8-11-2017

Bayesian Network Analysis for Diagnostics and Prognostics of Engineering Systems

Marc D. Banghart

Bayesian network analysis for diagnostics and prognostics of engineering systems

By

Marc D. Banghart

A Dissertation
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in Industrial and Systems Engineering
in the Department of Industrial and Systems Engineering

Mississippi State, Mississippi

August 2017

Bayesian network analysis for diagnostics and prognostics of engineering systems

By

Marc D. Banghart

Approved:

____

Linkan Bian
(Major Professor)

____

Lesley Strawderman
(Committee Member)

____

Andreas Tolk
(Committee Member)

____

Kari Babski-Reeves
(Committee Member)

____

Stanley Bullington
(Graduate Coordinator)

____

Jason M. Keith
Dean
Bagley College of Engineering

Name: Marc D. Banghart

Date of Degree: August 11, 2017

Institution: Mississippi State University

Major Field: Industrial and Systems Engineering

Major Professor: Linkan Bian

Title of Study: Bayesian network analysis for diagnostics and prognostics of engineering systems

Pages in Study: 105

Candidate for Degree of Doctor of Philosophy

Bayesian networks have been applied to many different domains to perform prognostics, reduce risk and ultimately improve decision making. However, these methods have not been applied to military field and human performance data sets in an industrial environment. Methods frequently rely on a clear understanding of causal connections leading to an undesirable event and detailed understanding of the system behavior. Methods may also require large amount of analyst teams and domain experts, coupled with manual data cleansing and classification. The research performed utilized machine learning algorithms (such as Bayesian networks) and two existing data sets. The primary objective of the research was to develop a diagnostic and prognostic tool utilizing Bayesian networks that does not require the need for detailed causal understanding of the underlying system. The research yielded a predictive method with substantial benefits over reactive methods. The research indicated Bayesian networks can be trained and utilized to predict failure of several important components to include potential malfunction codes and downtime on a real-world Navy data set. The research also considered potential error within the training data set. The results provided credence

to utilization of Bayesian networks in real field data – which will always contain error that is not easily quantified. Research should be replicated with additional field data sets from other aircraft. Future research should be conducted to solicit and incorporate domain expertise into subsequent models. Research should also consider incorporation of text based analytics for text fields, which was considered out of scope for this research project.

DEDICATION

I dedicate this research to Shannon Banghart. You have always inspired me and shown me to persevere and never lose sight of the bigger picture. You have taught me self-reflection and ever reminded me of the beauty of science – while ensuring I do not forget the endless possibilities our universe keeps hidden.

“You have your way. I have my way. As for the right way, the correct way, and the only way, it does not exist.” Friedrich Nietzsche.

ACKNOWLEDGEMENTS

This work would not have been possible without the support of several individuals and organizations.

First, I would like to thank my committee for all their mentoring and support. I extend a special thank you to Dr. Bian and Dr. Tolk for their mentorship and guidance throughout my doctoral studies. I would like to especially thank my committee chair, Dr. Bian. Thank you for the words of encouragement whenever I felt overwhelmed. To the rest of my committee I express my sincere thanks for sharpening my technical skills and the countless hours spent involved in my research.

I would like to thank the United States Air Force for the tremendous opportunities I have been given, which made this research possible. I would like to thank my family to include my mother, Quinta Banghart and sister Ella Banghart for the many hours they have spent listening to my research ideas. I would like to thank the forgotten judge critiquing my high school science project. Although I have forgotten your name, your words “always keep thinking” will remain with me forever. I would like to thank David Nelson who has always supported my professional development. David has always shown me what true leadership and professional engineering is.

Finally, I would like to thank all my unnamed friends for their support and always believing in me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

"To know what you know and what you do not know, that is true knowledge." Confucius

Engineering design and optimization relies on the application of knowledge of the natural world and sciences in order to solve problems, develop new products, or improve processes. Frequently these products or processes form part of larger complex systems. Complex systems pose significant challenges in terms of modeling and analysis due to their hierarchical nature, feedback loops, and failure propagation properties. Additionally, these systems typically involve both human and machine elements and are subject to environmental influences.

Uncertainty is inevitable. Uncertainty further introduces risk in terms of performance, safety, and cost. Uncertainty must be considered, both during engineering design as well as the remainder of the product or process lifecycle. Mathematical modeling of uncertainty is critical in order to gain knowledge and effectively develop or refine products, systems or processes.

The following dissertation is focused on applying Bayesian modeling and analysis to two important data sets in order to address several research gaps. Specifically, the research probes the question of uncertainty in engineering applications through the development of mathematical frameworks used to assist in performing statistical inference and prediction. We investigate both a military aircraft field data set as well as a human performance data set representative of an industrial process. We further focus on the evaluation of noise introduction within a training set and the associated algorithm performance and stability.

The dissertation is organized as follows. We start by broadly discussing complex systems theory and analysis approaches, prognostic and diagnostic systems along with current research gaps. Our discussion includes an overview of military maintenance management systems, since the research was conducted utilizing an associated data set. Next, we discuss the third main section of research covering the development of Bayesian networks on military aircraft for prediction and improvement of military readiness. We illustrate our method with a case study on the EA-6B aircraft. Next, we apply several machine learning algorithms, to include Bayesian networks, to a human performance data set. Finally, we perform a sensitivity analysis on the Bayesian Network analysis for a military readiness application. This is followed by concluding remarks.

1.1 Complex Systems Theory and Analysis

Complex systems are typically hierarchical in nature, contain feedback loops, and include large amounts of component interactions. Additionally, complex systems may be adaptive in nature, thus changing their behavior or outputs based on previous experience or measurements. These systems may also include significant human/machine interaction (Ladyman, Lambert, & Weisner, 2013).

Complex systems pose significant challenges in terms of analysis and prediction. This is due to the inherent nature of these systems, where cascading failures may occur and relationships are not linear. Thus, a failure of a component may lead to catastrophic events as it cascades through the system. Additionally, a small change may result in a large effect (Randall, 2011). Complex systems are challenging to model due to the unique properties of these systems, such as nonlinearity. Thus, individual components within complex systems are less important, with more focus placed on interactions.

Interactions within complex systems may be unintended or unanticipated. Many examples can be found where system complexity was at least in part a contributor to accidents or undesired outcomes. For example, in the nuclear industry, the task to plan for natural disasters, human error, and hardware failure is inherently complex. The planning encompasses risk analysis with the goal of identifying possible scenarios given one or more conditions along with their severity. One major component of this type of risk assessment also involves developing contingency or action plans given a set of conditions. In the nuclear industry this is called Severe Accident Management Guidelines (SAMGs) (Groth, Denman, Cardoni, & Wheeler, 2014). Development of these SAMGs involves forecasting potential scenarios and the associated actions taken in order to avoid a severe consequence, such as a reactor meltdown. Groth et al. (Groth, Denman, Cardoni, & Wheeler, 2014) proposes a novel application of Bayesian belief networks in order to promote a dynamic risk assessment strategy of high risk events. This allows greater flexibility and possibly better results during safety related events. Their analysis applies a Bayesian belief network to a nuclear power plant system. They considered the complex arrangement of several valves along with different levels of associated core damage.

1.2 Maintenance Management Systems and Field Data Collection

Computerized Maintenance Management Systems (CMMS) are utilized throughout both the commercial and military sectors to capture maintenance activities (i.e. field data) performed on a system(s). These data sets include narratives of the symptoms observed, codes intended to classify events, as well as when the maintenance was performed. These systems, when coupled with logistics demand, are utilized to ensure spare parts are available and track metrics related to the status of aircraft (ready for mission, in maintenance etc.). Additionally, these systems frequently track metrics providing insight into downtime (or readiness). For example, fields such as how many hours an aircraft waited for spare parts or maintainers can be tracked and utilized to calculate readiness metrics.

It has been well documented that CMMS data, to include military maintenance data, may contain errors. Typically, concerns within the data include:

- Discrepancies may be initiated by any number of reasons, not just failure. Pilots may report problems in flight, or ground crew may identify issues. Alternatively, discrepancies may be opened to simply perform an inspection or upgrade a system. Thus, failures must generally be identified by combining various codes associated with each record.
- Record coding likely contains large amounts of human error. As previously discussed, relying on coding to identify failures can be problematic. The data for military aircraft is captured globally from many different units and technicians. The technicians have different experience levels and familiarity with the data entry system, which may introduce error.
- The system is not closed loop in nature. The data collection process is not closed loop – meaning that documented failures are not verified as true failures. For example, a failure may be reported resulting in the replacement of a component.

The following day, the same failure, or one very similar, may be reported with another repair executed. Thus, it is not possible to ascertain if the first event was truly a failure or rather just inadequate troubleshooting.

- Data is gathered in a variety of environments. Data collection occurs both in combat and home station environments. The data does include a field indicating if collection was performed overseas; however, the impact of the environment is unknown. For example, less stringent quality assurance practices may be in place in a combat environment, coupled with a different aircraft duty cycle.
- Discrepancies are not described using a standardized grammar and nomenclature. The discrepancy and corrective action narratives frequently contain spelling errors or utilize different nomenclature and/or descriptions. Thus, keyword searches may or may not capture all the desired records.

Although these data sets are challenging to utilize and typically do not contain sufficient information to quantify remaining useful life of components, they may still contain useful information in terms of readiness determination, process optimization, and risk analysis. Additionally, in many cases other data sets that contain higher-fidelity information simply may not be available (or feasible) to collect.

Sensor technology continues to proliferate, along with the incorporation of diagnostic and prognostic capabilities. This trend of development of intelligent systems is driven by many factors, to include increased complexity, reduced manning, and the goal of optimizing uptime. Meeker and Hong discuss opportunities and challenges in regard to the proliferation of sensors, and the associated impact on field data (Meeker & Hong, 2013). They highlight that sensors are proliferating that also capture environmental variables, degradation of a component or system, as well other potential indicators of failure. Meeker and Hong further discuss that the short-term prediction of system failure, warning of emerging reliability concerns, remaining useful life prediction, and maintenance cost prediction are important applications of reliability data (Meeker & Hong, 2013). Meeker and Hong highlight several related research needs, to include the development of methods that can not only utilize field data, but also combine physics-based knowledge and expert opinion (Meeker & Hong, 2013).

1.3 Prognostics and Diagnostics

The underlying concept of Prognostic Health Monitoring, consisting of both diagnoses and prognosis, is illustrated in Figure 1.1, and is closely coupled with sensor proliferation.

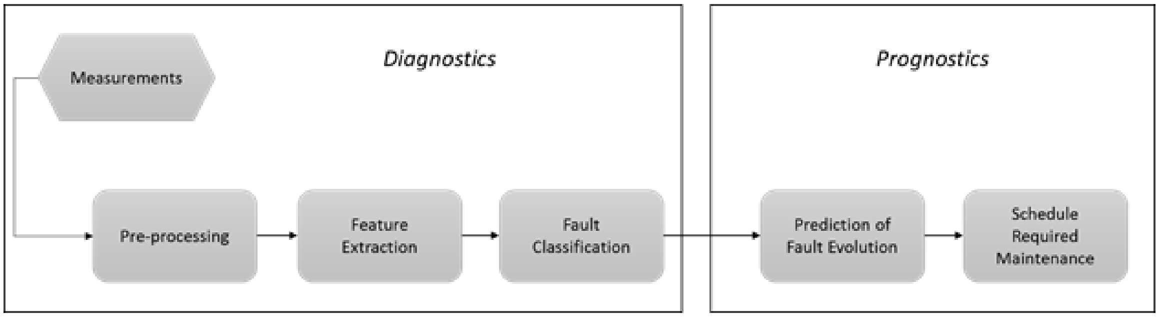


Figure 1.1 Diagnostic/Prognostic Process (Vachtsevanos, Lewis, Roemer, Hess, & Wu, 2006)

Diagnostic sensors and strategies have been around for decades. In order to understand the difference between diagnosis and prognosis, the failure progression timeline must be considered. Diagnosis typically refers to the identification of a part, component, or system that is either in a degraded or failed stated. Thus, the event has already occurred and, at best, activities to assess the impact can be pursued. Prognostics however aims to detect the event before it occurs. Thus, we aim to provide a sufficient window of opportunity where the decision maker can act before the high-risk event occurs. Naturally, such strategies have substantial benefit, to include improved safety, and reduced downtime and cost. In the context of maintenance systems, a prognostics algorithm would enable the prediction of future maintenance and readiness problems, as well as the associated maintenance actions. Additionally, such a system would have key benefits to include: (Vachtsevanos, Lewis, Roemer, Hess, & Wu, 2006):

- Improved planning of maintenance and associated manpower levels
- Targeting of training requirements
- Less reliance on time-consuming inspections
- Prepositioning of required resources

Several methodologies are currently utilized in order identify readiness degraders (or perform inference) in the military domain. These methodologies include post fielding supportability analysis, top degrader analysis, and reliability centered maintenance (P. & T., 2012; Lambert, Stock, & Ellram, 1998; Moubray, 1997; Blanchard, 2008). Although the approaches vary in methodology, in most cases field data is utilized along with anecdotal reports to estimate degraders in terms of readiness, safety or cost.

Several challenges must be overcome. First, data sets available for analysis are typically error prone due to the method and environment in which they are collected. Ideally, methods that can identify and remove the error should be developed. However, identification of the error may be difficult or impossible. The specific concerns with military field data will be discussed in subsequent sections within this dissertation. Additionally, the underlying systems representing these data sets are complex in nature with dynamic changes in missions and operating context.

1.4 Current Research Gaps

There are several drawbacks and limitations that common utilized analysis methods share. Methods such as top degrader analysis and Reliability Centered Maintenance pose challenges due to resource requirements needed to perform the analysis. In these analyses, large teams of domain experts must manually assess significant amounts of data and identify both the frequency and severity of failure modes. This task is problematic and potentially subjective (Banghart, Bian, & Babski-Reeves, 2016) (Shebl, Franklin, & Barber, 2012). Additionally, the task is further complicated when data sets contain large amounts of error, which is typical in field data.

Degraders to readiness can also be viewed through a risk assessment perspective. Many approaches have been proposed to perform risk assessment in the literature, however several problems remain. Frequently, methods rely on a clear understanding of causal connections leading to an undesirable event, which may not be known or readily identifiable. Kalman filters have also been proposed as a prediction algorithm due to their well-known predictive power in other applications. However, as illustrated by Villez et al, they suffer from a high false alarm rate, thus eroding confidence in the method (Villez, Srinivasan, Rengaswamy, Narasimhan, & Venkatasubramanian, 2011). Several research gaps remain in terms of human performance evaluation and prediction.

The existing approaches to predict human performance can be generally categorized into several categories: (1) qualitative approaches; (2) quantification of a human failure rate; and (3) machine learning based approaches. However, all of these methods have limitations. Qualitative approaches focus on the identification of factors degrading task performance and are potentially highly subjective. These methods are also difficult to generalize to multiple domains. Quantification of a human failure rate is problematic since performance-shaping factors must be well understood. Lastly, although several relevant machine learning techniques have been investigated in terms of human performance, none have been applied to an industrial setting.

In summary, this research will address the lack of proactive and predictive tools focused on process improvement and understanding of risk in human-machine engineering systems. The primary research objective is to develop system diagnosis and prognostic tools utilizing Bayesian networks in order to improve the performance of complex system.

CHAPTER II

PREDICTION OF READINESS UTILIZING BAYESIAN NETWORKS ON THE ON THE EA-6B AIRCRAFT

"Anyone who has to fight, even with the most modern weapons, against an enemy in complete command of the air, fights like a savage against modern European troops, under the same handicaps and with the same chances of success." Erwin Rommel

In 2010, the U.S. Office of the Secretary of Defense released a memorandum in an effort to highlight the state of weapon system reliability within the Defense Acquisition community. Specifically, the memorandum stated that a large percentage of systems have not been meeting requirements, resulting in new weapon systems not achieving operationally ready status on schedule. These delays coupled with budget cuts (GAO, 2015) have resulted in several aging military platforms to remain in service far longer than expected. This trend poses a significant challenge that must be overcome in order to operate the fleet both safely and cost effectively, while ensuring mission success.

Mission success can only be achieved by utilizing predictive methods to identify problems before they occur, and thus reduce risk. Methodologies to identify degraders to readiness in terms of logistics, manpower and reliability must be developed, applied and validated in this domain. These methodologies will allow proactive planning and risk assessment in order to anticipate readiness degraders before they occur, and thus take positive actions to reduce the consequences or prevent occurrence. These methodologies should also utilize current data collection systems if possible, in order to not impose additional data collection requirements in a fiscally restrictive environment.

Several methodologies are currently utilized to identify degraders to readiness within the military domain (P. & T., 2012; Lambert, Stock, & Ellram, 1998; Moubray, 1997; Blanchard, 2008). The approaches vary in methodology, however generally field data along with anecdotal reports are utilized within the process. Simplistic methods simply trend metrics such as downtime. These metrics are reactive in nature, since lagging indicators are trended. Thus, once an increase in downtime has been observed, the decision maker may not have a sufficient window of opportunity to act and reduce the overall impact. Methods such as reliability centered maintenance utilize a top down approach, first considering the function of a system or component, followed by development of failure modes that result in loss of function. The failure modes are quantified utilizing data sources such as field data. Part of the quantification includes assessment of the likelihood of occurrence and severity of the failure mode. Thus, the activity does support risk assessment. However, the challenge remains of sifting through large amounts of error prone field data, which imposes significant resource requirements. This task is problematic and potentially subjective (Banghart, Bian, & Babski-Reeves, 2016) (Shebl, Franklin, & Barber, 2012). Additionally, the task is further complicated when data sets contain large amounts of error, which is typical in field data. In complex systems interactions within failure modes will not be easy to capture – thus the analysis may miss several miss high risk cascading failures.

Utilization of field data poses a significant challenge. Specifically, the data frequently contains significant amounts of error and is fairly large in terms of the number of maintenance events. Metrics in order to assess readiness are frequently simple arithmetic averages of failure times (such as Mean Time Between Failure) or percentages calculated based on how many hours assets were available for use. In some cases, probability distributions are fit to repair, delay or failure times. These metrics are generally lagging indicators. For example, if a component exhibits a large logistics delay time readiness has already been impacted, and the weapon system is already in a failed state.

In order to overcome the limitations of commonly utilized methods for prediction of readiness degraders we applied a Bayesian network to a field data set obtained from the U.S. military. Bayesian networks have been applied to many different domains to include human performance and the nuclear industry (Ramana, 2011). Our methodology included contrasting several structure learning algorithms, with tabu search yielding the best results. By considering and utilizing field data representative of what is typically collected we developed a model that can be applied to noisy data, while yield accurate results. Additionally, our approach allows development of monitoring systems of field data to continuously set evidence within the Bayesian network allowing for real time assessment. Our method was tested utilizing a large data set representing the EA-6B aircraft.

The chapter is organized as follows. We provide a literature review covering methods of risk assessment currently being utilized during engineering design, or post design when systems are fielded. We then discuss applications of Bayesian networks to related problems. Next, we provide mathematical details of Bayesian networks and provide our research method. Lastly, we apply a Bayesian network to military data set in order to illustrate the concept along with concluding remarks.

2.1 Literature Review

Several approaches have been proposed in literature to improve or perform risk assessment. These include Bayesian methodologies as well as hybrid approaches incorporating concepts such as Kalman Filtering.

2.1.1 Bayesian Methods

Bayesian approaches have been widely applied in a wide variety of domains. For example, Kalantarnia et al. utilized a Bayesian Probabilistic approach to risk assessment. They consider risk in terms of three major steps: hazard identification, hazard assessment and risk estimation (Khan, Husain, & Abbasi, 2002; Crowl & Louvar, 2002; Kalantarnia, Khan, & Hawboldt, 2009). Kalantarnia et al. also recognize the importance of risk assessment within the process industry and the static nature of current approaches, which do not capture variations in risks induced by configuration or environmental changes. Kalantarnia et al. further highlight that often only high risk events are analyzed with near misses frequently not identified or ignored. They propose an algorithm that utilizes an underlying event tree with quantification of failure probabilities. They then apply Bayesian analysis to calculate a posterior probability of a respective end-state, which maps to a risk event. They utilize three methods to calculate prior probabilities. The methods include a deterministic approach (failure counts), probabilistic method (development of a failure distribution) and a Monte Carlo simulation approach to form failure probabilities. Their approach has a major problem. Firstly, the assumption is made that an event tree that will capture all paths that lead to a risk event can be determined. This is a daunting task for any complex system, and as other authors illustrate risks events are frequently missed during analyses (Kalantarnia, Khan, & Hawboldt, 2009). Their approach does introduce the concept of Bayesian updating of risks, however requires a clear mapping to event tree items in order to develop the network structure.

The development of a Bayesian network topology utilizing Failure Modes Effects Analysis (FMEA) has also been proposed in the literature (Akhlaghi, Naseh, Mirshams, & Irani, 2011). Mao and Canavero propose a Bayesian network model that unifies three tools to include Fault Trees, Event Trees and Electromagnetic Topology (EMT) to perform system-level vulnerability assessment in electromagnetic environments. They further propose that Bayesian networks can overcome limitations of Fault Trees and should be further applied (Mao & Canavero, 2016). Yuan et al. applies Bayesian networks to analyze very common dust explosions in industrial environments. They further illustrate the concept of risk updating by updating probabilities of root events and consequences within their formulation (as additional information becomes available) (Yuan, Khakzad, Khan, & Amyotte, 2015). Eliassi et al. formulate a Bayesian network model in order to assess the impact of protection system failures on a power grid. In their formulation the Bayesian network contains several layers to include components, minimal cut sets, system and an electrical bus layer (Eliassi, Seifi, & Haghifam, 2015).

Feng, Wang and Li (Feng, Wang, & Li, 2014) present a methodology that utilizes Bayesian network in order to facilitate risk assessment in an Information Technology (IT) security environment. The Bayesian network consists of factors that can be used to assess security risk, and is constructed based on expert opinion (via an algorithm developed by the authors). Then, a real time database of cases is utilized to update each observable node. If a certain probability threshold is reached the authors initiate additional analysis to ascertain the impact on the overall IT network (network vulnerability).

There are several limitations of existing applications of Bayesian networks. Many of the existing methods assume that it is possible to identify the underlying causal chain of events leading to a high-risk event. Additionally, the methods frequently rely on expert opinion which may be subjective.

2.1.2 Methods Incorporating the Kalman Filter

The Kalman filter uses system's dynamics model (e.g., physical laws of motion), known control inputs to that system, and multiple sequential measurements (such as from sensors) to form an estimate of the system's varying quantities (its state). It has been widely applied in radar and navigation problems, but also found use in economics and data fusion. Kalman filters have been applied in order to detect and diagnose fault conditions. As illustrated by Villez et al. in a buffer tank system several failure modes of a small system can be successfully predicted by a Kalman filter (Villez, Srinivasan, Rengaswamy, Narasimhan, & Venkatasubramanian, 2011). This is accomplished by continuously evaluating residuals between sensor measurements and Kalman filter predictions. This is accomplished by calculating the Mahalonobis distance of residuals to the origin given an expected covariance matrix of the prediction residuals. They further apply their formulation to a nonlinear system utilizing an Extended Kalman Filter.

Although their formulation was focused on fault identification (and not risk assessment) they illustrated that Kalman filters do not function well in nonlinear systems within this context (Villez, Srinivasan, Rengaswamy, Narasimhan, & Venkatasubramanian, 2011). Specifically, in their simulation they encountered too high of false alarm rate in the case of fault identification (Villez, Srinivasan, Rengaswamy, Narasimhan, & Venkatasubramanian, 2011).

Xu et al. developed an integrated physics-statistics-based model incorporating an adaptive Kalman filter in order to predict both reliability of components, to include cases where only a small sample size of failure data is available (Xu, Wei, Chen, & Kang, 2015). The authors utilize accelerated degradation test data as an input to their model. They integrate Physics-of-Failure (PoF) methods with statistical models. This allows them to capture underlying relationships between product life, material properties and environmental factors as well as describe the randomness of a degradation process among individual products. They further expand stochastic degradation models into two aspects. First parameters describing the characteristics of the product degrading process must be estimated, followed by parameters describing the variation between products. They start by estimating the parameters of a time dependent Physics-of-Failure model with accelerated test data. Next, they apply a Brownian motion stochastic process in order to capture inter-product randomness. The Brownian motion process includes a drift parameter which is updated utilizing a Kalman filter. Finally, the parameters of the Kalman Filter are quantified using Maximum-Likelihood-Estimation and accelerated test data (Xu, Wei, Chen, & Kang, 2015). This method is useful when accelerated test data is available, and the underlying Physics-of-Failure model is well-understood. Although novel, the method cannot be abstracted to risk management in a real-time sense since accelerated test data is not captured in this manner. Thus, this method cannot be directly utilized to measure risk in a complex system that is fielded.

Kalman filter based techniques are powerful methods to estimate the state of a system, and were investigated as part of this research. However, these methods typically require substantial input data, thus in cases where only a few observations are available they may not adequately predict risk.

2.2 Overview of Typical Military Field Data

The military collects large amounts of field data as part of normal operations. At the core of the collected data are individual maintenance events, illustrated in Table 2.1. These events describe a discrepancy or inspection and an associated action taken. Each record contains an associated date, aircraft serial number, system/subsystem identification, associated codes and meta-data fields.

Table 2.1 Example of Typical EA-6B Maintenance Records

Aircraft Serial Number	Detachment	Discrepancy	Corrective Action	Owner Org
161242	HOME	On takeoff Port Main Landing Gear showed barber poled after Gear Retraction. On extension, all Gear showed down and locked. No Gear transition light after Starboard Main showed up and locked.	Workcenter 220 repaired broken grounding wire causing the unsafe indication. Workcenter 120 performed operational checkout of the Landing Gear System IAW NA 01-85ADC-2-23.1A.1. System Checks fully operational. Area FOD and corrosion free.	FAG
161881	FRS	Port Main Landing Gear Forward Door does not close properly.	Verified rigging of Port Main Landing Gear Forward Door as required IAW NA-01-85ADC-2-3. Found that Port Forward Door Cylinder was outside of rigging limits. Adjusted as needed. Performed Operational check of Landing Gear and Emergency Landing Gear Systems.	FAE

Discrepancies, defined as an undesired physical condition, are initiated for various reasons to include failures. Pilots may report problems in flight, or ground crew may identify issues. Alternatively, discrepancies may be created to simply perform an inspection or upgrade a system. Each discrepancy includes several codes selected by the maintainer. These include codes identifying when the maintenance occurred, what type of maintenance was performed and what malfunction was observed.

Codes applied to each maintenance event or discrepancy likely contain large amounts of human error. As stated before relying on coding to identify failures can be problematic. The data for military aircraft is captured globally from many different units and technicians. The technicians have different experience levels and familiarity with the data entry system, which may introduce error.

2.3 Bayesian Network for Modeling Readiness Degraders

The goal of the research was to investigate probabilistic and predictive approaches to ascertain the readiness state of complex systems, where readiness is described by downtime, repair time and influenced by the respective component or malfunction code. Quantification of readiness will support risk assessment, since proactive actions can be taken to reduce the probability of adverse events (such as high downtime). Graphically a Bayesian network can be represented as a network with nodes and arcs. Nodes represent variables, while arcs represent probabilistic relationships. For example, supply availability may impact non-mission capable time thus an arc is drawn from the respective nodes to capture this relationship. The structure of these networks can be defined qualitatively utilizing expert opinion or quantitatively by several search algorithms.

The joint probability distribution for the Bayesian network formulation can be represented by considering that not every node is connected to all other nodes. Consider a Bayesian network that contains n nodes, X_1 to X_n, where $n = 10$ in this example formulation. A particular probability in the joint distribution can be represented by $P(x_1, x_2, \ldots, x_{10})$.

The joint probability can be factorized per the chain rule as such:

$$\begin{aligned} &P(x_1, x_2, \dots, x_n) \\ &= P(x_1)P(x_2|x_1)\dots, P(x_{10}|x_1, \dots x_9) \\ &= \prod_{i}^{10} P(x_i|x_1, \dots x_{i-1}) \end{aligned} \tag{2.1}$$

However, per the Markov Property the value of any particular node is only conditional on its respective parent nodes, resulting in:

$$\begin{aligned} &P(x_1, x_2, \dots, x_{10}) \\ &= \prod_{i} P\big(x_i \big| Parents(x_i)\big) \end{aligned} \tag{2.2}$$

Where $Parents(x_i) \subseteq \{x_1, \dots, x_{i-1}\}$

Several important variables within this data set were utilized in this formulation. Firstly, risk in a military environment is a function of reliability, maintainability, safety and readiness. Readiness is a function of both logistics constraints (delay times, stock levels) and maintenance actions (repair time, available resources).

In order to construct the predictive model, the structure of the Bayesian network must first be defined. Bayesian networks can be defined utilizing several methods to include manually constructing the network or utilizing search algorithms. The analysis investigated several search algorithms such as the hill climber and tabu algorithms. The analysis also considered utilizing a Naïve Bayes formulation since it has been shown to perform equally (or even outperform) more complex algorithms by several authors. Gacquer et al. applies six machine learning algorithms to an air pollution data set, and illustrate similar results in terms algorithm efficiency and accuracy when comparing

Naïve Bayes to k-nearest neighbors, Support Vector Machines and Decision trees (Gacquer, Delcroix, Delmotte, & Piechowiak, 2011).

The hill climbing algorithm is commonly utilized in order to identify the structure of a Bayesian network utilizing a training data set. The algorithm selects an arbitrary solution, which is then incrementally improved utilizing a local search procedure. This is accomplished by modifying a single element each iteration in such a way that the solution is improved each step. One strength of the tabu search algorithm is that moves are allowed, which select worse solutions as long as these moves are not contained within the tabu list. This algorithm has been shown to find a local optimum fairly well (Skiena, 2010). More advanced algorithms such as tabu search have been proposed and utilized. The tabu search algorithm utilizes local searches in order to identify an improved solution within immediate neighborhood of the current solution. Neighbors are defined as solutions that are similar with only minor differences.

Let S represent a set of moves that lead from one solution to another.

$$s: X(s) \rightarrow X, s \in S \tag{2.3}$$

Where $X\ is\ the\ solution\ space\ and\ S\ is\ a\ set\ of\ moves.$

Given Equation 2.3 we can formulate an optimization problem (Equation 2.4) and subsequent procedure for solving it (Bouckaert, 2004).

$$Minimize\ c(x): x \in X \tag{2.4}$$

Where $c(x)\ is\ he\ objective\ function$

We start the tabu search by selecting a random or arbitrary initial solution ($x \in X$) and setting the optimal solution (represented by x^*) to the current solution. A tabu list is maintained comprising the set of moves that would undo previous moves in t recent iterations, where t is defined as the tabu tenure. The tabu list is important in order to prevent the algorithm from cycling back to the same local optimum.

The tabu list is mathematically given by (Bouckaert, 2004):

$$TL = \{s^{-1}: s = s_i, i < k - t\} \tag{2.5}$$

Where $k\ is\ the\ iteration\ index, s^{-1}\ is\ the\ inverse\ of\ the\ move\ s$

During each iteration (until stopping criteria are reached) the counter is increment and a move, $s_k \in S - TL$ is selected such that $s_k(x_{k-1}) = OPTIMUM(s(x_{k-1}): s_k \in S - TL)$. Next, the best solution currently found is computed as follows. Let $x_k = s_k(x_{k-1})$. Next if $c(x_k) < c(x^*)$ let $x^* = x_k$.

Local score metrics allows for local search measures when identifying the underlying network structure given a training data set. Specifically, the overall network quality metric score can be calculated as the summation of the score of all individual nodes. Several score metrics were investigated and contrasted within this analysis to include the entropy metric, Akaike information criterion (AIC), minimum description length (MDL) metric, and Bayesian metric.

The concept of entropy is frequently utilized in information theory where a transmitter sends messages through a channel to a receiver. Entropy is utilized to refer to the expected value of the information contained in each message. This concept has been utilized within Bayesian network search algorithms where given precisely stated prior data a probability distribution is chosen that maximized entropy.

Mathematically the entropy metric, $H(B_S, D)$ where B_s represents the resulting network structure and D the input data is given by (Bouckaert, 2004):

$$H(B_S, D) = -N \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \frac{N_{ijk}}{N} \log \frac{N_{ijk}}{N_{ij}} \tag{2.6}$$

Where $q_i\ is\ the\ cardinality\ of\ the\ parent\ set\ of\ x_i$
$r_i\ is\ the\ cardinality\ of\ x_i$
$N_{ij}\ denotes\ the\ number\ of\ records\ in\ D\ where\ pa(x_i)\ takes\ its\ jth\ v$
$N_{ijk}\ denotes\ the\ number\ of\ records\ in\ D\ where\ pa(x_i)\ takes\ its\ jth$
$for\ which\ x_i takes\ its\ kth\ value$

Next the AIC metric is given by (Bouckaert, 2004):

$$Q_{AIC}(B_S, D) = H(B_S, D) + \sum_{i=1}^{n} (r_i - 1) q_i \tag{2.7}$$

The MDL metric is given by (Bouckaert, 2004):

$$\begin{aligned} Q_{MDL}(B_S, D) = {} & H(B_S, D) \\ & + \frac{\sum_{i=1}^{n} (r_i - 1) q_i}{2} \log N \end{aligned} \tag{2.8}$$

The Bayesian metric utilizes a prior on the network structure and the gamma function and is given by (Bouckaert, 2004):

$$\begin{aligned} & Q_{Bayes}(B_S, D) \\ & = P(B_S) \prod_{i=0}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ij} + N_{ijk})}{\Gamma(N'_{ijk})} \end{aligned} \tag{2.9}$$

Where $P(B_s) is\ the\ prior\ on\ the\ network\ structure$
$N'_{ij}\ and\ N'_{ijk}\ represent\ choices\ of\ priors\ on\ counts$

Algorithms performance can be measured utilizing several performance parameters. The main goal of these performance parameters are to ascertain which algorithm (if any) could be useful in predicting high delay times, component failure, maintenance characteristics and malfunction codes.

The first performance parameter considered was simply the percentage of correctly classified instances in a data set. Next, a kappa statistic was calculated. The level of agreement between the classification rule of the algorithm when compared against the observations can be expressed utilizing the kappa statistics. Other metrics such as the true positive rate and false positive rates also provide insight into algorithm performance. The true positive rate (or recall) is expressed as the proportion of instances classified to belong to class x, among all instances that actually do belong to class x. Precision was calculated as the proportion of instances that truly belong to class x, among all instances classified as belong to class x. The receiver operating characteristic (ROC) can be calculated by plotting the true positive rate and false positive rate followed by integration. The area under the curve close to 0.50 implies that the results lack statistical independence. Finally, the F-measure metric can be calculated by considering both algorithm precision and recall, as provided in Equation 2.10.

$$F - Measure = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \tag{2.10}$$

Overfitting of data to a model can result in poor predictive performance. Thus, in order to ensure overfitting did not occur a 10-fold cross validation scheme was utilized. The method estimates parameters via averaging after partitioning the data into equal sized samples equal to the number of folds. One of the samples is reserved for validation.

2.4 Case Study Introduction and System Description

The EA-6B Prowler is an electronic warfare aircraft originally introduced into service in 1971. Over 170 aircraft were built and the platform was the only dedicated electronic warfare platform that could be utilized in joint operations involving all United States (U.S.) military branches for a significant portion of its service life. The aircraft was flown by the U.S. Navy and Marine Core and finally retired in 2015. The EA-18G Growler replaced the Prowler.

The data set used in this analysis consisted of all maintenance work orders for the EA-6B landing gear system, cockpit warning/caution annunciator panel and the environmental control system turbine assembly. The data set was obtained utilizing the Freedom of Information Act (FOIA). These systems/components were selected from the larger data set for several reasons. First, they perform different but very important functions on the aircraft. These functions consist of warning the aircrew of unsafe conditions to providing shock absorption of landing and take-off loads. Additionally, they have different inherent designs ranging from electrical to hydro-pneumatic systems. Thus, the underlying failure mechanisms and duty cycles are vastly different. For example, landing gear accumulate damage from take-off and landing while electrical components wear out generally due to cycles.

Although specific details or pictures of the turbine assembly of the EA-6B are not available the principles of operation are described in terms of generic aircraft design. In jet aircraft, such as the EA-6B, air cycle air conditioning systems are frequently utilized within the Environmental Control System (ECS). These systems utilize engine bleed air in order to pressurize the aircraft cabin appropriately. Hot bleed air (from the engine) is routed through several heat exchangers and an expansion turbine that successfully cools the air down. The air is finally mixed with ram air in order to achieve the desired cabin temperature. The air cycle machine (ACM) forms the heart of such a system with the primary function of compressing air from the primary heat exchanger prior to routing to the secondary heat exchanger. The ACM is driven by a turbine assembly, which was considered in this analysis.

The EA-6B annunciator panel provides important warning messages to the aircrew via illuminated lights. Each warning light is illuminated by two incandescent light bulbs. Several failure modes apply to this panel to include burnt light bulbs, electrical connector damage, and corrosion to name a few. Cascading failure modes were not considered in this study since sufficient aircraft design information was not available. The annunciator is illustrated in Figure 2.1 for reference.

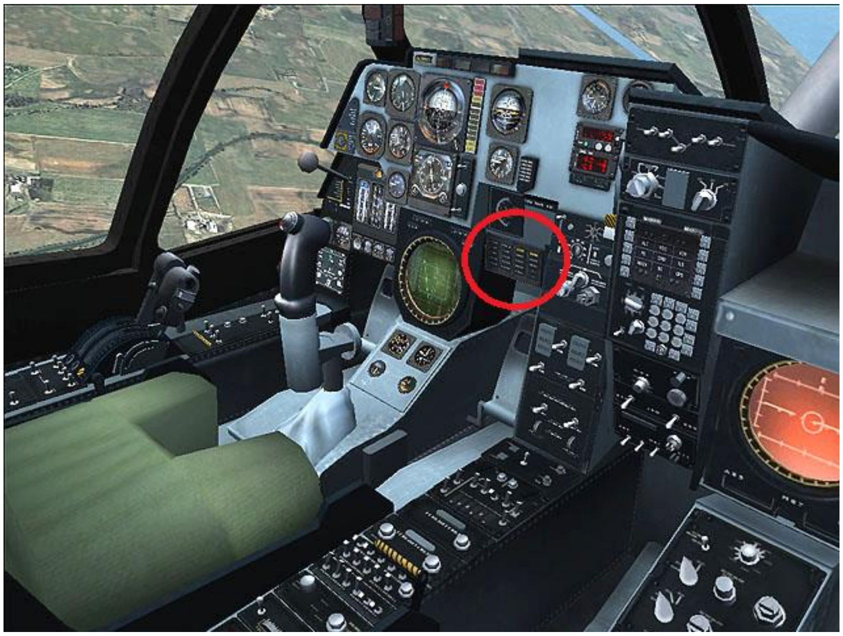


Figure 2.1 EA-6B Cockpit (obtained from http://www.fspilotshop.com)

The EA-6B landing gear system is illustrated in a block diagram as per Figure 2.2. The block diagram represents relationships or interfaces between parts. The system is operated by movement of a lever (to the up or down position) in the cockpit. Next, a main landing gear (MLG) sequence valve receives electrical signal to open the landing gear bay doors. The MLG Hydraulic door actuator opens (and holds doors open) using hydraulic power. Next, the MLG Sequence Valve electrically signals MLG actuator to retract/extend gear. Once the landing gear is extended/retracted, the MLG Sequence Valve signals doors to close. The doors close and the pilot receives indication of MLG status. Several switches are used in the system to determine gear status (up/down/in-transit) which are not shown here. Additionally, the gear is locked in the down position by a mechanical over-center mechanism.

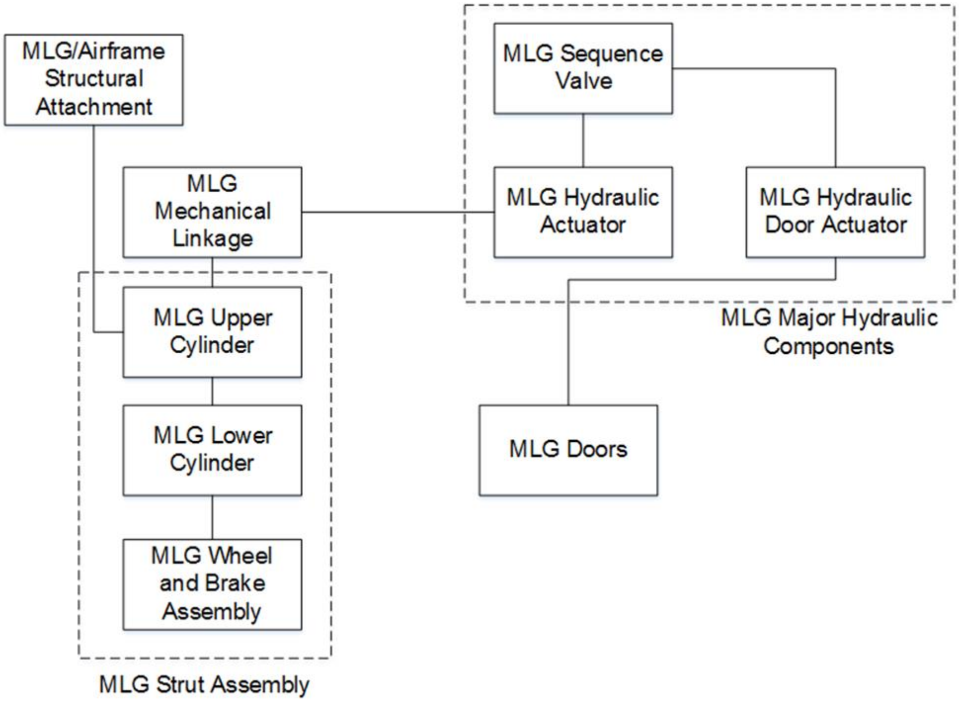


Figure 2.2 Main Landing Gear Components

2.4.1 Case Study Descriptive Statistics

The data set consisted of 1451 records and 18 variables. The first variable considered was the component name as given in the data set, which included 73 unique components. For each record the aircraft tail number was provided. Variables to include the location of the reporting squadron (deployed or U.S.), organizational codes for both the aircraft owner and operational unit were also provided. Maintenance information was provided to include the work center and maintenance level (field or depot). The date of the record and several codes were included. These related to when the discrepancy was discovered, what type of maintenance was performed and what kind of action was taken. Delay times and repair times were also provided. Finally malfunction code was included.

Data pre-processing was performed in order to convert several of the numeric variables to nominal scales. Numeric variables were mapped to a nominal scale utilizing percentiles with a negligible, low, medium, high and very high categories utilized. Categories mapped to numeric values as provided in Table 2.2 based on the 25th, 50th and 75th percentiles.

Table 2.2 Descriptive Statistics (zero values removed)

Variable	**25th Percentile**	**Median**	**75th Percentile**	**Maximum**
Repair Time hours	1.30	3.50	8.60	202.70
Awaiting Maintenance Time (AMT) hours	0.63	1.50	7.08	1,272.00
Awaiting Parts Time (APT) hours	0.30	1.30	33.70	3,944.60
Non-mission capable (NMC) hours	2.10	3.80	13.60	1,335.00
Corresponding Nominal Scale	Low	Medium	High	Very high

As illustrated in Figure 2.3, there was a large amount of variation observed in several variables. Boxplots were calculated based on inclusion and exclusion (marked with *) of cases were delay or repair time was zero. As can be seen, excluding these values significantly impacts calculation of the various percentiles. These data points were likely a combination of entry error, or in some cases very little delay time was actually incurred. These data points were not manipulated, and assigned a nominal scale value of “negligible”, since additional information was not available. Outliers were also present in the data set. However, these outliers were not removed from the analysis since they are likely actual observations (versus errors). For example, although repair times are typically low for the majority of components cases may arise where aircraft are in a down state for significant periods of time. This may be due to limited supply or manpower. The impact of the potential error within the data set is further investigated in chapter 4 of this dissertation.

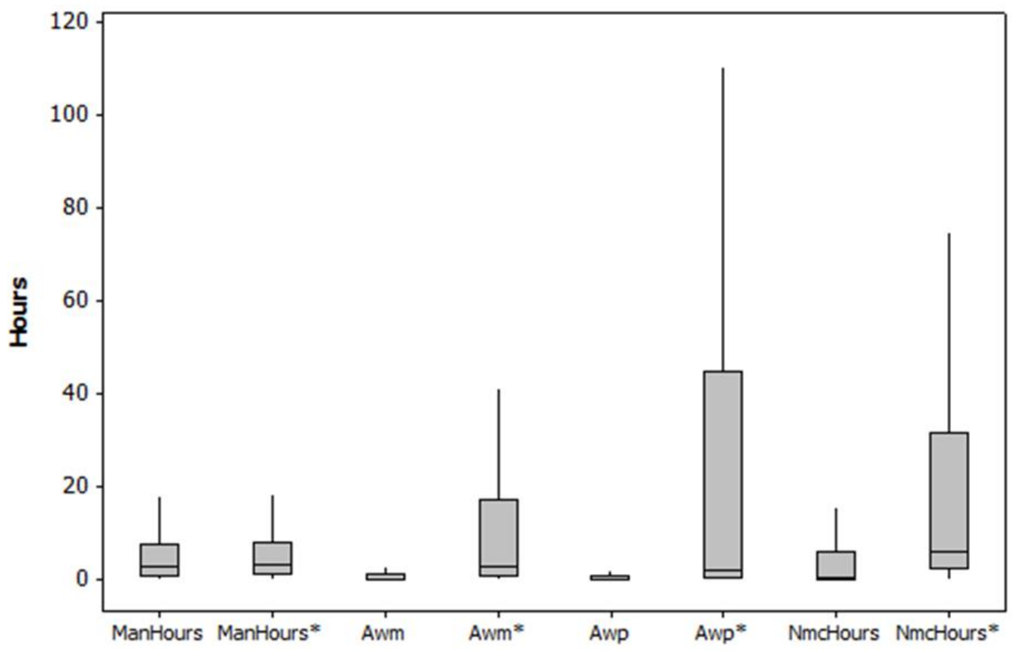


Figure 2.3 Comparison of Variables with zero-hour data point included and excluded

A Pareto chart of malfunction codes is also provided for the annunciator panel. As illustrated in Figure 2.4 malfunction code W62 and 374 were most prevalent (18.4 % of data set). W62 corresponded to defective fuse(s), switches, diodes, light bulbs or another consumable while 374 denoted internal failure of a component.

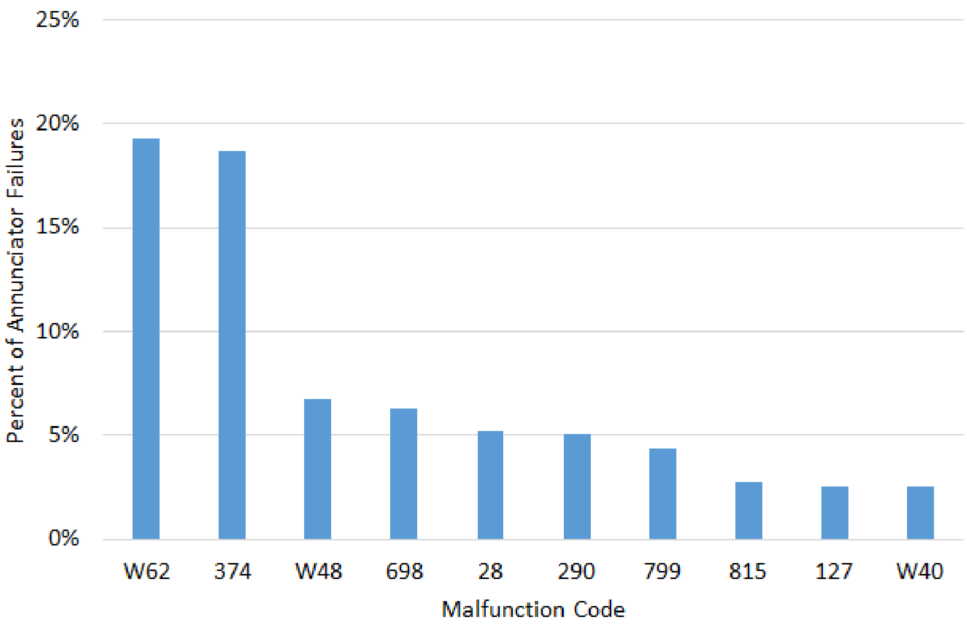


Figure 2.4 Pareto of Malfunction Codes of the Annunciator Panel

2.4.2 Case Study Results and Discussion

Several search algorithms were investigated and compared in terms of performance. Performance parameters were calculated when predicting three important variables. Component, malfunction code and NMC hours were chosen since they provide insight into potential readiness issues and are generally utilized to identify top degraders. In general, all the algorithms performed similarly, achieving the highest accuracy (>97%) when NMC hours was predicted. The tabu search algorithm performed slightly better than all others with an accuracy of 70% when predicting component, 62% for malfunction code and 98% for NMC hours. Overall performance of each search algorithm is provided in Table 2.3.

Table 2.3 Bayesian Network Algorithm Overall Performance

Variable to Predict	**Component**		**Malfunction Code**		**NMC Hours**	
Search Algorithm	**% Correctly Classified Instances**	**Kappa statistic**	**% Correctly Classified Instances**	**Kappa statistic**	**% Correctly Classified Instances**	**Kappa statistic**
Naïve Bayes	68.2%	0.58	62.3%	0.54	97.4%	0.96
K2	68.0%	0.59	61.1%	0.53	97.5%	0.96
Hill Climbing	69.7%	0.61	61.1%	0.53	98.0%	0.97
Tabu search	70.0%	0.61	62.0%	0.54	98.0%	0.97

Additional results when the tabu search algorithm was utilized is provided when predicting several variables next, and summarized in Table 3. As illustrated in Table 4 high accuracy (True Positive Rate > 0.85) was achieved when NMC hours, Action Taken and Type Maintenance were predicted. Awaiting maintenance, when discovered and the component itself were predicted with a true positive rate > 0.70. Lastly malfunction codes and awaiting parts hours could be predicted with a true positive rate > 0.60. The false positive rate was typically less than 5 % depending on which variable was predicted, and was deemed sufficiently low.

The tabu search algorithm can be further customized by modifying several input parameters. These include the maximum number of parent nodes for each child node, the maximum tabu list size and starting and stopping conditions for the algorithm. Several combinations of modifications were tested, which did improve the algorithm results. The initial solution or network structure can be set to utilize a Naïve Bayes formulation or completely random. This condition was set to false, which decreased the true positive rate when component was predicted by 2 %. Setting the Markov Blanket and Naïve Bayes properties to true, resulted in an increase of the true positive rate (for component prediction) to 73 %. The number of iterations were also increased from 10 to 20. This yielded improvement across the board in terms of true positive rates (Component = 73 %, Malfunction Code 66 %, Awaiting Maintenance Time 77 %, Awaiting Parts Time 65 % and NMC hours 98 %.

Table 2.4 Bayesian Network Algorithm Class Performance for TABU Search (Bayesian metric - Baseline)

Predictive Variable	**Weighted Average across all Classes**					
	True Positive Rate	**False Positive Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**
Component	0.70	0.04	0.69	0.70	0.69	0.95
Malfunction Code	0.62	0.04	0.56	0.62	0.57	0.93
NMC Hours	0.98	0.00	0.98	0.98	0.98	1.00
Awaiting Maintenance Time	0.74	0.02	0.81	0.74	0.76	0.94
Awaiting Parts Time	0.60	0.09	0.68	0.60	0.62	0.86
When Discovered Code	0.74	0.05	0.79	0.74	0.76	0.95
Action Taken Code	0.90	0.03	0.89	0.90	0.89	0.98
Type Maintenance Code	0.90	0.04	0.96	0.90	0.93	0.98

Sensitivity analysis was conducted in order to identify if any of the local search quality metrics resulted in superior results. The Bayes quality metric was deemed most accurate, with the entropy metric yielding several results. However, when the AIC and MDL metrics were utilized results plummeted below 50% accuracy.

The sensitivity analysis considered varying inputs into the algorithm such as the maximum number of parent nodes for each child node, tabu list size and number of iterations. It was determined that only marginal improvements were obtained, although algorithm computation time increased significantly.

The resulting Bayesian network was also graphically inspected in order to ascertain identified causal patterns by the search algorithm. The search algorithm identified a network that appeared to capture logical relationships between variables. For example, the failed component would impact variables such as repair time and NMC hours. Additionally, relationships between the organizational code and the location of the aircraft is appropriate.

Next, prediction results for specific components were investigated. As can be seen in Table 2.5, false positive rates remained low. The results indicated that several major components such as the caution lights, ECS turbine and MLG wheel assembly failure could be predicted with high accuracy (true positive > 0.80).

Table 2.5 Prediction results per class for “component”.

True Positive Rate	False Positive Rate	Precision	Recall	F-Measure	ROC Area	Component
1.00	0.00	1.00	1.00	1.00	1.00	Setscrew
0.96	0.03	0.40	0.96	0.56	0.99	Brake assembly
0.91	0.04	0.94	0.91	0.93	0.99	Caution lights
0.80	0.02	0.62	0.80	0.70	0.99	ECS turbine
0.77	0.07	0.79	0.77	0.78	0.95	Main Landing Gear wheel assembly
0.67	0.00	1.00	0.67	0.80	0.98	Panel assembly, land
0.61	0.07	0.59	0.61	0.60	0.92	Nose Landing Gear wheel/tire assembly
0.56	0.02	0.26	0.56	0.36	0.95	Landing Gear control selector valve
0.50	0.01	0.30	0.50	0.38	0.97	Nose Landing Gear retract/actuator cylinder

Identification of malfunction codes is important during the analysis process, beyond only identifying which component is likely to fail. This is tied to the inherent consequences. Certain malfunction codes, such as those related to overheated components may be indicative of higher risk. As can be seen in Table 2.6 several malfunction codes could be predicted with high precision. Interestingly, cannibalization of parts could be predicted with a true positive rate of 100 %. This finding is of great potential benefit to maintenance planners. Specifically, being able to predict when the supply system will not have a part available, and the maintenance strategy must resort to cannibalization can be utilized to pre-position components. Additionally, a malfunction code representing no defect was also predicted with high accuracy. This prediction is very useful in further analyzing the troubleshooting capability of a maintenance program.

Specifically, predicting when a no defect action likely will occur can reduce the burden on logistics and coupled with aircraft history flag additional investigation.

Table 2.6 Prediction results per class for "malfunction code".

True Positive Rate	False Positive Rate	Precision	Recall	F-Measure	ROC	Class	Mal code description
1.00	0.00	0.79	1.00	0.88	1.00	815	Cann. Action
0.93	0.05	0.91	0.93	0.92	0.99	787	Tire removal – normal wear
0.93	0.02	0.48	0.93	0.63	0.99	799	No defect
0.89	0.07	0.59	0.89	0.71	0.95	374	Internal failure
0.86	0.00	0.55	0.86	0.67	0.93	571	Magnetic particle inspection
0.77	0.06	0.50	0.77	0.61	0.96	W62	Defective fuses, switches, diodes and light bulbs.
0.75	0.00	0.50	0.75	0.60	1.00	801	No defect
0.71	0.06	0.26	0.71	0.38	0.96	290	Fails diagnostic test
0.54	0.01	0.41	0.54	0.47	0.98	28	Conductance incorrect

Prediction of wiring malfunction codes (denoted by W) did not perform well. As can be seen in Table 2.7 only one malfunction code could be predicted with any level of accuracy. The malfunction code of W62 corresponds to malfunction of fuses, switches, diodes and light bulbs. The poor performance in the wiring domain was expected by the researchers. Wiring malfunctions are inherently difficult to troubleshoot leading to high amounts of likely error within the data set. Additionally, malfunction codes for wiring are difficult to utilize in the field and may be assigned in a somewhat random fashion by technicians.

The lack of predictive accuracy for the majority of malfunction codes related to wiring was expected to some extent. The military currently utilizes a large amount of malfunction codes for wiring, assuming that the maintainer will select the most representative code. However, there are several concerns with the coding scheme that likely result in the maintainer selecting a limited subset. Firstly, the assumption is made that the maintainer has enough expertise and/or information of the underlying failure mechanisms in order to select the corresponding code. This is likely not the case, thus more generic codes are selected. Additionally, having hundreds of codes may result in the maintainer simply selecting one of the first codes listed versus evaluating the entire list.

Table 2.7 Prediction results per class for "malfunction code" (wiring failures only)

True Positive Rate	False Positive Rate	Precision	Recall	F-Measure	ROC	Class
0.77	0.06	0.50	0.77	0.61	0.96	W62
0.33	0.01	0.22	0.33	0.27	0.98	W40
0.06	0.00	0.29	0.06	0.10	0.96	W48

Bayesian networks can be utilized within a decision making process in several capacities. Statistical inference can be performed by setting evidence within the network and evaluating the resultant posterior probability across variables of interest. For example, given the validated network provided in Figure 2.5 we wish to determine the joint probability of the NMC hours being either negligible or very high. If we deem the probability to be negligible we can infer high downtime risk is low. The opposite would hold for the case of non-mission capable hours being very high. For brevity the associated conditional probability tables are not displayed in the figure.

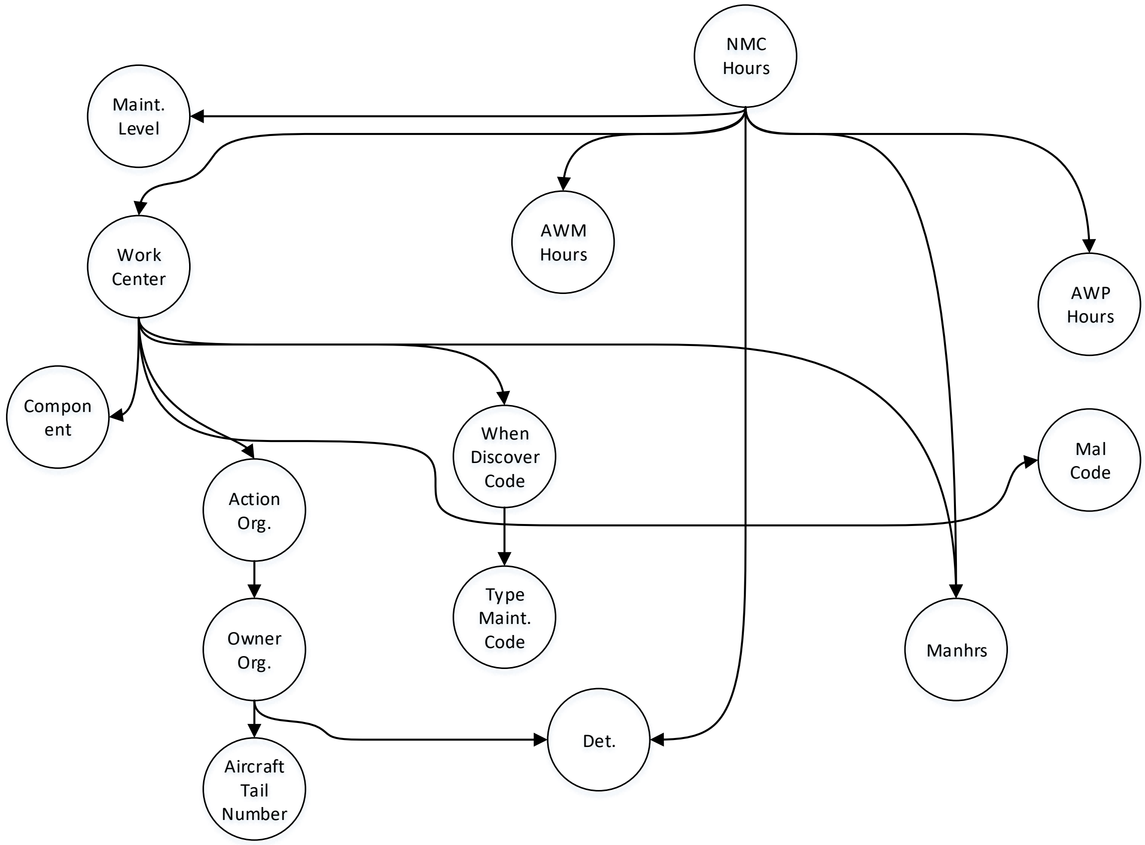


Figure 2.5 Bayesian Network for NMC Hours Target Feature

As an illustrative example consider the failure of two components; caution lights or the ECS turbine. We set evidence within the network that the failed component will be either the caution lights or ECS turbine and calculate the posterior probability for the different NMC categories. We can also set additional evidence regarding the specific failure mode if desired. As illustrated in Figure 2.6, the baseline model does not indicate a high probability of significant downtime (with no evidence set). This holds for failure of the caution lights. However, failure of the ECS turbine significantly increase the probability of significant downtime.

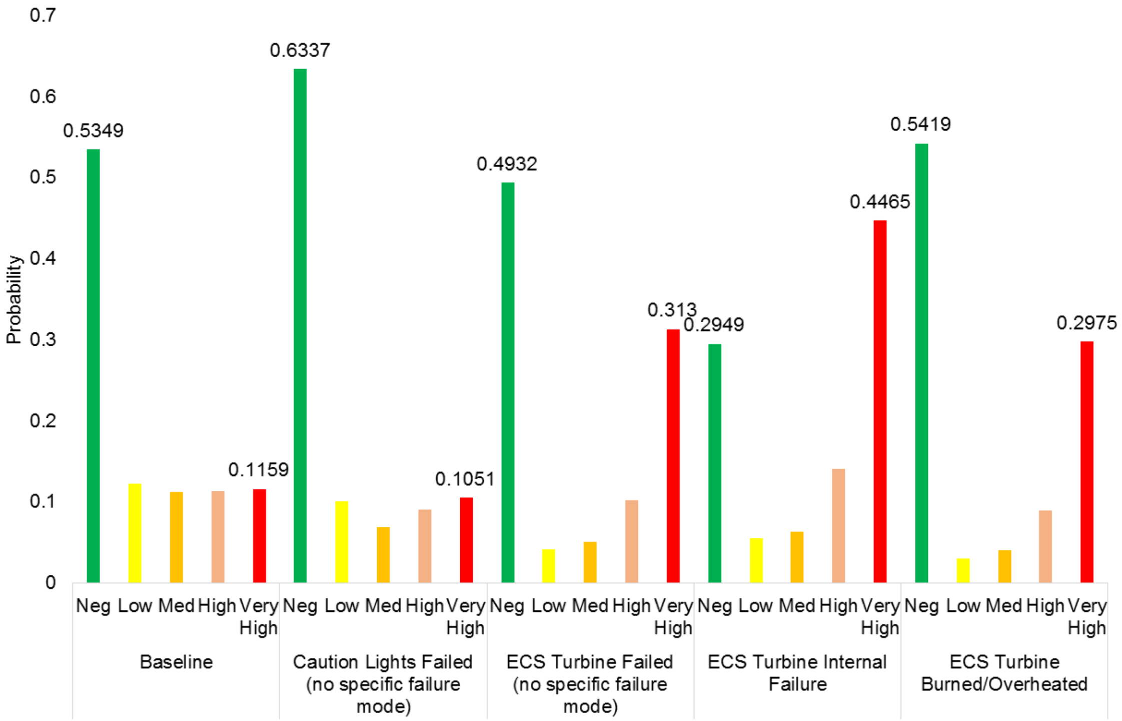


Figure 2.6 Calculated Posterior Probability given Different Failure Scenarios

2.5 Conclusion and Future Work

The research indicated that Bayesian networks are a viable method to support fielded systems to ascertain the likelihood of potential degraders to readiness. Although prediction accuracy of the associated component and malfunction code should be improved, the results are promising. Several variables within the Bayesian network exhibited large amounts of variation. The impact of this variation as well as coding scales utilized on prediction results must be further studied.

Specifically, the method was able to predict failure of several important components to include potential malfunction codes. The algorithm performed extremely well when comparing NMC hours – a critical metric for the military and measure of downtime risk.

The approach provides a predictive method that will yield substantial benefits over reactive methods, since the validated Bayesian network can be utilized to assess predict potential future outcomes. Additionally, the method does not rely on explicit understanding of causal connections within the system(s), nor identification of sequences of events leading to failure. This allows broad application of the technique. Future research is needed in order to identify the impact of qualitative variables from other analyses such as Failure Modes Effects Analysis on the associated network and integration with current data management and reporting systems.

CHAPTER III

HUMAN PERFORMANCE PREDICTION IN AN INDUSTRIAL WORKPLACE USING MACHINE LEARNING METHODS

"To err is human; to admit it, superhuman." Doug Larson

3.1 Introduction

Industrial process optimization is critical in order to minimize cost, while maximizing production. Typically, processes are designed with major emphasis placed on optimization of processing time, cost optimization and minimization of safety concerns to workers. Less emphasis is placed on human performance prediction, even though poor performance of workers may substantially reduce productivity (Copeland, 2015). In order to both design optimal and/or improve existing processes, greater emphasis must be placed on human performance within the system. Human performance must be examined from both demographic factors and their relationship to error rates. Quantitative models that can be utilized to predict human performance are powerful decision tools that can be used during process design and/or optimization. Several authors have utilized machine learning techniques in order to investigate human performance. Jantan et al. investigated if employees should be promoted based on several attributes such as gender, education level, knowledge and skill. Jantan et al. applied the C4.5 decision tree algorithm to human talent prediction in a human resources context (Jantan, Hamdan, & Othman, 2010). These studies were focused on prediction of human performance in a task domain that largely required problem solving by participants. Thus, these results can likely not be applied to very different task domains that require physical responses to cues by participants (i.e. move box from container 1 to 2). Additionally, these studies did not investigate the impact of feedback mechanisms on employee performance. Finally, the assessment of machine learning algorithms in order to predict human performance remains largely unstudied within the manufacturing and service industry. This chapter addresses this gap in literature by conducting an analysis of a dataset obtained in an industrial setting. Several mathematical models were utilized within this study in order to assess applicability of machine learning techniques within this domain. These models consisted of both unsupervised learning approaches as well as graphical methods.

The existing approaches can be generally categorized into several categories: (1) qualitative approaches; (2) quantification of a human failure rate; and (3) machine learning based approaches. However, these methods have limitations. Qualitative approaches focus on identification of factors degrading task performance and are potentially highly subjective. These methods are also difficult to generalize to multiple domains. Quantification of a human failure rate is problematic since performance shaping factors must be well understood. Lastly, although several relevant machine learning techniques have been investigated in terms of human performance none have been applied to an industrial setting.

The objective of this chapter is to assess the capability of machine learning algorithms to predict human performance in an industrial environment in order to support process improvement. Machine learning, originally defined by Arthur Samuel in 1959 as a “field of study that gives computers the ability to learn without being explicitly programmed”, has become a fascinating method of pattern recognition and predicting outcomes based on input data (Simon, 2013). The technique has been applied to a wide range of problems ranging from animal conservation, heart attack prevention and security. One major benefit of machine learning is that patterns are automatically identified along with classification rule(s). This classification rule can then be applied to new datasets for inference purposes, which can optimize processes, improve safety and reduce cost (Simon, 2013). The proposed methodology is demonstrated using real-world data collected from a study emulating a distribution facility. The data set included both human error rates and worker demographics. Based on subsequent causal links identified within the Bayesian network we can identify important demographics related to worker performance. Moreover, human error performance is predicted, and can be subsequently utilized to optimize industrial processes.

3.2 Literature Review

3.2.1 Qualitative Approaches

According to Koopmans et al., “there is no consensus on the definition and measurement of individual work performance” (Koopmans L. , Bernaards, Hildebrandt, de Vet, & van der Beek, 2014). They further state that various terms are utilized to describe work performance, often with unclear definitions. Individual work performance has been defined in certain domains such as organizational psychology as “behaviours or actions that are relevant to the goals of the organization” (Campbell, 1990). In this formulation individual work performance is constrained by behaviours under the control of the individual – with the environment not considered (Rotundo & Sackett, 2002).

Factors such as the environment clearly do play a role in human performance from a human factors viewpoint. For example, poor lighting would likely increase errors since individuals may miss visual cues. Koopmans et al. describe a heuristic framework (Koopmans L. , et al., 2011) that has recently been proposed within literature. This framework describes individual work performance in terms of four generic dimensions. These include task performance (individual task proficiency), contextual performance (employee behaviours that support social, organizational and psychological environment of workplace), adaptive performance (employee capability to respond to change) and counterproductive work (employee behaviours harmful to organizational goals). Koopmans et al. utilized the aforementioned heuristic framework in a study aimed at identification of indicators and their relative mapping to the framework dimensions. For example, prioritization is an indicator that can be mapped to task performance (Koopmans L. , Bernaards, Hildebrandt, de Vet, & van der Beek, 2014). Their study was multi-faceted, and compiled potential indicators from literature, medical databases and a survey of 253 participants. The authors then mapped these indicators to their most appropriate dimension. The participants were provided all the indicators and tasked identify the top six most important within each dimension. The results were then compiled by the researchers and the number of votes for each indicator was calculated. The authors acknowledge several potential issues with this study.

Firstly, the mapping of indicators may be subjective and context specific. Furthermore, even though participants were asked to consider a generic job (versus their specific job) when voting for indicators – it is not possible to measure if participants were able to transcend their own field of work. The study identified work quality, planning and organizing work, being results-oriented, prioritization and working efficiently as the top indicators of task performance (Koopmans L. , Bernaards, Hildebrandt, de Vet, & van der Beek, 2014).

3.2.2 Human Failure Rate Quantification

Human performance has also been studied significantly under the umbrella of Human Reliability Analysis (Mkrtchyan, Podofillini, & Dang, 2015). Several techniques to include the Standardized Plant Analysis Risk-Human Reliability Analysis (SPAR-H) have been developed in order to assess risk (Mkrtchyan, Podofillini, & Dang, 2015). These methods identify a set of factors which may influence behaviour, which are then further utilized in order to calculate a quantitative human failure rate. A plethora of approaches have been proposed and utilized in academia and industry to study human errors. These methods can be categorized as either task or context based. More recent research has also developed methods that combine tasks and context based approaches. In task based approaches the inherent task is deemed to have inherent failure mechanism – thus if performed by a human will have some probability of failure. This probability can be influenced by performance shaping factors and error-forcing conditions. Thus, this approach requires several data elements that must either be estimated or calculated in order for this approach to be utilized in a practical way (Mkrtchyan, Podofillini, & Dang, 2015).

Context based approaches, such as “A Technique for Human Error Analysis – ATHEANA” (Barriere, 2000) investigate in what context or domain the task must be performed. If a task is performed in an environment where the operator faces complex scenarios they may be more likely to commit an error. For example, a pilot may be more susceptible to human error while dogfighting versus level flight, which intuitively makes sense.

Techniques that incorporate both task and context based approaches can be found throughout literature. In 2009, Bell & Holroyd reported over 72 different techniques (Bell & Holroyd, 2009)! As Moura et al. (Moura, Beer, Patelli, & Lewis, 2015) points out that all these techniques are complicated endeavours due to the large uncertainties related to several variables (behavioural characteristics, technology aspects and organizational context), which leads to "reasonable concern about the accuracy and practicality of such probabilities". Furthermore, these approaches generally require an exhaustive list of tasks that an operator must perform and selection of performance shaping factors can be subjective.

3.2.3 Machine Learning Formulations

Software quality continues to plague both the commercial and defence sectors (Thakur, Gupta, & Gupta, 2015). Software quality and/or reliability is a function of human performance. Thus, the human element is considered one of the most important concerns within the IT sector (Thakur, Gupta, & Gupta, 2015). Researchers have proposed that one method to improve software quality requires development of an "ideal selection framework" during talent acquisition (Thakur, Gupta, & Gupta, 2015). One fundamental problem with development of such a framework is understanding which individual attributes may be good predictors of software quality. Additionally, relating these attributes to appropriate software performance metrics is also challenging.

Singh utilized machine learning algorithms to predict the number of defects in a software project (Singh, 2009). Chien and Chen (Chien & Chen, 2008) considered age, gender, marital status and education as predictors of employee performance in the semiconductor industry. They utilized data mining techniques and found that education along with work experience are indeed predictors of employee performance (Chien & Chen, 2008).

Thakur et al. (Thakur, Gupta, & Gupta, 2015) point out that previous research in this domain have not yielded highly accurate prediction models. They further comment that prior works heavily utilize decision trees, which are prone to overfitting. Lastly, they also considered more robust attribute variables such as programming skill, domain specific knowledge, communication and reasoning skills. Their performance parameter has three levels; good, average and poor. This parameter was quantified utilizing a brainstorming technique by a group of managers for each employee. In order to develop a higher accuracy prediction, Thakur et al. utilize the random forest algorithm. This technique utilizes bootstrapping and develops a large collection of individual prediction trees. Performance parameters are then calculated by averaging metrics such as true positive, false positive and the ROC across all the trees. Their model resulted in high true positive rates for all classes (good = .93, average = .85, poor = .93) and associated low false positive rates (good = 0.04, average = 0.04, poor = 0.08) (Thakur, Gupta, & Gupta, 2015). They further found that Grade Point Average (GPA) alone is not a good predictor of performance. Domain knowledge, analytical and programming skills appeared most important (Thakur, Gupta, & Gupta, 2015).

The study by Thakur et al. highlights that it is possible to develop quantitative methods to improve talent acquisition. Specifically, applicants can be screened/ranked utilized a machine learning algorithm along with skill-based testing. Although these results are important, the study has some limitations. Firstly, and maybe most importantly is the definition of the performance parameter utilized. The authors do not provide a clear definition or rating scale of what is considered a good employee. Specifically, is quality (# of defect/lines of code) most important, or was throughput considered most important? Instead this rating was left to manager's opinion – which likely introduces bias and subjectivity into the analysis. Additionally, the study assumed that all software projects are equally complex, which is likely not the case.

Although machine learning has been applied to human performance data sets, several gaps within the literature remain. Prior research has not incorporated feedback mechanisms within a formulation in order to investigate if error correction is applied by operators nor analysed environments human performance required physical responses to cues (versus only problem solving). Additionally, as mentioned by Mkrtchyan (Mkrtchyan, Podofillini, & Dang, 2015) Bayesian belief networks applications to human error probability estimation is starting to proliferate. However, they further highlight that there are several concerns with these formulations. The structure of the underlying networks are frequently developed and quantified by expert judgment, which raises concerns regarding validity (Mkrtchyan, Podofillini, & Dang, 2015). Additionally, much of the research fails to explain and map the underlying structure to current theories of human cognition, decision making and performance. We address these challenges in several ways.

Firstly, we utilized empirical data coupled with machine learning algorithms. In the case of application of a Bayesian network we do not utilize expert opinion to develop the network – rather relying on unsupervised learnings algorithms to accomplish the task.

This addresses the concerns of bias introduction from expert opinion as highlighted by Mkrtchyan (Mkrtchyan, Podofillini, & Dang, 2015). Next, we compare and contrast several machine learning algorithms. Finally, our formulation incorporates feedback mechanisms to the operator, individual attributes, experience levels, factors related to the capability of the operator to observe cues (hearing and eyesight) for different tasks. Thus, we provide a formulation that not only considers the human element, but also how their performance impacts an industrial process. The research is further important since it is the first paper (to our knowledge) that applies various machine learning algorithms to a human performance data set in a distribution facility. Our formulation is further inclusive of both environmental variables, operator characteristics as well as task attributes.

3.3 Research Method

3.3.1 Data Description

The data set was obtained from a previous study that investigated feedback mode preference and operator performance when utilizing different hand held scanning devices (Copeland, 2015). The data set consisted of 136 observations gathered over four trials from 36 participants. Participants were tasked with scanning labels on cardboard using an Intermec SF61B scanner programmed with various feedback modes. Once a participant scanned a label, they moved the box to place the box in a designated location based on the feedback emitted from the device. If the device emitted “good scan” feedback (single beep, green light, short vibration), the participant was instructed to place the box on top of a table. If the device emitted “bad scan” feedback (multiple beeps, red light, long vibration), the participant was instructed to place the box underneath the table. This task replicated a common package scanning and sorting task that is used in warehouse facilities. Participants were exposed to different feedback conditions during four trials of the experiment. Participant demographics are provided in Figure 3.1. Additionally, the median age of the participants was 23 years.

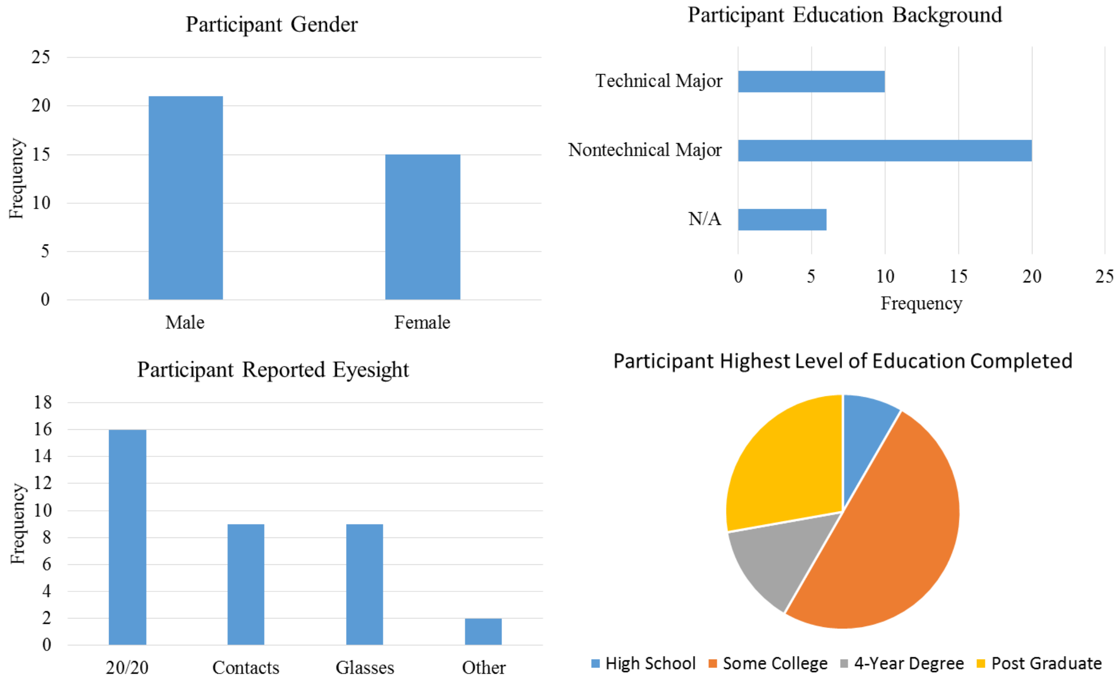


Figure 3.1 Participant Demographics.

For each of the observations several variables were captured. These variables were utilized as attributes or features within the analysis. Variables included participant demographics, number of packages scanned, total time taken to scan packages, and feedback mode. Variables such as time per package, hit rate and correct rejection rate were derived. Two variables were utilized to measure human performance. Hit rate can be defined an operator correctly placing the appropriate box into the correct shipping area. The correct rejection rate can be described as the operator correctly interpreting feedback from the scanning tool that a box is in the incorrect shipping area. A false alarm occurs when an operator receives no positive feedback from the scanning tool, however places the box in a shipping area. Overall the average false alarm rate was low (0.4%) regardless of which feedback mode was utilized. The average hit rate was 96.3% with a standard deviation of 5.8%. Tasks took an average of 267.0 seconds to complete with a standard deviation of 73.5 seconds. The longest task took 473 seconds while the shortest was 124 seconds. Task completion times appeared fairly consisted regardless of feedback mode utilized, gender or experience level.

Statistical analysis to include regression analysis, and analysis of variance (ANOVA) was utilized in order to determine if any of the independent variables had an effect on the hit or correct rejection rate. Regression analysis did not yield any significant results. Additionally, the effect of feedback mode was analysed. No statistically significant effect was found for feedback mode on completion time or hit rate. Thus, a model utilizing standard statistical methods could not be developed.

3.3.2 Machine Learning Analysis

3.3.2.1 Naïve Bayes Formulation

The Naïve Bayes classifier assigns class labels to problem instances, which are represented by vectors of features. The technique further assumes that the values of features are independent from another, thus each feature contributes independently to the final classification. Although this assumption may not hold in all cases, Naïve Bayes allows analysis with a small training data set.

Naïve Bayes can be represented as a conditional probability model. Specifically, we wish to calculate:

$$p(C_k|x_1, \ldots, x_n) \tag{3.1}$$

where the vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ represents the features or independent variables

Utilizing Bayes theorem Equation 3.1 can be rewritten as follows:

$$p(C_k|\boldsymbol{x}) = \frac{p(C_k)p(\boldsymbol{x}|C_k)}{p(\boldsymbol{x})} \tag{3.2}$$

Furthermore, we can utilize the chain rule and rewrite the numerator of Equation 3.2 as a joint probability model.

$$\begin{aligned}
&p(C_k, x_1, \ldots, x_n) = p(C_k)p(x_1, \ldots, x\_n|C_k) \\
&= p(C_k)p(x_1|C_k)p(x_2, \ldots, x_n|C_k, x_1) \\
&= p(C_k)p(x_1|C_k)p(x_2|C_k, x_1)p(x_3, \ldots, x_n|C_k, x_1, x_2) \\
&= p(C_k)p(x_1|C_k)p(x_2|C_k, x_1) \ldots p(x_n|C_k, x_1, x_2, x_3, \ldots, x_{n-1})
\end{aligned} \tag{3.3}$$

However, we assume that each feature is independent of every other feature given a category. Thus,

$$\begin{gathered}
p(x_i|C_k, x_j) = p(x_i|C_k) \\
p(x_i|C_k, x_j, x_q) = p(x_i|C_k) \\
p(x_i|C_k, x_j, x_q, x_l) = p(x_i|C_k)
\end{gathered} \tag{3.4}$$

and so forth

Thus, finally we can represent the joint probability model as:

$$\begin{aligned}
&p(C_k|x_1, \ldots, x_n) \propto p(C_k, x_1, \ldots x_n) \\
&\propto p(C_k)p(x_1|C_k)p(x_2|C_k) \ldots p(x_n|C_k) \\
&\propto p(C_k)\prod_{i=1}^{n} p(x_i|C_k)
\end{aligned} \tag{3.5}$$

Thus, given evidence **Z** the conditional distribution over the class variable C is:

$$P(C_k|x_1, \dots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \tag{3.6}$$

Finally a classification rule can be defined for Naïve Bayes.

$$c^* = \arg\max p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \tag{3.7}$$

Next, given a set of labelled examples, or training data the parameters of the model can be estimated. First, the probability for each class is estimated by:

$$\hat{p}(C_k) = \frac{N_j}{N} \tag{3.8}$$

$$where\ N_j represents\ the\ number\ of\ examples\ in\ class\ C_k$$

$$and\ N\ represents\ the\ total\ number\ of\ observations$$

Next, the probability of each value $x_k$of the attribute X_i and for the class C_k is estimated by:

$$\hat{p}(X_i = x_k|c_k) = \frac{N_{ijk}}{N_j} \tag{3.9}$$

3.3.2.2 Decision Tree Formulation

A decision tree is a simple, but yet effective method for classifying examples. The Naïve-Bayes decision tree (NBTree) is a hybrid approach that utilizes the advantages of decision trees and Naïve Bayes into a single algorithm developed by Ron Kohavi. The algorithm builds a decision tree by univariate splits at each node, however incorporating Naïve-Bayes classifiers at the leaves (Kohavi, 1996). The algorithm defines the utility *(denoted by u)* of a respective node split. This utility is calculated by performing 5-fold cross validation whenever a split is made, computing the difference in terms of prediction accuracy. The split is deemed significant if the utility is at least 5 % and the node has at least 30 instances. Attributes, denoted by X_i in terms of human performance included participant demographics, error rate, time per package and the scanning device utilized.

The complete algorithm as described by Kohavi is provided below (Kohavi, 1996).

> *For each attribute X_i, evaluate the utility, $u(X_i)$, of a split on attribute X_i. For continuous attributes a threshold is also found at this stage.*
>
> *Let $j = \arg\max_i(u_i)$, i.e., the attribute with the highest utility.*
>
> *If U_j is not significantly better than the utility of the current node, create a Naïve-Bayes classifier for the current node and return.*
>
> *Partition T according to the test on X_j. If X_j is continuous, a threshold split is used; if X_j is discrete, a multi-way split is made for all possible values.*
>
> *For each child, call the algorithm recursively on the portion of T that matches the test leading to the child.*

3.3.2.3 Bayesian Network Formulation

The Naïve Bayes method has a strong independence assumption. Thus, causal relationships between attributes or variables are not modelled, which may be an unrealistic assumption. Next, the analysis considered these causal relationships and thus utilized a Bayesian network. Bayesian networks are graphical models used to represent knowledge about an uncertain domain. It is a method that combines graph theory, probability theory, statistics, and computer science. Within a Bayesian network, each node represents a random variable. The edges between nodes represent probabilistic dependencies among random variables. The use of Bayesian network models does not necessarily imply that Bayesian statistics are being used. The nodes of Bayesian networks can sometimes be used to represent hypotheses, beliefs, and latent variables rather than random variables. A Bayesian network structure is ideal for prior knowledge in combination with observed data. Bayesian networks allow for an effective representation of knowledge mathematically, intuitively, and visually.

However, even with missing data, Bayesian networks can be used to predict future events and gain an understanding of problem domains (Ben-Gal, 2008). A simple example of a Bayesian network considers a back injury (Ben-Gal, 2008). The injury could have been caused by a couple things: sport or chair. If the chair is the problem, then a co-worker might report the same injury, which is where the variable "worker" comes from. Figure 3.2 shows this example represented by a Bayesian network.

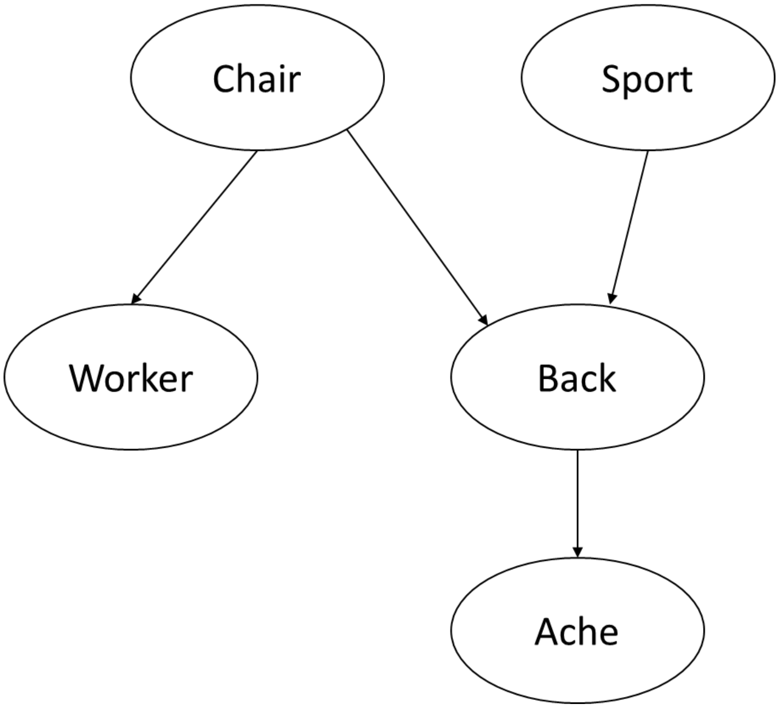


Figure 3.2 Backache example used in Bayesian Networks (Ben-Gal, 2008).

The joint probability distribution for the Bayesian network formulation can be represented by considering that not every node is connected to all other nodes. Consider a Bayesian network that contains *n* nodes, X_1 to X_n, where $n = 10$ in this example formulation. A particular probability in the joint distribution can be represented by $P(x_1, x_2, \ldots, x_{10})$. The joint probability can be factorized per the chain rule as such:

$$\begin{aligned} &P(x_1, x_2, \ldots, x_n) \\ &= P(x_1)P(x_2|x_1) \ldots, P(x_{10}|x_1, \ldots x_9) \\ &= \prod_{i}^{10} P(x_i|x_1, \ldots x_{i-1}) \end{aligned} \tag{3.10}$$

However per the Markov Property the value of any particular node is only conditional on its respective parent nodes, resulting in:

$$P(x_1, x_2, \dots, x_{10}) = \prod_i P\big(x_i \big| Parents(x_i)\big) \qquad (3.11)$$

Where $$Parents(x_i) \subseteq \{x_1, \dots, x_{i-1}\}$$

One common algorithm to determine the Bayesian network structure from a data set is the hill climbing algorithm. The algorithm initializes an arbitrary solution, which is then incrementally improved by adjusting a single element. The hill climbing algorithm is not guaranteed to find to a global optimum but has been shown to find a local optimum reasonably well (Skiena, 2010). The algorithm is not as advanced as algorithms as tabu search or simulated annealing, but may provide results equally useful. The tabu search algorithm utilizes local searches in order to identify an improved solution within immediate neighbourhood of the current solution. Neighbours are defined as solutions that are similar with only minor differences.

3.4 Case Study and Numerical Results

The classification techniques previously described were applied to the empirical data set. The provided data set was small, thus there was significant concern that overfitting could occur, and a 10-fold cross validation scheme was utilized. This method partitions the original sample into 10 equal sized subsamples. One of these subsamples is then utilized as the validation data set. Parameters are then estimated from each subsample and combined into a single estimator (via averaging).

Performance metrics were both calculated at the algorithm level and the class level. Overall performance metrics included metrics such as the percentage of correctly classified instances and the Kappa statistic. The Kappa statistic provides a measure of inter-observer agreement. A calculated test statistic greater than 0 implies that the classifier is doing better than chance alone.

Metrics such as the true positive rate or recall, which is the proportion classified as class x divided by the actual total in class x were also calculated. Next, the false positive rate was calculated which is defined as the proportion incorrectly classified as class x, divided by the actual total of all classes (except x). The ROC area was also calculated for each class. The curve was generated by plotting the true positive rate versus the false positive rate, and then calculating the area under the curve. An area close to 0.5 implies lack of statistical independence. Precision was also calculated which is defined as the proportion of examples which truly have class x divided by the total classified as class x.

A combined metric combining precision and recall was also calculated as provided in Equation 3.12.

$$F - Measure = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \tag{3.12}$$

As provided in Table 3.1, both the Naïve Bayes and Modified Decision tree methods only achieved a 65.9 % correctly classified instances for hit rate prediction. Results were much improved for correct rejection rate with a 95.5 % correctly classified instances. In both cases the Kappa statistic was also considered low, and could not sufficiently rule out agreement by chance. Thus, these classifiers were not considered a robust measure for prediction both to identify how often participants correctly place a box in the appropriate shipping area (hit rate) or how often they recognize an error based on feedback mode (correct rejection rate). Both algorithms also suffered from high false alarm rates.

Table 3.1 Naïve Bayes and Decision Tree Performance.

	Feature	**% Correctly Classified Instances**	**Kappa**	**Weighted Average**			
				True Positive Rate	**False Positive Rate**	**Precision**	**Recall**
Naïve Bayes	Hit Rate	65.9 %	0.01	0.66	0.65	0.44	0.66
	Correct Rejection Rate	95.5 %	0.00	0.96	0.96	0.91	0.96
Decision Tree	Hit Rate	65.9 %	0.02	0.66	0.63	0.44	0.66
	Correct Rejection Rate	95.5 %	0.00	0.96	0.96	0.91	0.96

The analysis utilized several search algorithms and a simple estimator in order to determine the structure and parameters of the underlying Bayesian network. The first algorithm considered was K2, which implements the hill climbing algorithm but add arcs based on a fixed ordering of variables. The analysis also considered the hill climbing with no fixed ordering of variables. The final variation on the hill climbing network was the repeated hill climber, which randomly generates a network and applies the hill climber algorithm repeatedly until a local optimum is returned. This algorithm forms a tree by calculating the maximum weight spanning tree utilizing the Chow and Liu algorithm (Chow & Liu, 1968). Lastly tabu search was also utilized. The tabu search utilizes hill climbing until a local optimum is found. Once this optimum is found the algorithm steps to the least-worst candidate. The algorithm does not consider points just visited.

The results for the overall performance of the algorithms is provided in Table 3.2, when hit rate was predicted. The TAN search algorithm significantly outperformed the other search algorithms utilized with a % correctly classified instances of 81 %, and a kappa statistic of 0.58. The kappa statistic provides a measure of agreement between the algorithm classifications of the algorithm and the observed classes. Utilizing a scale provided by Viera and Garrett (Viera & Garrett, 2005) the kappa statistic for TAN Search can be interpreted as having moderate agreement. Next, important metrics of algorithm performance such as true and false positive rates were calculated.

Table 3.2 Bayesian Network Algorithm Overall Performance (Hit Rate Prediction).

Classifier	**Search Algorithm**	**% Correctly Classified Instances**	**Kappa statistic**
Simple Bayes Network Classifier	K2	72.7 %	0.36
	Hill Climbing	68.2 %	0.15
	Repeated Hill Climber	68.2%	0.15
	TAN Search	**81.1 %**	**0.58**
	TABU Search	68.2 %	0.19

Several metrics were calculated for each algorithm utilized averaged across all classes, provided in Table 3.3. As expected, the TAN Search algorithm had the highest true positive rate and the lowest false alarm rate. The Precision and Recall and resulting combined measure (F-measure) appeared reasonable. Thus, it was concluded that the TAN search algorithm resulted in an adequate prediction.

Table 3.3 Bayesian Network Algorithm Class Performance (Hit Rate Prediction).

Search Algorithm	**Weighted Average**					
	True Positive Rate	**False Positive Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**
K2	0.73	0.42	0.64	0.73	0.67	0.86
Hill Climbing	0.68	0.55	0.50	0.68	0.57	0.77
Repeated Hill Climber	0.68	0.55	0.50	0.68	0.57	0.77
TAN Search	**0.81**	**0.30**	**0.80**	**0.81**	**0.78**	**0.93**
TABU Search	0.68	0.53	0.60	0.68	0.60	0.82

Next, the structure of the resulting Bayesian networks utilizing TAN Search was analysed. Eyesight and hearing capability, along with age, education and gender were all deemed important predictors. Additionally, completion time, scanning job and feedback mode also influenced the hit rate. Thus it appeared that task characteristics, environmental conditions and individual attributes were important during prediction.

Feedback has been shown as an important consideration in work environments in order to improve human performance. Not only is feedback itself important, the type of feedback utilized also plays a role. Past research has shown that feedback modes (auditory versus tactile for example) do impact human performance (Brewster, Raty, & Kortekangas, 1996; Scott & Gray, 2008). Auditory stimuli have been shown to improve overall productivity (Goomas & Yeow, 2010), while tactile feedback has improved reaction times (Scott & Gray, 2008). In high workload environments multiple feedback methods (thus redundancy) has been shown as effective (Haas & Van Erp, 2014). In contrast, low stress environments require simple single feedback modes. A magnitude of scanners are available today and selection of the appropriate device, more specifically feedback modes to be utilized is an important consideration during process design and optimization. Participants were provided four different types of feedback during performance of the task. The feedback provided an indication of which downstream location this specific box needed to go. Thus, if the participant received feedback that a box was not in the correct location, and subsequent moved it resulted in "correct rejection". The converse of this situation, where feedback was provided that the box was in the correct location and the participant did not move the box resulted in a "hit rate".

Thus, a high hit rate and correct rejection rate would indicate a well performing participant. Feedback modes were auditory, auditory and visual, auditory and tactile and a combination of all three. Thus, it was expected that that these feedback modes combined with characteristics of the participants (eyesight, hearing sensitivity) would form good predictors of performance (Menelas, Picinalli, Katz, & Bourdot, 2010; Spence & Lee, 2008).

It is important to note that only the TAN search algorithm provided significant results for this data set, when compared to other Bayesian network search algorithms. The TAN algorithm was also utilized in order to predict the correct rejection rate. As with decision trees, the correct rejection rate could be predicted with a high accuracy. This yielded 97 % correctly classified instances. The weighted averages of true positive, false positive, precision, recall, F-Measure and ROC area were 0.97, 0.64, 0.97, 0.96, and 0.98. Interestingly, in contrast to the decision tree technique where only feedback mode and gender were important the Bayesian network incorporated several additional variables (such as age, education etc.).

3.5 Discussion and Future Work

Human decision making can be analysed from an information processing or cognitive viewpoint. Specifically, the Wickens information processing model provides a framework to understand cognitive mechanisms involved during decision making (Wickens & Hollands, 2000). Additionally, the approach focuses on biases and processes utilized during decision making, limits of human attention, working memory and the use of heuristics that usually work well (but not always).

The main goal of this process is to map many-to-one information bits to the appropriate responses. The process consists of measuring cues from the environment, filtering these cues with the senses utilizing selective attention. This is followed by perception and diagnoses, closely tied to situational awareness. During diagnosis long-term and short term (or working) memory play important roles. Response selection is influenced by uncertainty, familiarity/expertise and time pressure. Additionally uncertainty of a consequence and familiarity with cues influence deliberation time and the overall process.

Selective attention is the mechanism by which we filter out cues. Thus, some cues may be ignored. Additionally, cues may be ambiguous or misinterpreted. Additionally, cue filtering is influenced by our past experiences stored in long term memory. The filtered subset of cues form the basis of situational awareness and allows reasoning about which hypothesis (or state of the world) we believe is true. Additionally, this process is typically iterative. Situational awareness is key during the diagnostic process. Our understanding of the situation is influenced by several factors to include our perception in terms of estimating cues, information fusion from all cues, background and beliefs stored in long term memory as well as our working memory capacity.

Cues play an important role when we reason about the state of the world. The cue diagnosticity, reliability (or credibility) as well as the physical features of the cue all play an important role. Additionally, during cue integration challenges such as missing information (or cues), overloading of working memory and salience differences between cues arise. For example, certain feedback modes may compete with environmental cues. In a loud environment more attention is required in order to hear an audible beep from a scanning device. Thus, it is possible to miss this cue more easily. Additionally, expertise influences weights assigned to cues during the decision making process. For example, experts may recognize a pattern of cues and make a subsequent diagnoses as described in recognition primed decision making. Feedback also plays an important role during the diagnostic process. For example, typically we learn from past mistakes thus ideally improving quality of decisions.

All the algorithms investigated exhibited much better results when predicting correction rejection rates versus hit rates. The analysis indicated that prediction of participants identifying cues related to boxes in incorrect locations ("correct rejection rate") was more accurate with predictor performance > 90 %. This is in sharp contrast to prediction of participant performance in terms of hit rate, with predictor performance < 75 %. Feedback cues related to correct rejection were stronger (blinking lights, vibration versus green light). Thus, the analysis may support that the physical features of the cue are important when prediction the correct rejection rate.

It was expected that the search algorithm would identify causal patterns that are supported when considering information theory. The algorithm identified that variables related to selective attention or cue filtering to include eyesight and hearing were important during prediction. Additionally, variables related to long term memory such as age, education level and prior experience were also identified as important within the Bayesian network.

However, differences were noted depending on which predictive variables (hit rate versus correct rejection rate) were selected. Causal connections between eyesight, hearing, age, education and prior experience were identified regardless of the predictive variable. These connections remained largely unchanged. However, when predicting the correct rejection rate causality was established between feedback mode and hearing. Research has shown that auditory feedback modes provide more focused attention (Vitense, Jacko, & Emery, 2003), which may explain why this connection was important.

This is based on the assumption that higher vigilance is applicable when identifying correct rejections.

The study showed that prediction of participant performance can be performed utilizing Bayesian networks. Additionally, the results were supported by our understanding of human information processing and cognition.

Future research is needed in order to identify if prediction performance for hit rate can be improved. Additionally, the underlying data set contained mostly inexperienced participants and the data was collected in a laboratory environment. An industrial environment was replicated to the greatest extent, however differences may remain. These differences may impact cue filtering and ultimately diagnoses.

CHAPTER IV

IMPROVING BAYESIAN NETWORK CREDIBILITY UTILIZING AN EA-6B AIRCRAFT CASE STUDY

"All models are wrong, but some are useful." George Box

4.1 Introduction

The application of machine learning, to include Bayesian networks, to various problem domains is proliferating and fast becoming one of the most important technical fields (Jordan & Mitchell, 2015). Bayesian networks are frequently utilized to perform both diagnosis and prognosis. It is important to assess model credibility to allow practical application of these methods to real problems (Averill, 2015).

Credibility can be established utilizing several methods, to include sensitivity analysis, formal model reviews, and model validation activities. The authors surveyed 23 journal articles published since 2015 which included the keywords Bayesian networks, and assessed how common practice sensitivity analysis is within related literature. The survey indicated that, although Bayesian networks have been widely applied, less focus (26 %) has been placed on the application of sensitivity analysis, data error, or uncertainty within the underlying network structure of derived models. The objective of this chapter is to assess the credibility of a Bayesian network derived from an EA-6B aircraft data. The assessment was performed by specifically evaluating the impact of purposefully introduced error within the training set on the prognostic capability of the network.

Broadly speaking, sensitivity analyses are categorized into either local or global methods. Local sensitivity methods are performed about a reference value (or baseline) within the model input space, while global sensitivity methods assign a probability distribution to model inputs. Hoshino et al., and Zhang et al. both perform sensitivity analysis of input variables to their Bayesian network case studies (Hoshino, van Putten, Girsang, Resosudarmo, & Yamazaki, 2016). Hoshino et al. apply Bayesian networks in order to model community-based coastal resources and calculate the posterior probability distribution of three performance indicators given different scenarios, or evidence. Zhang et al. consider the risk analysis of tunnel-induced pipeline damage and utilize a sensitivity measure coupled with domain expertise and setting evidence within the derived network (similar to the approach by Hoshino) (Zhang, Wu, Qin, Skibniewski, & Liu, 2016). Dadaneh and Qian consider application of Bayesian networks in the medical domain. They adopt a hierarchical model that utilizes various protein-protein interaction data sets that typically contain levels of noise. They place significant focus on identifying and integrating multiple networks to more accurately define the Bayesian network structure (Dadeneh & Qian, 2016). Introduction of purposeful error within training sets has not been considered within Bayesian networks, along with analysis of the associated model credibility. Real world data sets, such as the one utilized within this research, will likely contain error (or noise) which is not readily quantifiable in certain situations. Several challenges must be overcome in order to establish model credibility.

First, the potential impact of outliers or noise within the training data must be considered. Additionally, interactions between variables must be identified.

The research establishes model credibility by considering both the impact on model output based on variability within input parameters, as well as application of sensitivity analysis in order to quantify the interaction between variables, and their impact on overall model performance. Additionally, quantifying the variation in terms of the calculated posterior (or joint probability) across several variables under uncertainty is critical. High variability in the joint probability may influence decision makers when the model is applied is practice. For example, the nominal case may indicate a joint probability of 0.8 of a high risk event occurring. This high likelihood may spur action by the decision maker to proactively mitigate the high-risk event. However, if the joint probability changes significantly under noise conditions, the same decisions may not be taken. If, for example, the probability is merely 0.5, the decision maker may take no action. The joint probability is greatly influenced by the network structure and the associated conditional probability tables. This work has wide application and benefits. These results are important and provide credence to utilization of Bayesian networks in real field data, which will always contain noise or error that is not easily quantified.

This chapter is organized as follows. We start by providing relevant background for unfamiliar readers on machine learning techniques, then provide an example of how Bayesian networks can be utilized within a decision-making process. Next, we discuss sensitivity analysis techniques as a method to establish model credibility. We also provide a brief survey of techniques utilized by other Bayesian network researchers. We then discuss our analysis method focused on introduction of uncertainty within a controlled experiment. Finally, we apply our method to an EA-6B aircraft data set and discuss the results.

4.2 Machine Learning Background

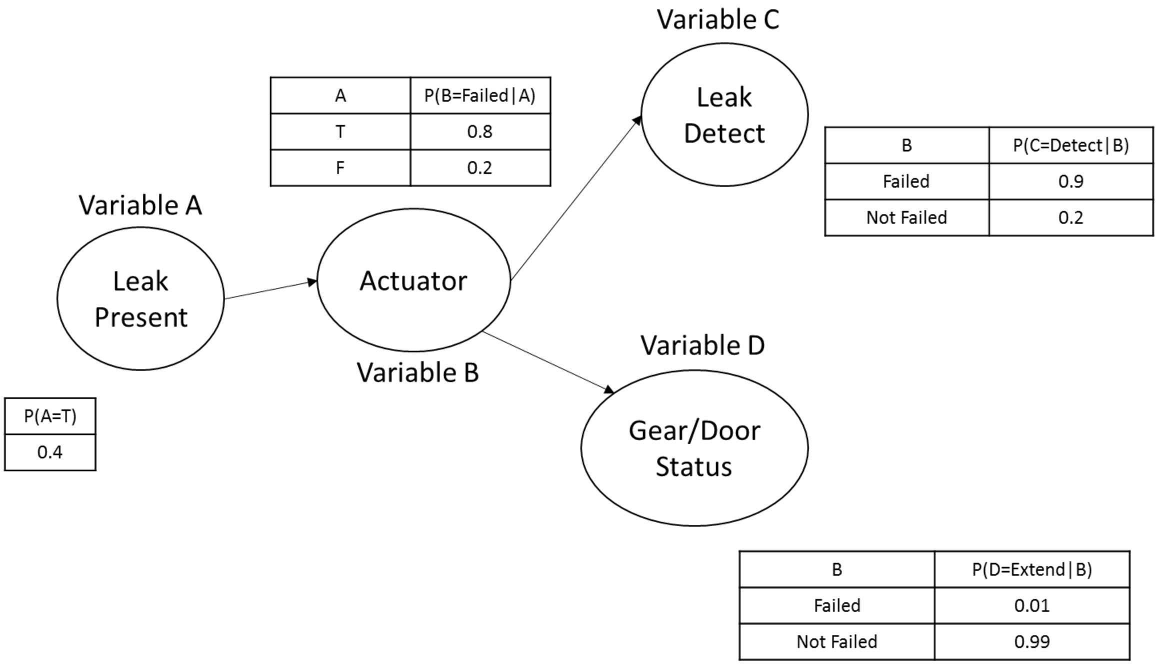
Application of machine learning, to include Bayesian networks, to various problem domains is proliferating. It allows automation in terms of pattern recognition and prediction and can be applied to a wide variety of problem domains. The classification rule(s) that is developed utilizing a training set and an associated algorithm can be applied to new data sets and serve several functions to include optimization, prediction and risk reduction.

Machine learning is predicated on the notion of automating the process of learning a mathematical model from a training data set made up of various features or variables. Specifically, the algorithms focus on understanding the relationship between descriptive features and a target feature (typically what we aim to predict). Machine learning methods can be broadly classified into information-based learning, similarity-based learning, probability-based learning, and error-based learning. Information-based methods utilize the notion of entropy and information gain to measure how informative various descriptive features are found to be. Decision trees are a common method of information-based methods. Similarity-based learning utilizes a feature space and relative measures of similarity. Techniques include the k-nearest neighbor algorithm. Probability-based learning includes Bayesian networks, and are focused on describing the probabilistic and causal linkages between features. Finally, error-based approaches are focused on minimizing the total error across predictions. Simple linear regression is an example of error based learning.

Bayesian networks are frequently utilized to perform both diagnosis and prognosis. Diagnosis typically refers to identification of a part, component, or system that is either in a degraded or failed stated. Thus, the event has already occurred and, at best, activities to assess the impact can be pursued. Prognostics, however, aims to detect the event before it occurs. Thus, we aim to provide a sufficient window of opportunity for the user to allow the appropriate decision making. The diagnostic and prognostic capability of Bayesian networks is a key benefit. Prediction is performed by first training the network and using the established conditional probabilities to compute the joint probability of an event. This allows statistical inference, illustrated in the following risk assessment example.

Bayesian networks can be applied to perform risk analysis in several different ways, and allow for the formulation of various alternatives or what-if scenarios focused on risk evaluation. The method may also be tailored to the specific domain and utilize different underlying assumptions and available information. Risk can be evaluated differently depending on the nature of the underlying Bayesian network topology. For example, one network can be constructed that focuses on the progression of failure from initial conditions through symptoms. Alternatively, a network can be constructed focused on the underlying functions within the system, which will be discussed next.

A causal network representing cause-and-effect, as well as the associated probabilities, can also be built. Consider a small, simple system consisting of a hydraulic actuator, connected to a landing gear door and the landing gear itself through some mechanical linkage. The mechanical linkage is not considered in this formulation. Consider that the system can fail when seals within the hydraulic actuator degrade, leading to loss of containment of hydraulic fluid. Further assume that the hydraulic system has some kind of leak detection sensor that provides an indication to the operator if a leak occurs. We can represent this scenario with the Bayesian network provided in Figure 4.1. Assume the various probabilities were either calculated from past data or subject matter experts.



P(A=T)
0.4

A	P(B=Failed|A)
T	0.8
F	0.2

B	P(C=Detect|B)
Failed	0.9
Not Failed	0.2

B	P(D=Extend|B)
Failed	0.01
Not Failed	0.99

Figure 4.1 Example Bayesian Network (Cause-Effect)

In this simple example, extending the gear with the actuator in a failed state may be considered a risky event, since, let's assume, it would result in gear collapse or failure to extend. According to Bayes theorem we can state:

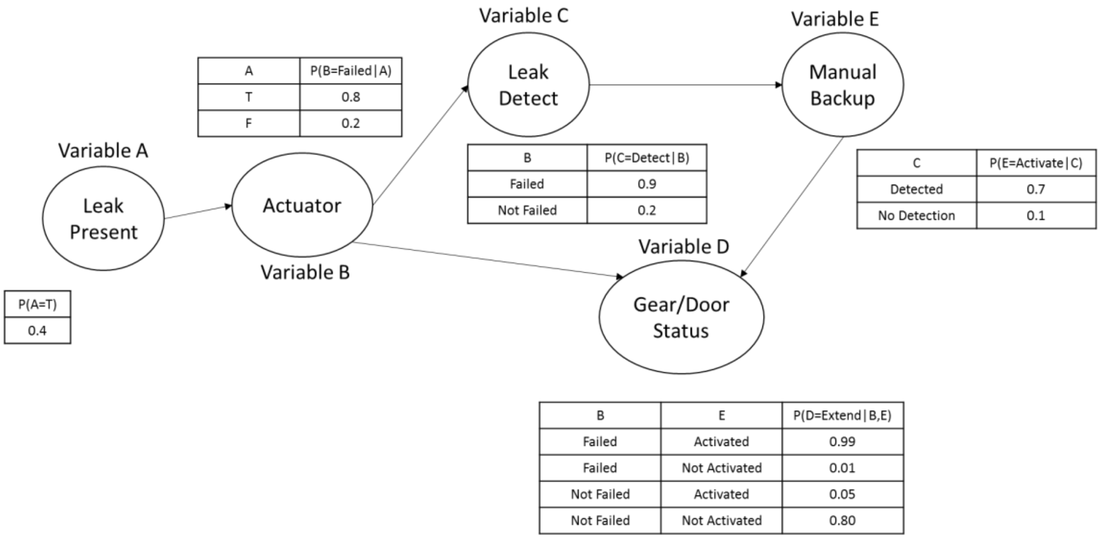
$$P(B, D) = P(D|B) \times P(B) \qquad (4.1)$$

We can then read the probabilities directly from the conditional probability tables and obtain the probability of the joint event (actuator failed, gear collapse):

$$P(B, D) = 0.01 \times 0.8 = 0.008 \qquad (4.2)$$

The calculated probability can then be utilized as a measure of risk by determining if this probability is low enough or can be tolerated. Additionally, the impact of design changes can be assessed. For example, if we could reduce the probability of the actuator failure mode occurring from 0.8 to 0.5 the resultant joint probability of a risk event would be 0.01 x 0.5 or 0.005. This allows for trade-off analysis and for the development of corrective action plans – critically important in the risk assessment process. Networks for different components or failure modes can be developed and the various high risk probabilities can be ranked for further investigation.

An alternative method to perform risk assessment is to consider multiple failure modes or backup systems in one network. Revisiting the previous example, we add a manual backup system. Specifically, if the operator receives indication that the actuator is leaking he may utilize a manual system to extend the landing gear. Graphically, this example is provided in Figure 4.2.



P(A=T)
0.4

A	P(B=Failed|A)
T	0.8
F	0.2

B	P(C=Detect|B)
Failed	0.9
Not Failed	0.2

C	P(E=Activate|C)
Detected	0.7
No Detection	0.1

B	E	P(D=Extend|B,E)
Failed	Activated	0.99
Failed	Not Activated	0.01
Not Failed	Activated	0.05
Not Failed	Not Activated	0.80

Figure 4.2 Example Bayesian Network with Backup System Included

Additionally, the number of states can also be expanded. Thus, instead of the actuator failure mode being in a true/false state, we can have states such as no failure, mode A, mode B and so forth.

Bayes theorem can be extended and we can calculate the probability of an event, $x_1,\ldots,x_n$ by (Kelleher, Namee, & D'Arcy, 2015):

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | Parents(x_i)) \tag{4.3}$$

Thus, the joint event of extending the landing gear, the actuator being in a failed state and the backup system not being activated:

$$\begin{aligned} &P(D,B,E') = P(D|E,B) \times P(E|C) \times P(C|B) \times PB|A) \times P(A) \\ &= 0.01 \times 0.7 \times 0.9 \times 0.8 \times 0.4 \\ &= 0.002 \end{aligned} \tag{4.4}$$

The interested reader is referred to Chen and Pollino for an overview of best practices in order to build and utilize Bayesian networks (Chen & Pollino, 2012). Additionally, several interesting applications of Bayesian networks are found in the literature within decision making. Johnson et al. show how an environmental scorecard can be updated utilizing Bayesian networks as additional information become available (Johnson, Logan, Fox, Kirkwood, & Pinto, 2016). La Morgia et al. utilize a decision framework based on Bayesian networks in order to investigate and reduce potential social conflicts while eradicating invasive species (La Morgia, Paoloni, & Genovesi, 2016). Ji and Tan develop a decision making model when large amounts of data is available in the food service industry (Ji & Tan, 2016). Colón-González et al. apply Bayesian networks in a public health decision making framework (Colón-González, et al., 2016), while Neapolitan et al. applies Bayesian networks to kidney transplant decisions (Neapolitan, Jiang, Ladner, & Kaplan, 2016). The breadth of these applications illustrates how widely Bayesian networks can be applied and utilized in decision making. Next, we discuss important methods to establish model credibility in order to successfully apply Bayesian networks to practical decision making.

4.3 Establishing Model Credibility through Sensitivity Analysis

Model credibility is in part established by considering the impact on model output (based on variability within input parameters), the application of sensitivity analysis (quantify the interaction between variables), and their impact on overall model performance (Renooij, 2014). In the case of Bayesian networks, we are specifically interested in the impact of outliers and variability on the structure of the trained network and the associated conditional probability tables. Additionally, the impact of changes to the underlying machine learning algorithm parameters is also important.

Broadly speaking, sensitivity analyses are categorized into either local or global methods. Botgonovo and Plischke provided a detailed overview of sensitivity analysis methods. For the sake of brevity, these methods will be highlighted here, but the reader is encouraged to reference if interested (Borgonovo & Plischke, 2016).

Local sensitivity methods are performed about a reference value (or baseline) within the model input space. Several approaches can be used, to include one at a time approaches (OAT). OAT approaches are frequently performed considering best case and worst case input scenarios. These scenarios are frequently derived using expert opinion, assuming alternative hypothetical futures. As pointed out by (Borgonovo & Plischke, 2016), it is critical that the scenarios must be “consistent, diverse, in a small number, reliable and efficient.”

Factorial designs are used in order to perform sensitivity analysis. Specifically, it is designed to select a set number of samples for each input parameter and run the model for all combinations. In contrast to the OAT approach, a factorial design allows us to investigate interaction effects between potentially dependent input variables. Thus, we can ascertain the importance of each factor on the overall prediction. Let's assume that we have *k* factors, each with two possible levels. Setting *k* equal to three and assuming two levels for each run we have 2^k different combination (or eight in this case). The full model can be described as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 + \beta_{123} X_1 X_2 X_3 + \varepsilon \tag{4.5}$$

Graphically, we can represent this design as cube with the eight runs representing the corners of the cube. Next, we can estimate the main effects, two-factor interactions and three-factor interactions, utilizing geometry. The main concern and barrier with utilizing a factorial design in this study are the number of factors and the associated levels (Hamby, 1994).

Global sensitivity methods assign a probability distribution to model inputs. Global sensitivity analysis measures can be broadly categorized in regression based methods, variance based methods, and density based methods. The reader is again referred to Borgonovo and Plischke for a detailed discussion (Borgonovo & Plischke, 2016).

The OAT approach is the simplest method and modifies a single input variable at a time, while keeping all others constant. Thus, changes in output can be monitored as input variables are modified. Variables can be modified utilizing their standard deviation in order to account for variability within the associated parameter (Hamby, 1994). Although the OAT approach has been extended to n-way sensitivity analyses, these results are often difficult to implement and interpret (van der Gaag, Renooij, & Coupe, 2007). Mathematically, one-way sensitivity analysis can be described as follows. First, let's denote a parameter we wish to study by $x = p(b_i|\pi)$ where b_i represents the value of variable B and π is the combination of values for all parents of B. Thus, we vary the parameter *x* while also co-varying the other parameters, $p(b_j|\pi), j \neq i$ to ensure that the parameters from the same distribution sum to 1. We further assume that they are varied in such a way to ensure their mutual proportional relationship is kept constant, represented by (van der Gaag, Renooij, & Coupe, 2007):

$$p(b_j|\pi)(x) = p(b_j|\pi) \cdot \frac{1 - x}{1 - p(b_i|\pi)} \tag{4.6}$$

for $p(b_i|\pi) < 1$.

Next, we discuss our analysis method in order to investigate the impact of uncertainty.

4.4 Research Method: Introduction of Noise

Sensitivity analysis evaluates the neighbored points to evaluate how resilient a solution is against slight changes. The distribution in one point can have several reasons:

- The underlying process is influenced by parameters not yet know to us, making it appear to be random
- The underlying process really is random
- The measurement system introduces the randomness
- There is additional noise of unknown origin

Uncertainties within Bayesian networks can stem from several sources, to include incomplete understanding of the underlying system, incomplete or imprecise data, as well as subjectivity if expert opinion was utilized (Chen & Pollino, 2012). The underlying data set likely contained significant noise, traceable to input errors. These could include measurement error or input errors by operators. One challenge with real world field data of this nature is the identification of the error. Typically, the data set may not always contain sufficient resolution or fidelity in order to ascertain the appropriateness of the various coding schemes utilized. Additionally, text narratives are written in natural language and vary significantly between technicians.

The analysis method consisted of several steps. A full factorial design was constructed in order to assess the importance of several factors (or features). In the case study utilized, NMC hours, AWM hours and Awaiting Parts (AWP) hours were included. Each factor had three levels corresponding to noise levels within the data, with the target feature assuming five levels. In this experiment, the target feature levels included AWM hours, Malfunction category, NMC hours, owner organization, and relevant system. Noise was defined as purposeful error and consisted of adding one standard deviation of each included variable. AWM, NMC and AWP hours were chosen as additional factors due to their likely perceived importance in formulating the Bayesian network. The factors were varied between three levels: no noise added, 10% noised added or 40% noise added. The addition of 40% noise corresponded to the likely worst case scenario by the researchers.

Next, the search algorithm was run and the resulting Bayesian network structure, conditional probability tables, and overall algorithm performance was captured. Several responses, both at the algorithm level and class level, were collected. At the algorithm level, the % of correctly classified instances in the data set (using the predictive algorithm) was the only response analyzed. At the class level, both the true positive and false positive rates were included. These were considered important to assess model credibility. Specifically, a model with high false positive rates would likely not be considered credible if practically implemented. Additionally, main effects and interactions were also investigated in the analysis.

In order to assess the impact of parameters of the search algorithm utilized, an OAT approach was utilized. Specifically, algorithm parameters, such as the number of parent nodes, iterations, and starting network, were all varied. The analysis also considered different local search metrics, to include the entropy, Bayes, MDL and AIC metrics. Additionally, global search versus local search was also considered along with utilization of validation techniques beyond cross validation.

The method was applied to a case study of the EA-6B wiring related maintenance actions and failures, which is discussed next.

4.5 Case Study and Results

The EA-6B Prowler aircraft has been an integral component to Navy deployments. The aircraft is primarily utilized as an electronic warfare platform, both offensively and defensively. In order to perform its mission, the aircraft has extensive electronic systems, to include wiring, externally mounted pods, and other jamming equipment. The EA-6B is depicted in Figure 4.3.

Figure 4.3 EA-6B landing on an aircraft carrier (U.S. Navy photo by Mass Communication Specialist 3rd Class Joshua Card/Released)

The case study is organized as follows. First, we provide an overview of the data set utilized as well as an explanation of the underlying maintenance process and variables considered. This is followed by an analysis of the input variables, to include a discussion of outliers. This is followed by a discussion of the data preparation performed in this research, such as converting numerical variables to categorical values. We then provide results in terms of algorithm stability, sensitivity analysis, to include main and interaction effects of input variables, and finally we provide results in terms of the predictive capability of the algorithm under noise conditions. This is followed by a discussion of our results and future work.

4.5.1 Data Set Description

The aircraft maintenance process starts with identification of a discrepancy by either the pilot or ground crew. Additionally, planned inspections may also be scheduled based on aircraft usage (such as flight hours). Once the aircraft is in a maintenance status, the time waiting for maintenance and parts is recorded. Additionally, the total time that the aircraft is in a down state is recorded. In some cases, the aircraft may be partially operational. In these situations, a partially mission capable time is captured (versus a non-mission capable time). Cannibalization of parts may also occur when the supply system does not have an associated spare available. Within the EA-6B data set, the awaiting maintenance time, awaiting parts, non-mission capable, and partially-mission capable hours were significant sources of variation.

Each maintenance record contained both continuous and categorical variables/features. Continuous variables consisted of man hours, elapsed maintenance time (EMT), awaiting maintenance time (AWM), AWP, NMC hours and partial-mission capable hours (PMC). All variables related to time were measured in hours. Table 4.1 provides a summary of all features included in the research along with their definition.

Table 4.1 Variables (features) within EA-6B data set

Variable	Variable Type	Description
Man-hours	Continuous	The total number of direct labor hours to include preparation time, inspection, troubleshooting or ordering of parts expended in order to perform the associated maintenance action.
Elapsed maintenance time	Continuous	Actual clock time associated with a maintenance action not including preparation, cure, or charging time. Subset of man-hours.
Awaiting maintenance time	Continuous	Total hours aircraft was in NMC or PMC status and awaiting maintenance resources.
Awaiting parts time	Continuous	Total hours aircraft was in NMC or PMC status and awaiting parts or supplies required in order to perform the associated maintenance action. Clock starts once parts/supplies have been ordered.
Non-mission capable hours	Continuous	Total hours aircraft was in a NMC status and unable to complete any mission. Includes all associated downtime accumulated from maintenance or logistics sources.
Partial-mission capable hours	Continuous	Total hours aircraft was in a PMC status and able to complete at least one (but not all) missions. Includes all associated downtime accumulated from maintenance or logistics sources.
Cannibalization flag	Binary	Flag indicating maintenance event is a cannibalization action. Cannibalization is defined as the removal of a serviceable part/component from a donor aircraft in order to restore another aircraft to serviceable condition. Typically performed when no parts are available in supply or during deployments.
Scheduled maintenance flag	Binary	Flag indicating maintenance event is a periodic inspection/servicing/replacement of a part/component performed utilizing mileage, operating hours or calendar time. Typically performed based on manufacturer recommendations or in order to mitigate failure modes.
Type maintenance code.	Categorical	A one-character numeric or alphabetic code identifying the maintenance personnel utilized.
Action taken code	Categorical	A one-character numeric or alphabetic code describing the type of maintenance/action accomplished.
When discovered code	Categorical	A one-character numeric or alphabetic code describing when the work order was identified.
Malfunction code	Categorical	Three character alphanumeric or numeric code identifying the malfunction that occurred. Wiring malfunction codes identified with a “W” in first digit. Examples include broken grounding strap, damaged relays, hard landing or loose.
Action organization code	Categorical	Three character alphanumeric code identifying the organization that actually performed the maintenance.

The data set utilized in this research was selected from a larger EA-6B fleet data set. The entire EA-6B data set was searched for malfunction codes that included a “W” as the first digit, thus signifying a wiring related malfunction. This resulted in a data set of 4,686 observations.

4.5.2 Analysis of Outliers

Several variables contained zero values. Zero values are allowed and did not necessarily raise concern. For example, an aircraft fault may be reported, however the ground crew may quickly perform troubleshooting and ascertain that no failure is present. This discrepancy may only take a small amount of time and the resulting NMC hours may have been considered negligible. The underlying distributions of the features did not appear normally distributed and were skewed. This held regardless of whether outliers were included or not and if zero points were removed. It was expected that the data would not be normally distributed, since typically maintenance or delay times are very short (less than 1 hour). Thus, the highest density of observations appear closer to zero. However, there may be instances where maintenance or delay times are substantial due to difficult maintenance actions, new personnel, or delays in part procurement.

In order to reduce variation within the data set (for aforementioned continuous variables), outliers were first identified. To ascertain if outliers were present and if they were valid or invalid the analysis considered both the minimum, maximum, and Interquartile Range (IQR) for each continuous variable. Invalid outliers were defined as noise within the data set and typically result from incorrect inputting of data by technicians. Valid outliers were identified as accurate observations that appear very different from other observations. For example, an aircraft may require a unique part that is no longer manufactured, resulting in an extreme amount of downtime. The upper threshold for outliers was calculated by removing all zero instances of each feature, determining the IQR and the upper outlier threshold (Q3 + 1.5*IQR). A lower threshold was not calculated since negative values were not present and Q1 was small.

The associated discrepancy and narratives were reviewed for each outlier. The review included consideration of what the discrepancy was and if the associated time (maintenance or delay times) appeared reasonable. The total days non-mission capable was manually calculated by subtracting the work order creation date from the completion date as recorded in the data set. The total days were then multiplied by 24 hours for comparison against the non-mission capable hours as reported in the data set. A large delta was observed in 45 % of the potential outliers. The large delta was due to the NMC hours field being zero in all cases. Four of the records did indicate a large amount of PMC hours, thus the delta could be explained. Three of the records did not include a corrective action narrative, were deemed likely clerical errors, and were removed from the data set. The majority of the remaining records did include awaiting parts hours, which did not clearly correlate to the derived total days NMC. The data quality is summarized in Table 4.2. In total, 697 records tagged as potential outliers were removed from the analysis, including several records with blank nomenclature fields. Thus, of the original data set that includes zeros (n = 4,686) approximately 14.9 % of the records were deemed invalid outliers or invalid observations and removed.

Table 4.2 Analysis of Data Quality and Potential Outliers for Wiring Data Set

Feature	**Count**	**% Blanks**	**% Zero**	**Min**	**Max**	**Q1**	**Q3**	**Outlier Upper Limit based on Q3+1.5*IQR**	**Number of possible Outliers**
Man hours	4,686	0.00 %	1.96 %	0.00	248.10	0.60	3.70	8.40	**580**
EMT		0.00 %	1.98 %	0.00	99.90	0.50	2.90	6.50	**584**
AWM		0.00 %	88.13 %	0.00	3,317.60	0.00	0.00	N/A	
AWP		12.16 %	72.11 %	0.00	12,172.50	0.00	0.00		
NMC Hours		0.00 %	86.26 %	0.00	8,381.00	0.00	0.00		
PMC Hours		0.00 %	95.01 %	0.00	14,623.00	0.00	0.00		
AWM*	556	N/A		0.10	3,317.60	0.80	22.40	54.80	**83**
AWP*	737			0.10	12,172.50	2.30	160.10	396.80	**89**
NMC hours*	644			0.20	8,381.00	2.40	28.70	68.20	**111**
PMC hours*	234			0.10	14623.00	1.8	37.20	90.30	**36**

4.5.3 Data Preparation

Next, binning was utilized in order to reduce the variation further while also converting the variables into categorical features as required by the algorithms utilized in this research. Range normalization techniques were considered; however, this technique is very sensitive to outliers (Kelleher, Namee, & D'Arcy, 2015). Thus, an equal-frequency binning technique was selected. The procedure first sorts instances in ascending order and then divides the data into approximately k-groups, with each group roughly containing the same number of instances.

Several categorical variables were also contained within the EA-6B data set. Of these, work center and malfunction codes had high amounts of variation. The work center variable is a three-character alphanumeric code identifying which functional area performed the maintenance. For example, maintenance officer, production control, or material control. The malfunction code is a three character alphanumeric or numeric code identifying the malfunction that occurred. Both the malfunction code and work center code measurement error introduce variation. Specifically, technicians are expected to identify the root cause of failure (for example high versus low cycle fatigue) and select the appropriate malfunction code. In the majority of cases, technicians likely do not have the required knowledge or information in order to select a highly accurate code. Technicians may not typically know the exact work center, thus introducing additional measurement error. For both codes, they likely only have a general idea of the appropriate malfunction code (for example using cracked) and work center (engines).

There is a plethora of malfunction codes (over 62 alone for wiring related events). However, these codes can be logically grouped, thus reducing the number of possible selections. For example, wiring codes W00 through W05 all describe different situations of wire chaffing. Thus, these can be grouped into a single category “harness/wire chaffing”. This approach was utilized by the researchers and significantly reduced the variation in the malfunction code, without losing any valuable information. The work center code variability was also reduced utilizing a similar approach. The first digit within the work center code describes a broad functional area. For example, all codes starting with a “4” were assigned to an engine functional classification.

Numerical variables were converted to nominal scales utilizing the descriptive statistics calculated previously with outliers removed and zero values not included. A nominal scale of LOW, MEDIUM, HIGH and VERY HIGH were utilized. The data was segmented into four equal parts (using quartiles) in order to identify the lower and upper limits of each level in the scale.

4.6 Results

4.6.1 Analysis of Learning Algorithm Parameters

As discussed in chapter 2 of this dissertation, tabu search was successfully utilized to construct a Bayesian network predicting several features within the EA-6B data set. Although, previous research did consider varying the parameters of the tabu search algorithm, a more extensive analysis was conducted with the wiring data set. Several parameters of the learning algorithm were varied, to include the number of parent nodes, iterations, and starting network conditions. The analysis also considered different local search metrics, to include the entropy, Bayes, MDL, and AIC metrics. Additionally, global search versus local search was also considered along with utilization of validation techniques beyond cross validation.

First, local search metrics were utilized and several algorithm parameters were varied. These included the maximum number of parent nodes, algorithm iterations, and the maximum size of the tabu list. The various local search metrics yielded similar results for each predicted variable. Marginal results (<60 % correctly predicted instances) were obtained for owner organization, relevant system (based on Work Unit Code), and malfunction category (based on MAL code). Increasing the maximum number of parent nodes from 1 to 2, setting iterations to 20 runs, and increasing the tabu list from 5 to 10 improved the results. Further increases to the tabu list and maximum number of parent nodes did not yield higher accuracy. Utilization of global search versus local search metrics, along with other validation schemes were also investigated.

Previous research performed by Isler et al. utilizing a congestive heart failure data set illustrated that the choice of cross-validation method utilized may impact classifier performance (Isler, Narin, & Ozer, 2015). In their work, they considered both the number of folds and the scheme (leave-one-out versus k-fold cross-validation), and found that increasing the number of folds reduced the classifier performance variance. Cross-validation was utilized as the preferred method of validation within this research, however "leave-one-out" and "cumulative fold" methods were also investigated. Global search, along with these additional methods, did not significantly improve the results. The impact of the number of folds utilizing local search (entropy search metric) was investigated. Specifically, the folds were increased from 5 to 25, while other parameters were kept constant. The prediction accuracy did not change significantly based on the number of folds utilized.

In order to investigate if the high prediction accuracy of NMC and AWM hours could be impacted by the records within the data set containing zero NMC hours, the analysis removed all these records. This resulted in a smaller data set of 597 records. Although the records with zero NMC hours were not deemed errors or outliers, the researchers could not clearly ascertain why such a large amount of records included zero NMC hours. It was assumed that there would be a resulting increase in NMC hours any time an aircraft is in maintenance status (thus man-hours greater than zero). However, several business rules may be in place within the Navy that were not available to the researchers. Utilizing the tabu search algorithm (max number of parents = 2; tabu list size = 15, iterations = 20) yielded an overall % correctly classified instances of 78.2 % for NMC hours. Cross-validation and the Bayes local search metric were utilized. As with previous analyses owner organization, relevant system and the malfunction category could not be predicted with any accuracy. Next, the class level results for NMC hours were reviewed. As can be seen in the class level results very LOW and VERY HIGH could be predicted with high true positive rates (0.93 and 0.71) respectively. For all classes except LOW, the false positive rates were low (< .10). Additionally, precision and recall metrics, along with ROC area, indicated overall good results.

4.6.2 Sensitivity Analysis Results

In order to ascertain the impact of noise a full-factorial Design of Experiments (DoE) was constructed. Results are provided both at the algorithm and class levels. Four factors were considered. Factor one (feature to be predicted) had five levels, while the remaining factors (NMC, AWM, and AWP) had three levels. The levels corresponded to increasing levels purposeful error, or noise added to the data set. The response at the algorithm level was % correctly classified instances, thus providing insight into if a resulting solution can be found. The class level results included false positive and true positive rates.

In both cases the residuals and model fit were deemed appropriate (algorithm level R-Sq = 97.5 %; class level R-sq = 97.9 %). The residuals indicated that the normality assumption was valid, and that no clear diagnostic pattern could be obtained. Thus, the models appeared adequate.

The results indicated that several factors, as well interactions between factors, were important. At the algorithm level the target feature was statistically significant (p-value < 0.000). Additionally, interactions were noted between the target feature and NMC, AWM and AWM hours. For the sake of brevity, only the interaction plot for the

% correctly classified instances is provided in Figure 4.4. The target feature appeared to have the only statistically significant main effect. Similar results were observed at the class level. The main effect plots at the class level for both the true positive and false positive rates indicated the target feature had a main effect, while varying the amount of noise within the other factors did not have a statistically significant main effect.

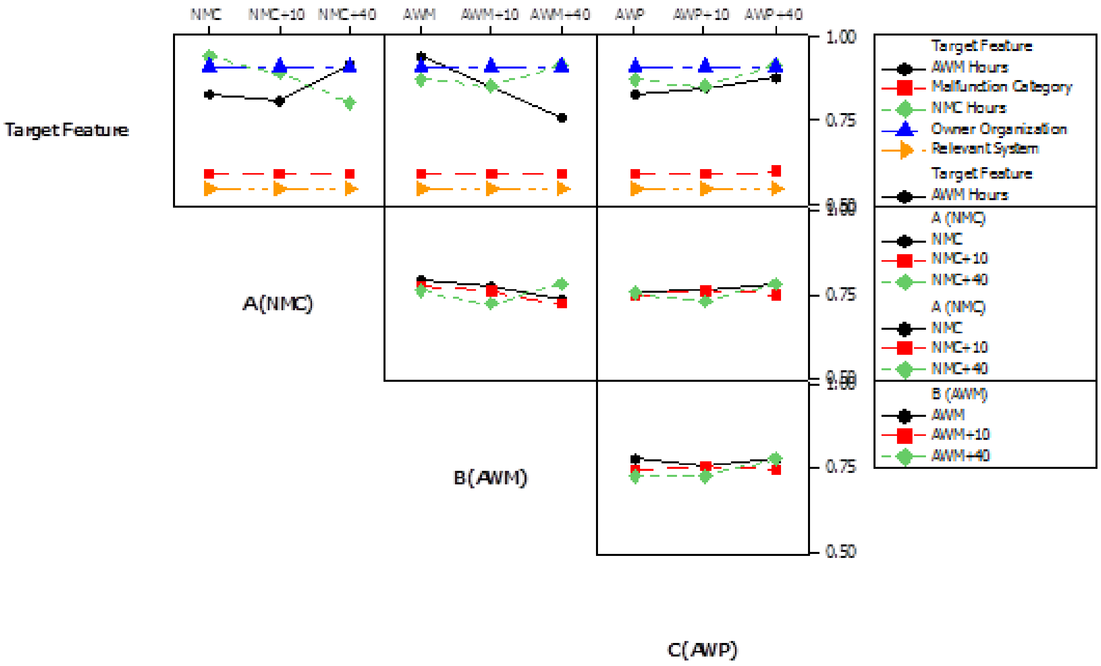


Figure 4.4 Interaction Plot (Algorithm Level)

Boxplots were also constructed of the true positive and false positive responses categorized by target feature. As illustrated in Figure 4.5, the false positive rate when predicting AWM hours varied significantly. Variation was also observed in the false positive rate for NMC hours, although to a lesser extent. The implications of these findings will be discussed later within this chapter.

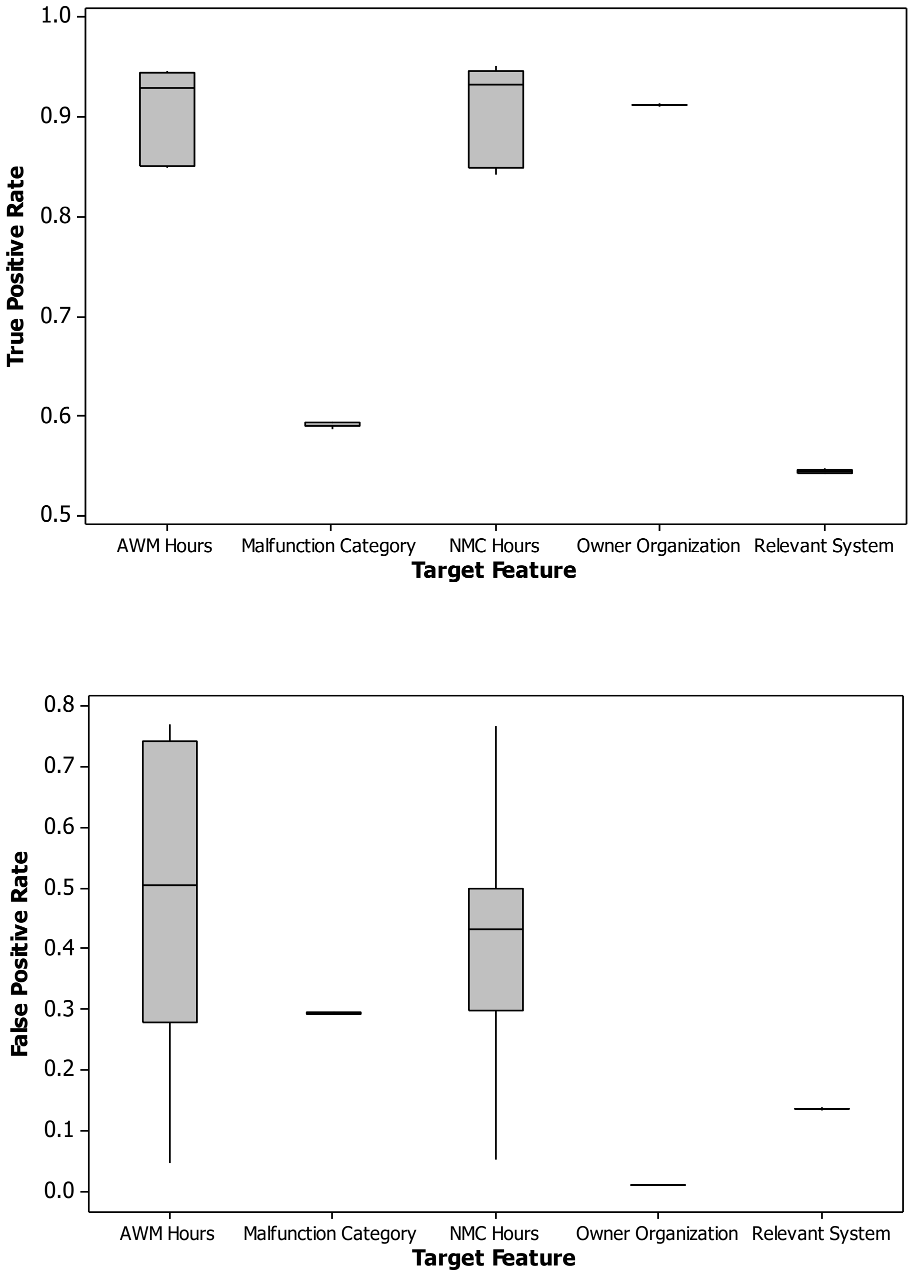


Figure 4.5 Boxplots of True and False Positive Rates over all Runs

4.6.3 Analysis of Predictive or Prognostic Capability under Noise

Bayesian networks are typically utilized as decision tools. They are employed as decision tools by calculating and comparing the joint probability across several variables, given evidence and a validated network structure. Thus, the analysis next focused on assessing how this joint probability may change given the introduction of more noise into the data set. Marcot applies a similar concept to three previously developed Bayesian networks related to forecasting polar bear and Pacific walrus population sizes as well as utilization of tissue samples to predict age of martens. In their analyses they considered the developed Bayesian network valid, and set various features within the model to extreme values while measuring the overall change in the calculated joint probability of a target feature (Marcot, 2012).

The results from the DoE indicated that interaction effects are present between the feature to be predicted and several of the delay times (AWM, AWP). Thus, 10 % and 40 % noise levels were added to the data set. NMC hours were utilized as the predictive variable since it is a critical metric utilized by the Navy in terms of readiness. The analysis considered several systems, to include countermeasures, Identify Friend or Foe (IFF), radar navigation, airframe, hydraulics, engines, instrumentation, and flight controls. Thus, the key systems within the EA-6B were included. The joint probability of NMC hours being LOW, MEDIUM, HIGH or VERY HIGH was calculated. Evidence was set to include AWP and AWM hours varied from VERY LOW to VERY HIGH. The results are plotted in Figure 4.6 in terms of the difference between the baseline case (zero noise) and the noise case. With 10 % noise added, the joint probability did not change significantly. Once 40 % noise was added, greater variation was observed (as expected).

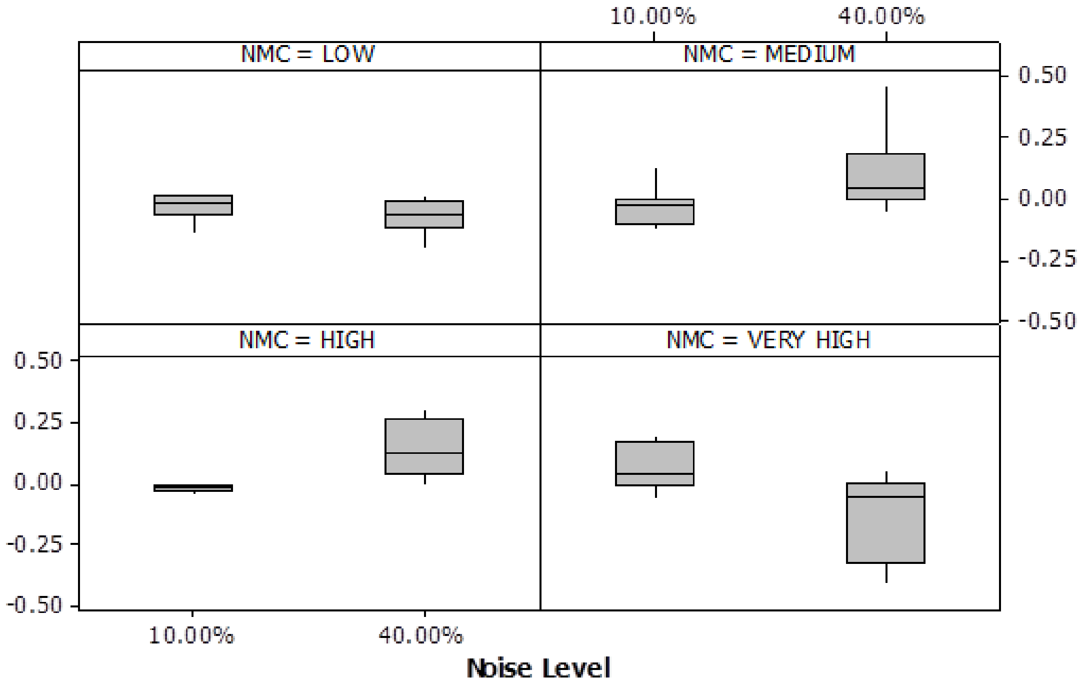


Figure 4.6 Impact on Calculated Joint Probability under Noise Conditions (change from baseline)

4.7 Discussion of Results

The analysis indicated that the tabu algorithm parameters did not result in improvement in terms of the prediction capability of the algorithm for this data set. Thus, the baseline case results were considered stable and adequate to utilize for comparison. The baseline case included both high and low prediction results, depending on the target feature.

The sensitivity analysis yielded interesting results. First, as evidenced by the low variation in the algorithm level results (% correctly classified), it appears that even in a noisy (or large unknown error) environment that a Bayesian network can be constructed successfully. Thus, when purposeful error was introduced the algorithm was still able to develop an underlying topological structure. This finding has both positive and negative aspects. True and false positive rates are critical to analyze to assess model credibility. A model with too low true positive rate will not provide predictive capability. A model with a very high false positive, or false alarm rate, will not retain the confidence of users.

The true and false positive rates remained stable for the owner organization, malfunction category, and relevant system. The true positive rate was low for malfunction category and the relevant system. The results may indicate that although Bayesian networks are tolerant to noise, the underlying predictive power of the algorithm may be significantly impacted. However, this appears very dependent on the target feature to be predicted, both illustrated by the main and interaction effects noted during the statistical analysis as well as by plotting the true and false positive rates.

The low variation under the 10 % noise case when predicting NMC hours will either be LOW or VERY HIGH was further supportive that Bayesian networks may be robust against noise. These categories corresponded to either the best case or worst case scenarios for NMC hours – thus a large deviation in the joint probability would likely result in different decision outcomes. For example, if the joint probability of NMC hours equal to VERY HIGH changed from 0.8 to 0.2 given the same evidence, the decision would change. In the 0.8 scenario the user may preposition equipment, supplies, or personnel since they are expecting high NMC hours. However, in the 0.2 scenario they would not make this decision.

4.8 Conclusion and Future Work

The research indicated that Bayesian networks appear to be robust against noise – however, not for all target features. In some cases, the results are drastically impacted. However, the results indicate that the impact may be minimal under low levels of noise. These results are important and provide credence to utilization of Bayesian networks in real field data – which will always contain noise or error that is not easily quantified. The researchers are not proposing that "garbage noisy data" can construct Bayesian networks with high predictive power, or that due diligence should not be performed in terms of design and validation of measurement systems. However, the research supports that Bayesian networks may be appropriate even in noisy data sets, where the error cannot be easily quantified – assuming a robust sensitivity analysis is performed.

Subsequent research is needed to identify if the results hold for other systems and aircraft, given similar input data sets. Research should also consider incorporation of text based analytics, since maintenance narratives were available. Further research should also be conducted investigating different sampling techniques coupled with cross-validation. An illustrative example can be found in Blagus and Lusa who investigated sampling and validation schemes for class-imbalanced data (Blagus & Lusa, 2015).

CHAPTER V

CONCLUSION

"It is better to conquer yourself than to win a thousand battles. Then the victory is yours. It cannot be taken from you, not by angels or by demons, heaven or hell." Buddha.

The capability to accurately predict military readiness and/or human performance in complex engineering systems provides an important decision tool. Additionally, quantification of the performance parameters of such a tool, to include false positive and true positives rates, is critical to ensure credibility.

Development of these predictive, or prognostic, tools is challenging. Two broad categories have been utilized. The first method utilizes system design knowledge to understand system operation and define causal relationships within a model. This method is challenging since detailed knowledge of the system is required along with associated measurements or observations. A second method utilizes data already collected, applies advanced algorithms, and attempts to predict an outcome based on a known training data set. This method is collectively known as machine learning.

The research performed utilized machine learning algorithms (such as Bayesian networks) and two existing data sets. The primary objective of the research was to develop a diagnostic and prognostics tool utilizing Bayesian networks and to assess its credibility when noisy data sets are utilized.

The research yielded a predictive method with substantial benefits over reactive methods. The algorithm could predict failure of several important components, to include potential malfunction codes and key drivers to military readiness (such as NMC hours). The research also considered potential error within the training data set, which is likely present in military data sets.

In order to ensure a credible model a DoE was designed to investigate model response under noise conditions. The research indicated that Bayesian networks appear to be robust against noise. For some target features the results were dramatically impacted, highlighting that sensitivity analysis is critical. However, the results indicate that, under low levels of noise, the impact was generally minimal.

These results are important and provide credence to utilization of Bayesian networks in real field data – which will always contain noise or error that is not easily quantified. The researchers are not proposing that “garbage noisy data” can construct Bayesian networks with high predictive power, or that due diligence should not be performed in terms of design of measurement systems. However, the research supports that Bayesian networks may be appropriate even in noisy data sets, where the error cannot be easily quantified, assuming a robust sensitivity analysis is performed.

The research also considered a human performance data set within an industrial setting. Although several authors have utilized machine learning techniques to investigate human performance, industrial workplace data sets have not been investigated nor have systems with feedback loops. The proposed methodology was illustrated using representative data of a real-world distribution facility that includes human error rates and worker demographics. Based on subsequent causal links identified within the Bayesian network, we are able to identify important demographics related to worker performance. Moreover, the human error performance is predicted, and can be subsequently utilized to optimize industrial processes.

Although the research was conducted utilizing a large military field data set, additional research should be conducted to replicate the results on other aircraft. Thus, investigation into how the results can be replicated for other systems and aircraft, given similar input data sets, will further improve model credibility. The researchers were not able to incorporate qualitative variables set by domain experts (for example, risk rating for a given component) into the developed Bayesian networks. Unfortunately, text fields within the field data did not provide sufficient discriminators in order to derive qualitative variables. Future research should be conducted in order to solicit and incorporate domain expertise into subsequent models. Research should also consider incorporation of text based analytics for text fields, which was considered out of scope within this research. Although the human performance data provided promising results, a larger data set would allow further refinement of the model.

REFERENCES

Akhlaghi, A. M., Naseh, H., Mirshams, M., & Irani, S. (2011). A Bayesian Networks Approach to Reliability Analysis of a Launch Vehicle Liquid Propellent Engine. *Journal of Aerospace Science and Technology, 8*(2), 107-117.

Averill, M. (2015). *Simulation Modeling and Analysis.* McGraw-Hill.

Banghart, M., Bian, L., & Babski-Reeves, K. (2016). Human Induced Variability during Failure Mode Effects Analysis (FMEA). *Reliability and Maintainability Symposium.* Tucson.

Barriere, M. (2000). *Technical Basis and Implementation Guidelines for A Technique for Human Event Analysis (ATHEANA).* Washington, DCC: Office of Nuclear Regulatory Research.

Bell, J., & Holroyd, J. (2009). *Review of human reliability assessment methods.* Suffolk: HSE Books.

Ben-Gal, I. (2008). Bayesian networks. In *Encyclopedia of statistics in quality and reliability* .

Blagus, R., & Lusa, L. (2015). Joint use of over- and under-sampling techniques and cross-validation for the development and assessment of prediction models. *BMC Bioinformatics, 16*(362).

Blanchard, B. (2008). *Systems Engineering Management.* Hoboken: John Wiley & Sons Incorporated.

Borgonovo, E., & Plischke, E. (2016). Sensitivity Analysis: A review of recent advances. *European Journal of Operational Research*(248), 869-887.

Bouckaert, R. R. (2004). *Bayesian Network Classifiers in Weka.*

Brewster, S., Raty, V., & Kortekangas, A. (1996). Enhacing scanning input with non-speech sounds. *ACM SIGACCESS Conference on Assistive Technologies* (pp. 10-14). Vancouver: Glasgoq ePrint Service.

Campbell, J. P. (1990). *Modeling the performance prediction problem in industrial and organizational psychology.* Palo Alto, CA: Consulting Psychologists Press.

Chen, S., & Pollino, C. (2012). Good practice in Bayesian network modelling. *Environmental Modelling & Software*, 134-145.

Chien, C. F., & Chen, L. F. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems and Applications, 34*, 280-290.

Chow, C. K., & Liu, C. N. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 426-467.

Colón-González, F., Lake, I., Barker, G., Smith, G., Elliot, A., & Morbey, R. (2016). Non-obvious correlations to disease management unraveled by Bayesian artificial intelligence analyses of CMS data. *International Society for Disease Surveillance Annual Conference.*

Copeland, E. (2015). *Exploring feedback mode redundancy in handheld scanning tasks (Master thesis).*

Crowl, D. A., & Louvar, J. F. (2002). *Chemical process safety: Fundamentals with application.* Prentice Hall Publication Inc.

Dadeneh, S., & Qian, X. (2016). Bayesian module identification from multiple noisy networks. *Journal on Bioinformatics and Systems*.

Eliassi, M., Seifi, H., & Haghifam, M. (2015). Incorporation of protection system failures into bulk power system reliability assessment by Bayesian networks. *IET Generation, Transmission & Distribution, 9*(11), 1226-1234.

Feng, N., Wang, H. J., & Li, M. (2014). A security risk analysis model for information systems: Causal relationships of risk factors and vulnerability. *Information Sciences, 256*, 57-73.

Gacquer, D., Delcroix, V., Delmotte, F., & Piechowiak, S. (2011). Comparative study of supervised classification algorithms for the detection. *Engineering Applications of Artificial Intelligence, 24*, 1070-1083.

Goomas, D. T., & Yeow, P. H. (2010). Ergonomics Improvement in a Harsh Environment Using an Audio Feedback System. *International Journal of Industrial Ergonomics, 40*(6), 767-774.

Groth, K. M., Denman, M. R., Cardoni, J. N., & Wheeler, T. A. (2014). "Smart Procedures": Using dynamic PRA to develop dynamic, context-specific severe accident management guidelines (SAMGs). *Probabilistic Safety Assessment and Management.* Honolulu, Hawaii.

Haas, E. C., & Van Erp, J. B. (2014). Multimodal Warnings to Enhance Risk Communication and Safety. *Safety Science, 61*, 29-35.

Hamby, D. M. (1994). A Review of Techniques for Parameter Sensitivity Analysis of Environmental Models. *Environmental Monitoring and Assessment, 32*, 135-154.

Hoshino, E., van Putten, I., Girsang, W., Resosudarmo, B., & Yamazaki, S. (2016). A Bayesian belief network model for community-based coastal resource management in the Kei Islands, Indonesia. *Ecology and Society, 21*(2).

Isler, Y., Narin, A., & Ozer, M. (2015). Comparison of the Effects of Cross-validation Methods on Determining Performances of Classifiers Used in Diagnosing Congestive Heart Failure. *Measurement Science Review, 15*(4), 196-201.

Jantan, H., Hamdan, A. R., & Othman, Z. A. (2010). Human Talent Prediction in HRM using C4.5 Classification Algorithm. *International Journal on Computer Science and Engineering, 2*(8), 2526-2534.

Ji, G., & Tan, K. (2016). A Big Data Decision-making Mechanism for Food Supply Chain. *13th Global Congress on Manufacturing and Management.*

Johnson, S., Logan, M., Fox, D., Kirkwood, J., & Pinto, U. (2016). Environmental decision-making using Bayesian networks: creating an environmental report card. *Applied Stochastic Models in Business and Industry*.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255-260.

Kalantarnia, M., Khan, F., & Hawboldt, K. (2009). Dynamic risk assessment using failure assessment and Bayesian theory. *Journal of Loss Prevention in the Process Industries, 22*, 600-606.

Kelleher, J., Namee, B., & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics.* London, England: MIT Press.

Khan, F. I., Husain, T., & Abbasi, S. A. (2002). Design and Evaluation of safety measures using a newly proposed methodology "SCAP". *Journal of Loss Prevention in Process Industries, 15*, 129-146.

Kohavi, R. (1996). Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid. *Second International Conference on Knowledge Discovery and Data Mining.*

Koopmans, L., Bernaards, C. M., Hildebrandt, V. H., de Vet, H. C., & van der Beek, A. J. (2014). Measuring individual work performance: Identifying and selecting indicators. *Work, 48*, 229-238.

Koopmans, L., Bernaards, C. M., Hildebrandt, V. H., Schaufeli, W. B., de Vet, H. C., & van der Beek, A. J. (2011). Conceptual frameworks of individual work performance - A systematic review. *Journal of Occupational and Environmental Medicine, 53*(8), 856-866.

Kruschke, J. K. (2015). *Doing Bayesian Data Analysis.* Elsevier.

La Morgia, V., Paoloni, D., & Genovesi, P. (2016). Eradicating the grey squirrel Sciurus carolinensis from urban areas: an innovative decision-making approach based on lessons learnt in Italy. *Pest Management Science*.

Ladyman, J., Lambert, J., & Weisner, K. (2013). What is a Complex System? *European Journal for Philosophy of Science*, 33-67.

Lambert, D., Stock, J., & Ellram, L. (1998). *Fundamentals of Logistics Management.* McGraw-Hill Higher Education.

Mao, C., & Canavero, F. (2016). System-Level Vulnerability Assessment for EME: From Fault Tree Analysis to Bayesian Networks - Part I: Methodology Framework. *IEEE Transactions on Electromagnetic Compatibility, 58*(1), 180-187.

Marcot, B. (2012). Metrics for evaluating performance and uncertainty of Bayesian network models. *Ecological Modeliing*, 50-62.

Meeker, W., & Hong, Y. (2013). Reliability Meets Big Data: Opportunities and Challenges. *Statistics Preprints*.

Menelas, B., Picinalli, L., Katz, B. F., & Bourdot, P. (2010). Audio Haptic Feedbacks for an Acquisition Task in a Multi-Target context. *IEEE Symposium on 3D User Interfaces* (pp. 51-54). Waltham: IEEE.

Mkrtchyan, L., Podofillini, L., & Dang, V. N. (2015). Bayesian belief networks for human reliability analysis: A review of applications and gaps. *Reliability Engineering and System Safety, 139*, 1-16.

Moubray, J. (1997). *Reliability-centered Maintenance.* New York: Industrial Press Incorporated.

Moura, R., Beer, M., Patelli, E., & Lewis, J. (2015). Human error analysis: Review of past accidents and implications for improving robustness of system design. *Safety and Reliability: Methodology and Applications*.

Neapolitan, R., Jiang, X., Ladner, D., & Kaplan, B. (2016). A Primer on Bayesian Decision Analysis With an Application to a Personalized Kidney Transplant Decision. *Transplantation*, 489-496.

P., M. D., & T., A. S. (2012, March-April). Designing for Supportability. *Defense AT&L: Product Support*, pp. 34-38.

Ramana, M. V. (2011). Beyond our imagination: Fukushima and the problem of assessing risk. *Bulletin of the Atomic Scientists*.

Randall, A. (2011). *Risk and Precuation.* Cambridge University Press.

Renooij, S. (2014). Co-variation for sensitivity analysis in Bayesian networks: properties, consequences and alternatives. *International Journal of Approximate Reasoning*, 1022-1042.

Rotundo, M., & Sackett, P. R. (2002). The relative importance of task, citizenship, and counterproductive performance to global ratings of performance: A policy-capturing approach. *Journal of Applied Psychology, 87*(1), 66-80.

Scott, J., & Gray, R. (2008). A Comparison of Tactile, Visual, and Auditory Warnings for Rear End Collision Prevention in Simulated Driving. *Human Factors, 50*(2), 264-275.

Shebl, N. A., Franklin, B. D., & Barber, N. (2012). Failure mode and effects analysis outputs: are they valid? *BMC Health Services Research, 12*(150).

Simon, P. (2013). *Too Big to Ignore: The Business Case for Big Data.* Wiley Press.

Singh, P. (2009). Comparing the effectiveness of machine learning algorithms for defect prediction. *International Journal of Information Technology and Knowledge Management*, 481-483.

Skiena, S. (2010). *The Algorithm Design Manual* (2nd ed.). Berlin: Springer Science and Business Media.

Spence, C., & Lee, J. H. (2008). Assessing the benefits of multimodal feedback on dual-task performance under demanding conditions. *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction - Volume 1* (pp. 185-192). Swinton, UK: British Computer Society.

Thakur, G. S., Gupta, A., & Gupta, S. (2015). Data Mining for Prediction of Human Performance Capability in the Software-Industry. *International Journal of Data Mining & Knowledge Management Process, 5*(2).

Vachtsevanos, G., Lewis, F. L., Roemer, M., Hess, A., & Wu, B. (2006). *Intelligent Fault Diagnoses and Prognosis for Engineering Systems.* Hoboken, New Jersey: John Wiley & Sons.

van der Gaag, L., Renooij, S., & Coupe, V. (2007). *Sensitivity Analysis of Probabilistic Networks.*

Viera, A. J., & Garrett, J. M. (2005). Understanding Interobserver Agreement: The Kappa Statistic. *Family Medicine, 37*(5), 360-363.

Villez, K., Srinivasan, B., Rengaswamy, R., Narasimhan, S., & Venkatasubramanian, V. (2011). Kalman-based strategies for Fault Detection and Identification (FDI): Extensions and critical evaluation for a buffer tank system. *Computers and Chemical Engineering, 35*, 806-816.

Vitense, H., Jacko, J., & Emery, V. (2003). Multimodal feedback: an assessment of performance and mental workload. *Ergonomics, 46*, 68-87.

Wickens, C., & Hollands, J. (2000). *Engineering Psychology and Human Performance.* New Jersey: Prentice Hall.

Xu, D., Wei, Q., Chen, Y., & Kang, R. (2015). Reliability Prediction Using Physics-Statistics-Based Degradation Model. *IEEE Transactions on Components, Packaging, and Manufacturing Technology, 5*(11), 1573-1581.

Yuan, Z., Khakzad, N., Khan, F., & Amyotte, P. (2015). Risk Analysis of Dust Explosion Scenarios Using Bayesian Networks. *Risk Analysis, 35*(2), 278-291.

Zhang, L., Wu, X., Qin, Y., Skibniewski, M., & Liu, W. (2016). Towards a Fuzzy Bayesian Network Based Approach for Safety Risk Analysis of Tunnel-Induced Pipeline Damage. *Risk Analysis, 36*(2), 278-301.