\omega), \overline{P}(\omega)$ *using equations 125 and 126*
4- *Get the next experiment outcome and increment n.*
5- *Repeat 2,3, and 4 until the difference between* $\underline{P}(\omega), \overline{P}(\omega), P(\omega)$ *is less than a predefined threshold like 0.05.*

Step number 5 defines the stop criteria for the algorithm, which we have not touched on yet. Since the purpose of the algorithm is to help represent knowledge only when little data is available, then the algorithm should pass control to the KP representation when enough data are available. How do we quantify that? One way is to measure the uncertainty gap or difference between the upper and lower bounds to see how close they are to the numerical value of relative frequency probability. Another way is to calculate the confidence in the value of the current probability using the Chernoff bound:

$$Conf \geq 2e^{\frac{-n\varepsilon_n^2}{2}} \tag{128}$$

Hence, the algorithm stops if the confidence (Conf) is more than 0.95. However, there are instances where the algorithm might fail, for example, if the decision-maker decided to start with a probability of 1 or 0, and it happened that the first experiment resulted in heads. This scenario is an example of an extreme case where a decision-maker is choosing inrational value, for if the probability of an outcome is believed to be 1 then there is no need for probability in the first place, because 1 denotes a deterministic outcome rather than a probabilistic outcome. However, the decision-maker can still use a value very close to 1, such as 0.99, to avoid such a situation.

In addition, we can use the same method to estimate conditional probability since conditional probability is also a probability, that is, it obeys the KP axioms. In such case, we can estimate $P(X_{n+1}/X_n)$ and thereby provide

a dynamic, online and real-time predictor of the next outcome based on the current one, with accuracy that increases with the availability of more data.

3.3.3 Other types of distributions

All the discussion and results of the previous section were under the assumption of Bernoulli variables applied to the very basic problem in probability, which is that of flipping a coin. Therefore, the algorithm developed so far would not be of much help to solve any scientific problem. How do we modify it in order to make it applicable to a wider range of random variable types?

For that end, two approaches are proposed. First, we can derive a new Chernoff bound for whatever variable type is in question, or we can modify our algorithm to make it applicable to the variable type. The second approach appears to be easier than the first are, although, literature is full of examples of Chernoff bound for various trails like the Poisson one [91]. Consider a random variable like X with outcomes (a,b,c and d). Clearly, X is not a Bernoulli variable, but we can make it look like a Bernoulli variable if we let P(success) = p = P(a), and P(failure) = q = 1 – p = P(b)+P(c)+P(d). Then we can use algorithm 1 to estimate p = P(a) and q, and we can repeat for the other variables and normalize them to 1 (see algorithm 2).

Algorithm 2: Probability update of discrete independent variables

1- ***With n =1, Get the next outcome and assign it to p.***
2- ***Estimate the probability of p using algorithm 1 leaving step 5.***
3- ***Repeat 1 and 2 until all variables are estimated.***
4- ***Normalize all probabilities so that they sum up to 1.***
5- ***Repeat until step 5 of algorithm 1 is satisfied.***

However, algorithms 1 and 2 assume independent trials, which is an assumption not always valid. If the trials are dependent then we can map them into a sum of variables that are not. One way to do so is detailed in reference [88].

3.4 Summary

The theory of probability is filled with rich concepts and results, from the philosophical debate on the nature and meaning of probability to the practices and theories of quantifying it in a given context. A good grasp of both ends is of great value in analyzing the needs and requirements for an adequate solution to a problem.

Although there are many interpretations of probability in the literature of philosophy, probabilistic logic and mathematics, CP stands out from the crowd as a relaxed approach to represent probabilistic statements when little or no

information is available about some situations. In this chapter, we have seen how CP can be made compatible with the quantative interpretation of probability and how it can be combined with KP to produce a simple formula to update the upper and lower probability bounds in real-time. Although the assumptions upon which the formula was built are simplistic, the resulting algorithm can be taken steps further to make it applicable to more complex and interesting problems. Such problems are introduced in chapters 4 and 5, where we will attempt to apply the CP approach to probability to two interesting problems: aviation safety and patient monitoring in ICU.

4. Application to aviation safety

Having introduced the proposed approach of using comparative probability to make sense of data as they evolve over time, it is time to apply the proposed approach to an active area of research and measure the benefits and drawback of the new algorithms. After all, a theory of probability is of no use if it cannot be applied to science in any constructive way. In this chapter, the CP approach to Bayesian networks will be applied to some interesting problems in aviation safety. Aviation is one of the highly active industry sectors, with millions of passengers transported every year. It is where safety is held at a high priority through state-of-the-art diagnosing equipment of potential faults and highly detailed procedures to ensure safe and comfortable journeys for the passengers. In such situations, high safety standards will be maintained through online and offline monitoring and analysis of aircraft equipment. In the online phase, the aircraft sub-systems are constantly swept for indications of faults. If a fault is detected, a diagnosis subroutine is initiated to identify the fault and isolate its source. The process of detecting, identifying and isolating a fault will ensure that the pilot is aware of the existence of the fault and that it is attended to before the situation worsens. On the hand, the offline phase is the process of analysing the flight data stored in the flight data recorders (FDRs) to look for abnormalities in equipment behaviours, to measure the pilot performance, and to categorize the type of flight to sets of

types. Such analysis of data, whether online or offline, presents researchers with unique problems that call for novel solutions. It would be safe to say that all of the techniques and algorithms used in analyzing data to detect anomalies involve a step in which the results have to be labelled either normal or abnormal. In other words, the algorithms have to make a decision regarding the normality of the data it is analysing. Researchers may use different terms to refer to such phases, such as if-then rules [92], threshold detection [93, 94], or classification [95], but such decision-making phases are the main theme of the thesis.

The previous chapter introduced several approaches to CP and detailed a hybrid approach to it that tried to combine the benefits of CP and KP. However, it is still unclear how the proposed techniques can be applied to any real world scenario beyond the basic coin and dice rolling examples. The purpose of this chapter is to investigate the application of CP to some interesting problems in aviation safety. The first problem is the real-time fault detection and the diagnosing of equipment onboard. The main challenge of this problem is how to identify a fault in an environment in which every piece of information is doubtful. The second problem is a DSS design where many cues collected from various pieces of equipment are combined to present the pilots with the bigger picture and to help them make better decisions through recommendations.

This chapter will start by establishing a context with current methodologies used in aviation safety's fault detection and diagnosis through a literature

review. It will then discuss the requirement for an online equipment readings validator / fault detector. The result will be used to build a decision tree to come up with recommendations to the pilot in order to draw better navigation.

4.1 Literature review

The current state of the art in data analysis methods can be loosely divided into two major categories:

1) model-based, and

2) data-driven (see Figure 20 [96]).

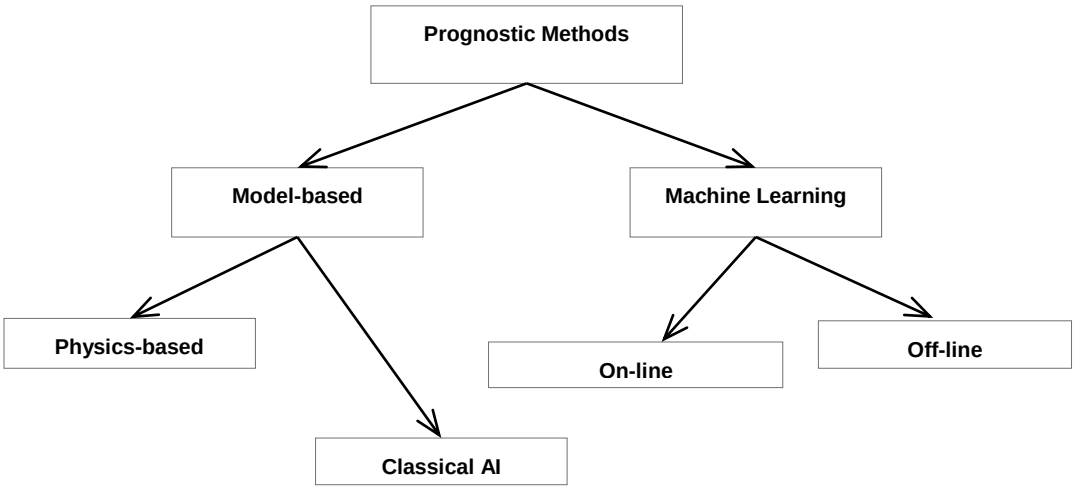


Figure 20. Classification of data processing methods

physics-based approach relies on first order depiction and/or modelling of the system, such as differential equations representation [97], whereas the AI-based uses classic AI techniques such as expert systems, finite state machine, or classical decision making. For example, thresholds can be defined for each parameter recorded in an FDR and the instances when these parameters exceed those thresholds can be used to trigger IF-THEN rules to

instantiate a consequential action or a report [98]. On the other hand, data-driven algorithms use historical records of data to model the process. They use either mathematical models or machine learning to achieve that goal. Some examples of these techniques are given in the following sections.

4.1.1 Model-Driven Data Analysis Approach

The model-based diagnosis approach is focused on finding relationships between system variables. These relationships can either be represented quantitatively, such as with mathematical equations, or qualitatively, such as with IF-THEN rules [94]. This approach dates back to the early 1970s and has been applied to jet engine diagnosis by Baskiotis and colleagues. They developed a general methodology for diagnosing a system in which one can mathematically represent the relationship between its internal mechanical state and its external performance [99]. Since then, various approaches have been implemented in this category. Reference [100] gives an introduction along with some examples of such approaches as applied to an actuator, a combustion engine, and a passenger car.

In general, model-driven diagnosis systems follow a two-step procedure. Firstly, they monitor for discrepancies (also referred to as residuals) between the actual and expected status of some measured parameters. These discrepancies can either be identified by an added redundant hardware such as sensors or analytically through functional representation connecting the inputs, states, and the outputs of the system together. Residuals can be thought of as features that need, ideally, to be triggered by only one fault type.

Secondly, they transfer residuals to their matching fault type through, for instance, a decision tree of IF-THEN rules [95, 101].

However, diagnosing systems are often hybrids of the above types. They may use quantitative and qualitative approaches to diagnose a residual. One recent approach in aviation safety is to verify the data obtained from an FDR with mathematically simulated data that is generated using, for instance, the 6 DoF representation of that aircraft. The deviation between the recorded data and the checking data is used for prognosis of potential faults in the aircraft [102].

In conclusion, the development of a prognostic algorithm requires the availability of high accuracy models of an aircraft. Such models are sometimes only available through the manufacturer. The greater the number of variables, states, and parameters to be modelled, the more complex the model becomes and the more it requires computational power. In addition, the growing complexity of avionics might put the modelling process beyond practical realization, let alone the increase in cost. Model-based diagnosis systems are generally limited to linear process, otherwise, they require the implementation of a piecewise linear approximation resulting in a possibility of poor performance. Model-based systems are also prone to errors due to uncertainties arising from parameter drift [95]. Finally, one obvious limitation is that they require all faults to be known beforehand. Otherwise, the system cannot detect residuals for labelling as unknown faults.

4.1.2 Data-Driven data analysis approach

In contrast to the model-based approach, the data-driven method, also known as the process-history-based method, does not require prior knowledge of a process, either quantitatively (such as the mathematical relation between the process variables) or qualitatively (such as rule-based interaction between inputs, states, and outputs of the process), but rather relies on extracting knowledge from historical data [103]. The process of extracting knowledge can be loosely classified into

a) parametric, and

b) non-parametric.

In the parametric approach, historical data is used as training examples to model the process into a parameterised model. Once that has been done, the training examples can be discarded as they are now represented by the model structure and the parameters. The model is then used to predict the next example. In contrast to this, the non-parametric approach does not generate parameters or learn from examples but rather uses the whole data or a selected sub-set of it as instant training examples to substitute the real process and/or predict the next one. Some of the approaches of that class do not require a training phase, parameters being generated, or models being built on the data. The system would be available instantly upon the availability of data, and it is hence called instant-based learning or memory-based learning [1,p. 737].

An example of the non-parametric approach is the k-nearest neighbour, which was employed by the so-called ORCA algorithm. ORCA is a data-driven anomaly detection algorithm proposed by S. Bay and M. Schwabacher in 2003. It was designed to overcome the requirement of high computational power for large high dimensional dataset and achieved near linear scaling performance [104]. C. Chiu and others proposed the use of the Case-Based Reasoning (CBR) to improve aircraft maintenance support [105]. CBR is a relatively new approach in machine learning, whereby similar past problems are used to solve the current problem on the basis of its similarity to the past ones. CBR has also been proposed for troubleshooting aircraft engines [106] and prediction of component replacement [107]. Support Vector Machine (SVM) is the most popular non-paramteric algorithm. It's attractiveness is related to the optimality property, its ability to construct maximum margin separators in which the distance from the boundary decision to the training examples is maximized, its ability to map the training data to a higher dimensional space where linear separation is otherwise not possible, and the sparse representation (its need to retrain only small proportion of the training examples rather than the whole data which lead to less memory utilization and less computational power [1,p. 737]). S. Das and colleagues have proposed an improvement over the ORCA and other algorithms for flight data anomaly detection through the use of the multiple kernel learning method. Their algorithm is applicable to both continuous and discrete data streams. However, their analysis was limited to flight levels below 10,000 ft [108].

On the other hand, the parametric approach can also be sub-divided into qualitative or quantitative. In the qualitative approach, the historical data of a process is used to extract expert rules such as the case with expert systems or to predict the trend of the process, such as the case with qualitative trend analysis (QTA) method. The quantitative approach can be either non-statistical or statistical. Neural Networks are examples of non-statistical methods whereas clustering and principle component analysis (PCA) are cases of statistical techniques [109] (see Figure 21 [109]).

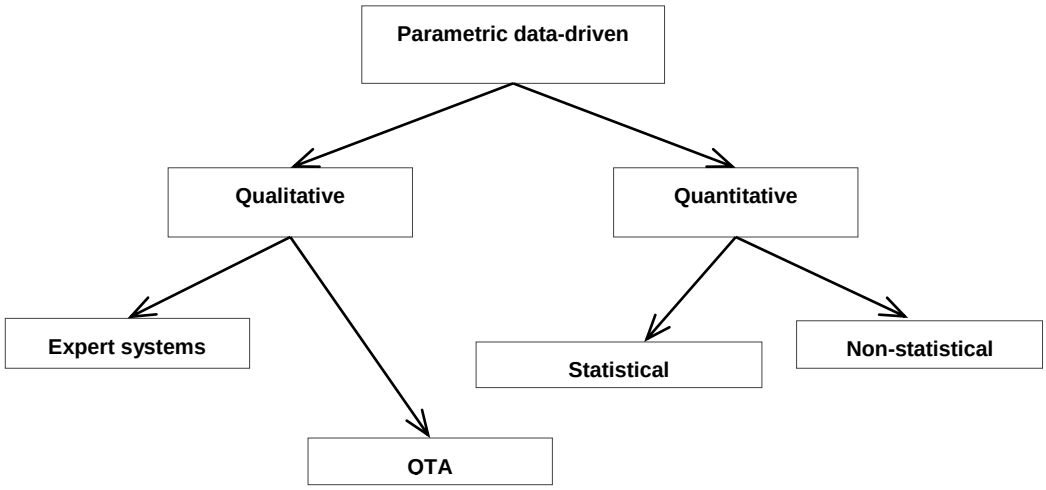


Figure 21. Classification of parametric data-driven approaches.

In expert systems used to define normal behaviour and, based on this, to extract anomalous behaviour that can be associated with a fault using, IF-THEN linguistic rules are usually used. Knowledge about the system can either be described as the state of a system, usually referred to as facts, or the relationship between these facts and the state of the equipment [110]. The diagnostic procedure can either use forward or backward reasoning. In forward reasoning, an observation or a fact, such as a residual, would form the antecedent part of some if-then rule to trigger a production rule, or a

consequence, which can be an alarm, or another if-then rule, as it has now become a new fact. In backward reasoning, an observation represents a hypothesis that is used to search the entire consequent parts of the knowledgebase for matches. The antecedent parts of these matches are then used as new hypotheses for new searches, and the process continues recursively until one hypothesis proves to be false or the entire knowledge base is verified [110]. The main problem with the expert system approach is the so-called ‘curse of dimensionality’, which is related to the fact that one needs a huge number of rules in order to describe all possibilities. Another issue with traditional expert systems is the subjective element and the fact that they lack adaptability. These problems have been addressed successfully by fuzzy logic [111], which works in a manner similar to expert systems but allows partial degrees of fulfilment and overlap. In this way, a small number of much more powerful rules (which may be partially satisfied – the degree of their activation is inversely proportional to the distance to prototype) can cover the whole data space [112]. These are in general very specific to a process and very difficult to update [109].

In qualitative trend analysis, the evolution of historical data is abstracted into symbols called primitives. It approximates the time development of each parameter by linear segments and then it labels them as:

a) increasing,

b) decreasing, or

c) steady.

Successive segments with similar primitives are aggregated together to form trends [113]. Trends are used to predict the future state of the system and consequently analytically to diagnose the system by comparing the analytical state of the system with the actual state. The power of QTA comes from its ability to represent development of a parameter in understandable terms, such as ‘steady’ or ‘increasing’, often used by experts in many fields, e.g. doctors monitoring patients’ conditions [113]. QTA has been applied to trend equipment malfunctioning in commercial aircraft using historical reports from the Service Difficulty Reporting (SDR) system [114]. The main problem with this approach is that the resulting model was too specific to one aircraft type and operator. No evidence was given as to how the resulting model can be used or even modified to become applicable on operators with mixed aircraft fleet. Additionally, the amount of computation power required to trend and analyze a process increases significantly as the number of parameters representing the process increases. However, PCA has been suggested to reduce the number of monitored parameters [115]. Finally, abstracting trends for non-linear time series requires a more complex segmentation method than a simple triangulation, if the number of resultant segments is to be reduced. Several approaches have been suggested, such as using neural network, or calculating the first and second derivative of the parameter changes over time [109].

Under the non-statistical category, we find Artificial Neural Networks (ANN) to be the most utilized approach to diagnosis. Artificial neural networks are attractive due to their rapid development speed and their ability to model

highly nonlinear processes. There is huge amount of literature that suggests ANN for fault detection and/or isolation. They have been used to detect aircraft sensor abrupt faults using the virtual sensor concept, in which ANN is used to reconstruct what the output of the system should be and, thereby, compare it with the actual output while monitoring for errors [116], self-calibrating sensors [117]. ANN has also been proposed for real-time control surface fault detection and isolation (FDI) using the same virtual sensor concept mentioned earlier [118]. A slightly different approach has been proposed by Ali and Tarter. In their method, instead of modelling sensors input/output pairs, the aircraft engine noise level during the flight profile is modelled and used for comparison [119]. Additionally, Savanur and others used an adaptive neuro-fuzzy approach to diagnose aircraft actuator faults [120]. However, ANN might be prone to misclassification error near the class boundary where the training data is sparse [121]. As is the case with the previously mentioned approaches, ANN has been applied to only a fraction of what a modern flight data recorder stores, and is also unable to detect unknown anomalies where no classes where defined during the training phase.

Whether a process is modelled through expert knowledge or by extracting knowledge from data, the resultant model is deterministic in a way that the future developments of parameters are uniquely and non-randomly dictated by their past states. This is not the case with every parameter recorder in an FDR. For instance, the sequence of pilot inputs involves elements of randomness. In such a case, the probabilistic/stochastic approach is more

reasonable when every observation is of a probabilistic nature and the majority of observations are assumed to be normal and a significant deviation from the normative is considered anomalous. Traditional quality control monitoring is one of the oldest users of statistical data-driven fault detection [109].

Samara and colleagues showed how to design a one-versus-one case statistical fault detector that was utilized for the angle of attack (AoA) sensor. Their statistical fault detector used a fixed-length sliding window to feed statistical calculation of the mean and standard deviation of the residual. However, it wasn't clear why some of their thresholds were chosen the way they were, apart from an ad hoc justification of reducing the number of false alarms [122]. Chu and colleagues used a least square (LS) regression approach to detect performance anomalies in flight data. They considered anomalous those samples that deviate from the scatter as a result of turbulence and system errors. The model requires the availability of huge amounts of historical data, which were generated artificially using one of NASA's medium fidelity flight simulators. Requiring thousands of flight examples of an aircraft type for training puts the method beyond practical consideration; in addition, the model has not been applied to real flight data [123]. Extended Kalman Filter (EKF) has also been proposed for fault detection [124]. EKF is the nonlinear counterpart of the widely distributed Kalman filter [125], which is a probabilistic state estimator that tries to estimate the states of a system when only noisy observations are presented. Since

data from FDR could contain both discrete and continuous parameters, it is unclear how accurate EKF can be when applied to continuous processes.

Cluster Analysis is one of the most used tools in the statistical analysis toolbox. Its popularity is probably the result of its powerful ability to organise data into groups (called: clusters) based on similarity in an unsupervised manner. There are generally two types of clustering algorithms:

a) on-line, and

b) off-line.

In hierarchical clustering, data are organized into nested groups of hierarchical fashion so that a data point is part of a sequence of nested partitions. The organization of data into hierarchical clusters can either follow a bottom-up (commonly known as: agglomerative) or a top-down (also known as divisive) approach [126]. On the other hand, partitional algorithms assign each data sample to a certain cluster. In fuzzy clustering [127], a data point can partially belong to more than one cluster. There are many approaches to achieving partitional clusters in the literature. They are usually divided into: centre-based, prototype-based, graph-based, and density-based, to name a few [126]. In the centre-based and prototype-based approaches, the clusters are represented in terms of centres (called: centeriods) and the data is assigned to that cluster where the distance to the centre is minimum. A centeriod of a cluster is the arithmetic mean of all data points within that cluster. The resultant shape of a cluster is of convex shape hyper-sphere (if Euclidean) and ellipsoid (if Mahalonobis) distances are used. The graph-based approach begins with a graphical depiction where data points can be

connected together, based on similarity, to form hyper-graphs. The approach works best if the data points are well separated. The density-based approach assumes that clusters are those special regions where the data points are denser that the other regions [126]. It is quite tolerant to noise and is mathematically efficient. The model of the data density distribution is often incorporated into the algorithm in terms of constraints or geometric properties of the co-variance matrix [126].

Clustering analysis has been widely used to detect anomalies. Only certain key-methods that have been applied to aviation safety, or that have the potential of being so, are mentioned. Thomas R. Chidester has applied the cluster-based approach to flight data collected from about 1300 flights. He used the resulting clusters to generate what he referred to as a ‘morning’ report, which measures the similarity of a flight data signature to the cluster obtained from the analysis. The similarity is then used as a score of how typical a flight is [128]. However, the analysis was limited to a given proportion of the flight data. Moreover, since the clustering method is centre-based, the resulting clusters are intolerant to noise. Finally, since the shape of the resulting clusters are hard determined by the algorithm and not by the distribution of the data itself, a flight signature could be misclassified into a wrong cluster, which may result in an increase in the number of false negatives (FN) or false positives (FP). Mark Ford reported an approach based on the use of clustering analysis that was conducted by the British Air Accidents Investigation Branch (AAIB), QinetiQ, and specialists from engine manufacturers to detect anomalous signatures in fuel flow to the engines of a

Boeing aircraft. They analysed 178,000 flight data and performed several experiments to understand the formation of ice and, thereby, augmenting the instances where the collected data is sparse. They started by cutting the recorded flight data into phases and then they focused on only two parameters: fuel flow and fuel temperature [129]. Although the report mentioned several tools for data mining, there was little discussion on which of them were actually used, what assumptions were made, and any other technical or mathematical processing of the parameters used. S. Budalakoti and colleagues developed what they called the SequenceMiner algorithm, which detects anomalous sequences of switch triggers inputted by a pilot in an aircraft's cockpit. They used a modified version of the k-medoids clustering algorithm by finding medoids within randomly selected regions of the entire dataset. They then used the Bayesian decision tree to model the differences and similarities of sequences within the clusters as a way of characterising the detected anomaly [130]. While the SequenceMiner works very well when applied to discrete sequential data, this is not the case on continuous ones [108].

There is also a growing research interest in using so-called artificial immune system to detect anomalies in flight data. Artificial Immune System (AIS) is a set of mathematical models that attempts to mimic the biological immune system (BIS) found in vertebrates. In recent years, there has been a rise of interest in AIS due to its adaptability, optimizability, and potential to detect anomalous behaviour. One important application of AIS are situations in which much information is known about the normal behaviour of a system

and a little or no information is known about the anomalous behaviour of a system. In computer security, for instance, the behaviour of a normal user connected to a server could be known through the analysis of raw access files, whereas the behaviour of a hacker trying to exploit the system is not always known, particularly with respect to the discovery of new vulnerabilities and methods to exploit a system. AIS are built on clustering, whereby one can use clustering to analyze the normal behaviour of a process. The resulting clusters space is referred to as the ‘self’ and one then uses the complementary space, known as the ‘non-self’, to generate detectors and apply them to the classification of new data as either the ‘self’ or ‘non-self’, i.e. an anomaly [131]. Jennifer N. Davis used an evolutionary algorithm to efficiently generate detectors that cover the complementary space where the clusters representing the normal behaviour of a system reside. The method was applied to data collected from flight data recorders [132, 133]. K. Krishna Kumar studied several potential models of AIS to be applied in aerospace applications and questioned the adequacy of these models to replicate the immune system metaphor [133]. In addition, there were further concerns regarding the representative power of the generated ‘self’ space over the normal behaviour of a system [134], and it has further been shown that the detectors generated over the Hamming-shape space are not well suited for online anomaly detection problems [135].

All of the methods reviewed thus far, except kNN, require a stage of offline training so as to extract knowledge from available historical data, except for the non-parametric approach, in which the system is available instantly once

the historical data is available. However, the computational power required for implementing a pure non-parametric system is tremendous given the huge dimensionality of the data recorded by an FDR. However, a recently proposed approach that has attracted much interest in online knowledge extraction algorithms where the parameters of the system are estimated and re-configured on the fly as the data is being passed to the algorithm. Some of these techniques can even “evolve” in the sense of introducing new clusters and rules to better describe the system [130]. One such evolving algorithm is based on the Takagi-Sugeno realization of fuzzy systems (commonly referred to as: eTS). eTS uses a density-based clustering algorithm called eClustering plus recursive LS (RLS). This results in a flexible structure rule-based model of the process that can be used to predict its next state. The structure of the system is able to evolve (add new rules or modify existing ones) according to the data density dynamic changes. Evolving systems have been used for anomaly detection, albeit in other industries such as detecting anomalies (or novelties) in video streams [136], machine health prognostic [137], and real-time characterization of car driver behaviour [138]. As this approach has not been applied to the aviation industry, it would be one of the aims of the team to estimate its potential for flight data processing, which is specifically the case for this project. One could think of an anomaly from a statistical point of view as those samples that statistically deviate from the normative represented by the majority of other samples. Thus, the detection of novelties boils down to estimating the density and defining the deviation from the mean density. A sample can be considered anomalous if this deviation is larger than two or three times the variance of the data, known as sigma [139]. For the estimation

of pdf to be computationally efficient, a recursive approach should be undertaken, such as the case with recursive density estimation (RDE) approach. RDE has been used as a novelty detector in video streams as opposed to the traditional kernel density estimation (KDE) approach [140].

4.1.3 Types of anomalies

An anomaly can be defined as a data-point, or a sequence of data-points, that does not conform to a well-defined perception of an expected behaviour [141]. Researchers use many terms to describe the task of detecting anomalous behaviour, often with different terminology. These include novelty detection, outliers detection, exception mining, or surprise detection [141]. In addition, the definition of anomalous behaviour is another area of debate, as research assumptions are often influenced by the availability of data, nature of application domain, and availability of validating model [142]. One of the earliest definitions of an anomaly comes from F. Grubbs, where it was defined as:

> *An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs [143].*

Hence, the random variations in observed values of a samples are assumed to be the result of a determined system of causes which, acting together, are considered normal, while anomalies are the result of another set of causes – such as human error or equipment malfunction – which cause the observations to further deviate from the normal distribution of the sample. If we assume a null hypothesis of: observation x_i conforms to the normal

Gaussian distribution of its belonging sample, then a simple way of testing that hypothesis is [143]:

$$T_i = \frac{(x_i - \bar{x})}{s} \tag{129}$$

where $\bar{x}$ is the arithmetic mean of the sample and s is the standard deviation calculated with $n - 1$ degrees of freedom given by:

$$s = \left\{ \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{n-1} \right\}^{1/2} \tag{130}$$

The value of Ti obtained from (129) is compared to a critical value that represents the likelihood of recording that observation by chance given the size of the sample. The likelihood is often referred to as: the significant level and common values of it are: 1%, 2.5%, and 5%. The critical value of T is often given in table format which lists a value of T for a given significance level and sample size (see reference [143] for an example). Once an anomaly is detected, it would be removed from the data-set and the mean and standard deviation values are re-calculated once again, and a search for anomalies is initiated again. The same procedure continues until no further anomalies can be detected [141]. However, if the number of anomalies is small compared to the size of the data-set, one could sacrifice the accuracy of the sample distribution for computation power efficiency [141]. Another approach is to use unsupervised classification, also known as clustering. Within the clustering approach, there are two types of anomalies:

1) An *outlier* can be thought of as a point in space lying outside those regions considered normal. Figure 22 shows a process represented by a space of two features. Points inside regions A and B belong to the class of normal behaviour. However, point X, which lies outside those regions, is considered anomalous. Hence the name: ***outlier***.

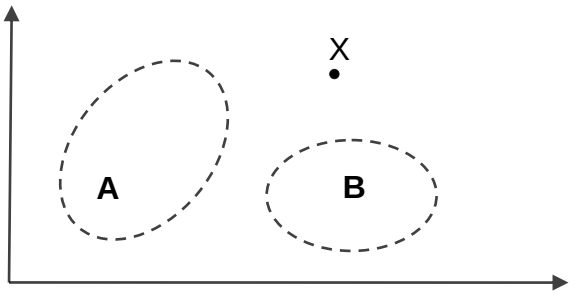


Figure 22. Point X is an outlier because it resides outside the normal region represented by A and B.

2) A surprise is a point assigned to a cluster where it was expected to be assigned to a different one due to the current sequence of events. Figure 23 shows that point X is assigned to cluster B where it should have been assigned to cluster C. An application for a surprise is removing misclassified nodes from a decision tree [144].

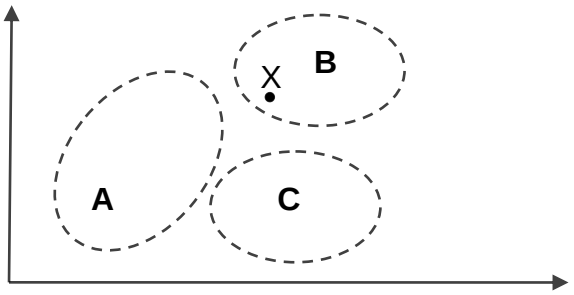


Figure 23. Point X is an outlier because it should truly be assigned to C not B.

A statistical approach to anomaly detection has the benefit of not requiring any prior knowledge of the application but it requires the availability of an adequate number of sample observations in order to estimate the normal distribution of that sample. This approach is sometimes referred to as the Type I approach [141]. In contrast to the Type I approach, Type II deals with the anomaly detection as a supervised classification problem. Data should be pre-labelled as normal or abnormal, but there could be several subclasses under the class of normal or abnormal. The system uses these labelled data to construct a model that can be used to classify a new data-point. Finally, the Type III anomaly detection approach requires the availability of data from one class, which is usually the normal class. The system would then use the complementary space of the normal class region to construct an anomaly detector [141].

4.2 Demonstrating a model-based diagnostic decision tree for validating aircraft navigation system accuracy

This section will detail the steps involved in designing a novel model-based fault detector and isolator to help pilots validate the accuracy of their navigation system. The designed system will be packed by a CP based Bayesian network to improve the performance of the system at times of less available information. In addition to proving the versatility of CP, the introduced system proposes a novel solution to fault detection and equipment

monitoring in the aviation industry. The reasons behind the proposed solution will also be discussed in the following subsections.

To assist the effort for aviation safety and increase navigation accuracy, large aircraft are required to use redundant measuring equipment. The accuracy of the navigation system can be verified by comparing the readings from two or more different equipment groups. For instance, an accurate altitude can be assumed when the altimeter reading of the pilot's panel is identical to that of the flight officer's panel. Otherwise, a search for the defective component is initiated which, in turn, might involve manual procedures such as switching to alternative air data or observing the status of the altimeter for visual defection cues such as a fluctuating pointer [8]. However, manual observations require the pilots to be in a high state of situational awareness where they would be able to comprehend the states of the aircraft and, in turn, make reasonable decisions [145]. This negates the purpose of a decision support system (or redundant measuring equipment) as they are supposed to raise pilot's situational awareness instead of the other way around.

One fault detection and isolation method that has received much research interest is the detection filter proposed by Beard in the early 1970s, where a fault is associated with a subspace of error state space called the detection space [146]. In this context, Caliskan and Hajiyev have studied four algorithms used to verify the co-variance matrix in a Kalman filter (KF) from a performance point of view [147]. However, since all KF-based algorithms follow signal-based modelling methods in which only the output signals are

monitored, these algorithms can only detect deviations from assumed normal behaviors. The enhancement to fault diagnosis and detection (FDD) proceeds on a strictly ad hoc manner, without any solid foundation to enhance generic applicability [148]. Other methods used to enhance fault detection are discussed in [149], including both the Multiple Model Adaptive Estimation (MMAE) and the Interacting Multiple-Model (IMM) algorithms.

Little attention has been given to establishing a framework to develop an FDD system that deals with navigation systems as a grid of mathematically and physically inter-related quantities in which the accuracy of a reading can be mathematically verified. Such verification could be worked out by a 6 Degrees of Freedom aircraft mathematical model. When the sensor states and mathematical states of an aircraft do not resemble each other, a search for a fault is initiated that involves qualitative fault isolation. In this demonstration, we will use the Bayesian diagnostic tree method to point to the most probable culprit of mismatching. The Bayesian diagnostic tree also serves as recursive Bayesian estimators to evaluate the probability density function of a given fault. Implemented with CP, we will also demonstrate the beneficial value of the implementation.

4.2.1 6 Degrees of Freedom Equations of Motion

When it comes to building a simulation model for a flying body with high fidelity, the 6 Degrees of Freedom (6DoF) is often the popular choice as it can be used to simulate displacement and rotation in three-dimensional space [150]. A rigid flying body (such as an aircraft) in free motion is able to move and rotate freely along any of the three perpendicular axes of a three-

dimensional space, hence providing the six forms of motion. The 6DoF equation of motion follows from applying Newton's second law of motion to a flying body subjected to aerodynamic and thrust forces $f_{a,p}$ and the earth's gravitational field. This can be written as [150]:

$$m\, D^{I}\, V_{B}^{I} = f_{a,p} + mg \tag{131}$$

Where m is the body's mass and V_{B}^{I} is the velocity of it with respect to the inertial frame (I). If the body to be modeled flies relatively close to the earth, the earth is often assumed to be the reference frame (E), and, for these purposes, assumed to be flat. To solve the previous equation, one needs to be able to access the forces applied to the body (B) with B taken as the reference frame. This change in reference frames is done through Euler transformation. Thus, equation (131) can be re-written as [150]:

$$m\, D^{B}\, V_{B}^{E} + m\, \Omega^{BE}\, V_{B}^{E} = f_{a,p} + mg \tag{132}$$

The calculation is best carried out using software packages that facilitate state-vector variable integration and matrix manipulation [150]. In this feasibility study, MATLAB was chosen as the simulation environment.

4.2.2 Aircraft Modelling

The use of high fidelity models to simulate aircraft motion in space is required for accurate validation in a non-simulation environment. The development of such a model requires extensive resources and modelling

time. A high fidelity model of a specific aircraft would require the knowledge of complete aerodynamic and thrust tables, flight control design, mass parameters, and the logic of the navigation and sensor operations. Only then could such a developed model be tested and its reliability thoroughly validated [150]. Unfortunately, such detailed considerations in modeling all onboard equipment would greatly affect the robustness of the model, and limit its application to other aircraft types.

As the focus of this demonstration is FDD/qualitative fault isolation, and given the time and resource constraints of this study, the decision was made to use a generic out-of-the-box model. The selection criteria for the model were primarily on their integration with academically-proven simulation environments, such as MATLAB, and trajectory visualizing software packages, such as FlightGear. AeroSim blocksets of MATLAB/Simulink block library developed by Unmanned Dynamics provide modules for rapid and fast aircraft modelling. A complete aircraft 6DoF model can be defined by generating a configuration script that specifies the aerodynamics and engine parameters for a specific aircraft type. It also provides a parser for importing FlightGear v. 0.9.2 models such as CESSNA-310. In the development phase of this study, the North American Navion was chosen to carry out the simulation. The models of the Aerosim block library are limited to only conventional aircraft with single piston engine and fixed pitch propeller. Nevertheless, this limitation was not deemed to affect the validity of the proof of concept, as the design of the system is modular and can be ported to use other aircraft types given an accurate mathematical model.

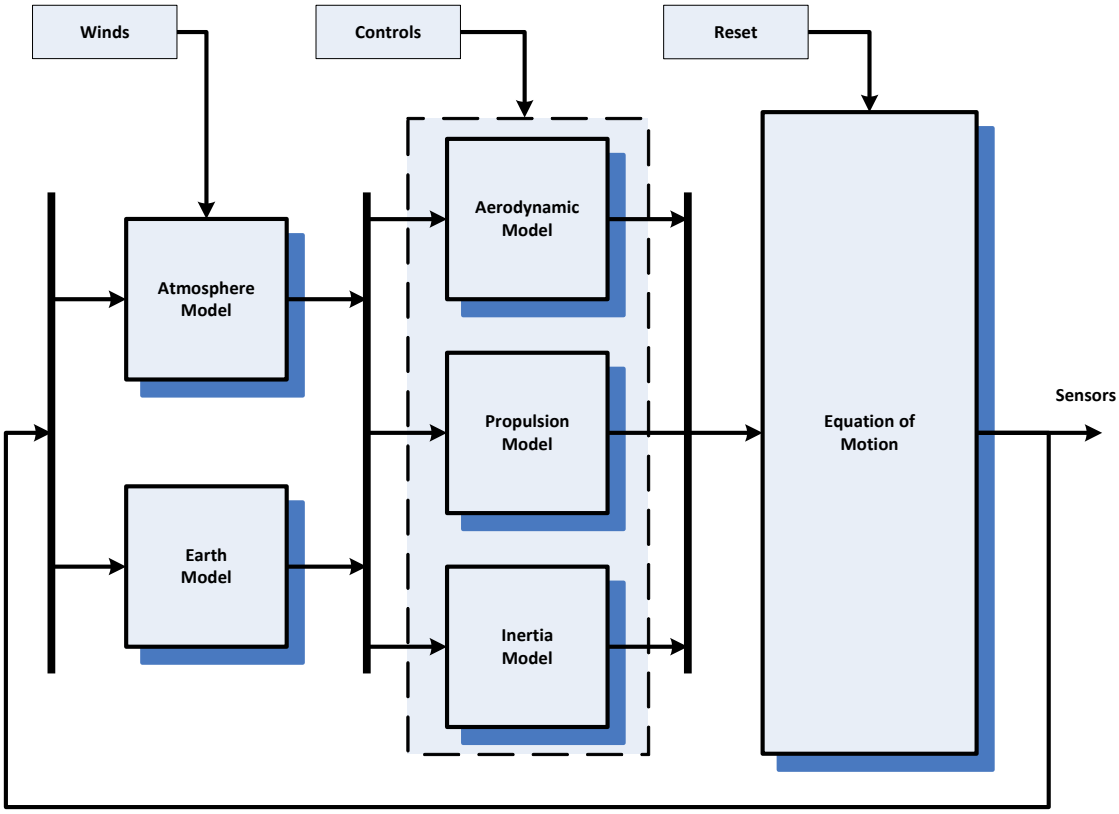


Figure 24. The internal structure of the complete aircraft block.

Figure 24 [151] shows a simplified block diagram of the internal structure used in Aerosim to simulate a complete aircraft. Controls from the pilot joystick are used by the aerodynamics, propulsion, and inertia models to calculate the total forces and moments applied to the aircraft giving the simulated atmospheric conditions and reference frame. These in turn are used to solve the equations of motion and obtain the aircraft position (altitude, latitude, and longitude), orientation (heading, roll, and pitch), and velocity. These vectors are used to update the atmospheric and earth model as a change in aircraft position might have an impact on the atmospheric conditions (e.g. pressure and gravitational forces). The sensor measurements are then derived directly from the calculated aircraft state. One drawback of

the Aerosim library is that there are no models developed to simulate appropriate aeronautical sensors [8], and generic analog and digital sensor blocks were thus used instead. This lack in specific modeling was also judged not to have a negative impact on the assumption and validity of this feasibility study as the sensors and vector states are treated as black boxes with variations artificially generated through the application of noise and scaling factors.

4.2.3 Current Functional Procedures

To compensate for sensor errors that equipment may encounter during operation, modern aircraft are fitted with redundant systems that work independently. The value of the measurements is then taken and the value displayed to the pilot is made through a majority rule or least square method. Since in most cases an aircraft in good condition might only experience a malfunction in a single piece of equipment, this error would be compensated for by the vote of the other redundant systems (assuming two or more redundant systems).

However, majority rule might fail if the cause of the malfunction is systemic in such a way as to affect the other redundant systems that are concurrently working out the same measurement. For example, the measurement of airspeed involves sampling pressure from outside the aircraft using special probes, called pitot probes. If an environmental condition such as icing could affect one pitot probe, it is not unreasonable to also expect some impact on the other, identical redundant probes.

Therefore, it is of great interest to be able to calculate the conditional probability of a malfunctioning sensor, given that another sensor has malfunctioned using the same process for measurement. Figure 25 shows a simple Bayesian network representation of two sensors (labeled S1 and S2) working out the measurement of a quantity in an influential environment (E). In this case, it is safe to assume the conditional probability of having a wrong reading given that E has occurred identically for both sensors, i.e.:

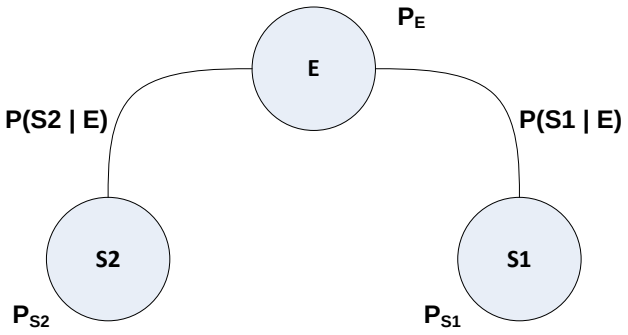


Figure 25. Bayesian network for two sensors S1 and S2 in environment E

$$\mathrm{P}(\mathrm{S1}|\mathrm{E}) = \mathrm{P}(\mathrm{S2}|\mathrm{E}) \ldots (3) \tag{133}$$

We are most interested in calculating the probability that the second sensor might be malfunctioning given that S1 has malfunctioned and E has occurred, i.e., we want to calculate: $\mathrm{P}(\mathrm{S2}|\mathrm{E},\mathrm{S1})$. One way of calculating this is:

$$\mathrm{P}(\mathrm{S2}|\mathrm{E},\mathrm{S1}) = \frac{\mathrm{P}(\mathrm{S1}|\mathrm{S2},\mathrm{E})\mathrm{P}(\mathrm{S2}|\mathrm{E})}{\mathrm{P}(\mathrm{S1}|\mathrm{E})} = \frac{\mathrm{P}(\mathrm{S1}|\mathrm{E})\mathrm{P}(\mathrm{S2}|\mathrm{E})}{\mathrm{P}(\mathrm{S1}|\mathrm{E})} = \mathrm{P}(\mathrm{S2}|\mathrm{E}) \tag{134}$$

The result of equation 134 means that if there are evidences for the occurrence of environment E, then the probability of one sensor malfunctioning has no statistical influence on the other. Both sensors would be influenced by E to the same probabilistic degree, whereas if E is assumed not to have a global influence, then the probability of having two wrong readings out of three is:

$$\mathrm{P}(2 \text{ out of } 3) = \mathrm{P}_{\mathrm{S1}}\ \mathrm{P}_{\mathrm{S2}} \tag{135}$$

The probability of two wrong readings out of three for sensors that have independent “ways” of working out a reading is dramatically lower than the probability of two wrong readings for those sensor types with a similar way of calculating a measurement. Thus, it is desirable to have a validating system that uses, to the maximum extent possible, independent methods of calculating the current states of an aircraft.

4.2.4 BADA and TEM

BADA (Base of Aircraft Data) provides performance operation data and aerodynamics parameters for about 151 types of aircraft. These parameters are the results of developing a mathematical model for a given aircraft using the total energy model (TEM) [152]. Consequently, the parameters can be used to check if an aircraft is operating within a set of recommended speed, rate of climb or descent (ROCD), or fuel flow. This could, in turn, provide a way of validating a current on-board situation in case the data being logged for any of these parameters goes beyond safe, recommended, or normal range. In this section, this process will be labeled BADA check or “Is exceedance”.

This type of validation could detect *exceedances* in real time rather than by the end of the trip, as is the case with the one described in [153]. As a side benefit, since the checking is performed against recommended operational data, commercial airliners would greatly benefit from the resultant fuel saving and maintenance, as pilots would be more likely to comply with recommended speed, ROCD, and so on. We will expand further on BADA usage as a DSS in section 4.4.

4.2.5 Assumptions and Proposed Design

The diagnostic decision tree network developed in this feasibility study is based on two assumptions:

1. The airplane is in good and airworthy condition such that the source of a problem could be traced back to one or two causes at most.
2. Mathematical aircraft model (called Math Engine) has some impact on the calculated parameters of the airplane vector state. This is borne out in aerodynamic theory:
    - Wind speed affects ground speed, position, and Euler angles.
    - Control surfaces (for pitch, roll, and yaw) affect ground speed position and Euler angles.

The reasoning of the proposed diagnostic decision tree is a natural extension from the two previous assumptions. If any sensor reading used as input to the Math Engine (ME) is affected by a malfunctioning, we would expect to find all the calculated parameters from the ME to differ from those of the onboard sensors (due to the second assumption). Since the probability that this

“disagreement” is due to malfunctioning of all equipment on-board is extremely low (due to the first assumption), the more logical explanation is that one of the ME input parameter is wrong.

Figure 26 shows the proposed algorithm for diagnosing differences in readings of different equipment/sub-systems. It starts with a simple check of whether every sensor’s reading of the Primary System (PS) is similar to that of the Redundant System (RS) and that of the ME. Readings are considered similar if the error is within a tolerated value that can be set appropriately. If all readings are similar, the confidence that everything is working fine increases. Nevertheless, this could be a false negative, as the network might have failed to detect anomalies in an equipment reading. Therefore, a more expensive test is performed to check for false negative which was accomplished by adding “is exceedance” checks, in which the readings from equipment are compared with the recommended operation levels taken from BADA.

However, if the readings differ, then the next observation that has to be noted is the proportion of disagreed cases that have been detected. If only one case of disagreement is detected, it is most likely that the ME output is true (this is evident by ME = PS = RS for the other parameters). The reading that is in disagreement with the calculated ME value is highly likely to be the culprit.

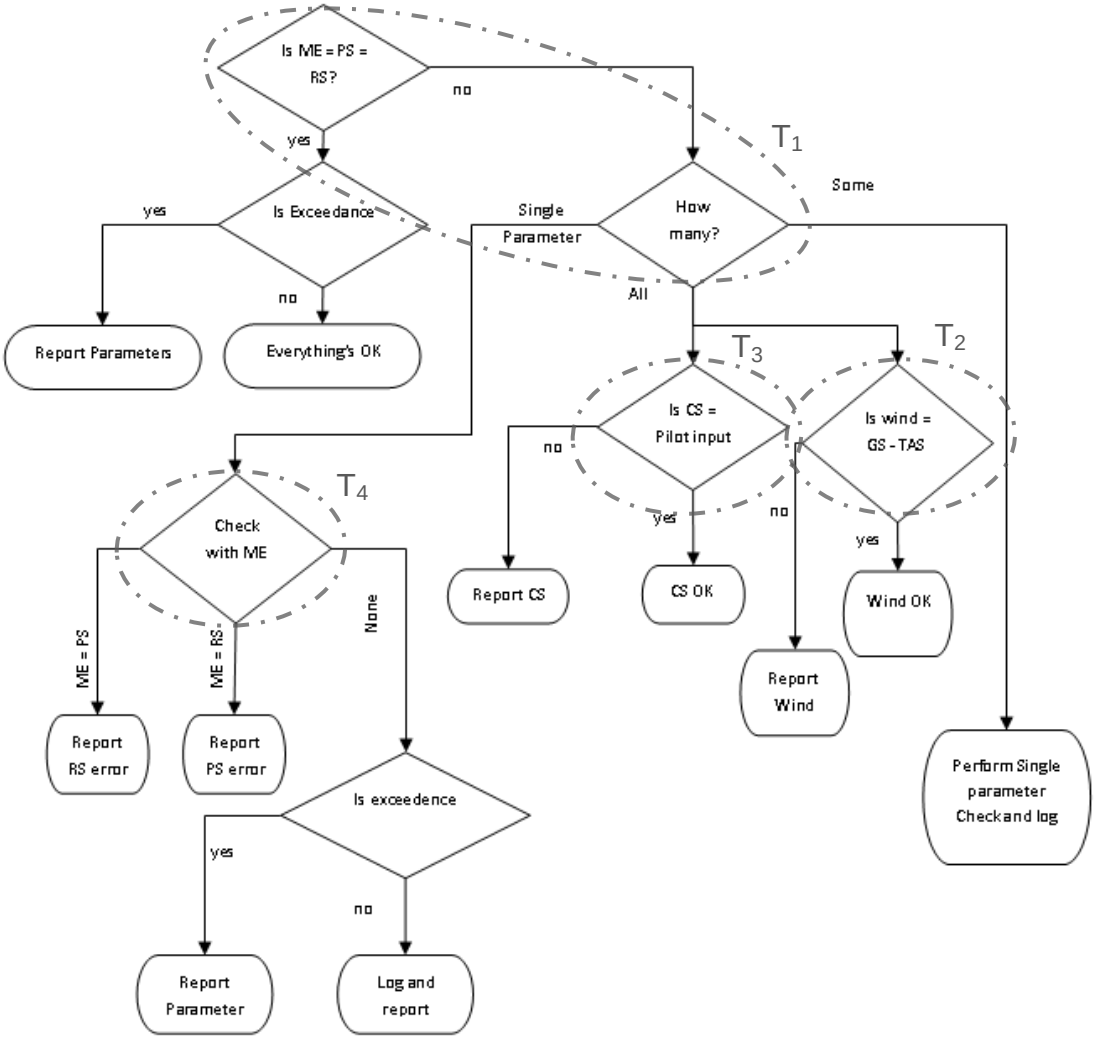


Figure 26. The proposed investigation engine

However, all of the readings differing from each other leads to the conclusion that one (or two) of the inputs to the ME are incorrect. A check of all the input parameters of the ME is needed. It is possible to check the Control Surface (CS = pitch, roll, yaw) by comparing its values with those extracted from the pilot CS input. Checking the other parameters to make sure they follow the same procedure is not implemented in this study. For example, it is fairly easy to check if the wind speed is measured correctly by this simple equation:

$$\text{Wind} = \text{GS} - \text{TAS} \tag{136}$$

Where GS is the ground speed, and TAS is the true air speed. Figure 27 shows a block diagram of the proposed system.

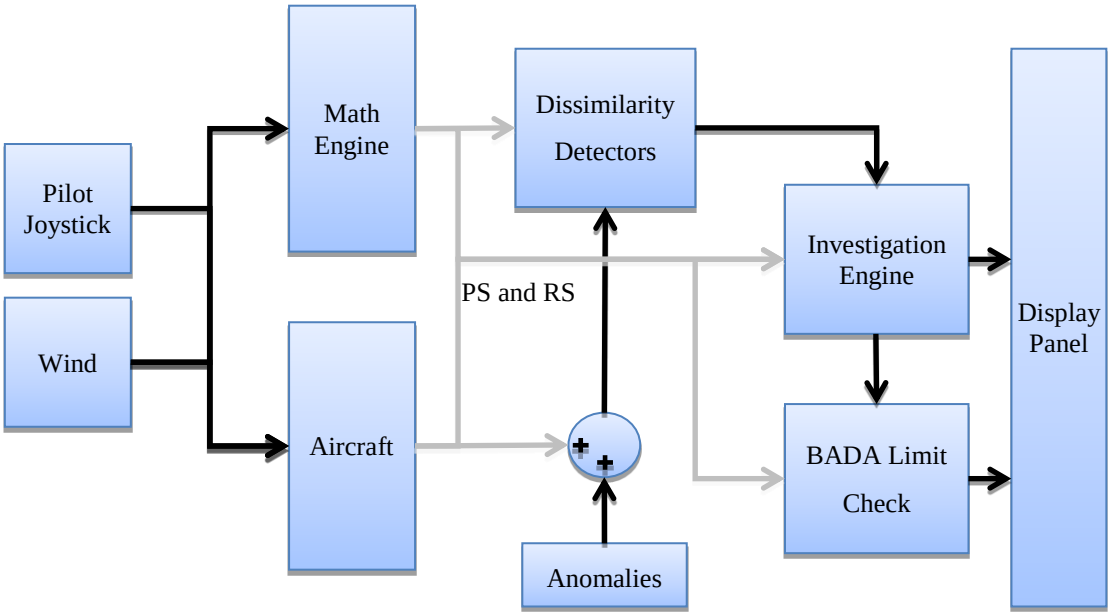


Figure 27. Block diagram representation of the proposed network

4.2.6 Mathematical formulation and analysis

In this section, the Bayesian network equivalent of Figure 26 is developed. In addition, we will derive the mathematical formulation of it. Since Figure 26 gives a diagnostic tree view of the network, all we need to do is convert it to its Bayesian network equivalent. However, it is much easier to think of the proposed design as an inference-type Bayesian network because constructing a Bayesian network from the point of view of cause-to-effect is easier and more straightforward than the other way around, from effect-to-cause.

Let the aircraft vector of states be denoted as Θ and an individual state number *n* from within the vector as Θ_n. Additionally, let θ be the vector of input variables to the Math Engine and the simulated aircraft in Figure 27. There are four tests to perform, as shown in Figure 26, labelled T_1 to T_4 and shown with dashed oval shapes. T_1 has four possible outputs depending on the similarity between the vectors of states from PS, RS and ME. These are summarized in equation 137.

$$T_1 = f(\Theta_{PS}, \Theta_{RS}, \Theta_{ME}) = \begin{cases} n \\ n-1 \\ \text{a where } 0 < a < n-1 \\ 0 \end{cases} \dots n = no.of\ states \quad (137)$$

where f is some function that outputs the number of similar states of PS, RS and ME. It can be as simple as a threshold detector or as complex as a clustering algorithm. In this demonstration, a threshold detector is used. The number of similar states can either be as high as n, which indicates that the vectors of states are identical, that all but one are identical, that some are identical and some are not, or that they may all differ. T_2 and T_3 have yes/no outputs, whereas T_4 has three possibilities as in 138:

$$T_4 = g(T_1, \Theta_{PS}, \Theta_{RS}, \Theta_{ME}) = \begin{cases} PS \\ RS \\ none \end{cases} \quad (138)$$

With these assumptions in mind, the Bayesian network can be plotted as in Figure 29.

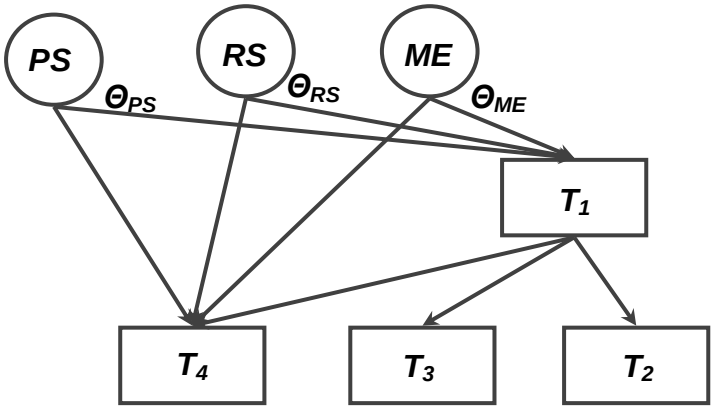


Figure 28. The Bayesian network equivalent of Figure 26.

The three circular nodes represent the output vectors of the primary system (PS), the redundant system (RS) and the math engine (ME). Their status can be either healthy or faulty and the way the status would be determined is through the four test nodes. Test nodes are action nodes, but since tests can have probabilistic results conveying the accuracy, or the error percentage, they could also be considered as probabilistic nodes. The Bayesian equation for Figure 28 is:

$$\begin{aligned} &\mathrm{P}(\mathrm{PS}, \mathrm{RS}, \mathrm{ME}, \mathrm{T}_1, \mathrm{T}_2, \mathrm{T}_3, \mathrm{T}_4) \\ &\quad = \mathrm{P}(\mathrm{T}_1|\mathrm{PS}, \mathrm{RS}, \mathrm{ME})\mathrm{P}(\mathrm{T}_2|\mathrm{T}_1)\mathrm{P}(\mathrm{T}_3|\mathrm{T}_1)\mathrm{P}(\mathrm{T}_4|\mathrm{T}_1, \mathrm{PS}, \mathrm{RS}, \mathrm{ME}) \\ &\quad \times \mathrm{P}(\mathrm{PS})\mathrm{P}(\mathrm{RS})\mathrm{P}(\mathrm{ME}) \end{aligned} \tag{139}$$

The evidences available to us are the results from performing tests T_1 to T_4 and the probabilistic query would regard the status of primary or secondary

systems. Taking a query for the status of the primary system as an example, we can write:

$$P(PS|\mathrm{T}_1,\mathrm{T}_2,\mathrm{T}_3,\mathrm{T}_4) = \frac{P(PS,\mathrm{T}_1,\mathrm{T}_2,\mathrm{T}_3,\mathrm{T}_4)}{P(\mathrm{T}_1,\mathrm{T}_2,\mathrm{T}_3,\mathrm{T}_4)} = \alpha P(PS,\mathrm{T}_1,\mathrm{T}_2,\mathrm{T}_3,\mathrm{T}_4)$$
$$= \alpha \sum_{\mathrm{RS}}\sum_{\mathrm{ME}} \mathrm{P}(\mathrm{PS},\mathrm{RS},\mathrm{ME},\mathrm{T}_1,\mathrm{T}_2,\mathrm{T}_3,\mathrm{T}_4) \quad (140)$$

Substituting equation 139 into 140 yields:

$$P(PS|\mathrm{T}_1,\mathrm{T}_2,\mathrm{T}_3,\mathrm{T}_4)$$
$$= \alpha \sum_{\mathrm{RS}}\sum_{\mathrm{ME}} \mathrm{P}(\mathrm{T}_1|\mathrm{PS},\mathrm{rs},\mathrm{me})\mathrm{P}(\mathrm{T}_2|\mathrm{T}_1)\mathrm{P}(\mathrm{T}_3|\mathrm{T}_1)\mathrm{P}(\mathrm{T}_4|\mathrm{T}_1,\mathrm{PS},\mathrm{rs},\mathrm{me}) \quad (141)$$
$$\times \mathrm{P}(\mathrm{PS})\mathrm{P}(\mathrm{rs})\mathrm{P}(\mathrm{me})$$

Equation 141 gives the exact inference formula to estimate the status of PS, giving the results of the four tests. The same procedure can be followed to infer the status of RS. Any of the approximate inference algorithms discussed in chapter 2 can also be used. However, the problem regarding the question as to where the numbers come from remains to be solved. The answer from a Bayesian point of view is to use subjective probability or, on the other end, to use statistical data to estimate the required probabilities. The approach adopted in this thesis is a hybrid of these two. It starts from a subjective degree of confidence to estimate an upper and lower bound that close down to the expected average of a variable as the amount of data increases. A decision-maker can initially assume that his/her tests are 100% accurate and

then revise the degree of confidence as collected data prove otherwise. In this demonstration, we follow the algorithms of determining the upper and lower bounds in chapter 3, algorithm 2 and that exact inference of equation 139 is used.

4.2.7 Experiment Set-up

The demonstration is conducted by setting up a simulation environment in MATLAB. The complete aircraft block from AerSim blockset is used to simulate an aircraft. The sensors and states outputs were labeled primary system (PS) and redundant system (RS), which represent a generalized way of identifying a reading from one of two independent sources, for example, a barometer or a GPS reading. Since the objectives of this study did not include investigating the systematic or environmental causes of malfunctioning equipment, but rather aimed to validate the readings, aircraft sub-systems (equipment) were treated as black boxes, and errors in equipment readings were simulated by the addition of random noise and/or by multiplying a reading by a scaling factor. The output of the investigation engine was monitored to determine if the faulty equipment in which the error was introduced was correctly detected.

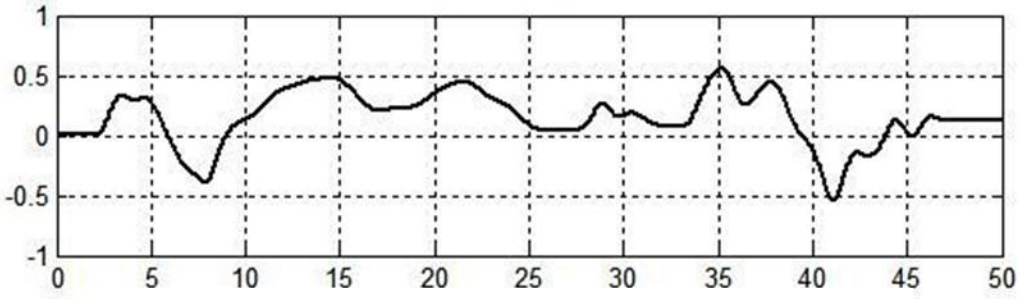
To test the operation of the network, scenarios have been created in which a malfunctioning equipment event was introduced while the output of the investigation engine was logged. The aim of these scenarios was to test the accuracy of the proposed investigation engine, and its ability to pin-point the faulty equipment whenever a fault was introduced. The first two simulations used a deterministic environment in which the investigation engine

operated at perfect (100%) accuracy and output was either 0 (for “no fault detected”), or 1 (for “fault detected”). The sampling frequency was 0.008 seconds. The purpose of the first two scenarios was to show the improvement of the developed FDD over the current procedures, whereas that of the third is to show the advantages of CP in comparison to the approaches of scenario 1 and 2.

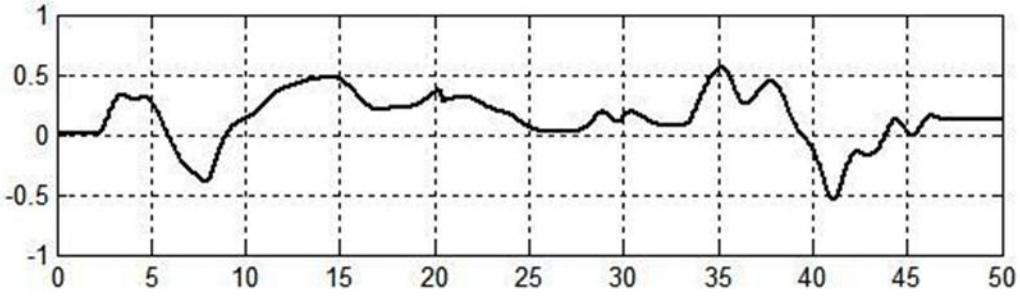
4.2.8 Scenarios 1: Fault in Primary System Pitch

The first scenario run was meant to test the operation of the network when simulating a malfunction in the sensor equipment responsible for

(a). Pitch (in rad) calculated by the Mathematical Engine (ME).



(b). Pitch (in rad) from the Primary System (PS) sensor.



(c). Output of the investigation engine, port: Fault in PS Pitch

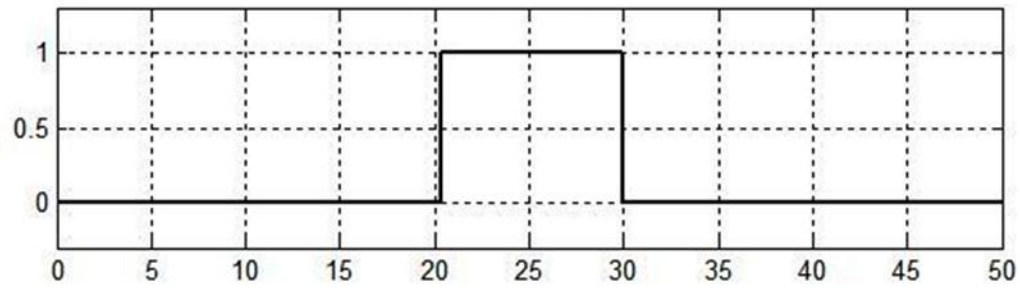


Figure 29. Results of scenario 1

displaying the current pitch attitude. The simulation time was set to 50 seconds, and throughout the simulation, the value of the aircraft pitch attitude was constantly changed by means of a pilot joystick. Figure 29 (a) shows the theoretically calculated pitch altitude, while (b) shows a graph of the pitch altitude sensor's reading. During time 0 to 20.5 seconds, the two values resembled each other and the investigation engine's port: Fault in PS Pitch was zero (figure 29 (c)). However, when a malfunction was introduced into the primary system's pitch sensor at time t=20.5 seconds, the investigation engine was successful in pin-pointing the faulty equipment. The fault was held for 10 seconds, during which the investigation engine's output port "Fault in PS Pitch" stayed at 1, producing a positive result for the test scenario.

4.2.9 Scenarios 2: Fault in Primary and Redundant Speed Sensors

The second scenario demonstrates the superiority of the network over current systems when using two sensors to independently calculate the same physical quantity. Once again, the simulation was set up to run for 50 seconds, and Figure 30 (a) shows the theoretically calculated airspeed values against time. Figure 30 (b) and (c) shows the airspeed's sensor reading on the Primary System PS (e.g. the pilot panel) and Redundant System RS (e.g. the co-pilot panel) respectively. Before time t=20 second there was no malfunction simulated, so the three graphs of ME, PS, and RS airspeed readings were the same. After time t=20 second, a malfunction was simulated on both the PS and RS sensors so as to indicate the same incorrect reading. Routinely, such faults might be hard to detect as, for example, both the pilot and the co-pilot would each confirm the same reading ignorant of the presence of a

malfunction in both systems. Since the mathematical engine relies on equations to estimate the correct airspeed value, it will report a dissimilar airspeed value which, in turn, is supplied to the investigation engine to identify

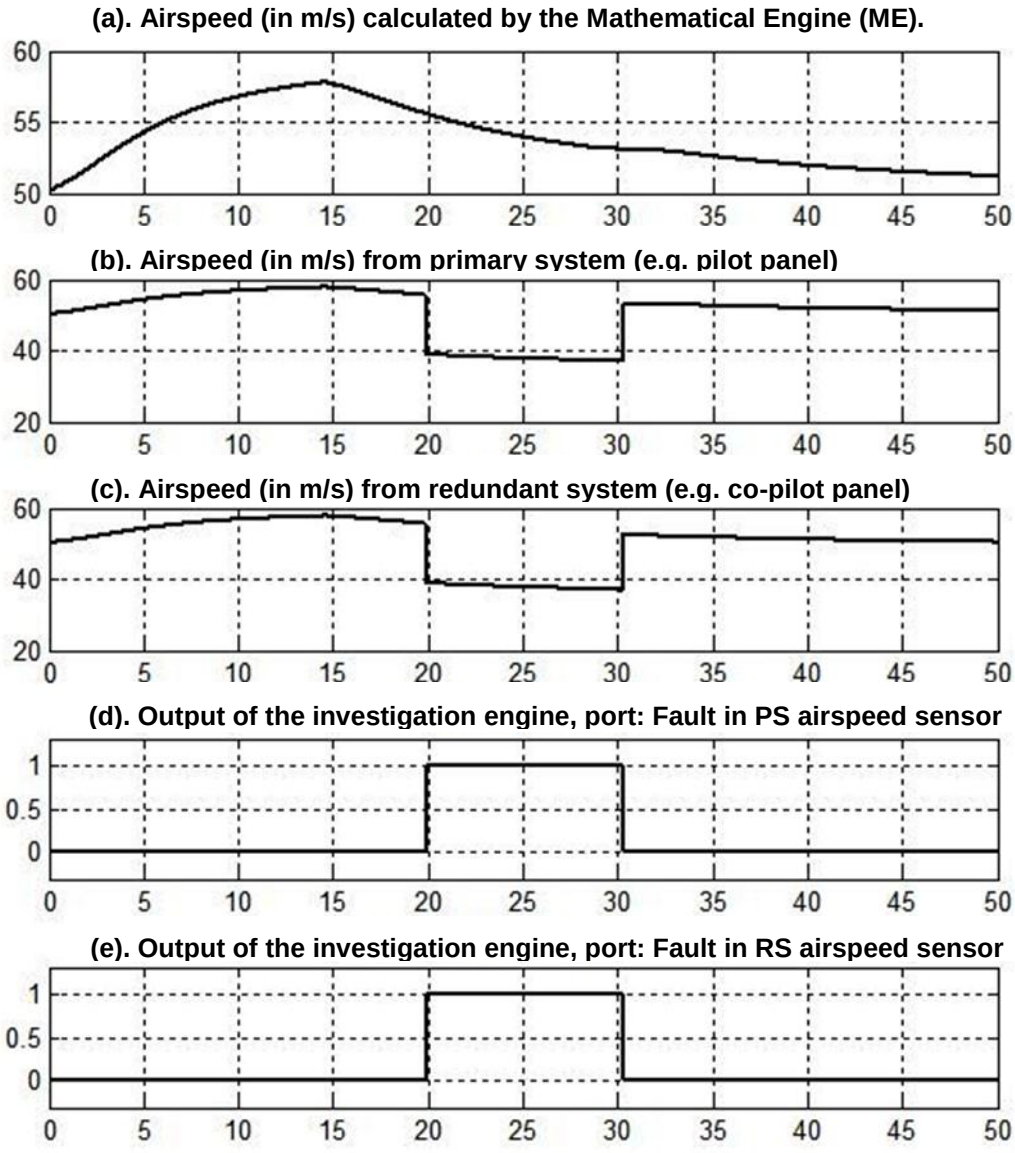


Figure 30. Results of scenario 2.

the source of the malfunction. Figure 30 (d) and (e) show output ports “fault in PS airspeed sensor” and “fault in RS airspeed sensor” as changing from 0 to 1 during the time of the fault indicating a successful diagnosis.

4.2.10 Scenario 3: Faults in more than single equipment

The two previous scenarios demonstrated the performance of the FDD algorithm but under single sensor/equipment fault or more than a sensor measuring the same quantity. The performance of the system from a deterministic and probabilistic point of view can be considered identical if the degree of confidence in the test results is high. The conclusion of at most a single fault followed our assumption that the aircraft is in good condition. However, modern aircraft have many complex systems measuring many quantities concurrently and it may be absurd to assume a single fault at any given time. Hence, the aim of scenario 3 is to record the performance of the network under two or more faults from two or more different equipment. We have already demonstrated, in scenario 2, the advantages of using the math engine as a third independent source of information in triple checking the health of equipment. However, if we assume a deterministic diagnostic tree, then we may not be able to deterministically pinpoint the faulty equipment. To see this, assume that the registered speed from the primary sensors was 250 m/s and from the secondary sensor was 250m/s, while the math engine indicated that it should have been 180m/s. In this case, we will be inclined to belief the data from the math engine. But suppose a second sensor group also registered the same phenomenon, where PS=RS$\neq$ME. Once more, we may be inclined to consider the math engine's calculated value as accurate, but if that continues then at a certain point we may start to realize that ME is the culprit and that PS=RS simply obtains because they are correctly measuring

what they are measuring. Hence, when we have two or more dissimilarities, we will not be absolutely sure about the source of the fault.

An ordinary probabilistic approach may not perform well because it requires the conditional/unconditional probability tables for all of the tests in the network. Since the proposed design is essentially novel, such tables are not available. In addition, equipment performance data are not standardized in any known way for the use of a sample and the labelling of typical data. Hence, CP approach seems more suitable than any other.

The simulation was carried out using the Bayes Net Toolbox for Matlab developed by Kevin Murphy [154] for its ability to make exact and approximate inferences. Firstly, the system was run for an hour without introducing any fault of any type to calibrate for any false positives that may result from processing discrepancies between Airsim, ME, and Bayes Net toolboxes. The run resulted in nine false positives from where the aircraft was at initialization and from times at which a manual hard roll was issued. The initialization phase resulted in false positives because the aircraft was initialized at a certain altitude, velocity and orientation, but the internal states, momentums, power settings, throttle, and control surface were not yet set to yield an aircraft in the initialization position, so the aircraft would oscillate for a short time until the simulation stabilized. The false positives were removed from the scenario results. Afterward, the aircraft was set to run in *holding* circulating above runaway for about four hours, during which faults were introduced and the degree of confidence in the reported fault location logged, as in Figure 26. Six pieces of equipment from each subsystem were monitored (speed, rate of

climb/ descend, altitude, pitch, roll and yaw). Table 2 shows a summary of the results.

Table 2. Simulation results of scenario 3

Time (HH:MM:SS)	No of simulated faults	number of Detected faults	Accuracy	Weighted average degree of confidence
00:10:00	1	1	100%	100%
00:15:00	2	2	100%	91.17%
00:16:00	3	3	66.67%	62.4%
00:17:00	4	2	50%	39.98%
00:18:00	5	1	20%	40.1%
00:19:00	6	6	100%	99.9%
01:10:00	1	1	100%	99.8%
01:15:00	2	2	100%	94.4%
01:16:00	3	3	66.67%	60.04%
01:17:00	4	4	75%	51.8%
01:18:00	5	3	60%	77.7%
01:19:00	6	6	100%	99.8%
02:10:00	3	3	66.67%	59.57%
03:10:00	3	3	66.67%	54.04%
04:10:00	3	3	100%	66.67%

The number of simulated faults increased from one to six at an interval of 1 minutes starting after 15 minutes of the simulation start time. An accuracy of 100% indicates that the FDD network has successfully pinpointed the source of the fault, whereas an accuracy of 66.67% when three faults were simulated

indicates that the FDD network managed to identify only two correct faults. It is quite clear that the network is weaker when the number of simulated faults is around three, which is half the total number of simulated sensors. However, the performance begins to increase slightly with time, as we see during the third and fourth hour, when the only focus of simulated faults was on the weaker case. Although scenario 3 provides a proof of concept for how CP behaves under a sparse amount of data, the result does not completely prove it. This is the case because the simulation was not run for days – or even weeks – to ensure that the system will still perform well, but, as the purpose of the proposed technique is short-term usage during the first initialization hours, the non-necessity of such a scenario is justified. However, due to the probabilistic nature of the proposed technique, the simulation should be re-run many times and the weighted average of the calculated probabilities should be recorded. Due to limitations of time and resources, however, this was left for a future extension of the demonstration.

4.3 Demonstrating an On-board Navigation Decision Support System using BADA

The second demonstration of this chapter is aimed at creating an online, real-time and onboard DSS that can help pilots navigate better, understand the bigger picture and enhance the results of the FDD of the

previous demonstration, on which it is built. Once more, the aim of the demonstration is to show successful and meaningful recommendations in real time as soon as the system initializes. The underlying DSS utilizes BADA operation data as a navigation support system in the sense of estimating efficient ranges of operation for speed, rate of climb or descent (ROCD), and fuel flow in terms of flight level, current speed, mass, and flight profile. It estimates such ranges of operations and uses probabilistic reasoning to calculate the beneficial value of the estimated ranges based on the reliability of equipment readings.

4.3.1 BADA Database Overview

BADA is a collection of text files that lays down operation performance parameters, airline procedure parameters and performance summary tables for more than 300 aircraft types [97]. It was developed and is maintained by the European Organization for the Safety of Air Navigation (EUROCONTROL). The information contained in these files was obtained using the mass-varying kinetic approach to aircraft modelling. It models an aircraft as a point along with the underlying forces acting upon it which causes its motion. Figure 31 [155] shows the structure of the BADA Aircraft Performance Model (APM).

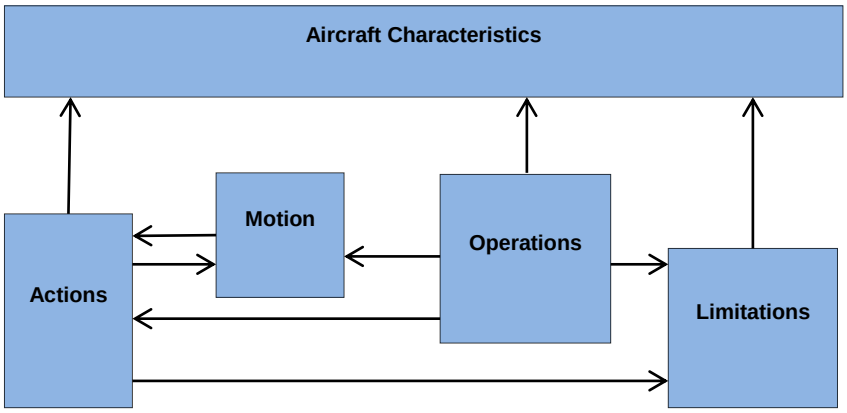


Figure 31. Structure of BADA APM

The model is organized into five sub-models, namely: characteristics, actions, motion, operation, and limitations. The arrows represent the dependencies between the sub-models. The actions sub-model is used to calculate various forces acting on the aircraft, whereas the motion sub-model accounts for geometrical, kinematics, and kinetic parts of motion using the Total Energy Model (TEM) method. The operations sub-model is used to simulate the different operation modes of an aircraft, such as flying with a constant Mach number. The limitation sub-model mimics the operational limit of the aircraft such as the maximum altitude, throttle limit, and maximum airspeed. Finally, the characteristics sub-model contains coefficients that characterize an aircraft such as the wing span [152]. Each modelled aircraft is parameterized into three text files. First, an Operations Procedure File (OPF) holds aerodynamic constants such as thrust, fuel, and drag coefficients. Second, an Airlines Procedures file (APF) contains parametric information about the recommended speed procedures during different flying phases, and third, a Procedure Table File (PTF) represents the recommended operation procedures in the form of look-up tables [152]. This demonstration utilizes the

look-up table of a specific aircraft that is contained within a PTF file as it gives the user direct access to performance data without the need to implement the complete TEM [97], which in turn reduces the complexity of the developed network.

The PTF file contains the recommended operating producers for airspeed, rate of climb/descent (ROCD), and fuel flow at different flight levels of a specific aircraft. (An example of a PTF file can be found in [156].) The header section of the PTF file specifies general information about the type of the aircraft, creation date, speeds, temperature data, maximum altitude and mass levels. This is followed by the table of performance data, where the operation information is organized into three sections: cruise, climb and descent [152]. In this demonstration, a script was written to subtract the performance table of a PTF file of a specific aircraft and organize the data in a look-up table that is more suitable for analysis by MATLAB. In addition, the script verifies the validity of the PTF file by checking for the presence of some permanent text within the header section of the file.

4.3.2 Assumptions and Proposed Design

We propose a framework that facilitates the base of aircraft data (BADA) as a navigation planning decision support system for pilots to make informed decision about navigation planning. The decision support system is implemented as a software tool to extract performance data of an aircraft type from BADA database, integrate with other on-board fault detection and isolation systems, and estimate the beneficial value of these recommendations. The designed network presented in this section is an

extension of the diagnostic decision trees described in section 4.2. Figure 32 shows the structure of the proposed design.

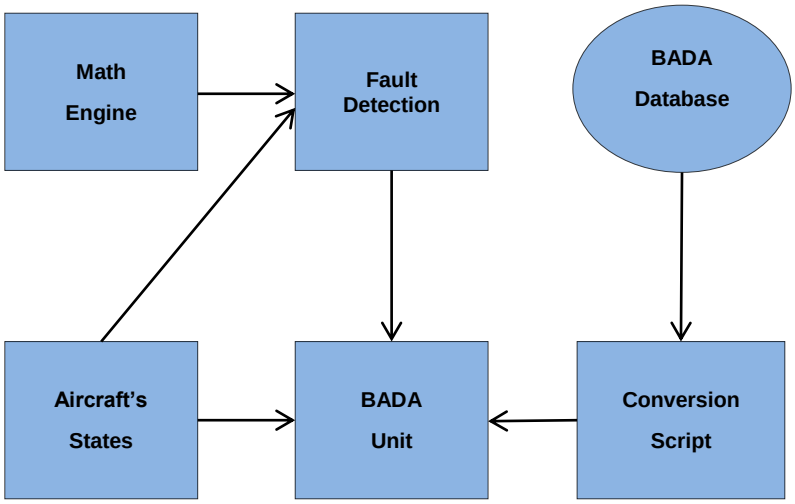


Figure 32. Structure of the proposed design

To ensure that the BADA unit receives navigation readings of high accuracy, a check for equipment faults is added to the network. The fault detection algorithm follows a model-based approach (see section 4.2) whereby the measured aircraft states are verified with a reading that has been calculated using a simulated aircraft running in parallel to the actual aircraft. The simulated aircraft uses the high-fidelity 6 Degrees of Freedom (6DoF) model to simulate displacement and rotation of an aircraft in three-dimensional space. The 6DoF model is contained within the Math Engine Unit (MEU). The states from the aircraft and the math engine block are applied to the Fault Detection Unit (FDU), where a state of no fault is assumed if the two data sources are identical. Otherwise, the FDU will start a diagnostic procedure to isolate the malfunctioning equipment using Bayesian diagnostic decision trees. As a result, the BADA Unit (BU) can select the most reliable source of

data to be compared to the recommended operation records, which were in turn obtained from the BADA database unit throughout the converter script.

The BU recommendation algorithm begins by checking for sources of reliable readings and the utility associated with making the recommendation. If no reliable information could be obtained, the BU will display a warning message informing the pilot about the situation. Otherwise, the BU begins by detecting the flight phase of the aircraft (i.e. cruise, climb, or descent). BU uses the value of the angle of attack, landing gear position and ROCD to detect the flight phase. Low ROCD and low angle of attack along with landing gear at the up position would indicate cruising phase. Otherwise, the aircraft is either climbing or descending. In order to detect which of these the aircraft is in, BU uses the ROCD values of the climb and descent from the PTF file as a feature search space. The nearest five neighbours to the current ROCD of the aircraft is calculated using the k-nearest neighbour algorithm, then the flight phase is determined based on basic majority vote. To reduce the amount of calculation, BU can be programmed to treat negative ROCD as descending indicator and positive ROCD as climbing indicator, which might be beneficial in situations in which no PTF file is available. When the flight phase is known, the corresponding look-up table for that specific phase is selected to obtain the recommended procedure data. If the measured aircraft states are not within the tolerated limit of the values recommended by BADA, the BU will inform the pilot about the situation and recommend changing his/her navigation parameters accordingly.

4.3.3 The Utility of the Recommendations

Maintaining high accuracy navigation requires the pilot to be in a state of high situation awareness where he or she can evaluate the fidelity of the aircraft's equipment readings and detect cases when an equipment reading is unreliable. It has been suggested that the highest level of situation awareness can be achieved by a thorough grasp of some key elements that, if put together, will synthesize the prevailing status of an environment [7]. Therefore, it is valuable to measure the benefit of giving some recommendations regarding operation procedures as the information used to derive the recommendation itself could be unreliable. The measured benefit would serve as a criterion to decide whether a specific recommendation would increase the pilot's situation awareness about his/her environment and, in turn, display that recommendation, or that the measured readings are unreliable to the degree that no recommendation is possible. To decide which situations would not be beneficial with respect to increasing situation awareness, we have used the principle of maximum expected utility in probabilistic theory, where each decision is associated with a utility function that represents the cost or benefit of making some decision. Figure 33 shows the structure of the decision network.

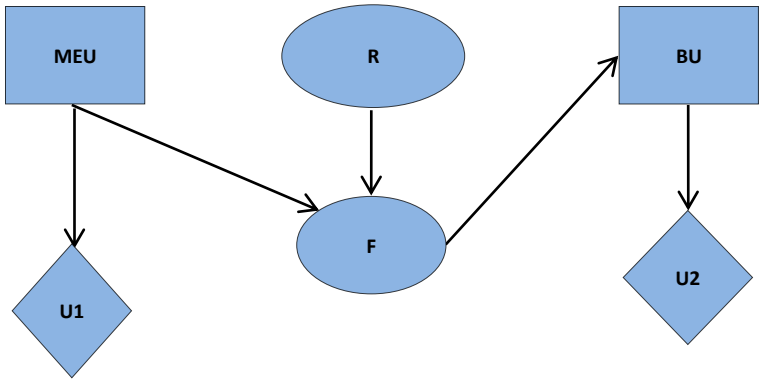


Figure 33. Structure of the decision network

Probabilistic variables are represented with the oval shapes labelled R and F. The variable R represents equipment reading whereas the variable F denotes the probability of detecting a faulty reading. Action (or decision) nodes are depicted in the rectangular shapes MEU and BU. The decision associated with each node is whether to go down the road of executing the block. Finally, the diamond shaped blocks U1 and U2 represent the expected utility associated with making decision MEU and BU respectively. Using equation 70 (from chapter 2), the expected utility of making the decision U1 is:

$$\mathrm{EU}_1(MEU|F) = \sum_{\mathrm{i}} \mathrm{P}(\mathrm{O}_{1\mathrm{i}}|, MEU)\mathrm{U1}(\mathrm{O}_{1\mathrm{i}}|MEU) \tag{142}$$

Similarly, the expected utility of making the decision U2 is

$$\mathrm{EU}_2(MEU|F) = \sum_{\mathrm{i}} \mathrm{P}(\mathrm{O}_{2\mathrm{i}}|, MEU)\mathrm{U2}(\mathrm{O}_{2\mathrm{i}}|MEU) \tag{143}$$

Using the product rule of probability calculus, equation 142 and 143 can be written as:

$$\mathrm{EU}_1(MEU|F) = \sum_{\mathrm{i}} \frac{\mathrm{P}(\mathrm{O}_{1\mathrm{i}}|, MEU)}{\mathrm{P}(\mathrm{F}, \mathrm{MEU})}\mathrm{U1}(\mathrm{O}_{1\mathrm{i}}|MEU) \tag{144}$$

$$\mathrm{EU}_2(MEU|F) = \sum_{\mathrm{i}} \frac{\mathrm{P}(\mathrm{O}_{2\mathrm{i}}|, MEU)}{\mathrm{P}(\mathrm{F}, \mathrm{MEU})}\mathrm{U2}(\mathrm{O}_{2\mathrm{i}}|MEU) \tag{145}$$

Thus, the overall expected utility is

$$\begin{aligned} \mathrm{EU} &= \mathrm{EU}_1 + \mathrm{EU}_2 \\ &= \sum_{\mathrm{i}} \frac{\mathrm{P}(\mathrm{O}_{1\mathrm{i}}|, MEU)}{\mathrm{P}(\mathrm{F}, \mathrm{MEU})} \mathrm{U1}(\mathrm{O}_{1\mathrm{i}}|MEU) \\ &+ \sum_{\mathrm{i}} \frac{\mathrm{P}(\mathrm{O}_{2\mathrm{i}}|, MEU)}{\mathrm{P}(\mathrm{F}, \mathrm{MEU})} \mathrm{U2}(\mathrm{O}_{2\mathrm{i}}|MEU) \end{aligned} \tag{146}$$

Since reliable recommendation follows the presence of reliable equipment readings, the utility of executing the BU will depend on the presence of a reliable source of speed, flight altitude, etc. In other words, the probability of detecting faulty reading F should be low. Since airspeed in modern aircraft is obtained from one or more inertial navigation subsystems and one or more navigational reference subsystems, we can assume U2 to be 0 if all airspeed data sources were unreliable and 1 otherwise. Furthermore, if there is (*n*) identical equipment (for instance: pitot probes), the subjective unconditional probability of faulty equipment (X) is:

$$\mathrm{P}(\mathrm{X} = \mathrm{fault}) = \frac{1}{\mathrm{n}} \tag{147}$$

4.3.4 Experiments Simulations

Once again, the demonstration is conducted through setting up a simulation environment in MATLAB. To test the operation of the network, scenarios have been created in which an event of equipment malfunctioning is introduced while the output of the BU is logged. Scenarios represent test data that can be used to validate system design requirements [157]. In the context

of aviation safety, the use of scenarios as test data has been widely proposed and well documented in the literature as a means of measuring the compliance of a design to the requirements of safety standards within civil aviation sector [158]. In order to generate a graphical presentation of the BU instead of a text-based recommendation, five output ports have been added to the BU representing the five possible state/recommendations that can be given by the unit. The first port states the detected flight phase represented in numeric format in which the numeric value 0 is used to represent the climbing phase, 1 to represent cruising, and 2 to represent descending. The second port states the availability of reliable airspeed reading, in which 0 represents that availability and 1 represents no reliable airspeed data. The following three ports are ROCD, True Airspeed (TAS), and fuel flow recommendations respectively. Each one of these ports can take the value 1 to represent information that is out of the BADA recommendation limit, 0 to indicate information that is within the recommended limit, or 2 to indicate that the information is not available yet. Figure 4 shows a block diagram of the experiment setup. All of the scenarios described below were for the Boeing 737 aircraft with an initial flight level (FL) of 75 and TAS of 300 knots simulated by the JSBSim library of aircraft models.

4.3.5 Scenario 1: Fuel Flow exceeding normal limit

The first scenario that was set up in this demonstration was to simulate events when fuel flow exceeded the value recommended by BADA. The aim of this scenario was to validate the ability of the network to detect anomalous fuel flow rate. Anomalous fuel flow rate could be a sign of a much more

hazardous engine problem or simply an unnecessary added cost. The events of this scenario were obtained by changing the aircraft angle of attack as to change the flight phase from climbing to descending. Anomalies to the amount of fuel flow to an engine were constantly introduced during the scenario while the outputs of the BADA unit were recorded. The simulation time was set to 20 seconds. Figure 34 shows the simulation results. Figure 34(a) shows the malfunctioning fuel flow graph of scenario 1. At FL of 75 and TAS of 300 knots, the recommended fuel flow should be around 210 kg/min during climbing phase and about 32 kg/min while descending. Since both values were significantly deviated from those recommended values, BU output port signalled the value 1 to indicate that the information fell outside of the recommended operation limit. This verifies the principle of operation of the unit.

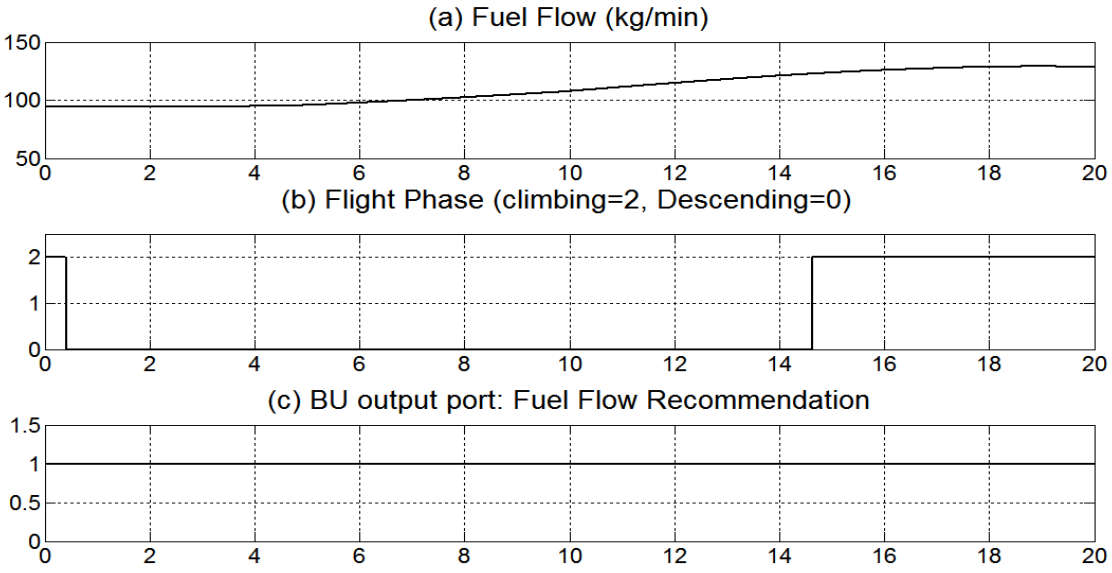


Figure 34. Simulation results of Scenario 1.

4.3.6 Scenario 2: No reliable Airspeed data

The second scenario demonstrates the case in which the airspeed measuring equipment is unreliable. The scenario was designed to show the BU response to that worst case setup in which the airspeed readings received from both the inertial and navigation sub-systems are defective and the fault detection unit has identified both systems as malfunctioning. Figure 35 shows the simulation results.

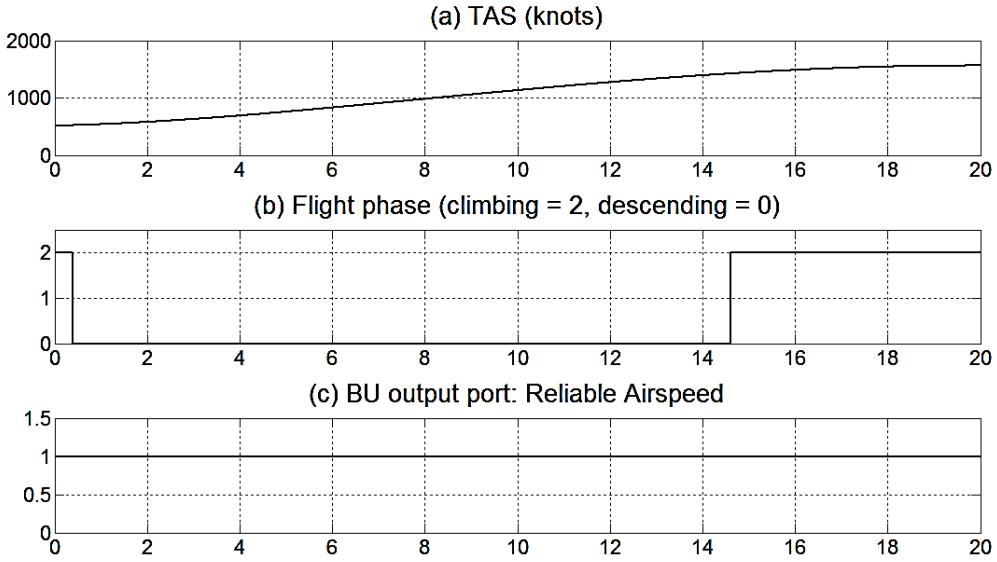


Figure 35. Simulation results of Scenario 2

Figure 35 (a) shows one of the malfunction airspeed equipment readings (the other equipment reading is identical). The value was dramatically higher than expected. Therefore, the BU indicated unreliable data to give a recommendation and halt operation. This result proves the validity of the decision-making network developed in this demonstration.

4.4 Summary

Ensuring a safe journey for passengers is of a high priority in the aviation industry. It is an aim that continues to motivate researchers to build more and more complex systems and procedures to enhance safety. Early detection of faults is a must-have in aviation, not just to ensure passengers' safety but also to decrease maintenance cost and to prevent faults from advancing to a stage where an intervention would be useless.

Chapter 4 presented a novel approach to FDD in aviation, which introduces an independent source of measurement that works concurrently with the aircraft systems to double check the validity of their status. The approach treats every piece of information it collects as doubtful until it is double-checked. We showed, with synthetic scenarios, how the proposed FDD compare to the state of the art methods and how it can be further enhanced using CP approach to a Bayesian network.

However, detecting faults is not all that is required for a safe flight. The pilot also needs to be aware of any problems, as well as their source and severity. In addition, too much information could overwhelm pilots and, in turn, slow down their responses. Hence, the information about faults and the status of equipment should be summarized and presented in an easily understandable form. Our second demonstration showed how a database developed for calculating trajectory could be modified to work as a DSS. Once

more, we showed the benefits of CP based Bayesian network in decision-making.

5. Application to Intensive Care Units

In the previous chapter, we saw how CP could be applied to aviation safety to enhance the performance of a fault detection and determination system. In addition, we saw an application of CP to decision-making by utilizing BADA as a DSS to draw better navigation plans. The results illustrated the feasibility of the CP for both problems. In this chapter, we will expand the application of CP to the field of medical informatics, more specifically to the monitoring of patients in ICUs.

Unlike the model-based approach of chapter 4, patients cannot easily be modelled. While it would be of great scientific value to model how the human body functions, this is often too complex to be feasibly accomplished. Not only do the functions and interconnections of organs need to be modelled, but also their interactions with foreign bodies such as germs. In addition, such requires modelling the response of an organ to a medicines, stimuli, or even surgical procedures. Moreover, the model should accommodate for all possible variations in the human genetic pool and racial traits. Consequently, a high fidelity model will be one that makes a suitable trade-off between representation power and flexibility. Experts will need to construct a model for

each age group, race, sex, and medical condition. Undoubtedly, such requirements are too broad to be achievable not only within the scope of the thesis but even for big research centres. Researchers have often chosen to follow the route of simplifying the model and accepting the loss of generality. The approach adopted in this thesis is a data-driven approach. In chapter 2, we have seen that the major setback of the Bayesian network was the fact that it is model-based and we saw how we could use examples from data to come up with a model. However, this approach requires some offline period in which the network learns to model itself from the examples before it is ready to make inferences. Another phase of training is also required to estimate the probability density functions of the various nodes within the network. The goal of this chapter is to design a novel system that is available immediately as the first data from a patient arrives and is consequently able to make decisions regarding patient care and to predict the future evolution of the patient condition while still in the ICU.

As with chapter 4, this chapter begins with a quick literature review of the available research in patient monitoring and patient state prediction. Then it will introduce the MIMIC II (the Multiparameter Intelligent Monitoring in Intensive Care) database. This is followed by a discussion of the proposed design and its mathematical formulations. Finally, the experiment setup and results will be provided.

5.1 Literature Review

The biomedical literature is full of research frameworks that adopted Bayesian networks to solve various problems. In fact, medicine is one of the most active application fields of Bayesian networks [6]. Since Bayesian networks are casually interconnected graphical models, they can be used to simplify the modelling process and to incorporate the experience of medical experts into the model through cause-to-effect interconnections. In ICUs, BN have been used to diagnose the cause of observed symptoms, to make future prediction about the state of a patient, and to monitor the stability of patients' vital signs [6]. Classically, the BN research frameworks in biomedical engineering have been dominated by an expert knowledge approach where the expertise of medical practitioners are used to construct the model [6]. One example of such an approach is the ALARM (A Logical Alarm Reduction Mechanism) network [159]. The ALARM network is a diagnostic BN designed as a DSS that outputs messages to provide information about possible problems. It has 8 connected diagnoses, 16 findings and 13 hidden variables [159]. Despite the popularity of the ALARM network, it does not provide a means by which it can be generalized or adopted to other problems [6]. In addition, it is not quite known how the network would perform when only a portion of the parameters is known or if they have been measured irregularly. Finally, since the ALARM network is a static BN, it cannot display the temporal evolution of patients' statuses over their staying period in ICU.

Newer approaches, such as the BN binary classifiers by Sierra and others, use data-driven or a hybrid model and data driven approaches [160]. They used a genetic searching algorithm to find the optimal structure of a Markov Blanquet BN that can classify ICU patients according to their survivability prognosis [160]. While the approach seems sound, the network would need an offline phase during which the training examples are batch applied to it until it converges to an optimal solution. This requirement sets this framework outside the objectives and aims of this thesis.

Ramon and others have compared four data mining algorithms to predict the progress of patients mortality risk in ICUs [161]. The four methods were Decision Tree Learning (DTL), First Order Random Forests (FORF), Naive Bayesian networks (NB) and Tree Augmented Naive Bayesian networks (TAN) [161]. Their approach was to use the change in a monitored parameter value rather than the absolute value at a given time [161]. As a result, it is not clear how any of the algorithms can distinguish between a normal steady value of a stable patient and an abnormal steady one. For example, a steady heart rate of 72 may indicate that a patient is in a good and stable condition whereas a steady heart rate of 50 may indicate a problem. Their results showed the superior performance of the BN [161]. In fact, NB scored an accuracy of 85% as compared to the risk level assigned by nurses and physicians [161], which makes sense as NB is naturally structured for prediction problems. Once again, however, the approach involves a phase of training in which the system is not able to make predictions.

Another example of using a Bayesian network for estimating the risk of mortality is provided by Mu, Jaglal and Nylon [162]. They used data collected from about 13,000 patients who underwent cardiac surgery in six surgery institutions in Ontario, Canada [162]. They then used many potential risk-indicating factors to construct the network, including age, sex, left ventricular function, type of surgery, urgency of surgery and repeat operation [162]. The novel aspect of the study may be the six risk factors with which the study concluded. However, the study is not of much help for the aim of the thesis because it does not provide a generalized algorithm that can be applied to other problems with the framework of BN in ICUs.

Nonetheless, the DBN approach is not by any means a vacant one. Many research frameworks suggested DBN for patient monitoring and assessment in ICUs. Charitos and others have used DBN to construct a diagnostic network for ventilator-associated pneumonia (VAP) in ICU patients [163]. A DBN slice has 30 variables, of which 6 are input variables, 8 are observed and 16 are hidden [163]. Their DBN slice is actually an extension of a static BN developed by Lucas and others through expert knowledge [164]. They proved the validity of the network using Brier scoring of 20 patients only [164], which figure seems very low. In addition, the improvement of the DBN over the BN in terms of the averaged receiver operating characteristics, from which it was derived, did not seem significant enough.

Other approaches include feature extraction and clustering into discrete risk levels [165], logistic regression [166], neural networks [167], and fuzzy logic [168]. Most of these approaches require a phase of training or modelling

using expert knowledge. Since the structure is fixed once the training phase is over, the amount of data would be great to allow for training a network to predict the future state of patients from different age groups, sexes, medical backgrounds, race, and geographical regions. One way around this limitation is to allow the network to reconfigure itself in real time. The use CP can potentially solve the issue of data requirements because it needs less information to make decisions than any other method. In addition, due to the law of averages, CP could reach the correct estimation of the probability distribution function of a variable with time. In turn, CP provides real time learning from data as they arrive to the system.

Overall, the purpose of this section was not to provide an extensive survey of algorithms and methods in analysing ICU data but rather to establish a context of the use CP as an approach to Bayesian Networks. The work presented in this chapter does not necessarily contradict the BN frameworks surveyed in this section, as all the CP does is support the decision-making and the quantification of probabilities when the amount of information is sparse. The import of the proposed system tends more towards the end of showing the versatility of the methods proposed in this thesis by mean of examples. However, this is not to say that there are no novel contributions in the DBN proposed in this chapter. In fact, the novelty comes from proposing an architecture that would work best in various situations, would adapt to solve various problems, and would present the results of analysis in the most meaningful way.

5.2 The MIMIC II Database

One of the main objectives of this thesis is to facilitate the making of decisions with little available information and the real time evolution of the process as more information becomes available. However, it is essential for any research to have as much data as possible to validate the predictions of a new proposal. In addition, data could be used to achieve a better estimation of the unconditional probability that would otherwise become the result of purely subjective speculation.

Usually, biomedical research relies on data collected from hospitals during the study period or through a third party. The time limitations of this thesis restrict the feasibility of the first choice. The acquisition of clinical data is an involved process that requires ethical permission, anonymizing the data and cleaning the data. A better choice, from the time management point of view, is to use a third party database where most, if not all, of the ethical and technical procedures have been carried out.

One of the widely used databases is the MIMIC (Multiparameter Intelligent Monitoring in Intensive Care) database [169]. It has records of more than 31,000 admissions of more than 25,000 patients, of which around 20,000 are adults and 5,000 are neonates [169]. The data was collected in the Beth Israel Deaconess Medical Centre for about 7 years and is currently managed by MIT [169]. Typical researches conducted using the MIMIC database include predication of mortality rate in patients with kidney disease [170], retrospective comparative analysis [171] and artificial vector modelling [172]. An extensive list of publications can be found at the physionet website [173].

Essentially, the MIMIC database is comprised of two types of data: clinical data that are stored in a relational database and waveforms data stored in flat files [169]. However, only 3,000 patients has waveform data and only about 2,500 patients have their waveform data associated with the clinical data [169]. In addition, many patients' cases have missing data, noisy values and typo errors. Some researchers have developed algorithms to deal with the missing data issue, in particular, the waveform data [174]. Others suggested applying rules derived from medical experience [170]. While dealing with missing data is essential to increase the size of the sample of patients under study, it is not essential to this research. Instead, we chose to discard any patient if any of his/her records is missing.

Once the system has calculated the conditional and unconditional probabilities of typical patients' cases using as accurate and clean data as possible, they would be used in cleaning and replacing missing data using a linear Kalman filter, for instance. In an online scenario, the system can be used to distinguish between data that does not make sense, such as a disconnected monitor probe that could result in an apparent heart rate of zero. In this study, we will be using both the waveform and the clinical data to predict various aspects and parameters that are usually monitored in ICUs.

5.3 System Overview

As is the case with chapter 4, the patients monitoring system uses Bayesian networks to predict, diagnose and analyse the clinical data collected during a patient's stay in a hospital. However, since we are interested in the evolution of the patient's state over time, DBN is used to accommodate for the dynamic nature of the problem. Since DBNs are recursive probability density estimators, the system can learn from past events to progressively enhance its own prediction representation ability of a patient case.

The novelty of the system comes from several improvements to the state-of the-art DBN. Firstly, it is available as soon as it is initiated and there is no need for prior knowledge, although prior knowledge can be used to enhance the performance and accelerate the learning process. Secondly, it does not require an offline phase during which the information is batch processed by the system to calculate its internal parameters. Lastly, the system is an open platform. By that, we mean new parameters can be plugged into the system without the need to redesign the system from scratch. This means that each parameter is modelled separately and assumed to be independent of the others. Since most collected clinical data, such as blood pressure, heart rate and temperature are independent, the assumption is valid and applicable in many situations. Figure 36 shows a block diagram of the proposed system. For clarity, each sub-block is labelled with a number. The following is an explanation of each sub-block:

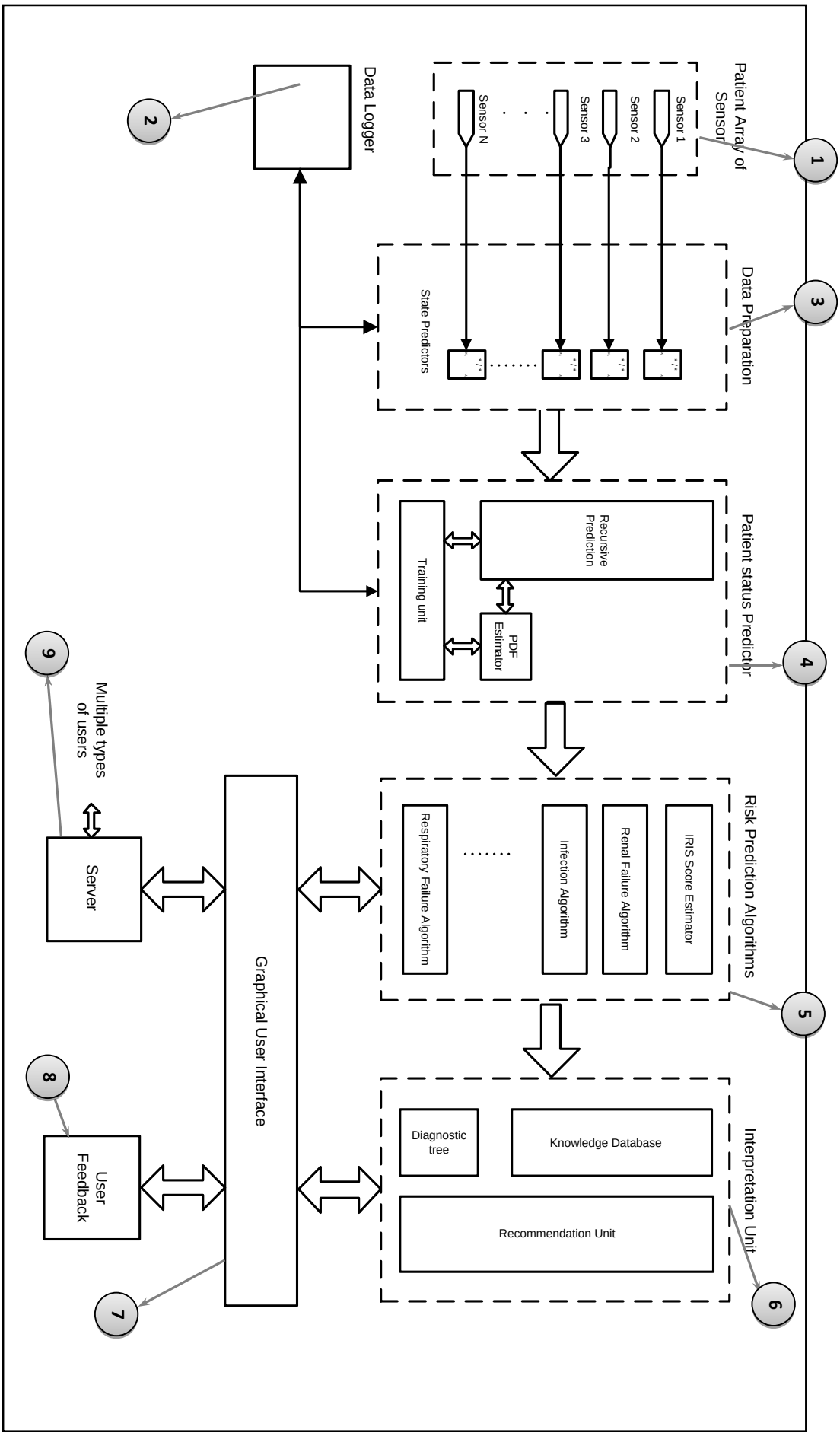


Figure 36. Overall block diagram of the system

1) **The patient's array of sensors** consists of several sensors that gather a patient's vital signs and/or other parameters into the system. Each sensor acts as an active listener tuned for a specific parameter. The sensors are to be seamlessly integrated into the current state of the art patients monitoring equipment, for example: heart rate, oxygen saturation, blood pressure…etc. However, in simulations, the data from patients are sequentially retrieved from the MIMIC database. The actual acquisition of patients' clinical data in realtime is left for future work in order to use the limited time of the study for building and refining the system itself rather than the acquisition of data.
2) **Data logger**. A good experimental set-up requires the gathering of as much information as possible to be analysed once the experiment concludes. Nonetheless, a system running in a production environment would also benefit from the logging of events and/or data for debugging purposes.
3) **State Predictor**. The output of the sensors unit is fed to a Bayesian state predictor, which works to find the most likely explanation for extraordinary band readings, which in turn reduces the rate of false positives. For instance, if the heart rate is logged as zero, it would be of great value to be able to tell whether this is because the sensor lid has dropped or because the patient's heart has stopped. The Bayesian state predictor uses a Dynamic Bayesian Network (DBN) algorithm to compare the current sequence of reading from a sensor with the most probable ones. If the readings do not match the most probable ones, then the sequence of readings is marked to be anomalous.

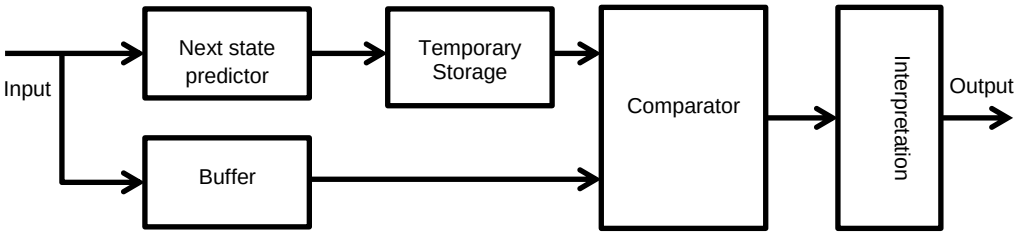


Figure 37. Data Preparation Unit

s by calculating the next most probable parameter value to be received from the sensors. This is done by calculating the conditional probability of all possible next parameter values given that the current one has occurred and taking the maximum of them. The same procedure is repeated (n) number of times and the results are stored in a temporary storage. Then the calculated probabilities for the next state parameter values are compared with the actual ones coming from the sensors. If the distance between the predicted next state and the actual next state for a sequence of (n) inputs is higher than the tolerance threshold, the state predictor will assume an abnormal reading, report it to the users, and log the events for further analysis. Otherwise, normal conditions are assumed.

4) **Patient Status Predictor**. The data acquired from the physical world is now ready for analysis. The block labelled “patient status predictor” uses the raw collected data to work out the probability density function (PDF) of every monitored parameter using a Bayesian recursive estimator in order to calculate the projected future value of each parameter after an adjustable amount of time.

5) **Risk Prediction Algorithms.** The system can automatically calculate the current and/or future score for different types of ICU scoring systems such as

IRIS, SAPS II, SAPS III, and APACHE II, although only the IRIS score is calculated during the simulations. The calculation of a score is conducted through two stages. A score predictor will calculate the most likely individual score of a parameter if the future score is required, and then these individual scores are combined in order to derive the overall score. The same procedure is carried out to predict other medical conditions such as renal failure, infection, or respiratory failure. Figure 38 shows a block diagram of the calculation of the IRIS score:

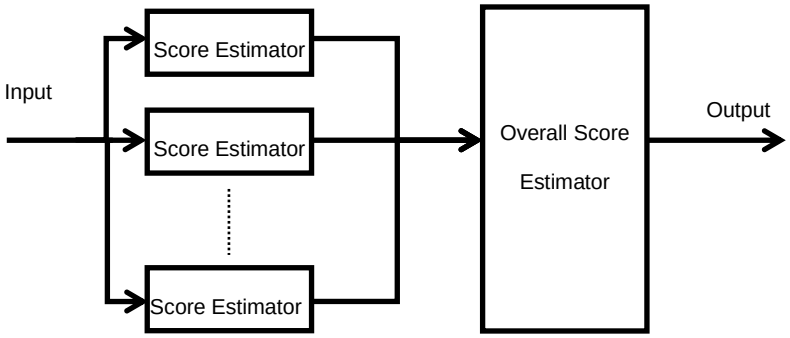


Figure 38. IRIS score calculation

6) **The Interpretation unit** uses the individual and/or overall score to display some recommendations based on a customizable lookup tables stored in a knowledge base. It can also be configured to display a recommendation based on a diagnostic tree derived from a database of extensive recorded patients' case studies. In addition, a user can define a set of recommendations on an individual patient basis, according to a medical class, or completely define custom messages from scratch. Finally, users can provide feedback from the feedback unit to better enhance the diagnostic tree accuracy in real-time.

7) **Graphical User Interface**. The patient's status can be monitored through the graphical user interface unit. It also provides the user with easy access to the configuration of the system and customization of the knowledge base and diagnoses as well as viewing of the raw logged date.

8) **User Feedback** allows users to provide feedback to the system. The feedback can be parametric, such as lab test results that could be needed to predict the future state of a patient, or nonparametric, such as the current diagnosis (or medical class) of a patient. All the information supplied by users is logged and used to enhance the operation of the network.

9) **Server.** All the results of the system such as the recommendation and predicted values are sent to a server where the data is stored in a secure database. Access to the data within the database is provided through a server-side script running under a web server. A patient's state can be accessed virtually from anywhere via the internet, if the user has the proper permission to do so. The results can be viewed on various types of devises including iPhones, iPads, tablets, PCs, and the like.

As stated previously, in order to reduce the computation power, not all of the blocks are simulated simultaneously. In the next sections, we will show two simulation setups showing the predictors in action along with their accuracy of predictions. The network shown in Figure 36 represents the extent of what could be accomplished by using CP-equipped Bayesian network, although CP is not necessary for the implementation of the system in Figure 36. If one can collect enough information to estimate the conditional probability tables required for the Bayesian network to operate, then CP is not required.

However, the very unavailability, or inadequacy, of such information justifies the use of CP.

5.4 Mathematical Analysis

The heart of the ICU monitoring system shown in Figure 36 is the predictor block, which calls for a good prediction algorithm. A good prediction algorithm is one that keeps a record of the current system estimates and updates it as new evidence is received [1,p. 571]. In this way, the algorithm becomes mathematically efficient, as it does not have to go back through time and do the calculations from the start all over again every time new evidence is gathered. Such an algorithm is referred to as a recursive estimator [1,p. 571]. DBNs can be used as recursive probability density estimators if a good temporal transition model is constructed. Let X be a hidden state variable that is to be estimated and let *e* be the available evidence on which X is to be estimated. Under the Markov assumption, it can be shown [1,p. 572] that X is given by:

$$P(X_{t+1}|e_{1:t+1}) = \alpha\, P(e_{t+1}|X_{t+1}) \sum_{X_t} P(X_{t+1}|X_t) P(X_t|e_{1:t}) \qquad (148)$$

Equation 148 is of the most importance because it shows that the current estimate of a variable is the product of the conditional probability of the current evidence times the likelihood of X on the basis of all the past evidences. The likelihood of the current X is simply a one-step prediction. Therefore, Equation 148 shows that the estimation of X involves updating its prediction by the

newly acquired evidence. Hence, the state of X at time (t+k) can recursively be predicted using the following equation [1,p. 573]:

$$P(X_{t+k+1}|e_{1:t}) = \sum_{X_t+k} P(X_{t+k+1}|X_{t+k})P(X_{t+k}|e_{1:t}) \quad (149)$$

Hence, the prediction of X at time (t+k) requires only the transition model of X [1,p. 573]. In addition, the arrival of new information will serve as a training hub that revises the current estimates of the model and keeps it up to date. This process is commonly known as filtering, which is the basis for estimating the likelihood of a sequence of evidence and for smoothing [1,p. 571]. One way to compute the likelihood of a sequence of evidences, shown in Figure 36 as state predictors, is to use Equation 148 to estimate X_t and then summing out X_t [1,p. 573]. However, it becomes mathematically inefficient as time passes [1,p. 573] so we opted to use the method described in the previous section, albeit without mathematical formulation.

Consider a monitoring system that utilizes four parameters and let the parameters be the heart rate (HR), the arterial blood pressure (ABP), the oxygen saturation (SO) and the respiration rate (RSP). These parameters will serve as the evidence on the basis of which the state of the patient (X) is inferred, which is represented by Markovian transition model. Figure 39 show how the DBN of such a system can be drawn. Assuming that E is the vector of evidence compromising HR, ABP, SO and RSP, the prediction of the state of the patient at time t+k can be found using Equation 149 as:

$$P(X_{t+k+1}|e_{1:t}) = \sum_{X_t+k} P(X_{t+k+1}|X_{t+k})P(X_{t+k}|e_{1:t}) \quad (150)$$

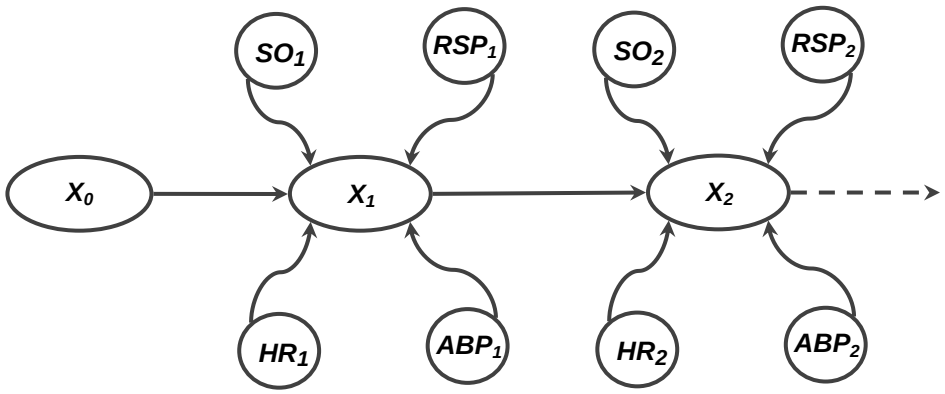


Figure 39. Four parameters DBN for monitoring patients' states

However, such a system violates our requirement to have an open system because, if we are to add a new parameter, then the current estimate of states will become invalid and will need a considerable amount of time to recalculate the estimate in the light of the new parameter. We have also seen that an essential requirement of DBN is that its structure should stay fixed. To overcome this limitation, we note that under most circumstances the four parameters are independent. Therefore, it will prove easier if each parameter has its own DBN model, which may be used to predict the future projection of its current and past values. Consequently, the prediction estimates coming from the individual models are used for further analysis such as calculating the future IRIS score or the probability of developing a complication. Hence, the DBN will boil down to a simple sensor model where the evidence represents the apparent measurement that should be used to infer the real measurement.

The practical benefit of such a model is in cases where the measured data are noisy, doubtful and/or irregular. We will denote the apparent measurement with a lower-case letter like (e) and the real measurement by an upper-case letter like (X). Hence, a typical individual sensor model is shown in Figure 40.

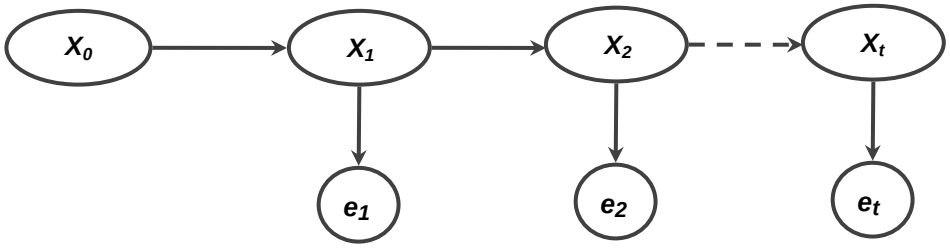


Figure 40. A typical individual sensor model using DBN

5.5 Experiment Set-up

Although all the data used in this study was real data obtained through the MIMIC database, the testing of the system operation and performance is conducted by simulation only. Two types of experiments have been carried out. The first is by using the MIMIC waveforms portion of the database to predict the evolution of patients' vital signs throughout time. The second utilized the clinical data portion of the database to infer the mortality risk of a patient about 24 hours in advance.

In each set-up, patients are randomly assigned into two groups. The first group is used to train the network. The training is the simple estimation of the conditional and unconditional probability tables of each variable. While this

step is not strictly necessary, as the network can initialize without prior knowledge, the result will become more accurate using probability tables that makes sense rather than starting from purely subjective speculations, as the researcher is not a trained physician. Once the training is done, the second group of patients is used to validate the accuracy of prediction. MATLAB is used as the simulation environment. The connection to the database is done locally through a JDBC driver. The MIMIC database itself is managed by postgreSQL. To reduce the latency resulting from database access time, a script is developed that retrieves all the required patients' data and converts them to MATLAB binary data. The accuracy of the system is measured using various techniques, as will be discussed in the following sections.

5.5.1 Predicating the IRIS Score

IRIS (Intensive-care Risk Identification System) is a lookup table used to profile the seriousness of patients' conditions in ICU [175]. It converts a physiological parameter to a score of, for example, between 0 and 3, with 0 representing a stable condition and 3 representing a deteriorating condition. An example of an IRIS lookup table is shown below [175].

Table 3. An example of IRIS lookup table

	Intensive-care Risk Identification System (IRIS) Value						
Variable	3	2	1	0	1	2	3
Respiratory Rate	<6	7-8	9-10	11-16	17-20	21-24	⩾25
Systolic Blood Pressure	<69	70-79	80-99	100-150	151-160	161-179	⩾180
Pulse (heart rate)	<40	40-49	50-59	60-100	101-110	111-130	⩾131

Sp02	<90	91-93	94-96	96-100			

The aim of the simulation is to predict the value of the overall IRIS score, which is simply the mathematical sum of the individual scores given by Table 3. The prediction is simply done using equation 148 to predict the projected future value of the respiratory rate, systolic blood pressure, pulse and oxygen saturation SpO2. A sample of 200 patients was used in the study. The patients were assigned randomly to training and testing groups. Each group comprised 100 patients. Then the system is run for a simulated period of 10,000 seconds (about 2 hours and 46 minutes). At every second, a prediction of the monitored parameter after k seconds is estimated and then compared with the actual one. Then the simulation is repeated with a different value of k. In this study, we started the prediction period with k= 30 seconds and then we incremented the prediction time by 30 seconds until k= 600 seconds. The average accuracy of predicting the heart rate for each run is shown in Figure 41. The average accuracy is calculated as:

$$\text{Accuracy} = \frac{1}{\text{N}}\sum_{\text{N}}\left(1 - \frac{\left|\text{IRIS}_{\text{real}} - \text{IRIS}_{\text{predicted}}\right|}{4}\right) \tag{151}$$

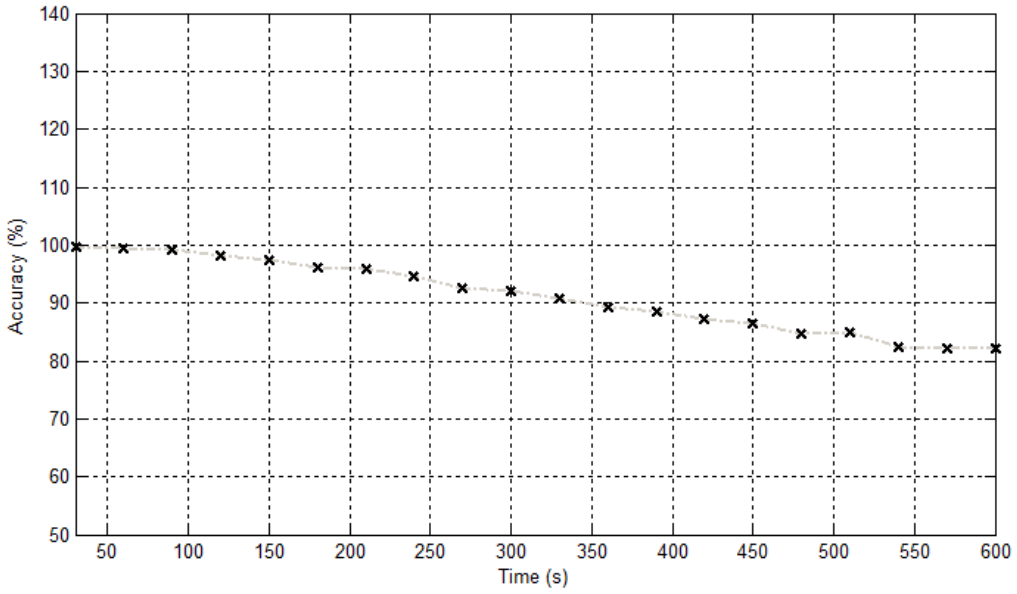


Figure 41. The average accuracy of predicting the IRIS score versus time

The time axis of Figure 41 starts at k=30 where the average prediction of heart rate is 99.9%. However, the accuracy of prediction drops as the predication period increases but stabilizes at k= 8-10 minutes at about 83%. There are two reasons for the drop of accuracy. Firstly, as the algorithm continues to predict in the future, the next estimated value starts to saturate and would stay fixed at a given value during the upcoming prediction cycles. In fact, the developed algorithm will halt the prediction as soon as it detects that the predicted value is saturated, that is, when it continues to be the same for a given number of cycles. The saturation state that is reached by the algorithm is known as the stationary distribution [1,p. 573]. The stationary distribution reflects the fact that, as the predicated time increases, the odds for every possible outcome become equally likely. Secondly, not all the possible outcomes of a parameter are equally estimated during training because there are always more data in the stable regions than the deteriorating regions. This will cause the algorithm to fall back to CP and estimate the upper bound of

probability of a parameter. For example, Figure 42 shows the predicted heart rate (in red) and the measured heart rate (in blue) versus time for a patient case when the prediction time is set to 30 seconds. The two curves coincide with each other almost everywhere except around t ≈0, where CP dominates. However, at k= 300, the effect of saturation becomes clearer (see Figure 43). Although the predicted heart rate diverges from the real measured one at these times when the patient heart rate starts changing rapidly, the effect on the calculated IRIS score is minimal because it only results in a ±1 error in IRIS calculation. An error of ±1 translates into an accuracy of 75%.

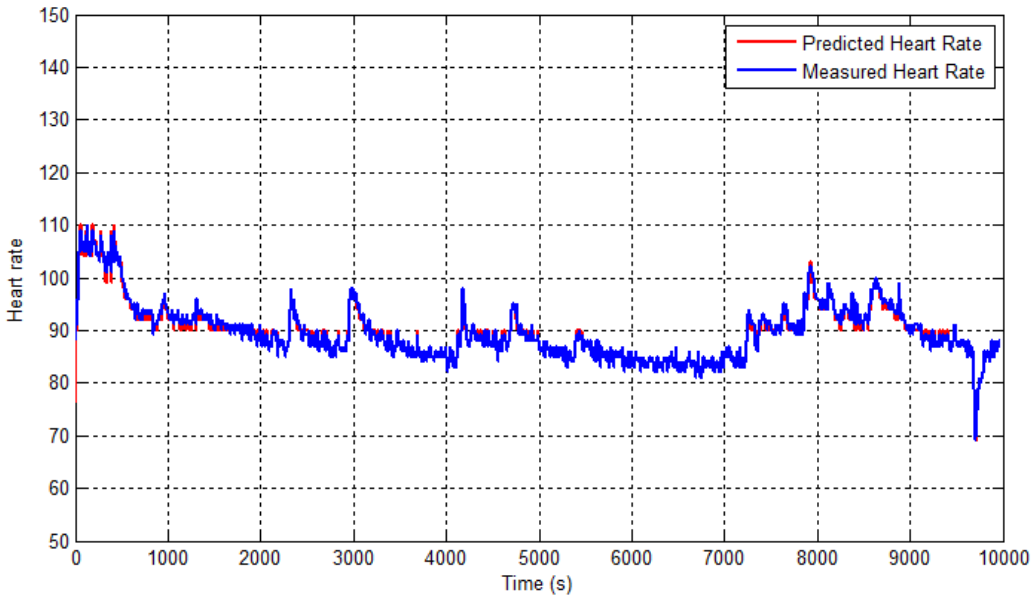


Figure 42. Predicted heart rate (in red) and the measured heart rate (in blue) versus time for a patient case when k= 30

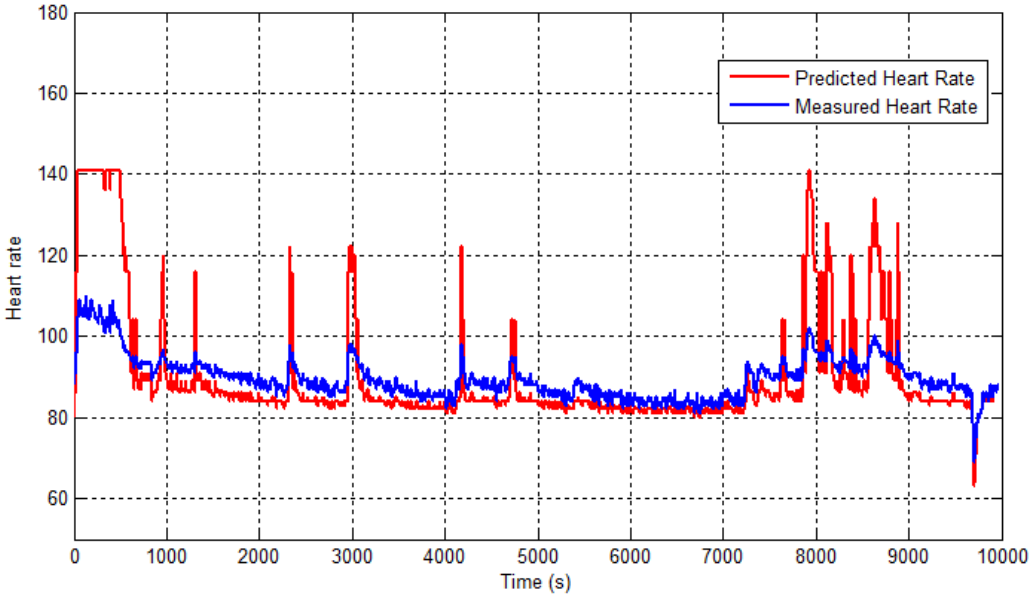


Figure 43. Predicted heart rate (in red) and the measured heart rate (in blue) versus time for a patient case when k= 300

The same prediction accuracy is obtained for the other parameters. In addition, the results stay valid even if we add a new parameter like the arterial blood pressure (ABP). Figure 44 and 45 show the predication of ABP at k=30 and k=300.

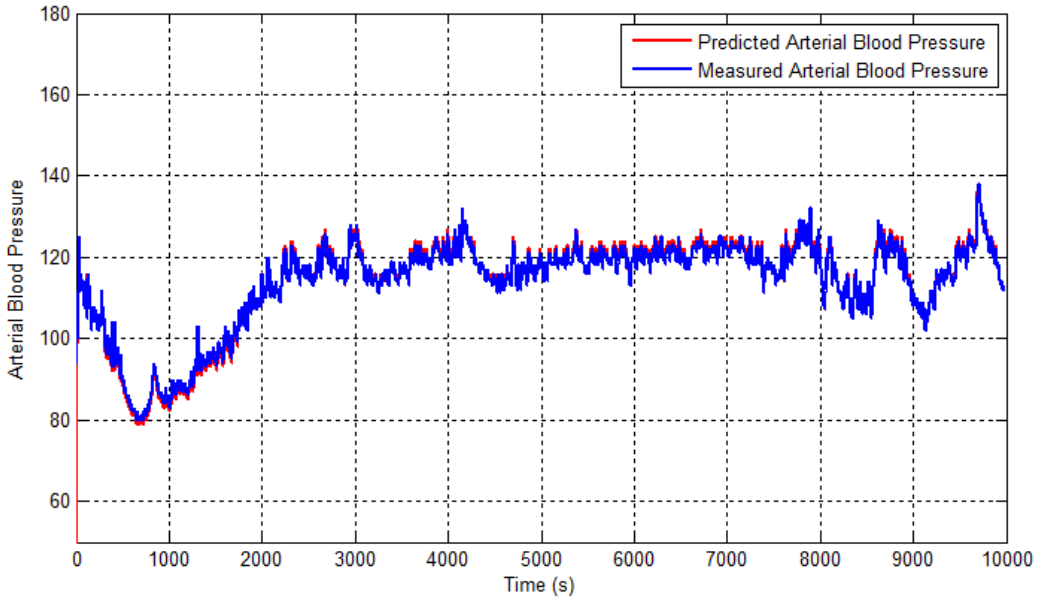


Figure 44. Predicted ABP (in red) and the measured heart rate (in blue) versus time for a patient case when k= 30.

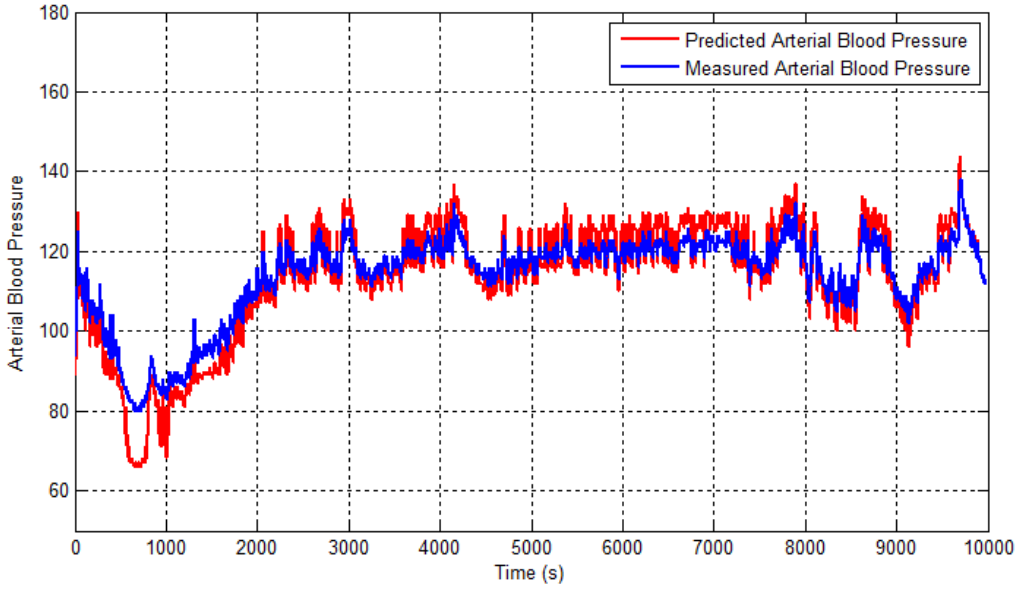


Figure 45. Predicted ABP (in red) and the measured heart rate (in blue) versus time for a patient case when k= 300.

Using Matlab's GUIDE (GUI Development Environment), a graphical user interface (GUI) was built to enable potential users to explore the features of

the demonstration and verify, in realtime, its accuracy. The GUI allows users to monitor up to four subjects simultaneously. Each subject can have his/her own predefinded IRIS score presets, monitoring time scale and the choice of which physiological parameter to plot. Figure 46 shows a snapshot of the developed GUI.

The versatility of the developed algorithm can be shown in different ways.

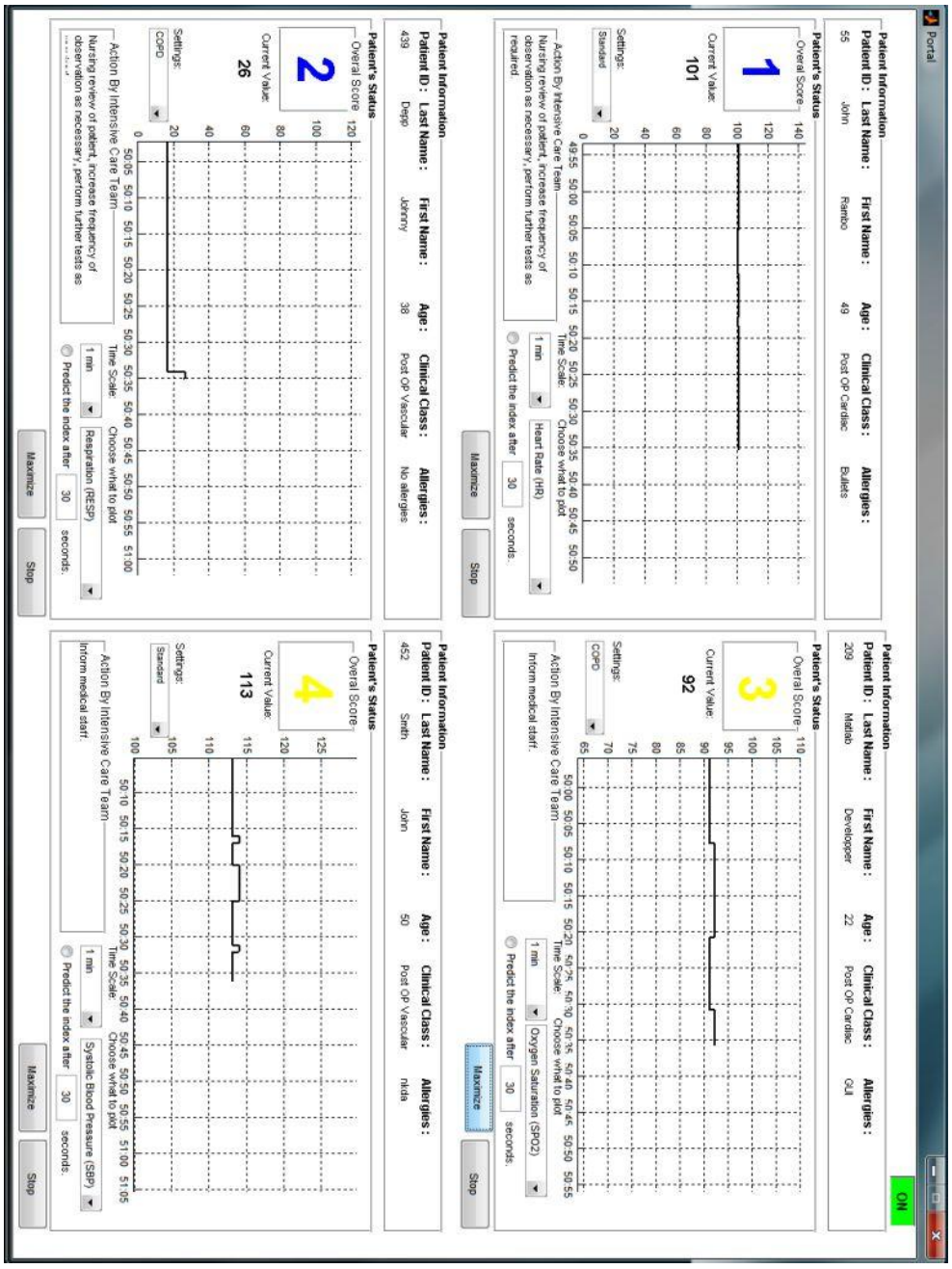


Figure 46. A snapshot of the developed GUI

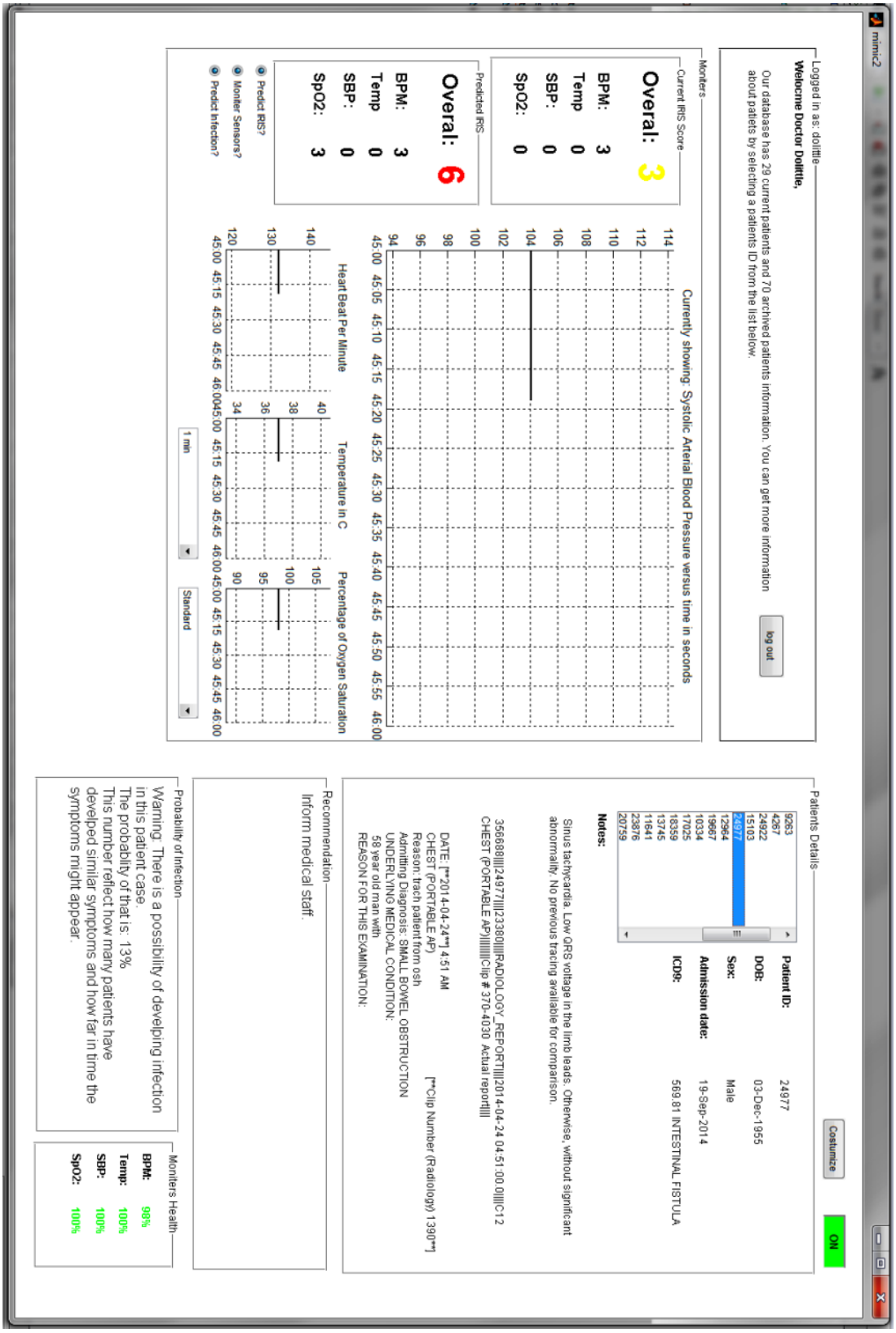


Figure 47. A GUI demonstration how the algorithm can be used to infer the probability of infection

With respect to GUI design, the algorithm can be wrapped with an interface

that requires users to login before they can use the system. Once logged in, they can choose which patient to monitor, switch a physiological parameter to the main axis plot, and estimate the IRIS score and the predicted IRIS score. Figure 47 show an example of such possibility. It also shows how the algorithm can be used to infer the probability of infection and the status of monitors.

5.5.2 Predicting Mortality Risk in Patients with a History of Cardiac Surgery.

The second scenario demonstrates the use of the clinical data portion of the MIMIC II database. As the clinical data is essentially data filled out manually by hospitals staff, it presents different challenges than the waveform data portion did. Firstly, clinical data are acquired less frequently than the bedside monitors are. While a bedside monitor may sample the data at a frequency of 125Hz, the clinical data may only be recorded once per an ICU admission, if at all. Secondly, the waveform data are acquired electronically whereas the clinical data is acquired from different sources, such as the hospital archives, lab test results, free text nursing notes and ECG reports [169]. While the challenges of processing electronically acquired data may be limited to dealing with noise and missing data due to equipment failure, disconnection, or synchronization, data acquired manually through archives and reports are more prone to typo errors, mistakes, irregular delay between

measuring and recording, or ignorance. In addition, not all the physiological variables are samples at the same rate even for the same patients. In some cases, a patient's blood pressure is measured every 15 minutes, and then the rate changes to every hour and so on. The way we dealt with irregularities is by neglecting patients' cases where not enough information is recorded or where there are many missing or empty variables. Then we re-sampled the data at a rate of one sample per hour through linear interpolation. This approach may not be the best since linear interpolation assumes the data to adhere to linear transition model without any justification of such a model. However, the use of other modelling and/or techniques is left for future research work.

The aim of the experiment is to estimate the mortality risk of a patient with a history of cardiac surgery about 24 hours before their date of the death. The MIMIC II version 2.5 has about 5,200 such patients, which we identified by running SQL queries that searched the nurses' notes for traces that indicate the existence of cardiac surgery within the records of the patients. However, not all of these patients have enough data to work with. After screening the patients with not enough or unclean data and randomly dividing them into testing and training groups, we had 1,106 patients for testing the algorithm and 2,580 patients for training. The physiological parameters chosen for predicting the mortality risk are blood pressure, oxygen saturation, heart rate, temperature and creatinine level. Figures 48 through 52 show the amount of records collected for the sample of patients within the last 24 hours.

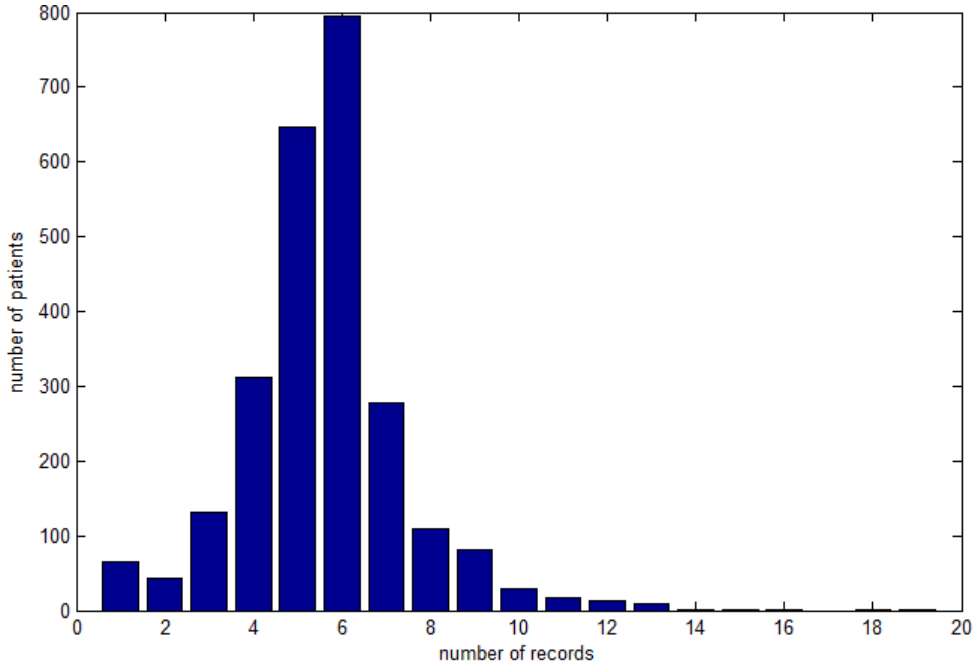


Figure 48. The number of records of temperature measurements of patients in the last 24 hours of their admission

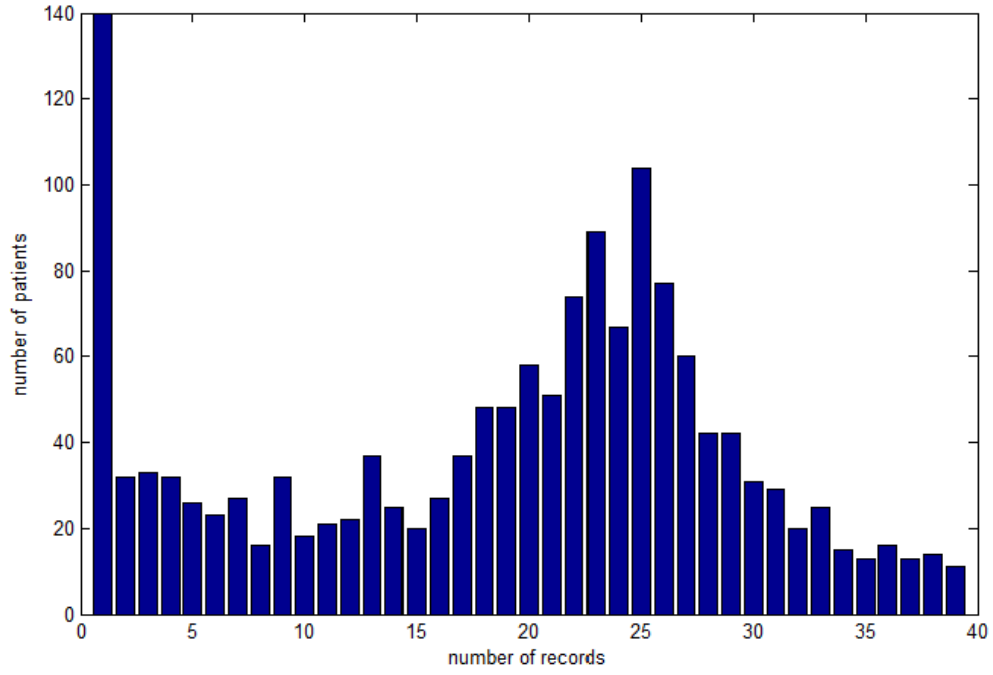


Figure 49. The number of records of blood pressure measurements of patients in the last 24 hours of their admission

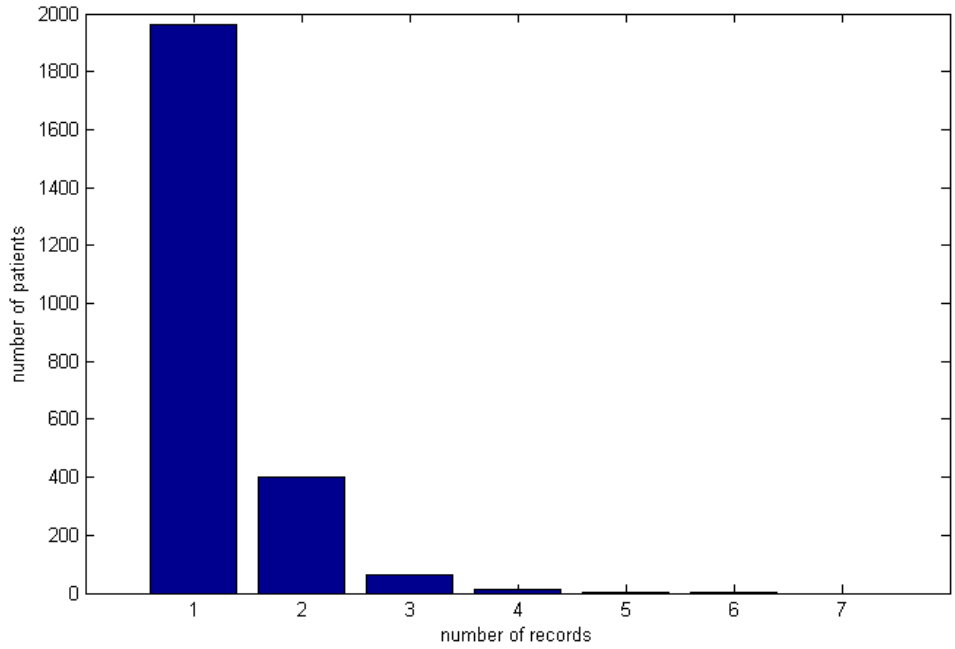


Figure 50. The number of records of creatinine level measurements of patients in the last 24 hours of their admission

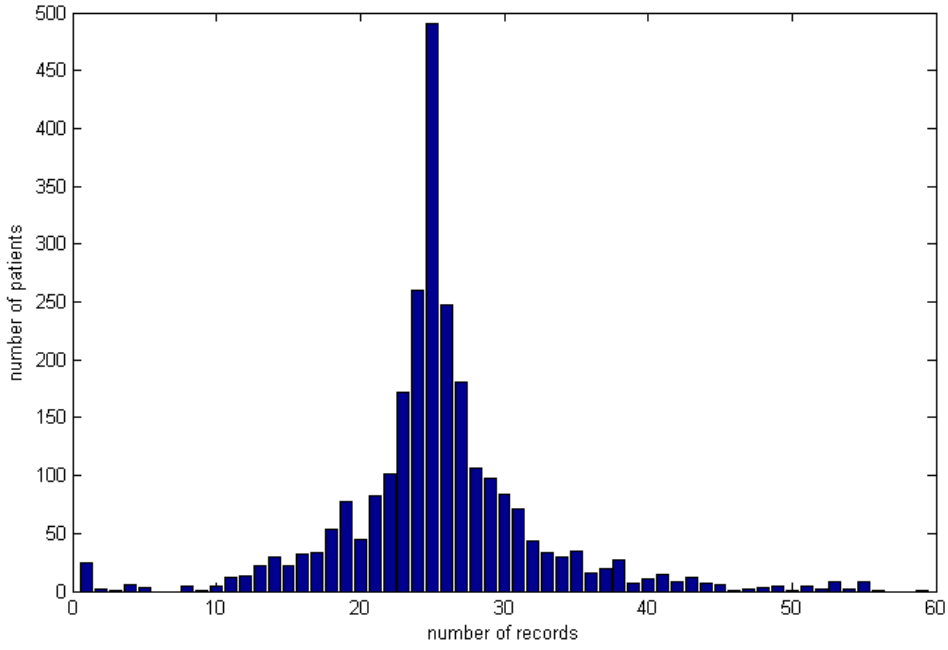


Figure 51. The number of records of heart rate measurements of patients in the last 24 hours of their admission

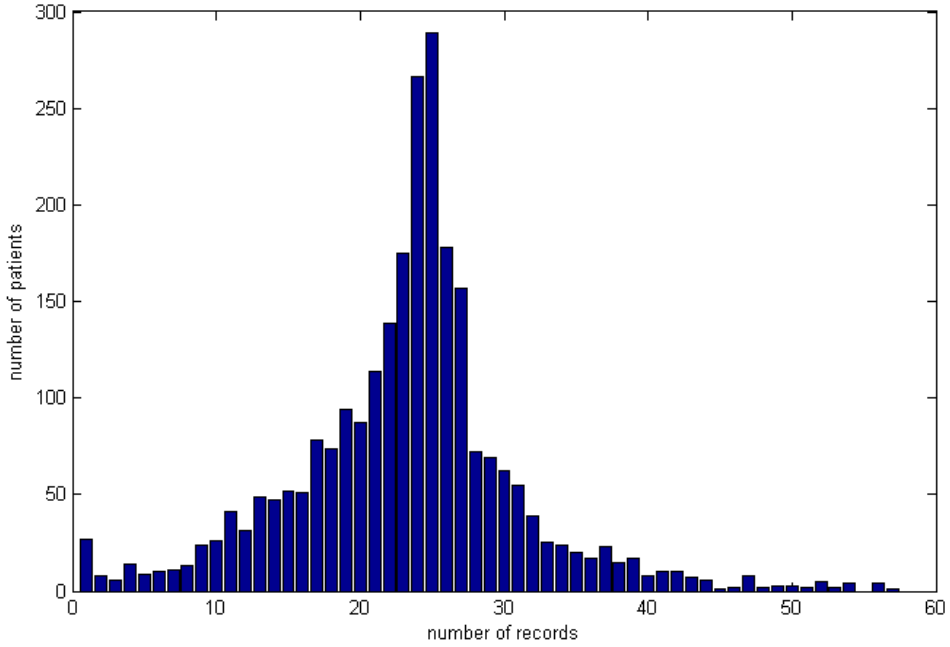


Figure 52. The number of records of oxygen saturation measurements of patients in the last 24 hours of their admission

It is evident from the five last figures that the number of records per patients or per parameters varies significantly. Creatinine levels are mostly measured once per admission. We have assumed that the creatinine level per patient did not change during the time of admission. Temperatures are measured 6 times per the last 24 hours in the patient stay in the ICU, whereas the heart rate, blood pressure and oxygen saturation are measured almost every hour. Hence, interpolation is not of high concern for the validity of the study since patients, on average, have already one record per hour for the potentially rapidly changing parameters. The patients from the training set were used to develop the conditional/unconditional probability tables necessary for prediction the evolution of these parameters during the last 24 hours of their admissions. Since the values of these parameters should be associated with a classification space of either survived or died, we used the predicted value to

search the space of the training sample for the classes of 10 nearest neighbours. If these classes are within the survived group, then the patient is considered a low risk. Otherwise, the patient is a high risk. The risk factor of a patient is calculated using the number of nearest neighbours to the class of deceased patients divided by the total Mahalanobis distance to them. Hence, the further away the predicted parameters are from the class of deceased patients, the smaller the risk will become, and vice versa. The algorithm is validated using the testing patients group and then the mathematical average of each prediction per the last 24 hours is calculated. Figures 53 and 54 show the results obtained during the validation of the algorithm.

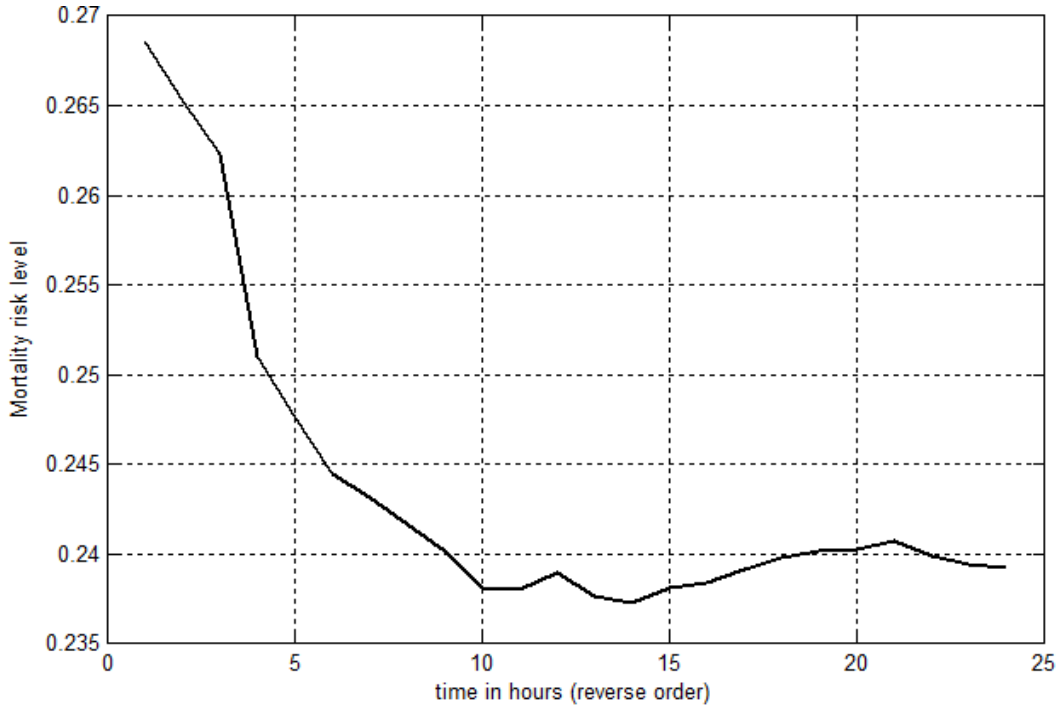


Figure 53. Average mortality risk of the portion of the testing patients group who were discharged from the hospital (survived).

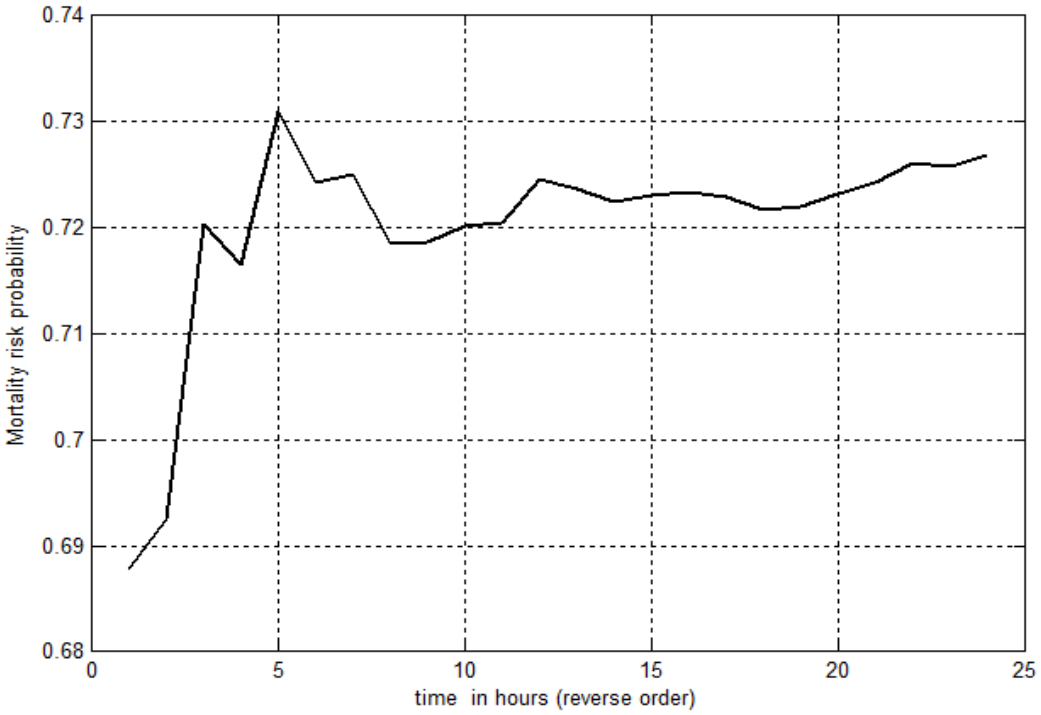


Figure 54. Average mortality risk of the portion of the testing patients group who did not survive

Figure 53 shows only the portion of patients from the testing group who were discharged from the hospital. The time is displayed in reverse order, that is, time 0 means 24 hours before discharge. It is evident from Figure 53 that the average risk is below 0.27. In addition, the risk level decreases as the patients approach their discharge time. Figure 54 shows the portion of patients who have died. The mortality risk associated with them are 3 times higher than that of those in Figure 53, which shows a very good isolation of the portion of patients with higher risk from those of lower risk. In addition, the mortality risk increases as the time of reported death of patients becomes more near. Despite the change in the risk level throughout the 24 hours not being significant, the difference in the risk level between the two portions (figures) proves the validity of the algorithm.

5.6 Summary

This chapter presented another application of the CP approach to BN. In contrast to chapter 4, which showed application to it using static BN, this chapter focused on dynamic BN. It presented a novel approach to the monitoring of patients in intensive care units.

Many research frameworks use artificial intelligence to analyze patients' vital signs in ICU and use these to predict their survivability, manage the admission of medicine or make other decisions. The extent of applications that researchers are currently proposing falls far beyond the scope of a single thesis chapter.

The focus of chapter 5 was on building systems and algorithms that predict the evolution of patients' physiological variables throughout their stay in an ICU, or hospital. The predictions can be utilized in several ways. They can be used to determine the likelihood of a sequence of measurements, to make decisions, or to estimate the stability of a patient.

We have presented two experiments using the MIMIC II database. In the first experiment, we predicated the evolution of patients' parameters up to 10 minutes in advance with an accuracy of up to 99%. In the second experiment, we predicted the mortality risk of a group of patients and showed the average evolution and levels of risk within each group.

6. Conclusion

The core objective of this thesis was to enhance the current procedure of designing decisions support systems when the available amount of information is significantly limited. In addition, it aimed to facilitate a better representation of information without falling back on the fallacy of extracting knowledge from ignorance or presuming situations in an ad hoc fashion without sound justifications. Nonetheless, the thesis took a practical approach to the matter at hand by applying the proposed theory to two interesting and challenging research areas: aviation safety and patient monitoring in ICUs.

In chapter 1, we showed that a decision-maker is, in essence, a gambler in the sense that every decision made involves an element of uncertainty. Unforeseen factors make the outputs of decisions uncertain. When dealing with uncertainty, decision-makers often need a method with which to quantify the likelihood of an outcome. The theory of probability provides a foundation for representing the doubt and trustworthiness of an outcome from both subjective and objective points of view. This has found a wide range of application in scientific research, from social science to engineering to quantum mechanics.

However, the mere representation of information is not sufficient to make decisions because the more decisions that are available at the hands of

decision-makers, the greater the chance of outcomes that are more favourable than the others are. Combining probability with preferences is the foundation for the modern theory of decision-making. Preferences are expressed in the form of utility functions. Utility is not a semantic equivocation of the notion of value but rather a transfer function that maps a decision to its relative usefulness. While deriving a good utility function for a given decision problem may be controversial and subjective to a certain extent, it is the analysis and estimation of probability that take most of the effort. From a Laplacian point of view, probability can be estimated objectively just by looking at the sample space of an event. However, an analytic approach to the event may not be the best way to infer matters of reality, as analytical judgements infer nothing more than the relationships between concepts, ideas, and meanings. Analytically, the odds of a coin toss landing on heads are the same a “yes” answer in an engagement proposal, while in reality we would consider such thinking absurd. The frequency interpretation of probability seems to us to provide the best answer, because it is objective and is estimated from the real world a posteriori. Nonetheless, estimating the probability of events in this way requires an extensive amount of data. Therefore, the theoretical framework of this thesis has sought to find the best interpretation of probability in the context of limited information.

6.1 Meeting the objectives

In order to meet the objectives of this thesis, both theoretical and practical approaches were adopted. The main objective is to find a better framework that can fully hold the expectations of decision-makers in making better decisions under sparse knowledge or in time-critical situations where the availability of information begs for more time than a decision-maker has.

Firstly, the common approach to decision-making, and in turn knowledge-based decision support systems, is to use probability theory backed by the utility functions to come up with the expected utility of making a decision. This approach was necessary if the designed DSSs in this research were to remain compatible with the current state of the art DSS. In addition, the research result proposed in this thesis should integrate to the repositories of science in a way that other researchers can make use of it. Hence, the approach should not deviate much from the direction of the current arrow of designing DSS.

Secondly, as the theory of probability is accepted as the main framework for representing knowledge with uncertainty, we analyzed many interpretations of probability in order to find the most suitable one that works with as little information available as possible without falling back on a strictly analytical approach or ignorance. A common criterion for assessing an interpretation of probability is given by Salmon (see chapter 3). It has three aspects, which emphasise the importance of usefulness, admissibility and ascertainability of an interpretation. We have analysed several candidates from many philosophical and mathematical approaches to the analysis of probability. These range from the Laplacian interpretation to the logical to the

comparative probability. The choice made was to use the comparative probability approach because it offered the best way to represent knowledge in circumstances in which little information is available, it could be made compatible with the Kolmogorov axiomatic probability, and it has many modelling options from which one can choose.

Thirdly, we surveyed the research done in the theory of comparative probability, its axioms, and application to computer science. We found that CP has been used for at least two purposes. The first purpose was as a standalone interpretation of probability that rivals all the quantitative probability theories. The second was a relaxed approach to quantitative probability and, to some extent, to provide a justification of the modern Kolmogorov axiomatic probability. The choice between the two approaches to CP was based on the requirements laid down in section 3.3.1. The requirements aimed at compatibility with other DSS and to utilize the strongest results of the Kolmogorov axiomatic probability, namely: the law of strong numbers and the central limit theorems. Hence, the second approach to CP proved more promising for the aims of the thesis.

Fourth, we strengthen the requirements of the best-fit theory with assumptions that will secure a place for the proposed theory in the current frameworks of both CP and KP research and ensure that probability continues to be considered the very guide to life. We assumed probability to be objective, just as the frequency interpretation of probability is. Probability should be inferred from data a posteriori, not from the space of possibilities. If no data exists, probability still exists but its objective value is unknown. That

means that CP is nothing more than a way of representing how much knowledge we attain about reality. This knowledge can be as high as an exact replica of reality or as low as a basic outline of it. As the amount of data acquired from an experiment increases, the probabilities of its outcomes are quantified using the frequency approach to probability. That will make the proposed theory compatible with the state of the art DSS as the probability calculated by it matches that of most frequently adopted approach.

Fifth, we used the Chernoff bounds to come up with a novel approach to updating probability bounds between successive experiment results. Chernoff bounds were used as upper and lower estimates of probability at a given experiment while taking into account all the previous experiment results. As the number of experiments increases, the gap between the upper and lower bounds becomes smaller until it approaches the expectation of the outcome of the experiment. The expectation of an experiment is nothing other than its probability. Hence, a mathematical foundation between CP and KP was established with a dynamic nature that puts CP as a foreground methodology to evaluate KP.

Sixth, we recognized that even with the availability of a simple approach to representing knowledge, the size of the joint probability tables may become too large to process, so we used a Bayesian network to simplify the processing of probabilistic queries and reduced the amount of mathematical backgrounds required to answer them.

Seventh, as probabilistic decision support systems work on averages, it would be unfeasible to attempt to justify the principles of the proposed

approach using an example or two. Instead, we adopted two approaches to tackle the issue. Firstly, we used scenario-based validation. Scenarios are ways of generating test data, which can be used to validate system design requirements. The second approach was the ability of the system to predict an output with high accuracy. We have shown examples of the first approach in chapter 4 and the second approach in chapter 5.

Eighth, we suggested two new enhancements to the detection and isolation of faults in aviation and to the optimising the navigation planning (see chapter 4). In the first experiments, we proposed a new method for detecting faults that should overcome any limitations that result from using majority vote coming from primary and redundant systems. Whereas, in the second experiment, we proposed a novel application to the BADA database as a DSS for navigation planning. Both experiments where implemented with CP to show the usefulness, admissibility and ascertainability of CP.

Ninth, an innovated ICU patient monitoring system was designed (see chapter 5). The novel system outperforms all current monitoring systems in terms of its versatility and prediction capabilities. We have shown how it can be used to predict the evolution of patients' physiological parameters over time and how it can predict the mortality risk in patients with a history of cardiac surgery even 24 hours before patients' date of death.

These nine points show the development of the reasoning according to which this research was conducted, starting from defining the research question to documenting the results. The research method dictates that a good theory should be able to predict some observations that can be

measured and compared to what the theory proposes. In the light of such requirements, it is the belief of the author that the thesis stands on very solid grounds with respect both to meeting the objectives and verifying the soundness of its theory.

6.2 Future Work

While an extensive amount of work has been put into this thesis in terms of both theoretical analysis and practical implementation, there are still some research questions and opportunities waiting to be fulfilled. The requirements and objectives of this thesis made it clear that the proposed theory should be integrable to science and that it should establish a context for the current frameworks of various areas in artificial intelligence, aviation and biomedicine. As such, it will become open to opportunities and criticism that extend far beyond the simple mean of two different applications and peer review process of all the papers published during the time of conducting this study.

On the opportunity side, the monitoring system described in chapter 5 has been filed for a patent in the UK. This has made possible a collaboration between Manchester University and Rinicom Ltd. In addition, the fault detection and isolation method described in chapter 4, along with the utilization of BADA network, enabled the School of Computing and Communication Systems at Lancaster University to secure funded research in the SVETLANA project to enhance the current procedures and performance of flight analysis programs. It was originally conducted in a contract to RNC

Avionics Ltd through the North West Development Agency Voucher Award. In addition, the work is continuing to apply CP to various methods in online clustering analysis, such as the Evolving Takagi Sugeno fuzzy model. In addition, it has been partially applied to noisy audio signal classification but the results are far away from complete.

One the criticism side, the major limitation of the work is the assumption of independent variables while using Chernoff bounds. This is, in fact, a limitation of Chernoff bounds. Proposed future directions of work would be to convert the dependent variables to independent, but no work has been done towards that yet. It will be of great value to find a way to extend the results of this thesis to dependent variables as well as to other types of random variables. Moreover, it will be of value to bring the MIMIC II up to its full potential by first finding a better way to clean up the data and replace missing information, and second to extend the open platform architecture proposed in chapter 5 for fast prototyping and deployment.

6.3 Final Remarks

As the case with any novel proposal, the comparative probability approach proposed in this thesis is not yet complete. The best way to show the power of it is through applying it to a wider range of applications and engineering problems while ironing out any issues that arises along the way. While this thesis worked as proof of concept for CP application to DSS and artificial intelligence in general, it is the belief of the author that it has achieved its objectives and still maintaining the de facto interpretation of probability intact. After all, it would not be of benefit to the scientific community to propose the seizure of their very best guide to life.

7. References

[1] S. Russel and P. Norving, *Artificial intelligence: A modern approach*, 3rd ed. New Jersey: Pearson Education, Inc., 2010.

[2] M. Wisniewski, *Quantitative methods for decision makers*, 4th ed. Essex: Pearson Education Limited, 2006.

[3] R. Ziemer and W. Tranter, *Principles of communications systems, modulation, and noise*, 6th ed.: John Wiley & Sons, 2010.

[4] J. Skurai and J. Napolitano, *Modern quantum mechanics*, 2nd ed. San Francisco: Pearson Education, 2011.

[5] R. Schlaifer, *Analysis of decisions under uncertainty*. New York: McGrew-Hill, 1969.

[6] K. Korb and A. Nicholson, *Bayesian artificial intelligence*. London: Chapman & Hall/CRC Press UK, 2004.

[7] M. R. Endsley, *et al.*, "Situation awareness information requirements for commercial airline pilots," MA: Massachusetts Institute of Technology International Center for Air Transportation1998.

[8] J. T. Luxhoj, "Risk-based decisionmaking for aviation safety using bayesian belief networks," in *Probabilistic safety assessment and management*. vol. 4, A. Mosleh and R. A. Bari, Eds., London: Spring-Verlog, 1998, p. 1530.

[9] E. Turban and J. Aronson, *Decision support systems and intelligence systems*, 6th ed. New Jersy: Printice Hall, 2001.

[10] D. Power, "Decision support systems: A historical overview," in *Handbook on decision support systems 1: Basic themes*, F. Burstein and C. W. Holsapple, Eds., ed Berlin: Springer-Verlog, 2008, p. 126.

[11] G. Forgionne, "Foundations and architectures of dmss," in *Decision making support systems: Achievements, trends, and challenges for the new decade*, M. Mora and J. Gupta, Eds., 1st ed London: Hershey, PA Idea Group Publishing, 2003, p. 2.

[12] V. S. Janakiraman and K. Surakest, *Decision support systems*, 1st ed. New Delhi: Printice Hall of India, 2004.

[13] J. T. Durkin and A. M. Greeley, "A model of religious choice under uncertainty," *Rationality and Society,* vol. 3, pp. 178-196, April 1, 1991.

[14] M. J. Machina, "Choice under uncertainty: Problems solved and unsolved," *The Journal of Economic Perspectives,* vol. 1, pp. 121-154, 1987.

[15] N. Miller. (2006, 22 July). *Chapter 6: Choice under uncertainty*. Available: http://www.hks.harvard.edu/nhm/notes2006/notes6.pdf

[16] P. J. H. Schoemaker, "The expected utility model: Its variants, purposes, evidence and limitations," *Journal of Economic Literature,* vol. 20, pp. 529-563, 1982.

[17] D. Schmeidler, "Subjective probability and expected utility without additivity," *Econometrica,* vol. 57, pp. 571-587, 1989.

[18] S. Ross, *A first course in probability*, 7th ed. New Jersey: Pearson Printice Hall, 2006.

[19] D. Wischi. (13 October, 2000). *The history of 'probability'*. Available: http://www.cs.ucl.ac.uk/staff/ucacdjw/Talks/histprob.pdf

[20] D. Koller and N. Friedman, *Probabilistic graphical models: Prnciples and techniques*. Cambridge: The MIT Press, 2009.

[21] K. B. Korb and A. E. Nicholson, *Bayesian artificial intelligence*. Florida: Chapman & Hall/CRC, 2004

[22] R. Hanna. (2011, 24.10). *Kant's theory of judgment (Summer 2011 ed.)*. Available: http://plato.stanford.edu/archives/sum2011/entries/kant-judgment

[23] G. Rey. (2010, 24.10). *The analytic/synthetic distinction (Winter 2010 ed.)*. Available: http://plato.stanford.edu/archives/win2010/entries/analytic-synthetic

[24] J. Joyce. (2008, 24.10). *Bayes' theorem (Fall 2008 ed.)*. Available: http://plato.stanford.edu/archives/fall2008/entries/bayes-theorem

[25] J. Oakland, *Statistical process control*, 6th ed. Oxford: Elsevier, 2008.

[26] V. Easton and J. Mccoll. (1997, 25.10). *Random variables and probability distributions*. Available: http://www.stats.gla.ac.uk/steps/glossary/probability_distributions.htm

[27] S. Vaseghi, *Advanced digital signal processing and noise reduction*. Singapore: John Wiley & Sons, 2008.

[28] H. Schwarzlander, *Probability concepts and theory for engineers*. Chichester: John Wiley & Sons, 2011.

[29] R. M. Feldman and C. Valdez-Flores, *Applied probability and stochastic processes*. New York: Springer, 2010.

[30] L. Jaisingh. (2000). *Statistics for the utterly confused*. Available: http://www.mhprofessional.com/product.php?isbn=0071461930

[31] G. Schay, *Introduction to probability with statistical applications*. Boston: Birkhäuser, 2007.

[32] Laerd Statistics, *Measures of spread*. Available: http://statistics.laerd.com/statistical-guides/measures-of-spread-range-quartiles.php

[33] Wikipedia, *Normal distribution*, 2011, Available: http://en.wikipedia.org/wiki/Normal_distribution

[34] T. T. Soong, *Fundamentals of probability and statistics for engineers*. Chichester: John Wiley & Sons, 2004.

[35] M. J. Hassett and D. Stewart, *Probability for risk management*, 2nd ed. Winsted: ACTEX, 2009.

[36] R. C. Jaffe, *Random signals for engineers using matlab and mathcad*, New York: Springer-Verlag, 2000.

[37] H. Fischer, *A history of the central limit theorem: From classical to modern probability*. New York: Springer, 2010.

[38] E. Lesigne, *Heads or tails: An introduction to limit theorems in probability*: AMS, 2005.

[39] D. Downing and J. Clark, *E-z statistics*. New York: Barron, 2009.

[40] Y. M. Suhov, *et al.*, *Probability and statistics by example: Basic probability and statistics*. Cambridge: Cambridge University Press, 2005.

[41] O. J. W. F. Kardaun, *Classical methods of statistics: With applications in fusion*. Berlin: Springer-Verlag, 2005.

[42] G. R. Arce, *Nonlinear signal processing: A statistical approach*. New Jersey: John Wiley & Sons, 2005.

[43] M. Otte and M. Panza, *Analysis and synthesis in mathematics: History and philosophy* Norwell: Kluwer Academic Publishers, 1997.

[44] A. Dasgupta, *Fundamentals of probability: A first course*. New York: Springer, 2010.

[45] F. Jensen, *Baysian networks and decision graphs*. New York: Springer-Verlag, 2001.

[46] A. Hájek. (2010). *Interpretations of probability (Spring 2010 ed.)*. Available: http://plato.stanford.edu/archives/spr2010/entries/probability-interpret/

[47] R. V. Mises, *Probability, statistics, and truth*. New York: Dover Publications, 1957.

[48] W. Sun, "Interpretations of probability," PhD, The University of Connecticut, Connecticut, 2003.

[49] W. Salmon, *The foundations of scientific inference*. Pittsburgh: University of Pittsburgh Press, 1966.

[50] R. Haenni, "Towards a unifying theory of logical and probabilistic reasoning," *4th International Symposium on Imprecise Probabilities and Their Applications*, vol. 1, pp. 193-202, 2005.

[51] K. R. Popper, "The propensity interpretation of probability," *The British Journal for the Philosophy of Science*, vol. 10, pp. 25-42, 1959.

[52] J. H. V. L.-V. Dis, "Stir in stillness : A study in the foundations of equilibrium statistical mechanics," Proefschrift Universiteit Utrecht, Tekst, 2001.

[53] D. Howie, *Interpreting probability: Controversies and developments in the early twentieth century*. Cambridge: Cambridge University Press, 2002.

[54] T. Fine, *Theories of probability*. London: Academic Press, 1973.

[55] D. M. Cifarelli and E. Regazzini, "De finetti's contribution to probability and statistics," *Statistical Science*, vol. 11, pp. 253-282, 1996.

[56] L. J. Savage, *The foundations of statistics*. New York: Dover Publications, 1972.

[57] B. O. Koopman, "The bases of probability," *Bulletin of the Americal Mathematical Society*, vol. 46, p. 763-774, 1940.

[58] G. Regoli. (1999, 24.11). *Comparative probability orderings*. Available: http://www.sipta.org/documentation/comparative_prob/regoli.pdf

[59] P. C. Fishburn, "The axioms of subjective probability," *Statistical Science*, vol. 1, pp. 335-345, 1986.

[60] J. M. Keynes, *A treatise on probability*. Toronto: Macmillan Co., 2007.

[61] P. C. Fishburn, "Ellsberg revisited: A new look at comparative probability," *The Annals of Statistics,* vol. 11, pp. 1047-1059, 1983.

[62] N. Haverkamp and M. Schulz, "A note on comparative probability," *Erkenntnis*, pp. 1-8.

[63] J. P. Burgess, "Axiomatizing the logic of comparative probability," *Notre Dame J. Formal Logic,* vol. 51, pp. 119-126, 2010.

[64] K. O. May, "Intransitivity, utility, and the aggregation of preference patterns," *Econometrica,* vol. 22, pp. 1-13, 1954.

[65] Amos Tversky, "Intransitivity of preferences," *Psychological Review,* vol. 76, pp. 31-48, 1969.

[66] A. Capotorti and A. Formisano, "Comparative uncertainty: Theory and automation," *Mathematical Structures in Computer Science,* vol. 18, pp. 57-79, 2008.

[67] A. Heifetz and P. Mongin, "Probability logic for type spaces," *Games and Economic Behavior,* vol. 35, pp. 31-53, 2001; R. Christian, *et al.*, "Flippable pairs and subset comparisons in comparative probability orderings," *Order,* vol. 24, pp. 193-213, 2007.

[68] B. L. Huber and O. Huber, "Development of the concept of comparative subjective probability," *Journal of Experimental Child Psychology,* vol. 44, pp. 304-316, 1987.

[69] T. Fine, "A note on the existence of quantitative probability," *The Annals of Mathematical Statistics,* vol. 42, pp. 1182-1186, 1971.

[70] P. Fishburn, *Utility theory of decision making*. New York: Wiley, 1970.

[71] C. H. Kraft, *et al.*, "Intuitive probability on finite sets," *The Annals of Mathematical Statistics,* vol. 30, pp. 408-419, 1959.

[72] R. D. Luce, "Sufficient conditions for the existence of a finitely additive probability measure," *The Annals of Mathematical Statistics,* vol. 38, pp. 780-786, 1967.

[73] D. W. Stroock, *An introduction to markov processes*. Berlin: Springer, 2005.

[74] F. Huber and C. Schmidt-Petri, *Degrees of belief* Berlin: Springer, 2009.

[75] G. Link, *One hundred years of russell's paradox: Mathematics, logic, philosophy*. Berlin: Walter de Gruyter, 2004.

[76] D. Ellsberg, "Risk, ambiguity, and the savage axioms," *The Quarterly Journal of Economics,* vol. 75, pp. 643-669, 1961.

[77] R. Sherman, "The psychological difference between ambiguity and risk," *The Quarterly Journal of Economics,* vol. 88, pp. 166-169, 1974.

[78] R. D. Luce, "On the numerical representation of qualitative conditional probability," *The Annals of Mathematical Statistics,* vol. 39, pp. 481-491, 1968.

[79] F. Peter C, "Interval models for comparative probability on finite sets," *Journal of Mathematical Psychology,* vol. 30, pp. 221-242, 1986.

[80] F. Peter C, "Finite linear qualitative probability," *Journal of Mathematical Psychology,* vol. 40, pp. 64-77, 1996.

[81] M. Kaplan and T. L. Fine, "Joint orders in comparative probability," *The Annals of Probability,* vol. 5, pp. 161-179, 1977; F. Terrence L, "Lower probability models for uncertainty and nondeterministic processes," *Journal of Statistical Planning and Inference,* vol. 20, pp. 389-411, 1988.

[82] W. Peter, "Towards a unified theory of imprecise probability," *International Journal of Approximate Reasoning,* vol. 24, pp. 125-148, 2000.

[83] P. Walley and T. L. Fine, "Varieties of modal (classificatory) and comparative probability," *Synthese,* vol. 41, pp. 321-374, 1979; P. Walley and T. L. Fine, "Towards a frequentist theory of upper and lower probability," *The Annals of Statistics,* vol. 10, pp. 741-761, 1982.

[84] P. Walley, *Statistical reasoning with imprecise probabilities*. London: Chapman and Hall, 1991.

[85] C. Fabio Gagliardi, "Graphical models for imprecise probabilities," *International Journal of Approximate Reasoning,* vol. 39, pp. 167-184, 2005.

[86] D. Thierry, "Reasoning with imprecise belief structures," *International Journal of Approximate Reasoning,* vol. 20, pp. 79-111, 1999.

[87] A. Capotorti and B. Vantaggi, "Axiomatic characterization of partial ordinal relations," *International Journal of Approximate Reasoning,* vol. 24, pp. 207-219, 2000.

[88] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association,* vol. 58, pp. 13-30, 1963.

[89] T. Hagerup and C. Rüb, "A guided tour of chernoff bounds," *Information Processing Letters,* vol. 33, pp. 305-308, 1990.

[90] R. Khardo. (2008, 17.12). *Computational learning theory*. Available: http://www.cs.tufts.edu/~roni/Teaching/CLT/LN/lecture4.pdf

[91] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomized algorithms and probabilistic analysis*. New York: Cambridge University Press, 2005.

[92] M. A. Kramer and B. L. Palowitch, "A rule-based approach to fault diagnosis using the signed directed graph," *AIChE Journal,* vol. 33, pp. 1067-1078, 1987; I. R, "Supervision, fault-detection and fault-diagnosis methods — an introduction," *Control Engineering Practice,* vol. 5, pp. 639-652, 1997.

[93] N. F. Thornhill and T. Hägglund, "Detection and diagnosis of oscillation in control loops," *Control Engineering Practice,* vol. 5, pp. 1343-1354, 1997.

[94] R. Isermann and P. Ballé, "Trends in the application of model-based fault detection and diagnosis of technical processes," *Control Engineering Practice,* vol. 5, pp. 709-719, 1997.

[95] V. Venkatasubramanian*, et al.*, "A review of process fault detection and diagnosis: Part i: Quantitative model-based methods," *Computers & Chemical Engineering,* vol. 27, pp. 293-311, 2003.

[96] M. Schwabacher and K. Goebel. (10 Jan). *A survey of artificial intelligence for prognostics*. Available: http://ti.arc.nasa.gov/m/pub-archive/1382h/1382%20(Schwabacher).pdf

[97] T. C. E. W. Team". (10 Jan). *Bada, the european organization for the safety of air navigation*. Available: http://www.eurocontrol.int/eec/public/standard_page/proj_BADA.html

[98] A. H. Ali, "Utilizing bada (base of aircraft data) as an on-board navigation decision support system in commercial aircrafts," *Intelligent Transportation Systems Magazine, IEEE,* vol. 3, pp. 20-25, 2011; Sagem, "Analysis ground station," 2008.

[99] C. Baskiotis, *et al.*, "Parameter identification and discriminant analysis for jet engine machanical state diagnosis," in *Decision and Control including the Symposium on Adaptive Processes, 1979 18th IEEE Conference on*, 1979, pp. 648-650.

[100] I. Rolf, "Model-based fault-detection and diagnosis – status and applications," *Annual Reviews in Control,* vol. 29, pp. 71-85, 2005.

[101] V. Venkatasubramanian, *et al.*, "A review of process fault detection and diagnosis: Part ii: Qualitative models and search strategies," *Computers & Chemical Engineering,* vol. 27, pp. 313-326, 2003.

[102] A. H. Ali, *et al.*, "Feasibility demonstration of diagnostic decision tree for validating aircraft navigation system accuracy," *Journal of aircraft,* vol. 47, 2010.

[103] J. Han, *et al.*, "Data-driven discovery of quantitative rules in relational databases," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 5, pp. 29-40, 1993.

[104] A. Stephen, *et al.*, "Mining distance-based outliers in near linear time with randomization and a simple pruning rule," *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 29-38, 2003.

[105] C. Chiu, *et al.*, "Intelligent aircraft maintenance support system using genetic algorithms and case-based reasoning," *The International Journal of Advanced Manufacturing Technology*, vol. 24, pp. 440-446, 2004.

[106] F. Mustapha, *et al.*, "Development of a prototype knowledge-based system for troubleshooting of aircraft engine and parts – a case study of cessna caravan," *International Journal of Mechanical and Materials Engineering*, vol. 5, 2010.

[107] S. Letourneau, *et al.*, "Data mining to predict aircraft component replacement," *Intelligent Systems and their Applications, IEEE*, vol. 14, pp. 59-66, 1999.

[108] S. Das, *et al.*, "Multiple kernel learning for heterogeneous anomaly detection: Algorithm and aviation safety case study," *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 47-56, 2010.

[109] V. Venkatasubramanian, *et al.*, "A review of process fault detection and diagnosis: Part iii: Process history based methods," *Computers & Chemical Engineering*, vol. 27, pp. 327-346, 2003.

[110] Y. Papadopoulos and J. Mcdermid, "Automated safety monitoring: A review and classification of methods," *International journal of condition monitoring and diagnostic engineering management*, 2001.

[111] L. Zadeh, "A fuzzy-set-theoretical interpretation of linguistic hedges," *Journal of Cybernetics,* vol. 2, 1972.

[112] P. Angelov, "An approach to on-line design of fuzzy controllers with evolving structure," in *Applications and science in soft computing series: Advances in soft computing*, J. M. G. L. Ahmad, pp. 63-68.

[113] S. Charbonnier, *et al.,* "Trends extraction and analysis for complex system monitoring and decision support," *Engineering Applications of Artificial Intelligence,* vol. 18, pp. 21-36, 2005.

[114] J. T. Luxhøj, "Trending of equipment inoperability for commercial aircraft," *Reliability Engineering and System Safety* vol. 64, 1999.

[115] M. R. Maurya, *et al.*, "Fault diagnosis by qualitative trend analysis of the principal components," *Chemical Engineering Research and Design,* vol. 83, pp. 1122-1132, 2005.

[116] P. A. Samara, *et al.*, "Detection of sensor abrupt faults in aircraft control systems," *2003. CCA 2003. Proceedings of 2003 IEEE Conference on Control Applications*, 2003, pp. 1366-1371 vol.2.

[117] P. Angelov and A. Kordon, "Evolving intelligent sensors in chemical industry," in *Evolving intelligent systems: Methodology and applications*, P. Angelov and D. Filev, Eds., ed: John Willey and Sons, 2010, p. 313.

[118] M. Azam, *et al.*, "In-flight fault detection and isolation in aircraft flight control systems," in *IEEE Aerospace Conference*, 2005, pp. 3555-3565.

[119] A. H. Ali and A. Tarter, "Developing neuro-fuzzy hybrid networks to aid predicting abnormal behaviours of passengers and equipments inside an airplane," presented at the Proceedings of SPIE, Orlando, USA, 2009.

[120] S. Savanur, *et al.*, "Adaptive neuro-fuzzy based control surface fault detection and reconfiguration," presented at the International Conference on Aerospace Science and Technology, India, 2008.

[121] S. J. Lou, *et al.*, "Comparison of fault detection techniques: Problem and solution," in *Proceedings of the 2002 American Control Conference*, 2002, pp. 4513-4518 vol.6.

[122] P. A. Samara, *et al.*, "A statistical method for the detection of sensor abrupt faults in aircraft control systems," *Control Systems Technology, IEEE Transactions on*, vol. 16, pp. 789-798, 2008.

[123] E. Chu and D. Gorinevsky. (Jan 10). *Detecting aircraft performance anomalies from cruise flight data*. Available: http://www.mitekan.com/images/PDF/aiaa2010_cgb.pdf

[124] F. Caliskan and C. M. Hajiyev, "Ekf based surface fault detection and reconfiguration in aircraft control systems," in *Proceedings of the 2000 American Control Conference*, 2000, pp. 1220-1224 vol.2.

[125] R. Kalman, "A new approach to linear filtering and prediction problems," *Transaction of the ASME - Journal of Basic Engineering*, 1960.

[126] G. Gan, *et al.*, *Data clustering: Theory, algorithms, and applications*. Philadelphia: SIAM, 2007.

[127] P. Witold, "Conditional fuzzy c-means," *Pattern Recognition Letters,* vol. 17, pp. 625-631, 1996.

[128] T. Chidester, "Understanding normal and atypical operations through analysis of flight data," in *Proceedings of the 12th International Symposium on Aviation Psychology*, Ohio, 2003, p. 239.

[129] M. Ford, "Use of data mining techniques during the investigation of boeing 777-236er g-ymmm," 2008.

[130] S. Budalakoti, *et al.*, "Anomaly detection and diagnosis algorithms for discrete symbol sequences with applications to airline safety," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews,* vol. 39, pp. 101-113, 2009.

[131] T. Stibor, *et al.*, "A comparative study of real-valued negative selection to statistical anomaly detection techniques artificial immune systems." vol. 3627, C. Jacob, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2005, pp. 262-275; F. González and D. Dasgupta, "Anomaly detection using real-valued negative selection," *Genetic Programming and Evolvable Machines,* vol. 4, pp. 383-403, 2003.

[132] J. Davis, "The design of an evolutionary algorithm for artificial immune system based failure detector generation and optimization," MSc, West Virginia University, 2010.

[133] K. Krishnakumar, "Artificial immune system approaches for aerospace applications," presented at the 41st Aerospace sciences meeting & exhibit, Nevada, 2003.

[134] Z. Ji and D. Dasgupta, "Applicability issues of the real-valued negative selection algorithms," in *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, New York, 2006.

[135] T. Stibor, *et al.*, "Is negative selection appropriate for anomaly detection?," in *Proceedings of the 2005 conference on genetic and evolutionary computation*, New York, 2005.

[136] P. Angelov, *et al.*, "An approach to automatic real-time novelty detection, object identification, and tracking in video streams based on recursive density estimation and evolving takagi–sugeno fuzzy systems," *International Journal of Intelligent Systems*, vol. 26, pp. 189-205, 2011.

[137] D. P. Filev and F. Tseng, "Novelty detection based machine health prognostics," in *Evolving Fuzzy Systems, 2006 International Symposium on*, 2006, pp. 193-199.

[138] D. Filev, *et al.*, "Real-time driving behavior identification based on driver-in-the-loop vehicle dynamics and control," *IEEE International Conference on Systems, Man and Cybernetics, SMC 2009*, 2009, pp. 2020-2025.

[139] R. O. Duda, *et al.*, *Pattern classification*, 2nd ed. Chichester: Wiley-Interscience, 2000.

[140] R. Ramezani, *et al.*, "A fast approach to novelty detection in video streams using recursive density estimation," in *Intelligent Systems, 2008. IS '08. 4th International IEEE Conference*, 2008, pp. 14-2-14-7.

[141] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, pp. 85-126, 2004.

[142] V. Chandola, *et al.*, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, 2009.

[143] F. E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11, pp. 1-21, 1969.

[144] G. John, "Robust decision trees: Removing outliers from databases," in *Proceedings of the first international conference on knowledge discovery and data mining*, 1995.

[145] H. A. Ali and T. Alex, "Developing neuro-fuzzy hybrid networks to aid predicting abnormal behaviours of passengers and equipments inside an airplane," 2009, p. 73520G.

[146] K. Yongmin and P. Jaehong, "On the approximation of fault directions for mutual detectability: An invariant zero approach," *IEEE Transactions on Automatic Control* , vol. 50, pp. 851-855, 2005.

[147] F. Caliskan and C. M. Hajivyev, "Innovation sequence application to aircraft sensor fault detection: Comparison of checking covariance matrix algorithms," in *Decision and Control, 1999. Proceedings of the 38th IEEE Conference on*, 1999, pp. 3956-3961 vol.4.

[148] Y. Zhang and X. R. Li, "Detection and diagnosis of sensor and actuator failures using imm estimator," *IEEE Transactions on Aerospace and Electronic Systems* , vol. 34, pp. 1293-1313, 1998.

[149] P. S. Maybeck and P. D. Hanlon, "Performance enhancement of a multiple model adaptive estimator," *IEEE Transactions on Aerospace and Electronic Systems* , vol. 31, pp. 1240-1254, 1995.

[150] P. Zipfel, *Modeling and simulation of aerospace vehicle dynamics*, 2nd ed. Virginia: American Institute of Aeronautics and Astronautics, 2007.

[151] U. Dynamics, "Aerosim: Aeronautical simulation blockset," 1.2 ed: Unmanned Dynamics, 2009.

[152] A. Nuic. (2009, User manual for the base of aircraft data (bada) revision 3.7. Available: http://www.eurocontrol.int/eec/gallery/content/public/document/eec/report/2009/003_BADA_3_7_User_manual.pdf

[153] A. Smith, *et al.*, "Feasibility demonstration of an aircraft performance risk assessment model," in *Digital Avionics Systems Conference, 2000. Proceedings. DASC. The 19th*, 2000, pp. 4D4/1-4D4/8 vol.1.

[154] K. P. Murphy. (2001, Jan 13). *The bayes net to olb ox for matlab*. Available: http://www.interfacesymposia.org/I01/I2001Proceedings/KMurphy/KMurphy.pdf

[155] A. Nuic, *et al.*, "Advanced aircraft performance modeling for atm:

Enhancements to the bada model," presented at the 24th Digital Avionics System Conference, Washington D.C., 2005.

[156] T. E. O. F. T. S. O. A. Navigation. *Bada performance file*. Available: http://www.eurocontrol.int/eec/gallery/content/public/documents/EEC_ACE_BADA_documents/B763__.PTF

[157] A. Sutcliffe, "Scenario-based requirements engineering," *11th IEEE International Proceedings* of *Requirements Engineering Conference, 2003,* 2003, pp. 320-329.

[158] K. Allenby and T. Kelly, "Deriving safety requirements using scenarios," in *Proceedings. Fifth IEEE International Symposium on Requirements Engineering*, 2001, pp. 228-235.

[159] L. Beinlich, *et al.*, "The alarm monitoring system: A case study with two probabilistic inference techniques for belief networks," in *Second European Conference on Articial Intelligence in Medicine*, 1989, pp. 247-256.

[160] B. Sierra, *et al.*, "Using bayesian networks in the construction of a bi-level multi-classifier. A case study using intensive care unit patients data," *Artificial Intelligence in Medicine*, vol. 22, pp. 233-248, 2001.

[161] J. Ramon, *et al.*, "Mining data from intensive care patients," *Advanced Engineering Informatics*, vol. 21, pp. 243-256, 2007.

[162] J. V. Tu, *et al.*, "Multicenter validation of a risk index for mortality, intensive care unit stay, and overall hospital length of stay after cardiac surgery," *Circulation*, vol. 91, pp. 677-684, February 1, 1995 1995.

[163] T. Charitos, *et al.*, "A dynamic bayesian network for diagnosing ventilator-associated pneumonia in icu patients," *Expert Systems with Applications*, vol. 36, pp. 1249-1258, 2009.

[164] P. J. F. Lucas, *et al.*, "A probabilistic and decision-theoretic approach to the management of infectious disease at the icu," *Artificial Intelligence in Medicine*, vol. 19, pp. 251-279, 2000.

[165] D. Apiletti, *et al.*, "Real-time analysis of physiological data to support medical applications," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, pp. 313-321, 2009.

[166] K. J. Tuman, *et al.*, "Morbidity and duration of icu stay after cardiac surgery. A model for preoperative risk assessment," *Chest*, vol. 102, pp. 36-44, July 1, 1992.

[167] R. Dybowski, *et al.*, "Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm," *The Lancet*, vol. 347, pp. 1146-1150, 1996.

[168] J. H. T. Bates and M. P. Young, "Applying fuzzy logic to medical decision making in the intensive care unit," *American Journal of Respiratory and Critical Care Medicine*, vol. 167, pp. 948-952, April 1, 2003 2003.

[169] G. D. Cliord, *et al.* (2011, 01 Feb). *User guide and documentation for the mimic ii database (2.6 ed.)*. Available: http://mimic.physionet.org/UserGuide/UserGuide.pdf

[170] L. Anthony, *et al.*, "A clinical database-driven approach to decision support: Predicting mortality among patients with acute kidney injury," *Journal of Healthcare Engineering*, vol. 2, pp. 97-110, 2011.

[171] C. Hug, *et al.*, "Clinician blood pressure documentation of stable intensive care patients: An intelligent archiving agent has a higher association with future hypotension," *Crit Care Med*, vol. 39, pp. 1006–1014, 2011.

[172] G. D. Clifford, *et al.*, "An artificial vector model for generating abnormal electrocardiographic rhythms," *Physiol Meas*, vol. 31, pp. 595–609, 2010.

[173] Physionet. (2011, 01 Feb). *Publications*. Available: http://mimic.physionet.org/publications.html

[174] A. Aboukhalil, *et al.*, "Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform," *Journal of Biomedical Informatics,* vol. 41, pp. 442-451, 2008.

[175] A. H. Ali and Others, "Monitering system," United Kingdom Patent, 2011.