Bushing Diagnosis using Artificial Intelligence and Dissolved Gas Analysis

Sizwe Magiya Dhlamini

A thesis submitted to the Faculty of Engineering and the Built Environment, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Doctor of Philosophy.

Johannesburg, 2007

Declaration

I declare that this thesis is my own, unaided work, except where otherwise acknowledged. It is being submitted for the degree of Doctor of Philosophy in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other university.

Signed this ____ day of _____ 20__

_____

Sizwe Magiya Dhlamini

Abstract

This dissertation is a study of artificial intelligence for diagnosing the condition of high voltage bushings. The techniques include neural networks, genetic algorithms, fuzzy set theory, particle swarm optimisation, multi-classifier systems, factor analysis, principal component analysis, multidimensional scaling, data-fusion techniques, automatic relevance determination and autoencoders. The classification is done using Dissolved Gas Analysis (DGA) data based on field experience together with criteria from IEEEc57.104 and IEC60599. A review of current literature showed that common methods for the diagnosis of bushings are: partial discharge, DGA, tan- (dielectric dissipation factor), water content in oil, dielectric strength of oil, acidity level (neutralisation value), visual analysis of sludge in suspension, colour of the oil, furanic content, degree of polymerisation (DP), strength of the insulating paper, interfacial tension or oxygen content tests. All the methods have limitations in terms of time and accuracy in decision making. The fact that making decisions using each of these methods individually is highly subjective, also the huge size of the data base of historical data, as well as the loss of skills due to retirement of experienced technical staff, highlights the need for an automated diagnosis tool that integrates information from the many sensors and recalls the historical decisions and learns from new information. Three classifiers that are compared in this analysis are radial basis functions (RBF), multiple layer perceptrons (MLP) and support vector machines (SVM). In this work 60699 bushings were classified based on ten criteria. Classification was done based on a majority vote. The work proposes the application of neural networks with particle swarm optimisation (PSO) and genetic algorithms (GA) to compensate for missing data in classifying high voltage bushings. The work also proposes the application of fuzzy set theory (FST) to diagnose the condition of high voltage bushings. The relevance and redundancy detection methods were able to prune the redundant measured variables and accurately diagnose the condition of the bushing with fewer variables. Experimental results from bushings that were evaluated in the field verified the simulations. The results of this work can help to develop real-time monitoring and decision making tools that combine information from chemical, electrical and mechanical measurements taken from bushings.

Preface

The work described in this thesis was carried out at the University of Witwatersrand from October 2002 until May 2007. I am grateful to Professor Tshilidzi Marwala, who was my supervisor, for his guidance and support throughout that time. Thank you to Mrs Moe Thida Htwe Hnin Hay Mar Dhlamini, my wife, for proof reading this work and encouraging me. I thank Mr Vincent Nelwamondo, Mr Thando Tetty, Ms Unathi Mahola, Mr Bheki Duma, Mr Lungile Mdlazi, Mr Brain Leke Betechuoh, Mr Shakir Mohamed, Ms Busi Vilakazi, Mr Obakeng Sebaeng and Mr Luke Machowski for valuable input. Mr Gregory Makathini inspired me to study, thank you.

To my parents Mawe Pauline Khanyile Dhlamini and Baba Vusimuzi Nkosi Dhlamini, I am eternally grateful for the sacrifices that you made and the support that you gave to all your children that allowed me to do this work. Ngibonga bokhokho, bo Dhlamini-Ndlangamandla-Mtungwa-Lokotwayo-Mhayise-Magiya, nine-baka-Nomembula. Ngini lenangiletsa emhlabeni nangicina. NgibonganeNkosi.

Except where reference is made to the work of others, this thesis is the result of my own original work. No part of this work has already been, or is currently being, submitted for any other degree, diploma or other qualification. This thesis is 172 pages in length, and contains approximately 37578 words.

S.M. Dhlamini

Johannesburg, May 2007

Contents

List of Figures

List of Tables

Nomenclature

AI: Artificial Intelligence
ANN(s): Artificial Neural Network(s)
ARD: Automatic Relevance Determination
CH4: Ethane
C2H6: Methane
C2H4: Ethylene
C2H2: Acetylene
CO: Carbon Monoxide
CO2: Carbon Dioxide
CSUS: California State University Sacramento (Guidelines for DGA)
DGA: Dissolved Gas Analysis
DP: Degree of Polymerisation
FST: Fuzzy Set Theory
GA: Genetic Algorithm
H2: Hydrogen
MLP(s): Multilayer Perceptron(s)
N2: Nitrogen
O2: Oxygen
PD: Partial Discharge
PSO: Particle Swarm Optimisation
RBF(s): Radial Basis Function(s)
SVM(s): Support Vector Machine(s)
TDCG: Total Dissolved Combustible Gases

Chapter 1

ARTIFICIAL INTELLIGENCE FOR BUSHING DIAGNOSIS

1.1 BACKGROUND

Bushings are a critical component in electricity transmission. They are used in transformers (Abdella & Marwala 2005), reactors, circuit breakers and switchgear. Bushings are used in all places where the clearance in a normally insulating medium such as air is insufficient to allow an energised conductor to pass through it without causing a flashover. Regardless of the root cause of failure, the result is often catastrophic (Dhlamini 2000). When a bushing fails catastrophically, there is a violent explosion propelling large broken pieces of porcelain several metres at velocities enough to embed the material in concrete walls and destroy neighbouring bushings. Often an explosion is accompanied by fire which is what destroys transformers.

Transformers are the single most important and most expensive item of equipment in the electrical network. van Wyk (1997) highlighted that more than 15% of transformer failures are due to bushings. Pukel et al. (2006) found that bushing failures on transformers account for 23% of failures. An Australasian reliability study of 2096 transformers over the period 1970-1995 concluded that bushings were second to tap changers as the component initially involved in failure and were amongst the top three contributors to costly transformer failures (Workgroup 1996), (Lord & Hodge 2003). More than 46% of transformer defects were found to be attributable to bushings, on load tap changers, and the cooling system (Sokolov 2001). Another study from the 63 members of the Edison Electric Institute listed bushings amongst the three most common transformer failure modes reported (Janick 2001).

A solution that targets bushings to prevent failure can increase the reliability of a power plant. Figure 1.1 shows how bushings look like and different areas where they are applied. The figure on the left shows three 132kV bushings connected to a transformer. The figure on the right, from Urban & de Villiers (2006), shows a medium voltage bushing used on an alarm system. Table 1.1, taken from Pukel et al. (2006), shows a frequency distribution of root causes of failures of high voltage transformers in Germany.

Table 1.1: Failure distribution of high voltage transformers

Root cause	Frequency	Root cause	Frequency
Windings	24%	Tank	10%
Core	20%	Cooling system	4%
Switching devices	14%	Safety devices	1%
Bushing	23%	Others	4%

1.2 LITERATURE REVIEW

Artificial intelligence has been studied and applied in many parts of power systems engineering. Dominelli & Lau (2004) were able to interpret the fault condition that caused transformer failure by using dissolved gas analysis (DGA) data and DGA criteria as presented in (IEEEC57-104 1991). Harpe & Kranz (1999) used artificial neural networks (ANNs) to monitor partial discharge in transformers. Wang (2000) used a multilayer perceptron (MLP) to diagnose the condition of a transformer using IEC60599 (1999) DGA criteria. Werle et al. (2001) used an MLP to locate partial discharge (PD) in a transformer. With a single layer MLP they were able to determine with 99% accuracy whether the PD was in air, oil or in the winding paper. Bhattacharya et al. (2001) used neural networks to improve the shape of an insulator shed profile. An adaptive fuzzy classifier was used by Purkait et al. (2003) for transformer impulse fault diagnosis. Purkait & Chakravorti (2003) also investigated using expert systems to analyse impulses on distribution transformers. Ghosh et al. (2006) used fuzzy logic to estimate inception and extinction voltages from partial discharge in a solid dielectric material.

A shortcoming of the previous studies is that they focus on transformers, particularly the windings and core. They ignore bushings, which are a root cause of many

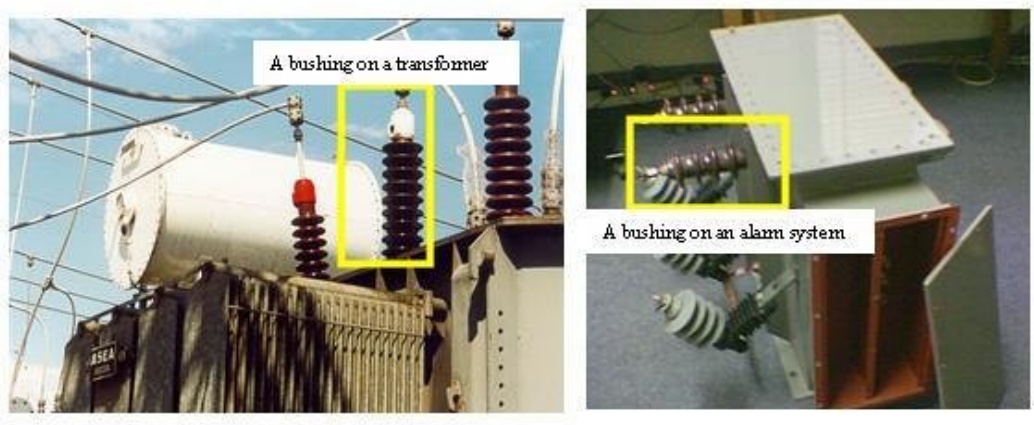

Figure 1.1: Different applications of bushings

transformer failures. Results from monitoring and testing the core and windings do not give information on the condition of the bushings

1.3 SCOPE OF THIS THESIS

This thesis evaluates machine learning tools that are based on human experience as well as standard methods of testing that are used for analysing data to determine the condition of high voltage bushings. This work also explores different optimisation techniques that are applied to compensate for data that is lost in the sampling and measurement process. Furthermore, it identifies and compares different techniques for extracting important features from bushing data. A method of representing abstract qualitative data in a precise quantitative manner is explored that will enable accurate comparisons to be made when evaluating bushings. This work used ten input criteria to test the efficiency of several ANN tools using expert criteria based on California State University Sacramento (CSUS) guidelines (Hubacher 1976), IEC60599 (1999), and IEEEC57-104 (1991) using Doernenberg & Gerber (1967) and Rogers (1975) methods.

1.4 FOCUS OF THIS RESEARCH

Questions that will be answered by the research are the following: How can a utility use machine learning to reduce bushing failures? Which machine learning techniques are most useful and accurate in the context of bushing condition monitoring? What measurements indicate defects and faults inside a bushing? What variables are most

important among the measured variables, i.e. those variables that give a unique signature of the condition of a bushing?

1.5 MOTIVATION FOR THIS RESEARCH

When a bushing fails on an equipment while it is operational in the field, the consequences are often catastrophic. In Figure 1.2 the image on the left shows a transformer a few minutes after failure. The fire that often accompanies a bushing explosion is fueled by oil from the bushing and the transformer. The image on the right is taken from Lau (2003). It shows a burnt out transformer, which was due to a bushing failure.

Figure 1.2: Destructive power of bushing failure

The material replacement cost of a large power transformer failing can be expensive. To quote a real incident: "A fault on a transformer at the Eastern Grid, Transmission substation Ariadne in KwaZulu-Natal resulted in a major loss. The transformer and associated equipment were totally destroyed. Passive fire protection and environmental measures worked well as the incident was limited to its point of origin. A formal investigation is underway" (Lennon 2006). The cost in lost revenue, legal costs, transport and repairs can be more than double the replacement cost. A reliable way of monitoring bushings using the data that is routinely taken by field service staff, and alarming in time to allow maintenance intervention can save money and improve the reliability of the power grid. There is concern about the fact that making decisions using any of the manual diagnosis methods, which will be listed

in Chapter 2, individually, is highly subjective. Also, the size of the database of historical information as well as the loss of skills due to migration and retirement of experienced technical staff highlights the need for an automated diagnosis tool that integrates information from the people, sensors and the database. The tool should recall the historical decisions and learn from new information. Such a tool will allow engineers to make more informed decisions about the plant that they operate and maintain.

1.6 PROCESS FOLLOWED IN CONDUCTING THIS RESEARCH

To answer the research questions and others that were identified in the research process, simulations were done on a Maxwell®, Electro® and Coulomb®, to produce electro-static models of a bushing. Matlab® simulations were done to understand the applicability of neural networks, fuzzy set theory and data reduction methods such as principal component analysis, multidimensional scaling and factor analysis and others for bushing diagnosis. Evolutionary computing methods in the form of genetic algorithms and particle swarm optimization within neural networks were compared to gradient methods such as scaled conjugate gradient and quasi-Newton and non-gradient methods such as the simplex method.

Different methods used for determining the condition of bushings were evaluated and compared. Such methods include partial discharge monitoring, visual inspection, noise monitoring, dissolved gas analysis (DGA) assessment, dielectric dissipation factor monitoring, temperature monitoring and mechanical assessment. Once a suitable method had been selected, data from healthy and failed bushings were analysed. Taking into account the availability of test equipment and reproducibility of the test results, a decision was made on which manual assessment method to use jointly with machine learning methods. The dissolved gas analysis (DGA) method was determined to be the most widely accepted tool within the electricity industry in South Africa and most reliable diagnostics tool for monitoring the condition of bushings and transformers (Mewalall et al. 2006).

Experimental results for manually classifying 60699 bushings based on dissolved gas information were compared to results of classification using artificial neural networks (Dhlamini & Marwala 2004). From this comparison a conclusion was made on whether or not neural networks can be used accurately to monitor and diagnose

a bushing's condition. Different types of neural networks were compared based on stability, accuracy and speed. Given the large number of bushings, together with the number of variables that are used to decide on the condition of each bushing using DGA, the machine learning tool should decide on the importance of the variables used in making the classification decision. Data relevance determination techniques were tested and compared.

During the work, it was found that many of the variables did not match the threshold criteria exactly to be classified as usable or unusable. There is a degree of tolerance within the DGA criteria that can be exploited using fuzzy set theory (FST). FST was integrated with ANN's to classify bushings and indicate the extent of degradation of each diagnosed bushing (Dhlamini et al. 2005).

The research ended off by comparing different machine learning techniques and highlighting the deficiencies that were identified. Recommendations were made on how the work can integrate into operating and maintenance practices, and what future research can be conducted in relation to bushing monitoring.

1.7 THE CONTRIBUTIONS OF THIS RESEARCH TO NEW KNOWLEDGE

Literature suggested many different methods of monitoring capacitive graded or high voltage bushings. A shortcoming of these studies is that they do not analyse how bushings fail, yet they propose methods of measuring and diagnosing the condition of bushings. This thesis identifies and explains how high voltage oil impregnated bushings fail. Then it identifies the important variables for monitoring the condition of bushings. The work then identifies a reliable method of measuring and monitoring the variables. It evaluates different types of machine learning tools in the context of condition monitoring for bushings.

Machine learning tools are compared based on speed, stability and accuracy in classifying. The networks are radial basis function (RBF) and multilayer perceptron (MLP). The kernel machine used is a support vector machine (SVM). In this work 60699 bushings' functional states and defects are classified based on ten criteria. The ANNs and SVM are configured to form a committee of classifiers. Classification by the committee is done based on a majority vote. The work proposes the application

of neural networks with particle swarm optimisation (PSO) and genetic algorithms (GA) to compensate for missing data in classifying high voltage bushings. The work applies fuzzy set theory (FST) to diagnose the condition of high voltage bushings (Dhlamini, Marwala & van Coller 2005). Relevance and redundancy detection methods are able to prune the redundant measured variables and accurately diagnose the condition of bushings with fewer variables. Experimental results from bushings that were evaluated in the field verified the simulations. The results of this work can help to develop real-time monitoring and decision making tools that combine information from chemical, electrical, optical, thermal and mechanical measurements that are taken from bushings.

1.8 STRUCTURE OF THIS THESIS

Chapter 1 explains why this research is necessary and how it will contribute to knowledge creation. The research process is explained in the first chapter, detailing the sources of information, tools that were used in the investigation and the evaluation criteria. Chapter 2 explains bushings and their applications in detail. The chapter also presents a review of the literature of existing work on the subject of bushing condition monitoring. The failure mechanism of bushings is explained. The variables that are important for condition monitoring are described. The meaning of the variables is also detailed with respect to condition monitoring. A suitable method of monitoring bushings is selected which is accurate, reliable and suitable for application of machine learning. Chapter 3 presents an explanation of artificial neural networks and support vector machines. The chapter presents results of radial basis function, multilayer perceptron, and support vector machine applied for diagnosis of transformer bushings. In this chapter different methods of optimisation are compared. In addition, different activation functions are used in different layers of the various machine learning tools.

Chapter 4 applies machine learning tools together with four dissolved gas analysis criteria to determine the condition of bushings. The criteria that are applied within the neural networks are Doernenburg, Rogers ratio, IEC60599 and California State University Sacramento (CSUS) Guidelines (Dhlamini et al. 2005). Chapter 5 looks at the impact that sensor failure has on the ability of machine learning tools such as neural network to diagnose the condition of bushings. A genetic algorithm and particle swarm optimisation algorithm are used to approximate the missing

variables jointly with an autoenconder. Chapter 6 presents an application of fuzzy set theory to diagnose the condition of bushings. A neuro-fuzzy classifier is also demonstrated in this chapter. Chapter 7 presents several data reduction techniques, which are used to identify the most important variables in the data set. The reduced data set is then used to diagnose the condition of bushings and make a comparison with a complete data set in terms of the accuracy of classification.

Chapter 8 compares different techniques and makes concluding remarks about each of the methods. Chapter 8 also makes recommendations and proposals about future research related to condition monitoring of bushings using machine intelligence. In addition the chapter makes recommendations about how the research can fit into the operating processes.

Chapter 2

CURRENT STATUS OF BUSHING CONDITION MONITORING

2.1 INTRODUCTION

The most prevalent maintenance practice in the electricity industry in South Africa used for monitoring and diagnosing the condition of bushings on transformers is to take a sample of the oil and perform dissolved gas analysis (DGA) using gas chromatography (Dalton et al. 2005). Oil is sampled and tested at six-month intervals for combustible gases, moisture and dielectric strength. The results are stored and updated in a database. Trending, IEC Ratios and the Modified Equilibrium Chart are the standard methods used to analyse the data (de Klerk 2005). If an increasing trend of gases is identified, employees with many years of work experience assess the severity of the defect on the bushing before making a decision as to whether a bushing should be removed from service for a detailed investigation and repair.

Mewalall, de Klerk & Moroaswi (2006) highlights that the above method has several problems. For instance, if the method of sampling is not correct and not uniform, sometimes the test results are useless. Turnaround time to diagnose the condition of power equipment is very long, so the plant operators and maintenance department cannot make informed decisions on time. The data is managed, analyzed and translated differently by each plant operator due to varying levels of skill and a multitude of approved international standards that are used to translate the data. Defective components have not been detected and removed from service until after they fail. The time interval is too long between sampling, testing, analysing, making a decision and then repeating the whole process. During the interval a fault or a

minor problem can develop into a serious fault or even failure of equipment. The monitoring and diagnosis process is costly due to its labour-intensive nature and also because of the equipment needed to do sampling and testing is expensive.

To address these problems portable gas analysers and online monitoring devices have been installed. The devices are shown in Figure 2.1. In addition to gas chromatography, gas separation columns and photo-acoustic spectrometry are used to increase the accuracy of identifying and measuring dissolved gases in oil. Data processing in these instrument relies on trending or IEC ratios or the Modified Equilibrium Chart. Decisions are still made based on multiple standards, by individuals with varying levels of skills. The decision making process is still the same as before, only that the laboratory has been moved to the field. These interventions do not address the problem of accurate and consistent decisions using accepted standards together with collective historical experience. The turnaround time from sampling to deciding has now been reduced from six months to approximately one month.

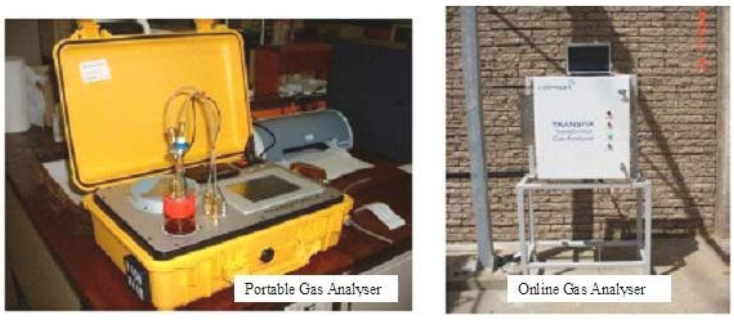

Figure 2.1: Dissolved gas analysis measuring instruments
(Mewalall et al. 2006)

A more complete description of methods that is used to monitor and diagnose transformers is that given by Pukel, Muhr & Lick (2006). Not all the diagnostics tools used on transformers can be applied on bushings, so a detailed discussion of the methods relevant to bushings is given later in this chapter. Figure 2.2 shows how monitoring and diagnosis methods can be grouped into five classes, namely optical, chemical, electrical, mechanical and thermal. Section 2.3 explains the classes in more detail.

Zhang et al. (2006) used an active Rogowski coil to monitor the leakage current at the bushing test tap, then used an ANN to make a decision on the health of the bushing. Stirl et al. (2006) used the tests tap to monitor the capacitance and dielectric dissipation factor or tan-delta (tanδ) of oil impregnated, resin impregnated

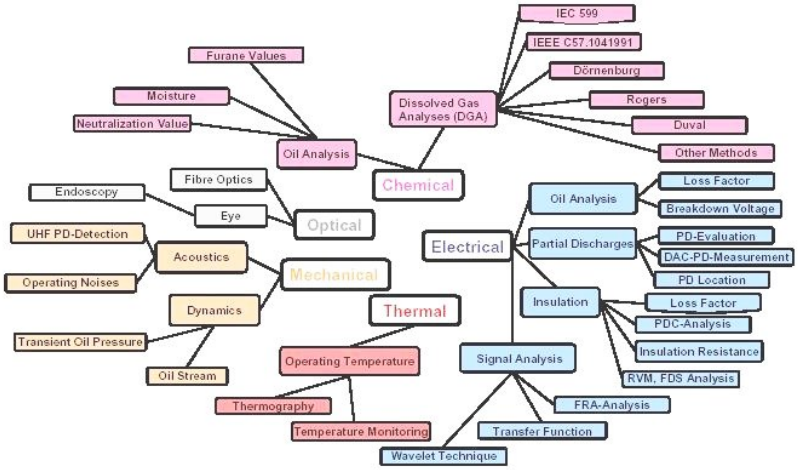

Figure 2.2: Monitoring and diagnosis tools for oil filled plant (Pukel et al. 2006)

and resin bonded bushings. They choose a capacitance sensor because the bushings they were monitoring do not all make use of oil insulation. Figure 2.3 shows an installation with an online tanδ monitor as installed by Avera (Stirl et al. 2006).

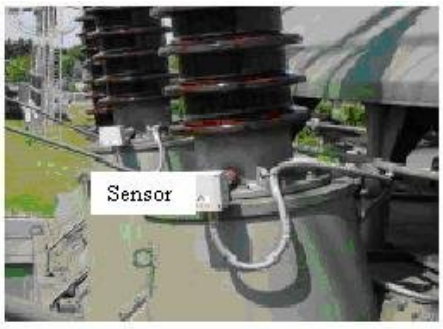

Figure 2.3: Online tanδ sensor on bushings

2.2 UNDERSTANDING THE FAILURE MECHANISM

It is important to understand the failure mechanisms of oil impregnated paper bushings in order to determine which variables are important to monitor. To explain the failure mechanism, one needs to understand the structure of a bushing. High voltage bushings consist of conducting screens that optimally distribute the electric field radially (E_r) and tangentially (E_t) around the bushing. The screens are wrapped in paper, submerged in oil and encased in a porcelain shell. Oil is the main internal insulating medium, used together with paper. Porcelain is the external insulating medium. Figure 2.4 shows the structure of an oil impregnated paper bushing.

Electromechanical forces caused by short-circuit fault currents can crack porcelain, deform the conductor and dislocate the screens. High voltage impulses from lighting and switching can cause flashovers internally and externally on the bushing. Overloading and fault currents generate heat as well as magnetic fields, which excite electrons and cause high temperatures that degrade oil and paper insulations. Contaminants such as water in the oil and surface pollutants result in partial discharges and leakage current flowing in bushings.

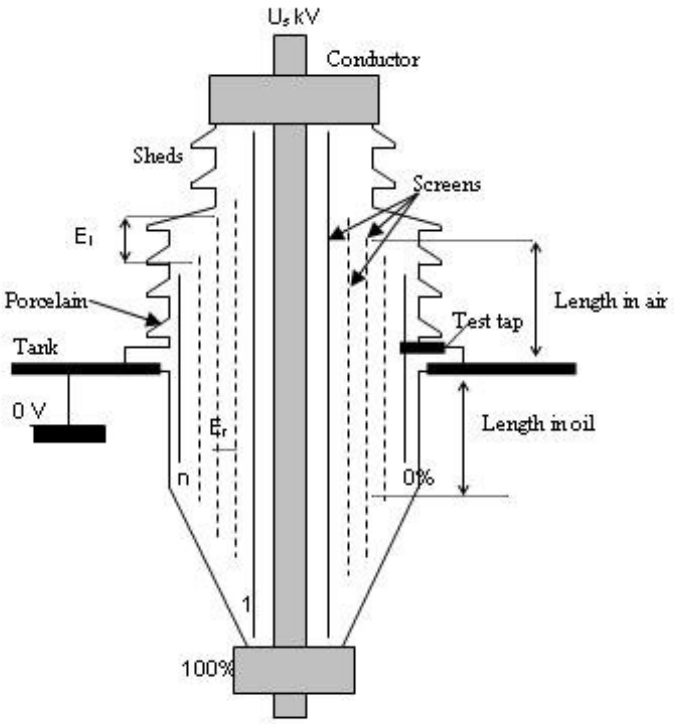

Figure 2.4: Components of oil impregnated paper bushings

If the geometry or chemical composition of bushing components changes because of induced stresses, the electric field distribution changes and the clearances may not be sufficient to sustain the normal operating voltage (U_s) when the system returns to its steady state. Details of the dielectric strength and minimum clearances is given in Appendix B.

The characteristics required in insulating oil used in bushings are defined in (IEC60867 1986). It contains hydrocarbon molecules linked together by carbon-carbon bonds. 338kJ/mol of energy will break the C-H bonds in insulating oil (IEC60599 1999). This scission of C-H bond will release hydrogen into the oil. Energy of 607 kJ/mol will break single bond C-C molecules. 720 kJ/mol will separate C=C double bonds.

960kJ/mol will break C≡C triple bonds. Changes in the chemical structure due to stress which is experienced by the bushing under normal operating and abnormal conditions can be detected by monitoring oil, leakage current, capacitance and partial discharge activity. To understand why dissolved gas analysis was selected, section 2.2 explains and evaluates condition monitoring methods before selecting a suitable method or methods on which to apply artificial intelligence.

2.3 METHODS OF FAULT DIAGNOSIS

Methods for diagnosing a bushing's condition can be classified into chemical monitoring, electrical monitoring, temperature monitoring, monitoring of optical parameters, and mechanical diagnostic methods.

2.3.1 Chemical condition assessment methods

Chemical condition assessment refers to detecting any change in the chemistry of the components of a bushing, such as oil, paper, porcelain, aluminium screen, conductor resin, gas or air around the bushing. In oil impregnated paper bushings, generally only the insulating oil and paper suffer detectable decomposition during the lifetime of a bushing. For bushings that are installed on a transformer the oil inside the transformer and that in the bushing are not able to circulate and mix. Generally the bushing conductor is a hollow cylinder to account for the skin-effect and to minimise material costs. With a hollow conductor the oil from the transformer is able to move into the conductor and cool down the bushing.

To ensure consistent diagnosis, it is important that correct procedures are maintained with regards to the sampling point, sample removal, sample storage and sample testing. Standard procedures for sampling are detailed in (IEC60475 1974) and (IEC60567 1992).

Fresh mineral insulating oil as defined in (IEC60867 1986) has a multitude of hydrocarbon molecules containing CH_3, CH_2 and CH linked together by carbon-carbon bonds. Changes in the chemical structure can be monitored and used as an alarm of possible failure. Foreign particles such as air, dust and water can enter the dielectric and form non-dielectric compounds with different thermal, physical and electrical properties. All these can be monitored using chemical processes catergorised into

gas detection monitoring and water contamination detection.

Gas detection and monitoring

Once the gases are formed inside the oil, they either dissolve in oil or accumulate as free gases, depending on the level of purity of the oil, the temperature and pressure inside the bushing, as well as the rate of production of the gases (IEEEC57-104 1991). There are numerous methods that are currently specified for sampling, and testing of oil-paper insulation for gases and chemical by-products of decomposition. (IEC60567 1992) specifies the method of sampling oil for analysis of free and dissolved gases in oil. Up to 97% of dissolved gases can then be extracted from oil for analysis using multi-cycle vacuum extraction called Toepler extraction method vacuum extraction, or single-cycle vacuum extraction called partial degassing, or by a method called stripping or another method called head space (IEC60567 1992). The concentration of gas is calculated as follows (IEC60567 1992).

$$C_i = V_i/V \tag{2.1}$$

where i is the dissolved gas, e.g.CH or H, etc. C_i is the concentration of gas, in $\mu l/l$, V_i is the volume of gas in ml, corrected to 20°C and 101,3 kPa. V is the exact volume of oil in ml.

Gas chromatography is a widely accepted method used to determine the gases dissolved in oil (Mewalall et al. 2006). IEC60813 (1985) recommends test methods to determine the stability of insulating oil to oxidation.

IEC60599 (1999) distinguishes six classes of faults that can be determined from analysing oil. They are quoted here in order of increasing severity from lowest impact to highest impact: discharges of low energy (D1), partial discharges (PD), discharges of high energy (D2), thermal faults below 300°C (T1), thermal faults above 300°C yet below 700°C (T2) and thermal faults of temperatures above 700 °C (T3). PD includes cold plasma and spark type discharges. It causes X-wax to be deposited on paper insulation if PD is of the cold plasma type, or pinhole punctures in paper if PD is of the spark type. D1 causes paper to carbonise, a process called tracking, or punctures through paper or appears as carbon particles suspended in oil. In D2 degradation power flows through the insulating oil and paper, so there

is extensive destruction and carbonization of paper and oil as well as metal fusion. Some times the equipment may trip due to sensitive earth fault detection when it has reached D2 degradation. T1 degradation is present if the insulating paper appears brown in colour. If the paper is black and brittle, i.e. carbonised, then degradation is classified as T2. Three levels of high temperature degradation (T3) can be distinguished. Carbonised or black oil indicates above 700°C temperatures, discoloured metal indicates temperatures above 800°C, fused metal indicates temperatures higher than 1000°C. Tables 2.1 and Table 2.2 show the ratios.

Table 2.1: Translation of DGA gas ratios - Part I
(IEC60599 1999) and (IEEEC57-104 1991)

	IEC 60599	IEEEC57-104		
		Doernenburg		Rogers Ratio
Gases	Normal limit	Normal	Faulty	Normal limit
Hydrogen(H_2)	140	100	200	$0.1 \leq CH_4/H_2 \leq 1$
Methane(CH_4)	40	120	240	
Ethane(C_2H_6)	70	65		$C_2H_4/C_2H_6 < 1$
Ethylene(C_2H_4)	30	50	100	
Acetylene(C_2H_2)	2	35	70	$C_2H_2/C_2H_4 < 0.1$
CarbonMonoxide(CO)	1000	350		
CarbonDioxide(CO_2)	3400			

Gas volumes of dissolved gases are then compared to each other and the root cause of fault is determined based on the gas ratios. Table 2.2 gives the ratios. Values exceeding the indicated magnitudes mean the unit is in danger of degradation. IEC60599 (1999) does not specify the absolute values of different gases to use as indicators of likely failure. The standard recommends that a utility develops this criteria by monitoring failures of plant and determining the limiting values for each gas. Rates of change in gases is recommended as a way of monitoring the condition of oil impregnated paper bushings. More information about rates of change appears in Appendix C.

IEEEC57-104 (1991) distinguishes three types of faults. They are: thermal discharges (TD), low intensity electrical discharges (LIED) and high intensity electrical discharges (HIED). Thermal discharges are in the temperature range from 150°C to 500°C. Oil degradation due to TD generates large quantities of hydrogen (H_2)

and methane (CH_4) and small amounts of ethylene (C_2H_4) and ethane (C_2H_6). Insulating paper contains cellulose, whose degradation due to TD produces carbon monoxide (CO), carbon dioxide (CO_2) and water vapour. LIED include partial discharges and very low level intermittent arcing. These produce hydrogen, methane and acetylene. If the intensity increases, then acetylene and ethylene are produced in higher concentrations. Arcing is a HIED because it produces temperatures in the range from 700°C to 1800°C. Large quantities of acetylene indicate HIED. One of the evaluation criteria recommended in IEEEC57-104 (1991) is the production rate per day of dissolved gases, where a rate greater than $0.1 ft^3/day$ or $2.832 \times 10^7 mm^3/day$ of total dissolved combustible gas indicates an active internal fault.

A disadvantage of using any of the ratios methods is that there are conditions when the combination of ratios yields a fault decision without indicating the cause of the fault. Another concern is that the decision flow process used in conjunction with the ratios methods is complicated. Appendix C shows the process diagram from IEEEC57-104 (1991).

The rate of gas production is calculated as shown in Equation 2.2 (IEEEC57-104 1991). The imperial unit method of calculating the rate of change is included in Appendix C

$$R = \frac{(S_T - S_o)(\frac{231}{16.39 \times 10^3})V \times 10^{-6}}{T} \tag{2.2}$$

where R is the rate of production of total dissolved combustible gas (mm^3/day), S_o is the concentration of the sample taken first (ppm), S_T is the concentration of the sample taken second (ppm), V is the volume of oil in the bushing (mm^3), T is the time in days. Alternatively (Dalton et al. 2005):

$$R = \frac{(y_2 - y_1).m}{\rho.(d_2 - d_1)} \tag{2.3}$$

Table 2.2: Translation of DGA gas ratios - Part II
(IEC60599 1999) and (IEEEC57-104 1991)

	IEC 60599		IEEEC57-104			
			Doernenburg		Rogers Ratio	
Gases	Normal limit		Normal	Faulty	Normal limit	
	Ratio	Cause	Ratio	Cause	Ratio	Cause
$\frac{C_2H_2}{C_2H_2}$	> 0.2	D	< 0.75	TD	< 0.1	Normal
	0.6 to	D2	> 0.75	HIED	< 0.1	LIED
	2.5	D1			0.1 to 3	HIED
	> 1	T			< 0.1	TD
	< 0.2	T2			< 0.1	TD$<$ 700° C
	< 0.1	T3			< 0.1	$TD >$ 700° C
	< 0.2					
$\frac{CH_4}{H_2}$	< 0.1	PD	> 0.1	TD	$> 0.1, < 1.0$	Normal
	0.1 to	D1	< 0.1	LIED	< 0.1	LIED
	0.5	D2	$> 0.1, < 1$	HIED	0.1 to 1.0	HIED
	0.1 to 1.0	T1			$> 0.1, < 1.0$	TD
	> 1	T2			> 0.1	TD$<$ 700° C
	> 1	T3			> 0.1	TD$>$ 700° C
	> 1					
$\frac{C_2H_4}{C_2H_6}$	< 0.2	PD			< 1.0	Normal
	> 1	D1			< 0.1	LIED
	> 2	D2			> 3.0	HIED
	< 1	T1			1.0 to 3.0	TD
	1 to 4	T2			1.0 to 3.0	TD$<$ 700° C
	> 4	T3			> 3.0	TD$>$ 700° C
$\frac{C_2H_2}{CH_4}$			< 0.3	TD		
			< 0.3	LIED		
			< 0.4	HIED		
$\frac{C_2H_6}{C_2H_2}$			> 0.4	TD		
			> 0.4	LIED		
			< 0.4	HIED		
$\frac{O_2}{N_2}$	< 0.3	T3				
$\frac{CO_2}{CO}$	< 3	T2	< 10	TD		

where y_1 is the reference analysis, y_2 is the last analysis, $(y_2 - y_1)$ is the increase in *microlitre* per *litre*, m is the mass of oil, in *kilograms*, ρ is the mass density in *kilograms* per cubic *metre*, d_1 is the date for y_1 and d_2 is the date for y_2.

A rate of change of more than 10% per month indicates an active fault, while a change of 50% per week suggests a serious fault (IEC60599 1999). All concentration values are at least 10 times the detection limit specified in (IEC60567 1992). Because under these conditions DGA values have a precision of 5% gas ratios have a precision of up to 10%, whereas, if the gas values are less than 10 times the detection limit, then the precision on DGA values decreases rapidly, to approximately 20% at 5× detection limit and up to 40% on a gas ratio (IEC60599 1999).

IEEEc57-104 also defines limits for the individual combustible gases and relate these to an expected condition of insulation oil and paper. Table 2.3 shows some of the safe limits for operating oil filled equipment including bushings, tap changers, transformers and instrument transformers. Conditions 1 to 4 are criteria developed to classify risks to transformers with no previous record of DGA. The same criteria are used for bushings, instrument transformers and oil filled switch gear. The only difference is that the limiting values of the dissolved gases are reduced for bushings in comparison to transformers for the same fault condition. Condition 1 (C1): TDCG below this level indicates the bushing is operating satisfactorily. Any individual combustible gas exceeding specified levels should prompt additional investigation. Condition 2 (C2): TDCG within this range indicates greater than normal combustible gas level. Any individual combustible gas exceeding specified levels should prompt additional investigation. Action should be taken to establish a trend. Faults may be present. Condition 3 (C3): TDCG within this range indicates a high level of decomposition. Any individual combustible gas exceeding specified levels should prompt additional investigation. Immediate action should be taken to establish a trend. Faults are probably present. Condition 4 (C4): TDCG within this range indicates excessive decomposition. Continued operation could result in failure of the bushing. Recommended actions based on the condition and total dissolved combustible gases are given in the specifications (IEEEC57-104 1991), (Hubacher 1976).

Other methods of analysis include those specified in IEC60628 (1985) which calculate the gassing tendency (G) of insulating oil at a temperature of 80°C, in the presence of hydrogen. The gassing tendency is calculated as follows (IEC60628 1985):

$$G = \frac{K.(B_{130} \; or \; B_{110} - B_{10})}{t} \tag{2.4}$$

Table 2.3: Concentrations of gases and the CSUS and IEEEc57-104 interpretation

	California State University Sacramento Guideline			IEEE C57-104			
	Normal	**Elevated**	**Abnormal**	C1	C2	C3	C4
Dissolved key gas concentration limits (ppm)							
Hydrogen(H_2)	< 150	> 150, < 1000	> 1000	< 100	> 100, < 700	> 700, < 1800	> 1800
Methane(CH_4)	< 25	> 25,	> 80	< 120	> 120, < 400	> 400, < 1000	> 1000
Ethane(C_2H_6)	< 10	> 10, < 35	> 35	< 35	> 35, < 50	> 51, < 80	> 80
Ethylene(C_2H_4)	< 20	> 20, < 150	> 150	< 50	> 50, < 100	> 100, < 200	> 200
Acetylene(C_2H_2)	< 15	> 15,	> 70	< 65	> 65, < 100	> 100, < 150	> 150
Carbon(CO) Monoxide	< 500	> 500, < 1000	> 1000	< 350	> 350, < 570	> 570, < 1400	> 1400
Carbon(CO_2) Dioxide	< 10000	> 10000, < 15000	> 15000	< 2500	> 2500, < 4000	> 2500, < 10000	> 10000
Nitrogen *(N_2)	> 10000		< 10000				

where G is the gassing tendency ($mm^3/minute$), B_{130} or B_{110} is the burette reading, in millimetres after 130 or 110 minutes of test, t is the test time in minutes, i.e. 130 -10 = 120 min if 50 Hz or 110-10 = 100 min if 60 Hz, B_{10} is burette reading, in millimetres, after 10 minutes of testing and K is the burette constant in mm^3 per millimetre burette reading, resulting in the value of G being positive if gas is evolved and negative if gas is absorbed.

Note: In Table 2.3 * refers to Dalton et al. (2005).

An alternative method of calculating gassing in the presence of nitrogen is shown in Equation 2.5 (IEC60628 1985).

$$G = \frac{p.(a-b)}{101.3} \tag{2.5}$$

where G is the gassing tendency ($cm^3/minute$), a is the burette reading at the

beginning of the test (cm^3), b is the burette reading at the end of the test (cm^3) and p is the barometric pressure (kPa).

Water contamination detection

Water causes the formation of water trees in solid insulation. Water trees develop into electrical trees along which partial discharge activity occurs until the component arcs and complete failure occurs (Dissado et al. 1999). In oil insulation, water contamination decreases the dielectric strength of oil and can ultimately lead to failure of the bushing. A standard way of measuring water content in oil-impregnated paper is by using the coulometric Karl Fischer titration method (IEC60814 1997). In this method the original reagents used are pyridine and methanol, and the reactions may be expressed as follows (IEC60814 1997):

$$H_2O + I_2 + SO_2 + 3C_5H_5N \rightarrow 2C_5H_5N.HI + C_5H_5N.SO_3 \tag{2.6}$$

$$C_5H_5NSO_3 + CH_3OH \rightarrow C_5H_5NH.SO_4.CH_3 \tag{2.7}$$

In coulometric Karl Fischer titration, the sample is mixed with a base solution of iodide ion and sulphur dioxide. Iodine is generated electrolytically and reacts with water in a similar way to that shown in the reactions Equations 2.6 and 2.7. Iodine is generated in proportion to the quantity of electricity according to Faraday's law, as shown by the following reaction: (IEC60814 1997)

$$2I^- - 2e \rightarrow I_2 \tag{2.8}$$

One mole of iodine reacts with one mole of water stoichiometrically as shown in reaction of Equation 2.8 so that 1 mg of water is equivalent to 10.72 Coulombs. Based on this principle it is possible to determine the amount of water directly from the number of Coulombs required for the electrolysis.

Meaning of gas content and water content

Standard characteristics required of bushing oil are detailed in IEC60867 (1986). Criteria for interpreting the condition of bushings and transformers using gases dissolved in mineral insulating oil are detailed in IEC60599 (1999) and IEEEC57-104 (1991). Dissolved gas analysis (DGA) provides extremely useful information on the condition of bushings, because the oil taken from the bushing gives information about the nature, causes and location of the defect, i.e. whether the defect is in oil or in paper insulation. DGA also tells the user the extent of the defect and degradation. Van Wyk (1997) suggests that DGA is superior to partial discharge as a condition monitoring tool, because PD normally verifies what DGA has already found. High levels of dissolved gases in oil within bushings can indicate insulation fatigue or other developing defects, like arcing, sparking due to bad electrical connection, arcing ground connections, arcing across oil barriers, arcing inside paper barriers, or arcing inside the bushing insulation. Loose high voltage connections and internal surface discharges at the bottom of a bushing also generate dissolved gases.

There are three widely accepted methods of gas analysis (DiGiorgio 2005), namely, total combustible gases (TCG) method, detection and identification of individual fault gases in the gas blanket above the oil method, and lastly dissolved gas analysis (DGA). The first method determines the total combustible gases that are present in the gas above the oil, individual gases are not identified. TCG method gives only a single value for the percentage of combustible gases, so one cannot diagnose the cause of the fault. The disadvantage of the first two methods is that most faults occur under the surface of the oil, the gases must first saturate the oil and diffuse to the surface before accumulating in the gas blanket above the oil. The advantage of DGA is that it detects the gases while they are in the oil, giving the earliest possible detection. Another advantage is that DGA identifies the individual gases so that the possible cause of the fault can be estimated.

It is important to note that the extent to which the gases dissolve in oil is dependent on the pressure and temperature inside the oil. All values used as limits are quoted at standard temperature and pressure, i.e. 101kPa and 25°C. Figure 2.5 shows the gas solubility in oil as a function of temperature. The more soluble the gas, the larger will be the volume of gas found in oil. Degradation indicators are commonly clustered into three groups, namely, paper degradation indicators, water degradation indicators and other indicators.

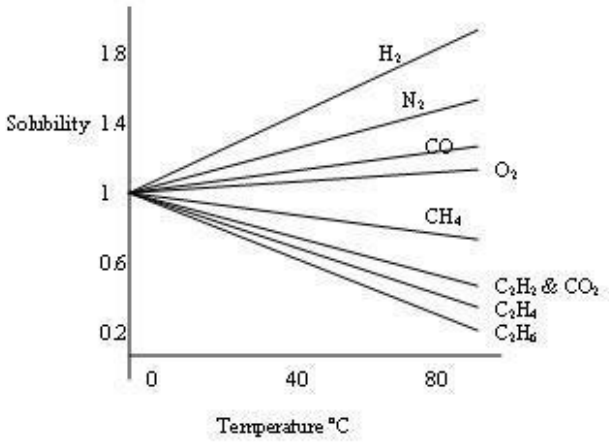

Figure 2.5: Approximate solubility against temperature at 101kPa (DiGiorgio 2005)

Paper degradation indicators

The cellulose in the paper breaks down over time and releases glucose, carbon monoxide, carbon dioxide, water, and acids. The glucose further breaks down into furans. Heat, water and oxygen are the primary causes of paper degradation. Evidence of combustion is the accumulation of carbon on the paper surface due to tracking, shown in Figure 2.6. Tracking is the process that produces a partially conducting path of localized deterioration on the surface of an insulating material as a result of the action of electric discharges on or close to an insulation surface. Damaged paper does not regenerate; it must be replaced by rewinding the bushing. Damaged oil can simply be drained and replaced at a lower cost than paper replacement.

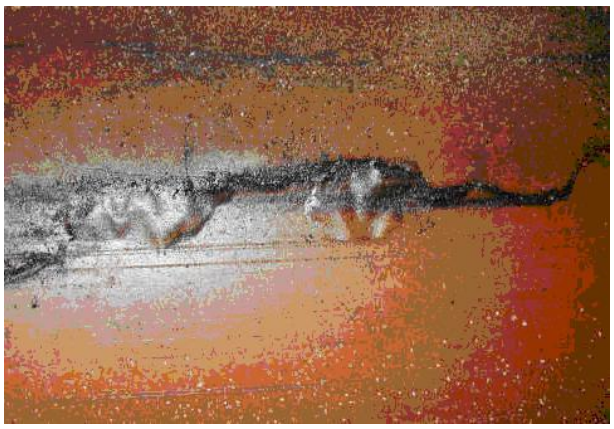

Figure 2.6: Carbonisation due to tracking on bushing paper insulation

Oil degradation indicators

The quantity and types of gases reflect the nature and extent of the stressing mechanism in the bushing (Lord & Hodge 2003). Oil breakdown is shown by the presence of Hydrogen (H_2), Methane (CH), Ethane (C_2H_6), Ethylene (C_2H_4), and Acetylene (C_2H_2). High levels of only hydrogen show that the degeneration mechanism is due to partial discharge at temperatures below 300°C. High levels of hydrogen and acetylene show that the degeneration mechanism is due to arcing at temperatures above 700°C. High levels of hydrogen methane, ethane and ethylene show that the degeneration mechanism is due to thermal faults at temperatures between 300°C and 700°C.

Other indicators

A high moisture content, i.e. more than 23ppm (Lau 2003), would also warrant the removal of the bushing from service because it can initiate PD activity inside the bushing and result in water treeing and electrical tree formation and growth. The ambient temperature together with the transformer loading will affect the normal operating temperature of the bushing. Elevated temperatures will increase the rate of degradation of the bushing insulation. The pressure of the bushing may reflect the extent of ingress of air and moisture into the bushing if the bushing pressure is less than the atmospheric pressure even if no leaks are visibly detectable.

Commercial tools for detection and measurement

Commercially available systems for bushing condition monitoring systems that are presented by de Klerk (2005) provide trending functions, with limited learning capability. Trending tools are referred to as expert systems, a grouping that is not included in the definition of artificial intelligence.

2.3.2 Electrical condition assessment methods

Measurable electric parameters are electric field, magnetic field, resistance, current, a.c. voltage, voltage impulses and partial discharge activity. If the magnitude of these variables under normal operating conditions is known, then changes can be monitored to give early warning of problems in the bushing

Electric field measurement

Three popular types of instrumentation for measuring electric field strength work on one of three principles, namely (IEC60833 1987): (1) the free-body principles; (2) the ground-reference principles and;(3) the electro-optic principles. Electric field strength meters have two parts, a probe or field sensor and a detector. The detector consists of signal-processing circuitry with a display. Important considerations on the conditions of use of electric field measurement instruments are given in (IEC60833 1987). Appendix E has images of some of the electric field sensors. Electric field measurement should take into account errors due to distortion of the electric field which are not accounted for in the model assumptions, which are used when designing the measuring equipment. A study by Dhlamini, Crowdy & van Coller (2000) which was based on a 132kV cyclo-alyphatic epoxy resin-silicon rubber (CER-SR) bushing, found several sources of error in finite element models used to calculate electric field strength at different points in a bushing. In a separate study, (Markey & Stevens 2003) found that voids with diameters in the range of 40nm to 50nm concentrated near the screen of XLPE cable are normal in new insulation. Failure is highly probable if voids of 1µm diameter in clusters of $100/mm^3$ (Markey & Stevens 2003) or conductive particles $\geq 100\mu m$ diameter occur inside solid insulation (Zheng & Boggs 2005). Volume ratio and area ratio based on void density are given by Equations 2.10 and 2.11 respectively.

Error 1

To model $1\mu m$ objects the drawing will be in the hundreds of nanometre domain. The resolution of the FEM software limits the size of objects to the *smallest element* : $total\ area = 1{:}10^3$, because the size of the molecules in the insulation is more than 1000 times bigger than the voids. This will automatically make any model with void defects inaccurate. Calculations shown in equations 2.9 to 2.11 show the size of the errors (Dhlamini et al. 2000).

$$Ratio_V = \frac{Volume\ of\ Voids}{Volume\ of\ Insulation} = 5.24 \times 10^{-8} \tag{2.9}$$

$$Ratio_A = \frac{Area\ of\ Voids}{Area\ of\ Insulation} = 7.85 \times 10^{-5} \tag{2.10}$$

$$Ratio_{limit} = \frac{smallest\ element}{largest\ element} \geq 10^{-3} \tag{2.11}$$

Error 2

Model scale nanometre and micrometre, but simulator has millimetres as highest resolution. Calculations of dielectric strength for materials used in the tools are for samples thicker than 1 mm. Electrical behaviour of 1mm and 1μm samples is not the same under the same voltage stress.

Error 3

The size of triangles used to calculate energy and the electric field in the material is not uniform. The distribution of triangles inside the dielectric is also not uniform, as shown in Figure 2.7 .

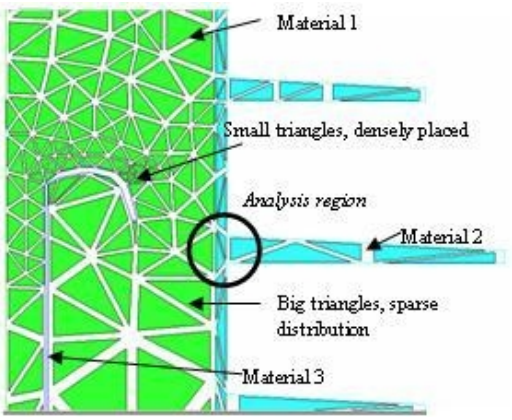

Figure 2.7: FEM triangles used to calculate electric field

The finite element model (FEM) and boundary element model (BEM) that were tested calculate accurate values of the electric field (E) based on the value of the voltage (V). V is calculated from the Equation 2.12 (Kraus 1991).

$$V_{21} = \frac{\lambda}{2\pi\varepsilon_r}.ln\frac{r_2}{r_1} \tag{2.12}$$

V_{21} is the peak voltage difference between point 1 and 2 ($r_2 > r_1$) (V_{peak}), and λ is the linear charge density (C/m), ε_r is the relative permitivity of the dielectric.

In any simulation of the bushing, the voltage of the conductor relative to the ground is entered together with the voltage of at least one other point, e.g. the outer screen at 0V. The distances (r) between different points in the simulated area are known because the dimensions of the drawing are entered. So λ can be calculated for any point (triangle apex) and then V can be calculated backwards from the above equation until the voltage is known for the entire space, i.e. at the apexes of all the triangles.

The electric field is calculated from the energy equation that relates the voltage to the electric field. Energy (u) stored by a capacitor is given by Equation 2.13 (Sears, Zemansky & Young 1987).

$$u = (1/2).\varepsilon_r\varepsilon_o Ad(\frac{V}{d})^2 = (\frac{1}{2}).\varepsilon_r\varepsilon_o AdE^2 \tag{2.13}$$

Using the triangles where the voltage is known at each of the 3 apexes, the energy of each triangle is determined using the average voltage of the three points. The voltage is divided by the perpendicular distance (d) from one side of the equilateral triangle to the apex. The result is squared and then multiplied by the same distance. This result is then multiplied by the area of the triangle (A), multiplied by the relative permittivity of the triangle material (ε_r), multiplied by ε_o , then divided by 2. The energy value of the individual triangle is stored. The total energy of a simulated region is calculated by adding the energy of each triangle, and then that total value is stored. The number of triangles is increased and the energy of each triangle is calculated and stored as before. The total energy is calculated again (u_{new}) and then compared to the previous value (u_{old}). Each time a new total energy

Table 2.4: Convergence data

Pass	Triangles	Total Energy (J)	Energy Error(%)
1	22	5.60150×10^{-1}	11.2659
2	65	4.03541×10^{-1}	1.9461
3	80	3.97800×10^{-1}	1.5381
4	103	3.92662×10^{-1}	0.9628
5	142	3.89000×10^{-1}	0.6263

value is calculated, this is defined as a pass in the FEM. A percentage difference is then calculated as shown in Equation 2.14.

$$\%\ Energy\ error = \frac{u_{new} - u_{old}}{u_{old}} \times 100\% \tag{2.14}$$

The percentage energy error in Table 2.4 is the mean of all the errors in all the triangles. If the percentage energy error after a pass is closer to the target error than the previous percentage energy error, then FEM will continue to increase the number of triangles in the simulated area until the selected number of passes is completed or the error is less than the target, e.g. 1%. If the opposite happens, and the percentage energy error after a pass is further from the target error than the previous percentage energy error, i.e. a bigger percentage energy error, then FEM will increase the number of triangles or will stop if the maximum number of passes is exceeded even if the error is not less than the target error, indicating no convergence.

If a small section of the simulated area calculates an incorrect energy value, this error is not detected if most of the balance of the simulated area is calculated correctly. Such an error in a small section results when an area of one material has more triangles than an equivalent area elsewhere in the same material. For the same average voltage in individual triangles, the section with small triangles will have a larger $(V/d)^2$ term and therefore a higher energy value (u) than that with bigger triangles. This means that the energy calculated in a region with more triangles will appear to be higher than the energy where there are fewer triangles. So, to get an accurate value of energy (u) in a particular material there must be a uniform distribution of triangles throughout that material, i.e. the number of size and the number of triangles per unit area must be constant.

The triangles along the sheath in Figure 2.7 are approximately the same size. Figure 2.12 shows the high values in electric field at the sheaths of the insulation. At the corners where the sheds are attached to the sheath, and along the screen curve, the triangles become very small. The area was measured to be $(6.5mm)^2 = 42.25mm^2$. The triangles in the corner area are very much smaller than the triangles elsewhere on the sheath, so the distance used to calculate energy is much smaller in the corners than along the sheath, in fact $\frac{d_{sheath}}{d_{corner}} >> 10^3$. This error in the energy value is reflected in the electric field values calculated from it. The fact that $\frac{d_{sheath}}{d_{corner}} >> 10^3$, means that the values calculated by the FEM in these different regions cannot be compared accurately. This is because the FEM can only compare lengths that do not differ by an order of magnitude greater than 10^3. The error in the corner area is thus a cumulative error of the two sources of inaccuracy stated above.

Impulses

Even though the frequency is very low during the life of a bushing, it will experience voltage or current impulses due to lighting impulse or switching. The impact of these surges is significant. A lightning impulse can by itself destroy the insulation of the bushing or leave the bushing with localised damage that results in partial discharge under normal power frequency operating voltage. IEC60897 (1987) is based on the work of Wald (1947) who explains in detail that the concepts of probability of lightning induced insulation failure. The likely time when liquid insulation will fail when subjected to lightning impulses is given by the Equation 2.15 (Wald 1947).

$$L_1 = \left[\begin{array}{c} n \\ d \end{array} \right] P_1^{d_n} (1 - P_1)^{(n-d_n)} \tag{2.15}$$

where L_1 is the probability of obtaining a set of observed values with $P = P_1$, P is the breakdown probability under standard lightning impulse of a dielectric liquid for a given value of the test voltage, P_o is an arbitrary fixed breakdown probability which may be related to accepted quality criteria, P_1 is the breakdown probability lower than P_o, the value of which is chosen to take into account the precision of the test voltage applied, n is the number of impulses applied and d_n is the number of breakdowns observed after n impulses have been applied. What this means is that the more impulses the insulation experiences, the higher is the probability of failure.

Partial discharge

Insulating oil inside the bushing must withstand very high stress to avoid the inception of partial discharges (PD). Partial discharges are micro sparks that occur at stress enhancement points such as voids in electrical insulation, ends of the screens and at multi-dielectric material interfaces. Table 2.5 shows the factors of field enhancement under various conditions as discussed by Nyamupangedengu & Reynders (2003).

Table 2.5: Five stress enhancement factors

Void Defect Type	Stress Enhancement Factor
Long Flat perpendicular to E	ε
Spherical	$3\varepsilon/1+2\varepsilon$
Cylindrical	$2\varepsilon/1+2\varepsilon$
Long and parallel to the E	1
Long flat and perpendicular to E	ε

Here ε is the relative permittivity of the insulation and E is the electric field in the insulation.

The classical PD model is shown in Figure 2.8 (Zheng & Boggs 2005), the voltage across the defect capacitance is given by the equation below (Kuffel & Zaengl 1984):

$$v_c = v_a \varepsilon_r \frac{t}{d} \tag{2.16}$$

where t, the cavity width, is such that $t << d$ and d is the insulation thickness, v_a is the applied voltage on the electrodes and ε_r is the insulation dielectric strength. Partial discharges emit electromagnetic radiation up to greater than 500MHz (Orton 2002).

In Figure 2.8, the discharge across the cavity is equivalent to short circuiting the capacitance c. The total capacitance across the insulation increases, resulting in a corresponding fall of the voltage across the electrodes. The difference between the voltages on a and that on the electrodes causes charge to flow. The apparent charge transferred to the electrode upon initial discharge can be approximated by Equation 2.17 (Orton 2002):

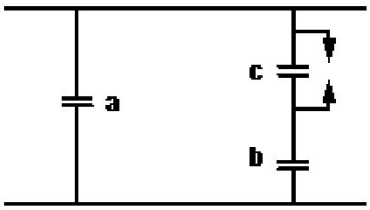

Figure 2.8: abc Analogue partial discharge model

$$Q = \frac{1.64 \times 10^{-8}.a^{5/2}.p^{1/2}.\varepsilon_r}{rln(\frac{R_o}{R_i})} \tag{2.17}$$

Where Q is the charge in Coulombs, r is the radial position of the cavity (m), p is the pressure of the air in the cavity (Pa), ε_r is the dielectric constant, R_i is the radius of the conductor-semiconductor interface (m), R_0 is the radius of the dielectric and ground-semiconductor interface and a is the cavity radius (m).

A time lag exists between the application of inception voltage and occurrence of the discharge (Orton 2002). The time lag comprises of a statistical time t_s and formative time t_f . The former is the time it takes for a seed electron to appear in the cavity and the latter is the time required for the breakdown to develop once the process has been initiated. Parameter t_f depends on the mechanism of avalanches in the discharge process. Parameter t_s can be up to 5 seconds in virgin defects (Nasser 1991) as it is dependent on the statistical probability of pre-ionisation in the cavity. Consequently PD inception would be at a higher voltage than the theoretically predicted value (Bartnikas & Novak 1993).

When the electric field in the defect reaches the breakdown threshold, and if a seed electron is available at an appropriate position, the void gas ionisation process is initiated. The ionisation process is initiated through Townsend primary and secondary avalanches (Kuffel & Zaengl 1984). The threshold is either in terms of the number of electrons n_e where $10^8 \lg n_e < 10^9$ as determined by Penning (Morhuis & Kreuger 1990) or in terms of the electric field in the cavity when $E_c = KE$ and $K = 0.1 \rightarrow 1$ according to Meek (Kuffel & Zaengl 1984). E is the electric field in the insulation and E_c is the electric field due to the space charge behind the avalanche head given by Equation 2.18 (Boggs et al. 1987):

$$E_s = \frac{5.3 \times 10^{-7}.\alpha.e^{\alpha x}}{(x/p)^{0.5}} \tag{2.18}$$

α is the first Townsend coefficient, x is the distance (cm) the avalanche has progressed and p is the gas pressure ($Torr$). On a 50Hz a.c. voltage waveform a partial discharge would be indicated by a partial collapse in the voltage at the PD inception voltage (V_{inc}), on both the positive and negative cycles, shown in Figure 2.9.

The flow of charge during a PD has a duration ranging from a few nanoseconds to microseconds (Zheng & Boggs 2005). If the voltage across the void is maintained above the PD inception voltage even after the PD event, other PDs will happen until the voltage drops below the PD inception voltage. This is the reason for the multitude of pulses in the voltage waveform, superimposed on the 50Hz frequency.

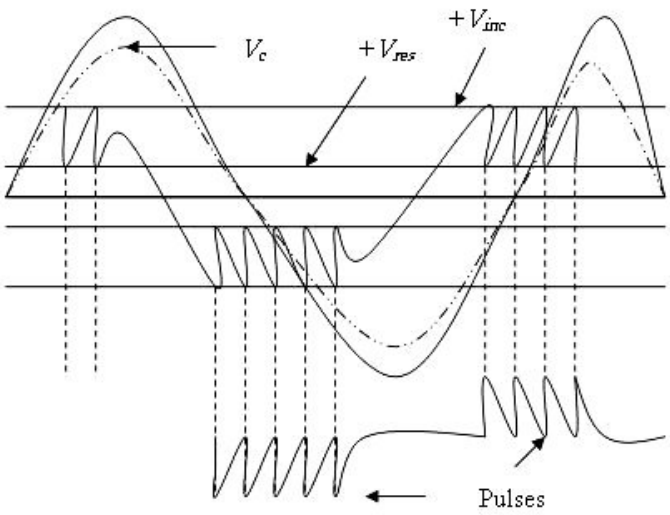

Figure 2.9: Repetitive partial discharge under a.c. voltage
(Zheng & Boggs 2005)

According to IEC61294 (1993) partial discharge inception voltage test results for oil are more important than electric field strength or lightning impulse breakdown voltage tests (IEC60897 1987). The IEC60833 (1987) standard observes no advantage in tests made at higher sensitivity than $100pC$ compared to those made at $100pC$.

Resistance monitoring

The leakage current (I_r) across measured on the bushing is the sum of two components, a surface component (I_s) and a volumetric component (I_v). Measurement of the total leakage current under alternating current fields is an indicator of the condition of the insulation. Instruments that measure the leakage current normally express it as a ratio of the leakage current (I_r) divided by the capacitive current (I_c) through the same insulation. This ratio is called the loss factor or dissipation factor or tan delta (tanδ). Figure 2.10 shows the equivalent circuit for a bushing. The relationship between the dielectric dissipation factor (tanδ), the conductivity (σ) and the permittivity (ε) of a liquid with negligible dipolar losses is given by Equation 2.19. Insulating oil is an example of insulation that has negligible dipolar losses (IEC61620 1998).

$$tan\delta = \frac{\sigma}{\varepsilon.\omega} \tag{2.19}$$

where ω= $2\pi f$ and f is the frequency of the voltage (Hz) and σ is the conductivity. The unit of ω is radians per second. For liquid insulation σ is a characteristic measured at thermodynamic equilibrium. To fulfill this requirement, there should be no high electric stress and/or prolonged voltage applied to the liquid at the time of measurement.

According to IEC60247 (1978) older methods gave inaccurate results, because thermodynamic equilibrium was not maintained during the measurements of the d.c. resistivity (ρ). While performing measurements of resistivity in accordance with IEC60247 (1978), electric stress up to $250V.mm^{-1}$ is permitted for 1 minute. By applying an electric field to the sample, one would excite the electrons and increase the temperature of the sample. The extent of temperature increase in relation to applied voltage is unknown. Halliday et al. (1988) state that the variation of resistivity with temperature is given by Equation 2.20. The result is that the tanδ calculated from that resistivity value is not the same as that calculated in IEC61620 (1998) (Halliday, Resnick & Merrill 1988).

$$\rho - \rho_o = \rho_o.\alpha.(T - T_o) \tag{2.20}$$

where T_o is a selected reference temperature (K), ρ_o is the resistivity at that reference

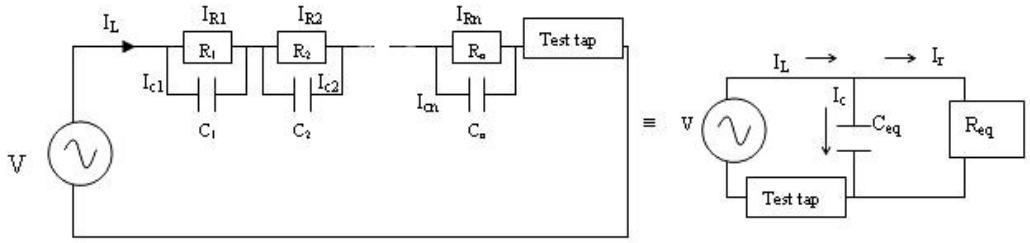

Figure 2.10: Electrical equivalent circuit of bushing

temperature $(\Omega.m)$, α is the thermal coefficient of resistivity $(K-1)$, T is the temperature at which resistivity is measured or calculated and ρ is the resistivity at that arbitrary temperature $(\Omega.m)$. The resistance is calculated from the resistivity as shown in Equation 2.21. Conductance (σ) is the inverse of resistivity (ρ) (Sears et al. 1987).

$$R = \rho.\frac{L}{A} \tag{2.21}$$

where L is the length of the conductive path (m) and A is the cross-sectional area of the conductive path (m^2). The unit for conductance is $(\Omega^{-1}m^{-1})$.

In Figure 2.10, V is the applied a.c. voltage (V), I_L is the leakage current (A), C_i is the capacitance between successive concentric cylinders (F) and R_i is the resistance of the insulation between successive concentric cylinders (Ω). The impedance (Z) of the circuit is given in Equation 2.22 (Sears et al. 1987).

$$Z = \sqrt{R_{eq}^2 + (-X_c)^2} \tag{2.22}$$

$$R_{eq} = \sum_{i=1}^{n} R_i \tag{2.23}$$

$$X_c = \frac{1}{\omega C_{eq}} \tag{2.24}$$

$$\frac{1}{C_{eq}} = \sum_{i=1}^{n} \frac{1}{C_i} \tag{2.25}$$

where R_{eq} is the total resistance (Ω), X_c is the total capacitive reactance (F) and

C_{eq} is the total of all the series capacitances (F). The voltage drop across the entire impedance is then $V_{drop} = I_L.Z$ (V), the voltage across an individual capacitor (C_i) is $V_{ci} = -I_{ci}.X_{ci}$ (V). The expression means that there is a voltage rise across each capacitor. This is the imaginary component of voltage. The voltage drop across an individual resistance is $V_{Ri} = I_{Ri}.R_i$ (V). The voltage drop across the resistor is the real component of the voltage drop. For the bushing the loss factor (tanδ), shown in the phasor diagram in Figure 2.11, is given by Equation 2.26 (Sears et al. 1987).

$$tan\delta = \frac{I_r}{I_c} = \frac{1}{R_{eq}\omega C_{eq}} \tag{2.26}$$

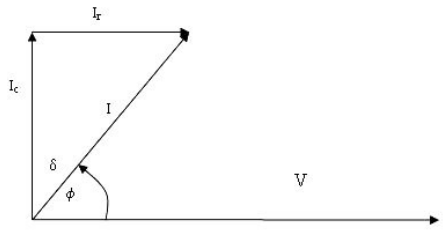

Figure 2.11: Phasor diagram showing tanδ

2.3.3 Mechanical condition assessment methods

All bushings have a solid external insulation of porcelain or glass or resin or silicon rubber. Externally a bushing resembles a cylinder with fins. The fins assist with cooling and electrical creepage. High voltage bushings have a condenser which consists of aluminium screens which are wrapped in paper and impregnated with oil or resin to provide even distribution of the electrical field radially inside the bushing and tangentially along the bushing surface (Dhlamini 2000). Figures 2.12 and 2.13 show an example of the electric field around a bushing. Inside the condenser is a cylindrical conductor. There are many types and applications of bushings; examples include those with resin impregnated paper solid condenser, and gas insulated condensers. IEC60137 (2003) lists and describes all the commercial types of bushings. Bushings experience mechanical shocks due to transportation, or earthquakes or electrical faults or impacts from birds landing on them. When the bushing is energised, with dislodged paper or foil in the condenser, partial discharges occur inside the bushing and short circuits between foils can also happen. When the latter happens the foils melt and a path for leakage current and circulating currents is created inside the bushing, resulting in the failure of the component. IEC61463 (1996) makes recommendations on specifications and tests for bushings that are subjected

to seismic incidents. In most cases the stiffness and mass of the bushing system are known. So the natural frequency can be determined for a healthy bushing.

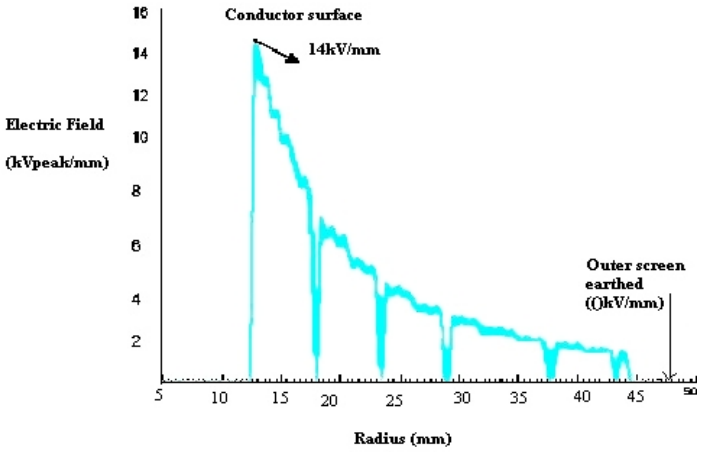

Figure 2.12: Radial electric field distribution inside a bushing

If one compares the healthy vibration signature of a bushing to one of an unhealthy bushing, changes in the signature can be used to determine the extent of degradation. Such a method would apply a similar principle to the frequency response analysis (FRA) model proposed by Pleite et al. (2003). Acoustics monitoring which includes ultra-high frequency partial discharge (UHF-PD) detection and audible frequency band operating noise monitoring are other methods of mechanical condition monitoring.

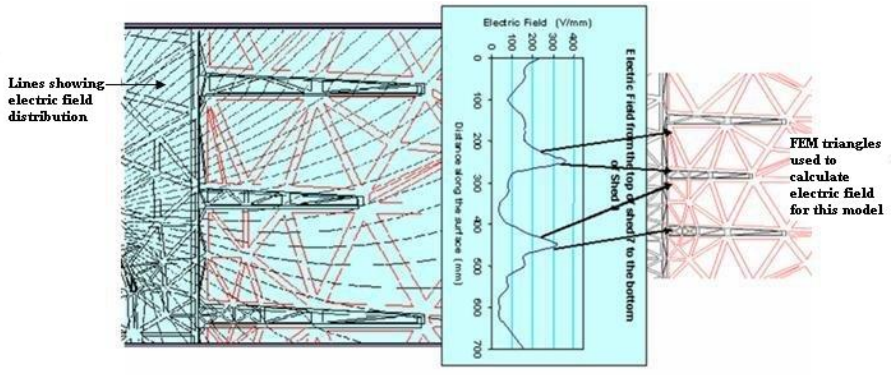

Figure 2.13: Electric field and FEM triangles around a bushing

2.3.4 Temperature and optical condition assessment methods

Pukel et al. (2006) found that the frequency band used to monitor transformers for optical and thermal indicators of failure is from $10^4 Hz$ to $10^{16} Hz$. The same band of the electromagnetic spectrum would be used to monitor oil impregnated paper bushings, because they operate under similar conditions within the same temperature limits as a power transformer. Figure 2.14 shows an example of a thermal-optical image taken using an infrared camera. It shows a bushing with a hot connection at the high voltage lead. The bushing may fail due to melting of the conductor at the poorly connected location, which generates heat. The camera can record temperatures up to $1500°C$, which is useful because a pure aluminium conductor melts at $660°C$ and copper melts at $1085°C$. Heat is transmitted through the conductor by means of conduction. Once the dielectric has been degraded by partial discharges and heating, circulating currents inside the bushing generate more heat which further destroys the insulation (IEC60599 1999).

2.3.5 Combination of condition assessment methods

In practice bushings are monitored using a combination of the methods mentioned. The cost of the method used depends on the importance of the bushing and availability of test equipment.

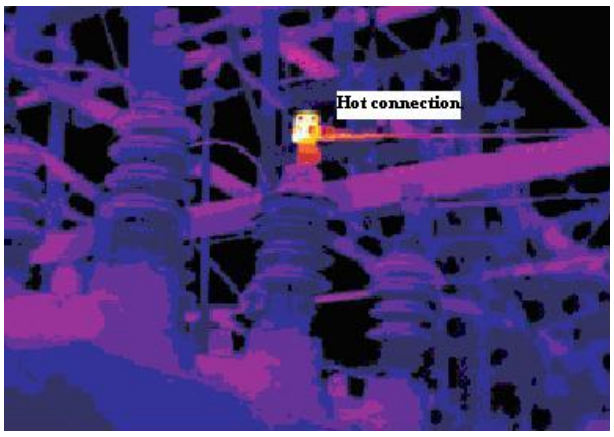

Figure 2.14: Thermal image of bushings in operation

2.4 SELECTION OF METHOD AND EVALUATION CRITERIA

After considering various commonly available diagnostics methods the author chose dissolved gas analysis as the condition monitoring tool for oil impregnated paper bushings. The decision to select dissolved gas analysis as the criteria on which to apply artificial intelligence is based on three factors.

(1) The progression of most failure mechanisms observed in bushings can be monitored using DGA tests before failure happens, so DGA is an effective non-destructive diagnostics tool.

(2) The electricity industry in South Africa uses DGA to assess bushings, so historical data is available to test ANN.

(3) There are many well established international standards that support interpretation on DGA test results.

The interpretation criteria used to evaluate the condition of each bushing was IEC60599 (1999), IEEEC57-104 (1991) for total combustible gases and California State University Sacramento (CSUS) Guidelines for individual combustible gas content (Hubacher 1976). The electricity industry in South Africa uses these standards to decide on the condition of its bushings (Dalton et al. 2005). The IEC60599 guidelines are more conservative than CSUS and IEEE C57.104, so a decision based on this standard promotes plant safety, by alarming when gassing reaches the lower limits of dangerous levels. The CSUS includes all gases classified as well as N_2 and O_2 which are not included in IEC60599 (1999) and IEEEC57-104 (1991),standards. CSUS guidelines form a subs-category within the specifications of IEC60599 (1999) and IEEEC57-104 (1991). All the models are suitable where no historical data for rates of change is available, as was the case for the bushing.

CSUS guidelines were developed based on the work done by Hubacher (1976) . IEC60599 (1999) is based on the work of Doernenberg & Gerber (1967) and Duval (1974). IEEEC57-104 (1991) is based on the work of Doernenberg & Gerber (1967), Hubacher (1976) and Rogers (1975).

The science of DGA interpretation using the limits in the standards is not exact. Discrepancies in the DGA limits for diagnosing using different gases and gas ratios

are due to variations in bushing design, the applied voltage, the ambient temperature, loading condition, age of equipment, ratio of oil to paper, whether the bushing is applied on a bolted or welded type sealed transformer, whether the oil is taken from a power transformer, instrument transformer, bushing or tap changer, whether the bushing contains only mineral oil or mixtures of mineral oil and dodecylbenzene, the ratio of the mixture, as well as many other field conditions. The utilities that developed IEC60599 (1999) found that the limiting values are lower on bushings than on transformers and an example is shown in Table 2.6.

Table 2.6: DGA limits for different equipment

Equipment	H_2	CO	CO_2	CH_4	C_2H_6	C_2H_4	C_2H_2
Bushing	140	1000	3400	40	70	30	2
Transformer	100	200	5000	50	50	50	5
Transformer*	150	850	1200	130	70	250	270

Note: In Table 2.6 * denotes transformers with separate oil tap changer

2.5 FAILURE PROCESS OF OIL IMPREGNATED PAPER BUSHINGS

The initial cause of breakdown could be by a lightning surge, or fault that distorts the condenser, so that under operating voltage conditions partial discharges occur inside the bushing. During degradation the oil breaks down into combustible gases which dissolve in oil. Paper breaks down into carbon, water vapour, carbon monoxide and carbon dioxide. Carbonised surfaces on the paper insulation together with the decreased resistivity of the oil dielectric allow a steady leakage current to flow from the conductor, through the oil, screens and paper to the earthed tank. Current will heat up the oil and paper. This defect will be sustained in a bushing until maintenance personnel detect the defect by routine maintenance or until a person is injured by leakage current or until the sensitive earth fault protection detects the leakage current and the equipment trips. During the degradation process combustible gases will accumulate inside the oil. If a bushing is on a transformer with sensitive earth fault protection, after tripping, the auto-recloser will re-energise the transformer after a few seconds. A high current arc, approximately 20 times the rated current, will pass through the bushing from the conductor to the tank. The arc will heat up the oil, creating more gas, and cause the already present gases to

expand and in some cases to ignite. This will cause the pressure inside the bushing to exceed the tensile strength of porcelain, and the porcelain shell will explode. The oil from the transformer will leak out through the broken bushing, and may be ignited by the burning oil form the bushing. The transformer will then trip on earth fault protection and the fire will burn until there is no more fuel or until it is extinguished.

2.6 CONCLUSIONS

After considering various commonly available measurement and diagnostics methods used to monitor bushings the author chose to apply dissolved gas analysis as the criteria for condition assessment. The interpretation criteria used to evaluate the condition of each bushing was IEC60599 (1999), IEEEC57-104 (1991) for total combustible gases and California State University Sacramento Guidelines (CSUS) for individual combustible gas content. The IEC60599 guidelines are more conservative than California State University Sacramento Guidelines (CSUS) and IEEE C57-104 for bushing diagnosis, so a decision based on this standard ensures that plant is operated safely. However, CSUS is more conservative than the other standards for diagnosis of other oil filled equipment, such as transformers and switchgear. The CSUS specification includes all the gases classified by both IEC60599 and IEEE C57-104, as well as N_2 and O_2 that are not included in each of the other standards.

Chapter 3

NEURAL NETWORKS AND SUPPORT VECTOR MACHINES

3.1 INTRODUCTION

Two artificial neural network (ANN) architectures and a kernel machine are compared in this chapter. The ANNs used are radial basis function (RBF) and multilayer perceptron (MLP). The kernel method used is called the support vector machine (SVM). The analysis in this chapter seeks only to determine the best performing neural network or kernel machine in terms of speed, accuracy, and stability on bushing condition monitoring. This section does not seek to identify suitable criteria to use for DGA with artificial intelligence. This is done in the next chapter. In this work 60699 bushings were classified based on ten criteria. The classification was done in two different ways. Firstly, stand-alone networks classified the bushings independently. Secondly, an election system was introduced, where the neural networks worked as a committee of classifiers, each classifying the bushings, and then electing the best classifiers to make a decision, on behalf of all the other networks. This section also compares different criteria that are commonly used to determine the condition of a bushing and diagnose the cause if a bushing is found to be defective. The methods are then used to develop a computational intelligence tool that is used to autonomously classify a bushing as defective or not, and then diagnose the cause of any defect.

3.2 FUNDAMENTALS OF ARTIFICIAL NEURAL NETWORKS

The basis of the neural network (NN) architecture is the neuron, which approximates the human brain neuron. If the input to the neuron is defined as x, then the neuron is an activation function $f(x)$. The extent to which each neuron reacts to an input depends on the weight of that input. The activation function or neuron is equivalent to transfer functions in control systems. The output (y) is related to the input by the activation function, as shown in Equation 3.1 (Hines 1997).

$$y = f\left(\sum_{k=1}^{n} x_k w_k + b\right) \tag{3.1}$$

where w_k represents the weights, the inputs are x_k, the activation function is f, b is the biases, and n is the number of neurons in the hidden layer.

The output is calculated by multiplying the input by a weight vector, adding a bias term, summing the results and applying an activation function to the sum (Hines 1997). A schematic of a single neuron and a neural network is shown in Figure 3.1. An artificial neural network (ANN) is a collection of parallel and series interconnected neurons with multiple inputs and one or many outputs. The neurons are classified into three groups called layers. Each neuron has an input layer, an activation function and an output layer.

3.3 WHY USE MLP, RBF and SVM?

ANNs and support vector machines are referred to as artificial intelligence because they can generalise. In other words they can solve unseen problems using training experience. Artificial neural networks are useful for classification and regression problems for which data is available. One advantage of ANN is that if the rules governing the interaction of different variables contained in a set of data are unknown, neural networks can be used to explain any interaction between the input variables by ranking the variables in order of importance to obtaining a desired solution. Also if the equations that represent the relationship between the input variables and the output are not known or are difficult to derive analytically, one can use ANN to determine a result for a set of input variables. Another advantage of neural

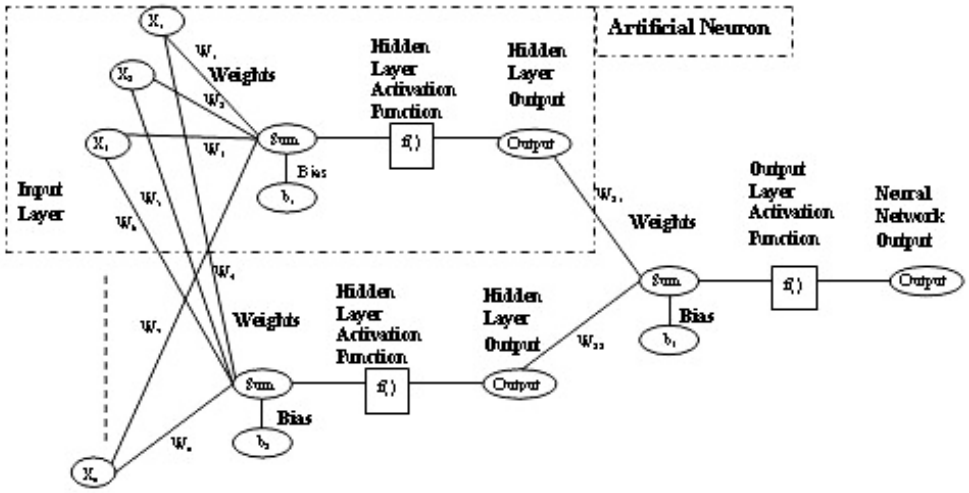

Figure 3.1: Hidden layer single output three neuron artificial neural network (Hines 1997)

network based diagnostics methods over other methods such as expert systems and fuzzy logic is that they learn from previous data and update their knowledge as they learn. Neural networks are also flexible because they allow the user to add inputs, as and when additional criteria become available from field personnel. The disadvantage of neural networks is that they do not work if there is no historical information of the problem. Another disadvantage is the danger that historical information used to train an ANN can be completely different from current information for the same component, so the user needs to ensure that the training data always reflects the current status of the problem. To compensate for the limitations of traditional neural networks, adaptive neural networks are trained using online. The data needed to train a neural network must be complete, i.e. each input must have a corresponding output, and all the input variables must be available for output.

Hines (1997) mentions eight types of artificial neural networks. (1) The MLP is one type of artificial neural network which belongs in a class of Hebbian networks (Hebb 1949). Hebbian learning is based on the postulate that to affect learning in connections between two neural cells, the strength of the connections are increased when both sides of the connection are active (Willshaw & von der Malsberg 1976). (2) Instar learning network is another type of network (Fiesler & Beale 1997). Instar has learning rate equal to forgetting rate. (3) The outstar network is another Hebbian type network with a memory (Bandemer & Gottwald 1990), (Palgrave 2000). (4) The crossbar structure network is another type of neural network (Pacheco 1997). (5) Competitive networks, are a type of neural network which includes those called

by the same name as well as self organising feature maps or Kohonen networks (Behnke & Karayiannis 1998). (6) There is also the probabilistic neural network (PNN) or Bayesian Neural Network (Bishop 1995). (7) Radial Basis Functions are another type of Hebbian network. The main difference between MLP and RBF is that RBF are designed so that only inputs close to the optimum solution result in an output. This is called local mapping, as opposed to global mapping, where all inputs activate an output, as is the case with MLP (Nabney 2003). (8) Yet another network is called a generalised regression neural network (GRNN). The GRNN is a feed-forward ANN suitable for prediction problems and approximating functions (Hines 1997).

Nelson & Illingworth (1991) describes artificial neural networks as adaptable and tolerant. Tolerant to faults means that if there are small errors in some of input data, they will not change the overall system behaviour because of reinforced learning from historical data that is stored in the weights of a trained network for a known component or system. Adaptable means that ANN can learn from new data when the system changes. ANN can generalise historical training data by applying it on unseen examples. ANN can learn new conditions and consequences on a system. If data is noisy, inaccurate or incomplete and estimates are available, ANNs can predict an accurate value for the erroneous data. The last characteristic is very important in practical applications where data is not often exact.

Looney (1997) found feed-forward neural networks to be the most versatile, most reliable, and most suitable for regression and classification of non-linear data. Han & Kamber (2001) highlight several uses of AI in data mining. These applications include data modeling, data fusion, data visualisation, data selection, pattern evaluation and knowledge presentation.

MLP, RBF and SVM are popular machine learning algorithms (Kecman 2001). Work by Marwala (2001), Haykin (1994) and Cybenko (1989) proved that the MLP is very effective for classification problems. The machine learning algorithms were compared, in the context of bushing condition monitoring, using the following criteria: time efficiency, complexity , generalization ability and memory size, i.e. number of neurons used and CPU memory usage.

3.4 MULTILAYER PERCEPTRON

3.4.1 Architecture of multilayer perceptron

The MLP used for the classification problem had 2 layers of neurons, i.e. one hidden layer and one output layer. The work tested from 3 to 100 hidden neurons on an MLP. The single hidden layer structure was selected, as opposed to a multi-hidden layer structure because research has proved that the single hidden layer MLP is accurate enough for most approximations (Haykin 1994), (Cybenko 1989), (Hornick, Maxwell & Halbert 1989). One outer neuron was used because only one number was required as a decision. The input was a vector with 10 variables. Weights were used between the inner and output layer as shown in Figure 3.2. MLPs are commonly drawn as nodes lines connecting the nodes. Each node represents a neuron and each connecting line represents a weight. The weight indicates the strength of connection between each neuron. The weights are numbers in the range between 0 and 1 (Haykin 1994). The network uses back propagation with a momentum function to reduce the probability of the neural network getting stuck in a local minimum. Bias terms are also used, even though none of the inputs is zero. The bias terms allow the inputs to influence the output even when the inputs are zero (Hines 1997), (Haykin 1994).

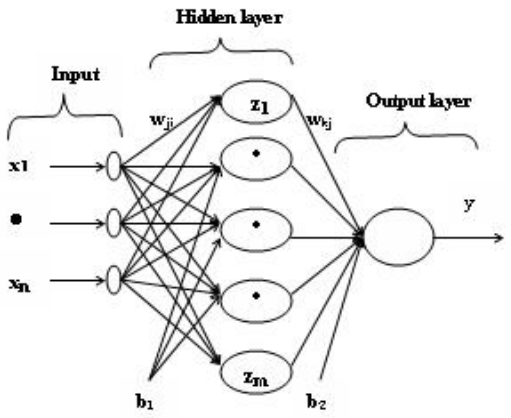

Figure 3.2: Multilayer perceptron architecture

A logistic activation function for the output neuron (y_k) for a classification problem produced a more repeatable result than a softmax or linear function. The functions are definded in Table 3.1.

3.4.2 Mathematical description of multilayer perceptron

Mathematically a single hidden layer MLP is represented by Equation 3.2 (Hines 1997):

$$y = f_i \left(\sum_{j=1}^{m} w_{2j}.f_h \left(\sum_{k=1}^{n} x_k w_{1k} + b_1 \right) + b_2 \right) \tag{3.2}$$

where w_{1k} and w_{2j} are the weights in the hidden layer and output layer respectively, f_i is the output layer activation function, f_h is the input layer activation function and b are the biases. The MLP used *tanh* hidden layer and *logistic* output layer activation function. In activation functions with exponentials, a is called the slope parameter and is normally 1 (Hines 1997). The input layer does not process any data. It sends the input, modified by a weight(w_{1k}) to all the neurons in the hidden layer. The bias allows the activation function to be offset from zero, so that the output of the activation function is defined, and is useful. The functions and their derivatives are shown in Table 3.1. The inner bracket of Equation 3.2 represents the hidden layer. It can be written as in Equation 3.3 (Nabney 2003):

Ten input variables (x_i) were used in this case for each of the 60699 bushings. The first set of bias values (b_j) was randomly generated. In this instance there were weights, (w_{ji}), linking the input to the hidden layer so the relation between the input and output layers is given by Equations 3.3 to 3.6 (Bishop 1995).

$$a_j = \sum_{i=1}^{d} w_{ji} x_i + b_j \tag{3.3}$$

$$z_j = tanh(a_j) \tag{3.4}$$

$$a_k = \sum_{j=1}^{m} w_{kj}.z_j + b_k \tag{3.5}$$

$$y_k = \frac{1}{1 + e^{(-a_k)}} \tag{3.6}$$

Table 3.1: Activation functions and kernel functions
(Nabney 2003), (Gunn 1998)

Activation Function	**Equation**	**Derivative**	**Range**
Constant	$f(x) = a$	$\dot{f}(x) = 0$	
Linear	$f(x) = a.x$	$\dot{f}(x) = a$	n.a
Logistic	$f(x) = \frac{1}{1+e^{-a.x}}$	$\dot{f}(x) = \left(a.\frac{1}{1+e^{-ax}}\right).\left(1 - \frac{1}{1+e^{-ax}}\right)$	x=0 and x=1
Tanh	$f(x) = \frac{e^{ax}-e^{-ax}}{e^{ax}+e^{-ax}}$	$\dot{f}(x) = a.\left[1 - \left(\frac{e^{ax}-e^{-ax}}{e^{ax}+e^{-ax}}\right)^2\right]$	x=−1 and x=1
Exponential	$f(x) = e^{ax}$	$\dot{f}(x) = ae^{ax}$	
Softmax	$f(x) = \frac{e^{ax}}{\sum_{j=1}^{n} e^{a.x_j}}$	$\dot{f}(x) = \frac{e^{ax}\left(\sum_{j=1}^{n} a.e^{ax_j}\right) - \left(\sum_{j=1}^{n} e^{ax_j}\right).e^{ax}}{\left(\sum_{j=1}^{n} e^{ax_j}\right)^2}$	n.a
Threshold	$f(x) = 0 : x < 0$ $f(x) = 1 : x > 0$	$\dot{f}(x) = 0$	n.a
Signum	$f(x) = -1 : x < 0$ $f(x) = 1 : x > 0$	$\dot{f}(x) = 0$	n.a
Gaussian	$f(x) = \sum_{i=1}^{n} a_i.e^{-(\frac{x-b_i}{c_i})^2}$	$\dot{f}(x) = -2\sum_{i=1} n.\left(\frac{x-b_i}{c_i}\right) a_i e^{-(\frac{x-b_i}{c_i})^2}$	
Thin plate spline or(rbf)	$f(x) = x^2.log(x^2)$	$\dot{f}(x) = \frac{2x^2}{xln10} + 2xlog(x^2)$	$x > 0$
x^2tps	$f(x) = x^4log(x^2)$	$\dot{f} = \frac{2x^4}{xln10} + 4x^3log(x^2)$	$x > 0$
Polynomial	$f(x) = a + b.x + c.x^2 + \Lambda k.x^n$	$\dot{f}(x) = b + 2.c.x + \Lambda n.k.x^{n-1}$	

where a_j is the input to each hidden layer neuron, z_j is the output from the hidden layer after activation with the hidden layer function, a_k is the input to each output layer neuron y_k is the output after application of the output neuron.

How multilayer perceptrons process data

Multilayer perceptrons are an example of supervised artificial neural networks (ANNs). Supervised ANNs learn from examples, like the brain. The process of learning from examples is called *training*. If the ANN sees enough examples the network can generalise what it has learnt and accurately apply it to samples that it has not seen before. MLP is a feed-forward network. It makes use of weights that are adjusted by different methods of optimisation during supervised training. *Supervised* refers to the fact that a target is used in training. A sketch of the feed-forward process is shown in Figure 3.3. The supervised training requires that a set of Q example input data vectors ($x^{(q)}$) and a set of K output targets associated with those inputs ($t(k)$) such that the inputs and outputs are paired as ($x^q, t^{k(q)}$). Thus each particular input is mapped onto a specific target, i.e. $x^{(q)} \rightarrow t^{k(q)}$ for $q = 1, .., Q$ (Hines 1997).

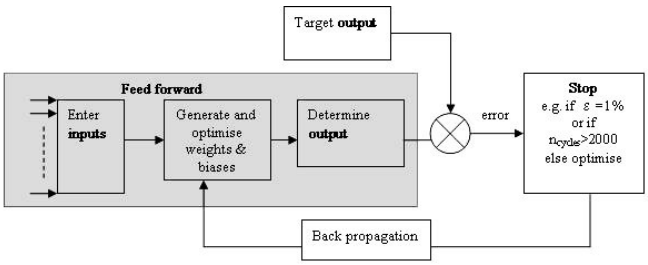

Figure 3.3: Feed forward architecture with back propagation (Hines 1997)

3.4.3 Training methods of MLP

The process of minimising the error by adjusting the weights is called training. Sequential training is where all the weights are updated each time a target and input is presented. Figure 3.4 shows the flow chart illustration of sequential training. The starting point of the weight and bias values are randomly selected to reduce the

probability of the optimisation getting stuck in local minima. Due to the random criteria, it is classified as stochastic. Another method of training is called batch training. A different training technique is called adaptive learning. The sequential training method was used for training during this work. Adaptive learning rate techniques in the form of genetic algorithms and particle swarm optimisation were also applied.

In all methods of training when the sum-of-squares error (SSE) is less than or equal to a desired value then the optimisation stops. Alternatively, if the SSE diverges to a large value then the optimisation stops after a predetermined number of attempts, meaning that the optimisation failed. The MLP neural network gave better accuracy of classification with a smaller SSE.

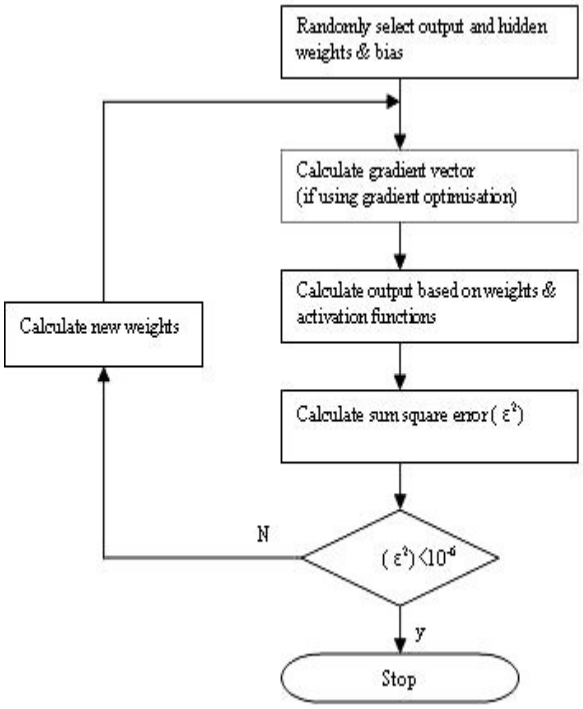

Figure 3.4: Sequential training process

Optimisation methods

The optimisation methods used to adjust the weights within the neural networks included gradient methods such as scaled conjugate gradient (SCG) (Moller 1993),

conjugate gradient (CG) (Bishop 1995), Quasi-Newton (Nabney 2003), simplex (Haykin 1994), particle swarm optimisation (PSO) (Shi & Eberhart 1998) and genetic algorithms (GA) (Michalewicz 1996). The gradient methods use both the function and gradient of the function to locate a minimum point.

The Nelder-Mead simplex method optimises using only the objective function. Search methods such as simplex that use only function evaluations are most suitable for problems that are nonlinear or have a number of discontinuities (Haykin 1994). Gradient methods are more efficient when the function to be minimized is continuous in its first derivative. Gradient methods use information about the slope of the function to dictate a direction of search where the minimum is thought to lie. All the above methods are numerical optimisation methods which make use of trust-regions (Dhlamini & Marwala 2004), if the cost function $f(x)$ is at a point x in n-space and wants to improve, i.e., move to a point with a lower function value, then the numerical methods approximate (f) with a simpler function (q) which attempts to reflect the behavior of function (f) in a neighborhood (N) around the point x. This neighborhood is the trust region. A trial step (s) is calculated by minimizing over N. This is the trust-region sub-region $min_s\{q(s)\Lambda s \in N\}$. The current point is updated to $(x+s)$ if $f(x+s) < f(x)$; otherwise, the current point remains unchanged and N, the region of trust, is shrunken and the trial step computation is repeated.

Optimisation errors generated in the neural network can be due to selecting the initial trust-region far from the optimum region, so that by the time the number of pre-selected iterations is completed the optimization is nearing the optimum point. Common factors that can affect the optimisation result are:

(1) The choice and computation of the approximation function (q), defined at the current point (x);

(2) The choice and method of changing the trust region N; and

(3) The accuracy with which the trust-region sub-problem is solved.

A comparison of the techniques in terms of speed in CPU time (s) and accuracy based on 200 iterations is shown in Table 3.2. Figure D.1 in Appendix D gives details of the CPU as well as other computer hardware used during the experiment.

Table 3.2: Comparison of optimisation methods

Criterion	**SM**	**SCG**	**CG**	**QN**	**QD**
Accuracy	100%	100%	100%	100%	66%
CPU Time(s)	2.5040	0.992	22.66	2.324	72.32

Back-propagation

Back-propagation is an optimisation process that is used during training. It is used to minimise the objective function. The objective function is normally the squared error, defined as in Equation 3.7 (Hines 1997).

$$\varepsilon^2 = (T_i - f(x)_i)^2 \tag{3.7}$$

where $f(x)_i$ is the value of the target or output calculated by the network and T_i is the target or the actual output value measured on the system.

Accuracy of the network is calculated using a mean value of the error function (ε). The mean error is called the sum-of-squares error, and is calculated as shown in Equation 3.8. To eliminate the negative the absolute value of the error is used (Hines 1997).

$$\varepsilon = \frac{1}{n}\sum_{i=1}^{n}\sqrt{(t_i - t_{pi})^2} \tag{3.8}$$

where t_i is the actual target, t_{pi} is the target predicted by the neural network and n is the total number of targets.

Various methods of optimisation were used to minimise the error. These include scaled conjugate gradient (SCG) (Moller 1993), conjugate gradient (CG) (Bishop 1995), Quasi-Newton (Nabney 2003), simplex (Haykin 1994), particle swarm optimisation (PSO) (Shi & Eberhart 1998), genetic algorithms (GA) (Michalewicz 1996), and the Levenberg-Marquardt method of minimization of the least squares norm (Levenberg 1944), (Marquardt 1963). If a gradient based method is used to minimise the error, then the activation function needs to be differentiable (Hines 1997). During each cycle of the optimisation the weights and bias terms are both changed until the error is less than the desired value or the number of optimisation cycles is exceeded.

Updating weights

The output is calculated and changed in proportion to the negative gradient of the sum-of-squares error with respect to the weights. Output layer weights update as follows (Hines 1997).

$$\Delta w_i = \eta_i . \frac{\partial \varepsilon^2}{\partial w_i} \tag{3.9}$$

$$= -\eta_i . \frac{\partial \varepsilon^2}{\partial f(x)} . \frac{\partial f(x)}{\partial x} . \frac{\partial x}{\partial w_i}$$

$$= -\eta_i . \frac{\partial (T - f(x))^2}{\partial f(x)} . \frac{\partial f(x)}{\partial x} . \frac{\partial x}{\partial w_i}$$

$$= -\eta_i (-2)(T_i - f(x)_i) . f(x)_i . \dot{f}(x)_i . f(x)_h$$

$$= 2\eta_i . (T_i - f(x)_i) . f(x)_i . \dot{f}(x)_i . f(x)_h \tag{3.10}$$

where $f(x)_i$ is the value of the target calculated by the network, T_i is the actual target, $\dot{f}(x)_i$ is the derivative of the function, including the activation function, η is a momentum function which amplifies the change, $f(x)_h$ is calculated from the input variable and the hidden layer activation function and the new output layer weight is then $(w_{i+1} = w_i + \Delta w_i)$.

So if the change in the error function is very small, then the weight will change minimally. The process of identifying the minimum value is iterative, so it is slow if one is to search the entire range of the activation function. Choosing an efficient optimisation technique that matches the activation function and its derivative is important. The hidden layer does not have a target so the error in the output must be back-propagated to the hidden layer, so that the hidden layer weights can also be optimised (Hines 1997). The updating process in the hidden layer follows the same steps as the output layer. Recall that (Bishop 1995):

$$T_{pi} = f_i \left\{ \sum_{j=1}^{m} w_{2j} . f_h \left[\sum_{k=1}^{n} x_k w_{1k} + b_1 \right] + b_2 \right\} \tag{3.11}$$

$$\Delta w_h = -\eta_h . \sum_{q=1}^{r} \frac{\partial \varepsilon^2}{\partial w_h}$$

$$\Delta w_k = -\eta_k . \sum_{k=1}^{n} \frac{\partial \varepsilon^2}{\partial f(x)_i} . \frac{\partial f(x)_i}{\partial x} . \frac{\partial x}{\partial f(x)_h} . \frac{\partial f(x)_h}{\partial x} . \frac{\partial x}{\partial w_k} \tag{3.12}$$

Note that the terms in Equation 3.12 can be rewritten as follows:

$$\frac{\partial \varepsilon^2}{\partial f(x)_i} = (-2).(T_i - f(x)_i).\dot{f}(x)_i, \frac{\partial f(x)_i}{\partial x} = \dot{f}(x)_i, \frac{\partial x}{\partial f(x)_h} = w_i$$

$$\frac{\partial f(x)_h}{\partial x} = \dot{f}(x)_h, \frac{\partial x}{\partial w} = x_h$$

This means that

$$\Delta w_k = -\eta_k . \sum_{k=1}^{n} (-2).(T_i - f(x)_i).f(x)_i.w_i.f(x)_h.x_h \tag{3.13}$$

where x_h is the input or x_k, $f(x)_h$ is calculated from the input variable and the hidden layer activation function and $f(x)_i$ is calculated from the input variable and the output layer activation function, the new hidden layer weight is $(w_{h+1} = w_h + \Delta w_h)$, T_i is the target and η is a momentum function which amplifies the change but need not have the same magnitude as before.

3.4.4 Preprocessing data

Before data can be processed by an MLP, the data needs to be preprocessed. Preprocessing is the formatting of the data into information that can be used within the neural network (Han & Kamber 2001). Bishop (1995) emphasises that preprocessing is often the most important stage in the development of a solution, because the choice of preprocessing steps has a significant effect on the performance of the generalization. Preprocessing includes data cleaning, data integration and transformation, data reduction, data discretisation and concept hierarchy generation (Han

& Kamber 2001). Removing the text from numerical data, identifying and removing outliers, normalising, compensating for missing data, transferring the data from a multiple databases into the classification software platform, are all part of preprocessing.

Data cleaning

Data cleaning algorithms attempts to fill missing values, smooth out noise, identify outliers, and correct inconsistencies in the data.

Missing variables

Han & Kamber (2001) summarised six methods used to compensate for missing values as follows:

(1) Ignore the missing value: This is done when a class label is missing. This method may be useful if the classification task is using categories linked to the labels. This method is ineffective, especially when the percentage of missing variables per attribute is not constant.

(2) Fill in the missing value manually: This approach is time consuming, and not feasible for large data sets with many missing values.

(3) Use a global constant to fill missing values: This method is not recommended, especially if values such as $infinity$ or $zero$ are used. This is because it distorts the original data set by creating interest in attributes that may be unimportant.

(4) Use the attribute mean value to fill the missing value: This option is better than all the above, even though it still biases the data.

(5) Use the attribute mean value for all samples belonging to the same class: If for example a risk rating criteria is available for bushings, then all bushings with the same risk rating may use the same magnitude for a missing value in a particular attribute.

(6) Use the most probable value to fill the missing value: The missing variable

can be estimated by regression techniques such as neural networks (Bishop 1995), inference based Bayesian methods (Bishop 1995), (Han & Kamber 2001), decision trees (Han & Kamber 2001) or evolutionary computing (Dhlamini, Nelwamondo & Marwala 2006).

Option 6 is the recommended method to replace missing data (Carreira-Perpin 2001), (Friedman 1977) and (Quinlan 1989) . Noise smoothing is often done using regression techniques (Barbara et al. 1997). Noise smoothing is often achieved using methods that are also used for dimension reduction, such as binning techniques, decision tree induction and concept hierarchies (Han & Kamber 2001).

Inconsistencies in data

Data inconsistencies can be corrected manually by a paper trace. Inconsistencies may result when the same variable in multiple database has different names. Known functional dependencies between variables can be used to detect inconsistencies in data. Data integration and removal of redundant data can remove inconsistencies (Han & Kamber 2001).

Integration and transformation

Data integration refers to merging of data from various sources. The main concerns in integration are *schema integration*, *redundancy* and *data value conflicts* (Han & Kamber 2001). Schema integration is concerned with verifying that data from the same attribute is stored correctly in the integrated database, even if it is called different names in the source databases. Using *meta data*, which is data about the data, generally addresses the schema integration problem. A variable is redundant if it can be derived from another variable in the same database (Han & Kamber 2001). Redundant variables must be removed from the data base. Redundancy determination is commonly done using correlation analysis shown in Equation 3.14 (Reyment & Joreskog 1993). The correlation coefficient (r) is the measure of correlation. This measure is defined as the ratio of the covariances of the two variables to the product of their standard deviations (Reyment & Joreskog 1993):

$$r_{ij} = \frac{s_{ij}}{s_i s_j} \tag{3.14}$$

where r_{ij} is the correlation coefficient, s_i and s_j are the standard deviation of variables x_i and x_j, N is the sample size and s_{ij} is the covariance.

The correlation coefficient is more generally written as in Equation 3.15:

$$r_{ij} = \frac{\sum_{n=1}^{N} y_{ni} \cdot y_{nj}}{\left(\sum_{n=1}^{N} y_{ni}^2 \cdot \sum_{n=1}^{N} y_{nj}^2 \right)^{1/2}} \tag{3.15}$$

where y_{nj} is the difference between a variable x_{nj} and the mean of that variable $\bar{x}_j$, y_{ni} is the difference between a variable x_{ni} and the mean of that variable $\bar{x}_i$, shown in Equations 3.16 and 3.17 respectively.

$$y_{nj} = x_{nj} - \bar{x}_j \tag{3.16}$$

$$y_{ni} = x_{ni} - \bar{x}_i \tag{3.17}$$

The mean is the arithmetic mean given by (Holman 1989).

$$\bar{x}_i = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{3.18}$$

where x_i is the input variable and N is the total sample size. The median can also be used for normalising. The median is the point which divides the data in half (Holman 1989), i.e. the halfway mark. The biased standard deviation or population standard deviation was used because the sample was large, i.e. 60699 samples. The standard deviation or root mean square deviation was defined as shown in Equation 3.19 (Holman 1989):

$$\sigma = \left[\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x}_i)^2 \right]^{1/2} \tag{3.19}$$

The deviation is $s_i = x_i - \bar{x}_i$ and the variance is the square of the standard deviation (σ^2). Variance is a measure of scatter of individual values from the mean. For the

j^{th} variable, the variance is as in Equation 3.20 (Han & Kamber 2001).

$$s_j^2 = \frac{1}{N}\sum_{i=1}^{N} y_{nj}^2 \tag{3.20}$$

Covariance is a description of the relationship between two variables shown in Equation 3.21. So for variables x_i and x_j, the covariance is defined as the average of the sum of the products of the deviation scores (Han & Kamber 2001).

$$s_{ij} = \frac{1}{N}\sum_{n=1}^{N} y_{ni} \cdot y_{nj} \tag{3.21}$$

Data transformation

In data transformation the data is changed into forms which are suitable for machine analysis. Han & Kamber (2001) lists five types of data transformation, namely:

(1) Smoothing
Smoothing is the removal of noise from data. Techniques include binning, clustering and regression

(2) Aggregation
An example of aggregation is to convert daily data to monthly data. The aim of aggregation is to construct data cubes that allow analysis the same at different levels of granularities, so that it can be visualised differently.

(3) Generalisation
Use of concept hierarchies is the most common form of data generalisation. An example is converting numerical age values to classes such as $young$, $middle-age$ and old.

(4) Normalisation
Normalised data is scaled to fall within a small range, such as -1.0 to 1.0, or 0.0 to 1.0.

(5) Attribute construction
Attribute construction is sometimes called feature construction. New attributes are

created and added to the data set's attributes to aide the transformation process.

Normalising is the most important aspect of data transformation (Han & Kamber 2001), so this section is discussed it in more detail. Application of neural networks, clustering and nearest neighbour classification all must be preceded by normalising. The reason why the range 0 to 1 is selected is because most the activation functions are defined, and are not saturated in that range. Figure 3.5, Figure 3.6 and Figure 3.7 show a hyperbolic tangent (*tanh*) function, *Gaussian* function and *thin plate spline* respectively.

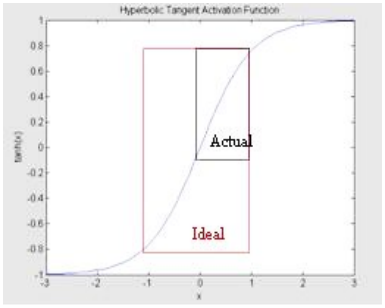

Figure 3.5: Hyperbolic tangent activation function

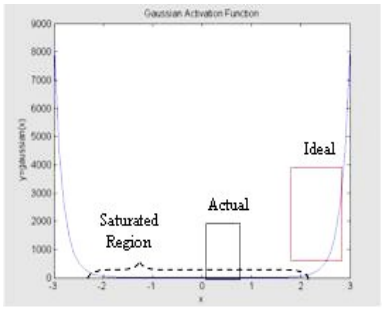

Figure 3.6: Gaussian activation function

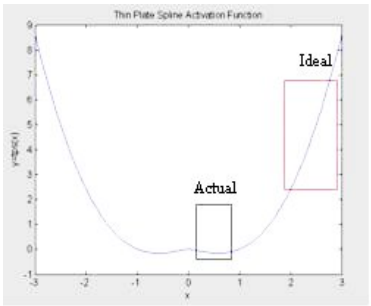

Figure 3.7: Thin plate spline activation function

The input is x-axis and the y-axis is the output. The ideal operating point on the

activation function curves is where the slope of the curve is highest so that the activation can have highest impact on the input. The box marked *ideal* shows the desired operating range, while the box marked *actual* shows the actual operating range. This work tested only the popular methods of normalisation, namely: *min-max* normalisation, $z - score$ normalisation, *absolute maximum* normalisation and *decimal scaling* normalisation (Pyle 1999), (Weiss & Indurkhya 1998), (Holman 1989). Below are details on the methods.

(1) Absolute maximum normalisation
Absolute maximum normalisation is a linear transformation of the data using the maximum value of each attribute as the normaliser (Holman 1989). Symbolically it is defined as follows:

$$Input_{norm} = \frac{Input}{Input_{max}}$$

mathematically written as:

$$x_{norm} = \frac{x_i}{x_{max}} \tag{3.22}$$

where, $Input$ is the actual input value, $Input_{max}$ is the maximum input data for that parameter and $Input_{norm}$ is the normalised input data for that parameter.

This method produced poor results because the smaller input values were drowned by division with large inputs. It distorts the input data. Even after outlier removal, the maximum value of each attribute was very much larger than most of the data set, so it was not representative of the rest of the data. Using this method resulted in the data becoming clustered at one point close to zero. The method produced zeros for some input values, even when the inputs were not zero, but were significantly smaller than the maximum value.

(2) Min-Max normalisation
Min-max normalisation is a linear transformation method that preserves the relationships among the original data. Symbolically it is written as follows:

$$Input_{norm} = \frac{Input - Input_{min}}{Input_{max} - Input_{min}}$$

mathematically written as (Holman 1989):

$$x_{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{3.23}$$

where $Input$ is the actual input value, $Input_{min}$ is the minimum input data for that parameter, $Input_{max}$ is the maximum input data for that parameter and $Input_{norm}$ is the normalised input data for that parameter.

(3) Z-score normalisation
Z-score normalisation is also called *zero mean* normalisation (Han & Kamber 2001). The value of each attribute is normalised based on the mean and standard deviation as follows:

$$Input_{norm} = \frac{Input - Input_{mean}}{Input_{std}}$$

or

$$Input_{norm} = \frac{Input - Input_{median}}{Input_{std}}$$

$$x_{norm} = \frac{x_i - \overline{x}}{\sigma} \tag{3.24}$$

where $Input$ is the actual input value, $Input_{mean}$ is the mean input data for that parameter, $Input_{median}$ is the median input data for that parameter, $Input_{std}$ is the standard deviation of input data for that parameter and $Input_{norm}$ is the normalised input data for that parameter.

The $z-$ *score* method of normalising worked best because it takes into account the deviations of the data relative to a central point. After outlying data had been pruned, this method did not give emphasis to maxima and minima values. Giving equal weight to all the variables was important in this work because, most of the data for used bushings that are not damaged lies close to the mean (Dhlamini, Nelwamondo & Marwala 2006). Han & Kamber (2001) recommends to use $z - score$ normalisation for when outliers dominate the min-max normalisation or when the minima and maxima are not known.

(4) Decimal scaling normalisation
Normalisation by decimal scaling works by moving the decimal point of the attributes (Weiss & Indurkhya 1998). Mathematically it is written as:

$$x_{norm} = \frac{x_i}{10^j} \tag{3.25}$$

where, x_i is the actual input value of an attribute, x_{norm} is the mormalised input and i is the smallest integer such that $x_{max} \times x_{norm} \leq 1$.
The decimal scaling normalisation method is very similar to absolute maximum normalisation, and has all the associated disadvantages.

Data reduction

Performing classification, regression, clustering or other forms of data mining on large data sets can be impractically time consuming if there is a huge data-base. Data reduction methods retain the integrity of the data while representing the data in a smaller volume (Barbara et al. 1997). Classification of the smaller set should produce the same result as the complete set, but should be more efficient, i.e. use less resources in time and computation. Common data reduction techniques are the following:
(1) Data cube aggregation, where for example days can be aggregated to months, without losing accuracy of data.
(2) Dimension reduction, where irrelevant or redundant data is detected and removed.
(3) Data compression, using encoding methods to reduce data size.
(4) Numerosity reduction, by parametric (model estimation) or non-parametric methods, i.e. using sampling, clustering or histograms.
(5) Discretization and concept hierarchies, where a range replaces individual data points.

3.4.5 Data postprocessing

After a result is derived by an MLP, the data needs to be postprocessed. Postprocessing is the action of transforming the data into values that provide useful information contextualised to where the information is applied (Bishop 1995). Postprocessing may involve all or some of actions done in the preprocessing step.

3.5 RADIAL BASIS FUNCTION

3.5.1 Architecture of radial basis function

Unlike in an MLP where each input is summed at each of the hidden neurons, each radial basis function (RBF) network input is summed only at some of the hidden neurons. Figure 3.8 shows the structure of an RBF. The input layer does not have weights or biases in the RBF. In other words the inputs are multiplied by 1 prior to summation at the hidden layer. Weights are only on the output layer (Hines 1997). The RBF used a thin plate spline hidden layer activation function and a linear output layer activation function. Activation functions used within RBF and SVM hidden layers are also called kernels (Hines 1997).

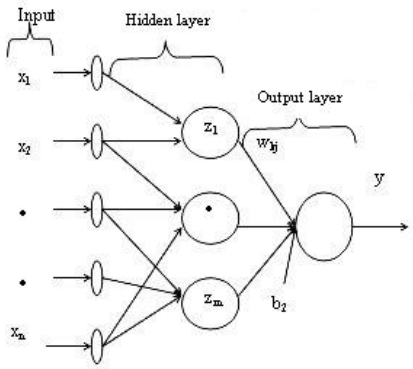

Figure 3.8: Radial basis function architecture

3.5.2 Mathematical description of radial basis function

The thin plate spline was used for this classification exercise, because it produced the smallest training set error when tested on the classification problem. Generally the Gaussian is the preferred function for RBF (Hines 1997). The name radial basis function is due to the fact that the network calculates a radial distance in the function shown in Equation 3.26 (Hines 1997).

$$r = \|x_i - x_m\| = \sqrt{x_i^2 - x_m^2} \tag{3.26}$$

where r is the radial distance between the input data point x_i and the mean of all the input data x_m.

The activation function for the thin spline RBF is given by (Nabney 2003).

$$f_k(x) = x^2 log_e(x) \tag{3.27}$$

$$T_{pi} = f_i\left\{\sum_{j=1}^{m} w_{2j}.f_k\left[\sum_{k=1}^{m} x_k w_{1k}\right] + b\right\} \tag{3.28}$$

where w_{1k} are the weights in the input layer, w_{2j} are the weights in the output layer. A solution is obtained by adjusting the weights until all the inputs give the target output using the same weights. The activation functions are not changed during training. Each input to the neural network does not necessarily activate the output. It only has an effect on the output if it lies close to the output in the optimised local minimum. There is also no bias in the input layer. The results of the RBF simulation are shown in Figure 3.9. Table 3.3 shows the accuracy in relation to the number of hidden neurons.

3.5.3 Training of radial basis functions

Radial basis functions use backpropagation to train, the same as MLPs. The radial basis function (RBF) is faster to train than the MLP (Nabney 2003), (Looney 1997); (Dhlamini & Marwala 2004). The RBF generally needs more hidden layer neurons

than the MLP to achieve the same accuracy, and works best when hidden neurons are equal to, or more than, the inputs. During the optimisation the denominator of the RBF weights is optimised, and can tend towards zero, resulting in an undefined value being generated. RBF work best with few optimisation cycles.
RBF's are unstable. The reason being that in the calculation of the differential equation of the activation function, the value for denominator is approaching zero in the thin plate spline (Hines 1997).

$$\dot{f}(x) = \frac{2.x^2}{x.\ln(10)} + 2x.\log(x^2) \tag{3.29}$$

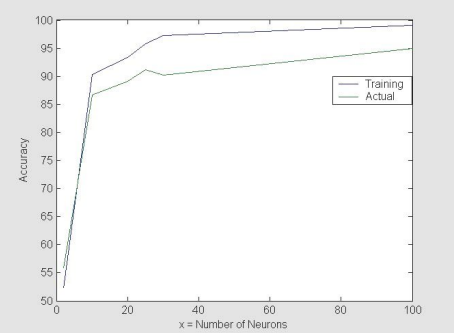

Figure 3.9: Accuracy in relation to number of neurons

Table 3.3: Accuracy in relation to number of neurons

Training	Actual	Neurons
99.04	94.97	100
97.27	90.21	30
95.80	91.19	25
93.34	89.11	20
90.31	86.69	10
52.30	55.77	2

3.6 SUPPORT VECTOR MACHINE

3.6.1 Architecture of support vector machine

The support vector machine (SVM) can be applied to classification as well as regression problems (Cristianini & Shawe-Taylor 2000). SVM operate based on the

structural risk minimization principle. Gunn et al. (1997) found SVM to be superior to conventional neural networks for classification problems. The SVM may include RBF as kernel functions nested within the activation function. Other kernel functions used are polynomial, Gaussian RBF, exponential RBF, Fourier series, splines, b-splines, or sum or product combinations of these. The theoretical criteria that are recommended for selecting a suitable function is the bootstrapping method, another is the cross validation method (Gunn 1998). The hidden layer kernel selected for SVM was the b-spline and the output activation function was a constant (a =1).

The SVM has the form shown in Figure 3.10 (Vapnik 1999). The SVM maps the input into a higher dimensional feature space and then constructs an optimal plane within the feature space (Vapnik 1999). For instance a 5 input problem may be mapped into a 200 dimension space within the SVM, to obtain a single output. If the plane separating data is not linear the SVM maps the input data onto a higher dimensional plane or feature space where an optimal hyper plane can be defined. This work observed that eleven support vectors were used with 14 inputs, to yield an accuracy of 88% (Dhlamini & Marwala 2004).

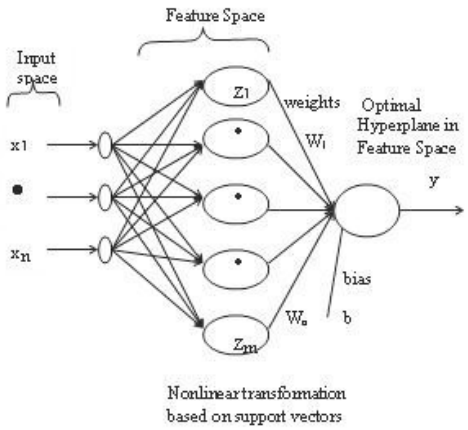

Figure 3.10: Support Vector Machine architecture
(Vapnik 1999)

3.6.2 Mathematical description of support vector machine

The SVM summarises the data into a limited number of support vectors. Equations 3.30 to 3.31 describe how support vectors are derived. The data points where the

Lagrange multipliers are non-zero are called support vectors. The support vectors will lie on the margins of the hyper plane if the data is linearly separable. If the number of support vectors needed is small the calculation requires fewer resources. The classification inner activation function is given by (Vapnik 1999):

$$z_i = f(x) = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*).k(x_i) \tag{3.30}$$

where, $k(x_i)$ is the kernel function, described as a convolution of the inner product for the feature space and α, α_i^* are the Lagrange multipliers (Vapnik 1999), (Gunn 1998). The hyper plane (h) that optimally separates the data is derived by minimizing the Lagrangian $(\phi\ )$ with respect to the weights (w), bias (b) and α (Gunn 1998),

$$\phi = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{t} \alpha_i y_i \left\{ \left[(w, x_i) + b \right] - 1 \right\} \tag{3.31}$$

The multipliers are constrained in the range, $0 \leq \alpha_i, \alpha_i^* \leq C$. C is a misclassification tolerance, which can be chosen to represent the noise in the data. C can be a value from 1 to infinity (Gunn 1998). The data points where the Lagrange multipliers are non-zero are called support vectors.

The output function is given by (Vapnik 1999):

$$y = f(z) = \sum_{i=1}^{m} (w_i.f(z_i) - b) = \sum_{i=1}^{n} \left[w_i.f(x_i, \alpha) - b \right] \tag{3.32}$$

3.6.3 Training of support vector machine

Training of SVM is by optimisation of weights using backpropagation (Gunn 1998). With 60699 bushings the SVM produced a training error of 52.30% in 1012.6 seconds, using 1000 support vectors, with a training set of 5000 bushings, each with 10 inputs. Vapnik (1999) describes the computationally expensive characteristic of SVM when transforming from the input space into the high dimensional feature space as the "curse of dimensionality".

3.7 RESULTS OF CLASSIFICATION USING STAND-ALONE ALGORITHMS

This section presents the performance results of radial basis function (RBF), support vector machine (SVM) and multilayer perceptron (MLP) with data from dissolved gas analysis (DGA) tests. The analysis is aimed only to determine the best performing neural network or kernel machine in terms of speed, accuracy, and stability. This chapter does not seek to identify suitable criteria to use for DGA with artificial intelligence. This is done in the next chapter.

3.7.1 Data used for classification

Ten variables were used to test the performance of MLP, RBF and SVM by classifying the condition of each bushing using a combination of criteria from CSUS (Duval 1974), IEC60599 (1999) and IEEEC57-104 (1991). The ten input parameters were concentrations of the following gases: Hydrogen, Oxygen, Nitrogen, Methane, Carbon Monoxide, Carbon Dioxide, Ethylene, Ethane, Acetylene and the total dissolved combustible gas (TDCG) content. The TDCG consists of: hydrogen, carbon monoxide, methane, ethane, ethylene and acetylene. TDCG is calculated as in Equation 3.33 (IEC60599 1999):

$$TDCG = H_2 + CH_4 + C_2H_6 + C_2H_4 + C_2H_4 + C_2H_2 + CO \tag{3.33}$$

The output variables were binary with 1 indicating 'accept the bushing' and 0 indicating 'reject the bushing'. The available data were taken at one instant for each of the bushings, so it is not possible to use the rate of change of a parameter as another criterion in the classification. The gases' concentrations are quoted in parts per million (ppm) in an oil sample. The concentration of the gases in oil as determined using gas chromatography, were used as inputs to the neural network. The criteria for deciding on the condition of the bushing are shown in the Table 2.3. As it is customary in neural networks, the variables were normalized to fall within 0 and 1 bounds. Normalizing ensures that the input to the neurons lies in the range where the activation functions are defined.

3.7.2 Multilayer perceptron results

An accuracy of classification of 98% was achieved with 60699 bushings. Table 3.4 shows the relation between accuracy and the number of neurons as well as the number of iterations and the time to simulate. Figures 3.11 and 3.12 are graphs showing accuracy in relation to number of neurons. From the results one can see a clear correlation between the number of neurons and the accuracy of classification, and another correlation between the number of iterations and the accuracy. The time to simulate increases as the number of iterations and the number of neurons increases. From Table 3.5 one can see that the selection of the correct optimisation method is core to the performance of the neural network. The scaled conjugate gradient was found to be the most effective. It is clear that training data affect the accuracy of the classification. A training set that is too small, i.e. not representative of expected actual conditions, yields incorrect results. A training set that is biased in favour of a particular condition also yields poor results. In an efficiently trained neural network, the maximum difference in the accuracy between the training and actual data is approximately 1%.

Table 3.4: MLP simulation results

Training	Actual	Neurons	x(Cycles	x(time)
97.40	97.69	100	100	53.163
96.65	97.74	100	50	26.788
88.96	90.08	100	10	7.711
84.12	86.5138	100	2	3.415
95.98	97.62	50	50	13.76
90.21	91.53	50	20	5.4705
88.64	89.94	50	10	3.696
97.39	97.72	30	100	17.543
92.28	97.37	30	50	8.552
92.79	93.27	30	20	3.746
89.50	90.88	30	10	2.063
97.70	97.79	10	100	6.119
96.20	97.15	10	50	3.065

Table 3.5: Accuracy in relation to optimisation method

c100n30	SCG	QN	GRADDESC	CONJGRAD	HMC
Training	97.80	98.84	43.16	98.02	47.22
Actual	97.63	97.20	44.22	97.60	47.14
Time	17.92	70.37	11.03	93.01	13.30
Iter error	0.00	265.01	121228.75	341.16	n.a

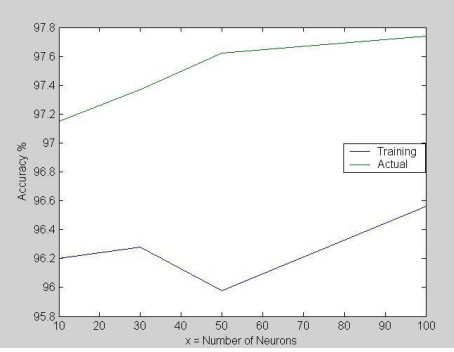

Figure 3.11: Accuracy against number of neurons

3.7.3 Effect of number of hidden layer neurons

An exhaustive search process revealed that a ratio 1:0.8 of hidden neuron to inputs gave the highest accuracy of classification. The work tested from 3 to 100 hidden neurons on an MLP and RBF. The single hidden layer structure was selected, as opposed to a multi-hidden layer structure because research has proved that the single hidden layer MLP is accurate enough for most approximations (Haykin 1994), (Cybenko 1989),(Hornick et al. 1989).

There was no evidence found to suggest that using fewer or more hidden neurons has detrimental effects on the performance of the neural network. However, overfitting destroys the networks ability to generalise. Overfitting results from using too many neurons. At Texas Instruments Defence Systems research and test center, Nelson & Illingworth (1991) consistently used twice as many hidden layer neurons as inputs. Leke-Betechuoh (2004) found that a ratio of hidden layer neurons to inputs of 1.5 was optimal for both MLP and RBF. Werle et al. (2001) found that the MLP they used classifying partial discharges in transformers performed poorly when more than one hidden layer was used. Marwala (2001) used a committee of MLP each with 10 inputs, and 10 hidden neurons to accurately detect faults in structures with accuracy of up to 98.3%. The activation function in the hidden layer was a hyperbolic tangent, while the output activation function was a logistic function. Mak, Allen

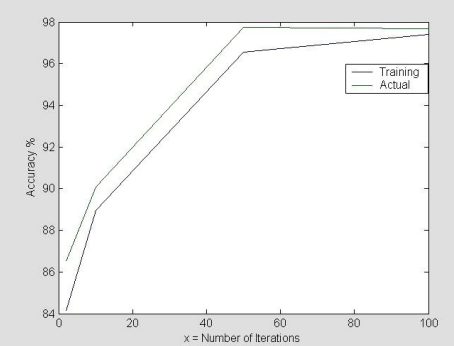

Figure 3.12: Accuracy against number of iterations

& Sexton (1994) compared RBF with MLP for recognizing 10 people's voices, each saying the numbers 0 to 9; they used 50, 100, 200 and 300 hidden neurons for RBF and 10, 20 and 40 hidden neurons on MLP. The RBF with 300 neurons performed best with 98.7% accuracy. The MLP with 20 neurons performed best with 90.2% identification accuracy. The RBF with 100 neurons achieved the same accuracy (90.0%) as the 20 neurons MLP.

3.7.4 Important factors when using MLP

This work identified five important considerations when using MLP.

(1) Normalise the data, and be consistent in the method of normalising and the factor that is used in the normalisation.

(2) Use a training data set with representative data of the real problem. The training data must have as many combinations and results as possible, so that the neural network can make a generalisation more accurately.

(3) Use enough hidden neurons, less is better e.g. ratio of inputs to hidden neurons of 1:0.8. A network with a large number of neurons provides very good results, due to overfitting, but is not able to generalise what it learned during training, i.e. unseen data gives results which are good.

(4) Use many cycles of optimisation, but do not over train.

(5) Use the most efficient optimisation algorithm to optimise the weights of the network.

When deciding on the normalising range, one needs to take into account the activation function that one will use. The normalised data should not lie in the saturation region of the selected activation function, nor should it all lie too closely clustered at one point. The author chose to normalise input data so that it is in the range between 0 and 1, and not -1 and 1, so that a comparison could be made using the different activation functions. The input is x-axis and the y-axis is the output. The ideal operating point on the activation function curves is where the slope of the curve is highest so that the activation can have highest impact on the input.

3.7.5 Causes of errors

Common errors are over-fitting, and under-fitting. Over-fitting or over-training occurs when too many neurons are used for the available inputs. When this happens, the neural networks work well with the training data, but give large errors when processing unseen data. During this work a ratio of neurons to inputs of 10:8 was effective. Under-fitting occurs when too little data are used for training the network. When this happens the neural network produces poor results because it cannot generalize results for data that is completely different from that used in training. The training data set should be able to represent as many real life conditions as possible. To prevent under-fitting Moody et al. (1992) recommends using at least 30 times as many training cases as there are weights in the network. Based on Moody's recommendations a 10 input, 8 hidden layer network with 1 output needs at least $30 \times (10 \times 8 + 8 \times 1) = 2640$ training data. Such a large data set is not always available. It is important to normalize data that is input to the network.

Dhlamini & Marwala (2004) found that the errors calculated during the neural network optimization could not provide adequate information about the network's performance. The errors are due to learning and generalization during the training cycles, i.e. sum square error. In other words, the work observed that a network can get an accuracy of classification of 95% while the optimisation error is 10^5. And inversely, one can get an accuracy of classification of 53% with an optimisation error of 10^{-5}. Mak et al. (1994) made a similar finding.

3.7.6 Comparison of standalone MLP, RBF and SVM

RBF is faster than MLP and SVM, and can produce the same level of accuracy as MLP. However RBF is not stable. The SVM was able to classify 67 data with no errors in its training. The time period for learning with 462 data points was 14 minutes compared with 8 seconds for RBF and 45 seconds for MLP. The findings of this work are in agreement with those of (Mak et al. 1994). Mak et al. (1994) compared RBF with MLP for recognizing 10 people's voices, each saying the numbers 0 to 9, they used 50, 100, 200 and 300 hidden neurons for RBF and 10, 20 and 40 hidden neurons on MLP.The RBF with 300 neurons performed best with 98.7% accuracy. The MLP with 20 neurons performed best with 90.2% identification accuracy. The RBF with 100 neurons achieved the same accuracy (90.0%) as the 20 neurons MLP.

3.7.7 Results of election of a classifier by a committee of networks

In this section a committee of classifiers independently classified 60699 bushings. The network with the most accurate classification was selected to represent all the networks in making a final decision on whether to accept or reject a bushing. This distributed decision making practice is useful in case one processor fails in the monitoring plant. The architecture of the classifiers appears in Figure 3.13.

To exchange information the classifiers used point to point communication, as opposed to broadcast communication. Each classifier sends its information to a central decision maker which evaluates it and then broadcasts a decision. In point to point communication, a classifier needs the address, such as a mail address, IP address, identification number or file name, of the other classifiers in order to communicate.

In this case, the classifiers can be trained with either the same set of data or with different sets of data. Training them with the same set of data allows a fair comparison to be made between the classifiers in selecting the best performing candidate. If, however, each classifier trains using different training data, then one ensures that the final classification of actual data classification is based on a larger sample, making it more likely to be representative of the actual data.

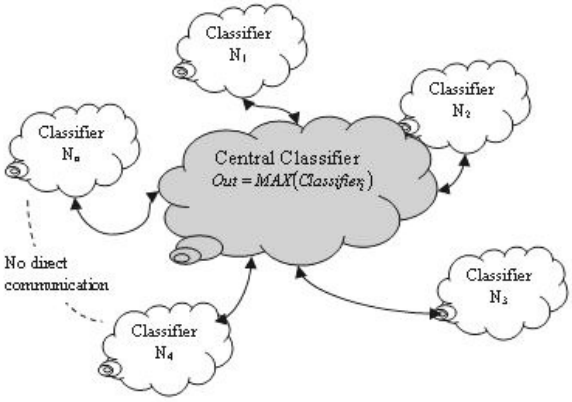

Figure 3.13: Structure of committee

Table 3.6: Classifier election results of accuracy

Elected	97.83	97.78	96.28	97.70
MLP	97.83	97.78	55.65	97.70
RBF	87.74	43.54	70.01	58.67
SVM	0.00	0.00	0.00	0.00
Time	60.11	61.07	60.86	60.76

3.7.8 Classifier election results

The overall classification accuracy of the classifier election was 97% as shown in Table 3.14. Figure 3.15 shows simulation results for several trials. From the graph it is clear that the MLP classifier displayed the highest accuracy. For that reason MLP was the most frequently selected classifier.

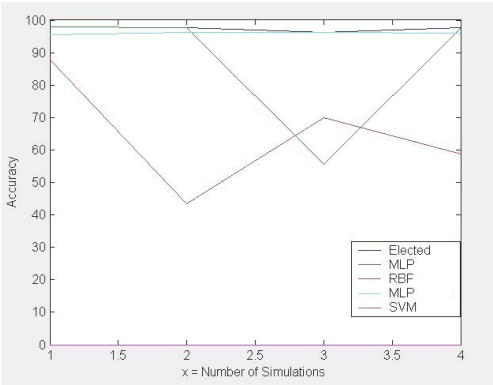

Figure 3.14: Classifier election results

3.7.9 Results of multi-classifier voting system

Each classifier trains using different training data in order to make sure that the final vote of actual data classification is based on a larger sample, which is more likely to be representative of the actual data. In the voting system the networks or the committee of networks can have a weighting factor.

The solution that was selected in this case is the mean or the median of the decision of each network for each bushing. For this classifier system to work all the networks must submit a decision. The architecture of the classifiers is shown in Figure 3.15. All the classifiers should be on the same processor, so if the processor fails, the system is at risk.

The arithmetic mean was used to determine the majority vote. The median was used to determine the democratic voting decision. The median value corresponds to a cumulative percentage of 50%, i.e. 50% of the values are below the median and 50% of the values are above the median. The position of the median is calculated according to Equation 3.34 (Holman 1989).

$$median = \frac{n+1}{2} \tag{3.34}$$

where median is the number in the middle and n is the number of values in a set of data. In order to calculate the median, the data must be ranked, meaning that the data must be sorted in ascending order first (Holman 1989). Six parallel neural networks were used for this exercise. Each of the network types had a different training data set.

The voting process without weighted networks gave very poor results with mean accuracies of 55% and medians in the same range. However, when the networks were weighted in the voting process the classification accuracy improved significantly to 99%. See Table 3.7 and 3.8 for more details. The median value was always marginally better than the mean accuracy.

Table 3.7: Voting results for unbiased network classification

Classifiers	**Weight**	**Accuracy**			**Mean Results**
MLP1	1	55.154	55.357	54.775	
MLP2	1	97.736	55.611	55.6	
RBF1	1	49.933	49.601	45.929	
RBF2	1	92.228	51.955	44.234	
SVM1	0	0	0	0	
SMV2	0	0	0	0	
Vote Accuracy using Median		96.093	53.997	54.75	68.28
Vote Accuracy using Mean		92.768	53.125	52.201	65.03

Table 3.8: Voting results for weighted network classification

Classifiers	**Weight**	**Accuracy**			**Results to Compare**
MLP1	0.05	96.87	54.124	97.676	
MLP2	0.11	97.727	55.636	54.481	
RBF1	0.04	47.596	34.511	51.845	
RBF2	0.06	51.242	79.776	52.213	
SVM1	0	0	0	0	
SMV2	0	0	0	0	
Vote Accuracy using Median		99.993	99.99	99.998	99.994
Vote Accuracy using Mean		99.938	99.723	99.951	99.871

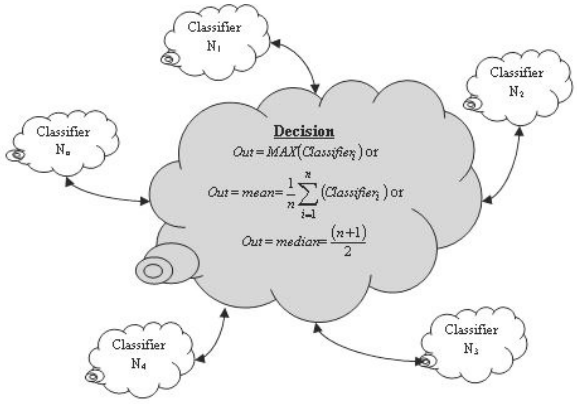

Figure 3.15: Structure of committee

3.7.10 Comparison of classification methods

To keep each classifier independent of the other, different data was used to train each network. Some networks used 1000 training data others used 5000 training data. All the classifiers used the same 60699 bushings as unseen input for classification.

Note that the mean and median in the above table cannot be bigger than the individual means and medians, if the final percentage accuracies were not used to determine the mean and median results. The reason why the overall accuracy value is larger is because the individual classification results were evaluated, then the mean decision was calculated using the individual results. So the final mean is the average of the individual classifications from each network.

One should bear in mind that a 1% error in the classification would leave 607 bushings which might be damaged, in operation in the field. This equates to 102 transformers which might be blown up and burnt beyond repair. The ranking in terms of performance of the networks is as follows.

(1) Weighted networks voting using median 99.99% accuracy
(2) Weighted networks voting using mean 99.87% accuracy
(3) Classification using independent MLP 97.74% accuracy
(4) Classification using independent RBF 94.97% accuracy
(5) Classification using independent SVM 88% accuracy with small data
(6) Voting using unbiased network selection gives 55% accuracy. An unbiased network selection criterion means that all the networks have the same weight. So the result for each bushing from each network is used equally to calculate the mean and

the median for each bushing. The result is then used in determining the final overall result.

The results demonstrate that by using an ensemble of neural networks one is able to create a more robust classification tool. Using different training data allows the collective networks to see a more complete representation of the data. The effects of instability of an individual network or a poor optimisation during one evaluation cycle can be overcome by grouping the neural networks. The best option is to use a committee of networks with a bias towards more efficient networks.

3.8 CONCLUSIONS

3.8.1 Effects of number of optimisation cycles

The number of cycles affects the optimisation calculation used to determine the weights of the neural network. The optimisation was set to stop if the number of cycles reaches the preset number or if there is a change equal to or less than 1% in the error of the weight optimisation. More iterations produce a smaller error in the weight calculation, and a higher classification accuracy in the neural network. Overtraining reduces the ability of machine learning to generalise.

3.8.2 Effect of number of neural networks

With more networks voting on the final result the resulting network is more robust. It is less dependent on the stability of the individual network and the training data used in that network. This group characteristic is important in hardware implementation of neural networks.

3.8.3 Effect of training set data distribution

Balanced data, which showed data with 47% of the bushings rejected and 53% accepted, produced the highest accuracy in training, i.e. 98%, compared to 97% for data that was 42% reject and 58% accept in the decision. In both cases there were 1000 bushings with 10 inputs and one output. The fact that different permutations of the input affect the output is evident in the fact that a training sample with

5000 bushings had a 97% accuracy compared to a 95% accuracy for a 1000 bushing training set, even though the smaller training set was 47% reject and the 5000 training sample was 41% reject. An ideal training data set should be 50% accept and 50% reject with all combinations of the input variable permutations that can produce an accept or reject decision.

3.8.4 Effect of size of training set

The size of the training data set definitely affects the accuracy of the neural network. The fact that different permutations of the input affect the output was shown in the work. A smaller training sample may not contain all the permutations of inputs that would produce a decision. From this analysis the author would recommend that the training set be at least 10% of the actual data. A larger training set is better.

3.8.5 Normalisation

Normalisation is a critical part of the success of the neural network. An effective method of normalisation uses the z-score normalisation. Scaled conjugate gradient was selected for optimising weights in the neural network because of its speed and accuracy.

3.8.6 How did MLP compare to SVM and RBF

MLP is the best option for classification, when compared to RBF and SVM. SVM is very slow. To prevent under-fitting it is best to use training cases that represents as many diverse conditions in reality as possible. To prevent over-fitting a ratio of inputs to neurons of 1:0.8 is effective for MLP. To prevent over-training 50 iterations was found to be optimal. Rather than using standalone MLP or RBF or SVM, the best option is to use a committee of learning machines, with a bias towards more efficient networks.

Chapter 4

DIAGNOSIS USING DISSOLVED GAS ANALYSIS CRITERIA

4.1 INTRODUCTION

Based on the benefits observed by using a collection of neural networks, the work proposes a network of MLP classifiers to perform the human expert role in deciding on the correct result using DGA interpretation procedures given in IEEE c57.104, IEC60599 and CSUS. Classifying using a selected criterion, as in the previous chapter is not enough to get complete information on the condition of the bushing and the cause of the developing defect as well as the extent of degradation. The aim of the previous chapter was to identify suitable neural networks to perform the diagnostics work. For this chapter the MLP was used in a committee of classifiers trained using four different criteria. The criteria were California State University Sacramento (CSUS) criteria, Rogers ratio criteria, Doernenberg criteria and IEC60599 criteria.

4.2 COMPARISON OF DIFFERENT DGA INTERPRETING METHODS

The aim of using a neural network is to introduce machine learning to perform human decision making supervision in the event that the different methods reach a deadlock. Assuming all the methods have equal weighting, when a deadlock (2:2) condition arises, the expert then has to make a decision. When the consolidated result from the individual networks together with the expert is compared to the

expert alone the percentage of correct results is 97.8% during training and 96.44% with unseen data. The deadlock situation arose 43 times in training data, and 1512 with the larger data set.

The training data set comprised 103 data of in the ratio of 35:68, not-faulty: faulty bushings. Table 4.1 compares the expert criteria to the different methods. Comparing the expert's classification to the individual methods found that CSUS criteria are closest to the experts view. CSUS compared to IEC60599 produced the same diagnosis 39% of the time. Unseen data was taken from 2926 bushings that were evaluated in the field.

Table 4.1: Fault classification using CSUS, IEC60599, Rogers and Doernenburg criteria

Criterion	Training Data		Actual	
	No faulty	Faulty	No faulty	Faulty
CSUS	56	47	85	2841
Doernenburg	71	32	1972	954
Rogers	7	96	256	2670
IEC60599	70	33	1813	1113
4 methods same result	2	4	9	609
3 same	26	28	135	661
2 same	43	43	1512	1512
Combined	38	65	162	2764
Epert	35	68	60	2866

The results of the analysis are shown in Table 4.2 and compared to works by Guardado et al. (2001), Akbari et al. (2005), (Morhuis & Kreuger 1990). The combined result refers to the mean taken from all the methods.

Table 4.2: Percentage correct diagnosis for 2926 bushings

Criterion	Dhlamini and Marwala	Akbari	Guardado	Moruis
CSUS	98.33%		100%	
Doernenburg	33.99%	80%	90.91%	83.02%
Rogers	89.96%	70%	90.91%	7.08%
IEC60599/NBR7274	40.11%	80%	93.94%	18.40%
Combined using ANN	96.44%	100%		69.81%
Combined ANN & Fuzzy	90%			81.60%

Ideally the expert's decision would result in removal from service of bushings that

were identified to be defective. In practice the network owner does not always remove those bushings that are identified as defective, because of the impact on the network of failure of a particular bushing may be low, or non-availability of spares may prohibit the removal of the bushing or a variety of other reasons affect the decision. Dalton et al. (2005) indicated that sometimes bushings did fail in the field even though they were monitored and diagnosed. This means that sometimes the expert decision is incorrect, and that the expert is sometimes overruled by operating decisions. So in practice the field expert criteria are not implemented consistently. In other words it is not useful to go out and look in the field to evaluate if what the bushing maintenance expert stated to be defective, has actually been removed from service, and then use his criteria for evaluating all the bushings. For this reason an expert criteria was developed for consistency. The criteria are shown in Table 4.3. The expert criterion proposed in this work when compared to CSUS produced the same results 98.33% of the time, and produced the same result as the combined methods 96.4% of the time.

For the purpose of this work the expert decision is based on conservative criteria from the findings of IEC60599, with added limit for oxygen and total dissolved combustible gases (TDCG). Oxygen (O_2) is found dissolved in oil inside the bushings even after the bushings is filled with oil and sealed. The volume of O_2 gives an indication of the extent of degradation of paper insulation because it is an ingredient in the formation of CO and CO_2, which both highlight paper degradation. The TDCG value is important because it indicates the overall extent of damage. The limits for hydrogen and the hydrocarbon gas concentration in oil samples taken from bushings were selected based on IEC60599. These limits are lower than those recommended in CSUS, IEEE c57.104 for Doernenburg and Rogers methods. It is important to note that the ratio of oil:paper in bushings is significantly smaller than in a transformer. Case studies presented by (DiGiorgio 2005) highlight the importance of O_2 and TDCG variables in an evaluation criterion. The expert criterion is not the same as IEC60599 criterion nor is it the same as CSUS criterion, because it does not give the same results as either of them, nor does it give the same results as for CSUS compared to IEC60599.

Table 4.3: Expert criteria used for condition assessment of bushings

	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	CO	TDCG	O_2
Expert limits	140	40	70	30	2	1000	1282	2000

4.3 DIAGNOSIS OF FAULTY BUSHINGS

Each method highlights several causes of degradation based on the types of gases present and the amount of each gas relative to the other gases. Figure 4.1 shows how IEEE C57-104 classifies fault types.

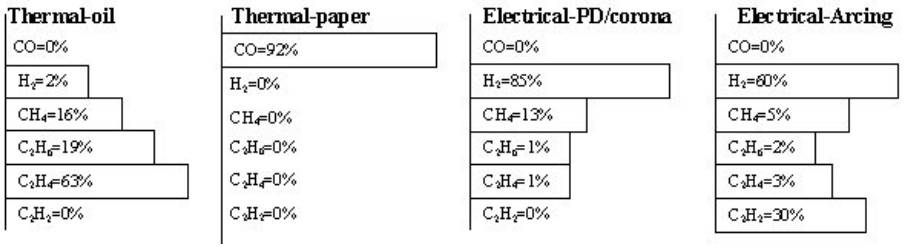

Figure 4.1: Gas composition for various fault conditions according to IEEE c57.104

For simplification three categories are used, namely arcing, thermal, and corona. Corona produces the least damage, and thermal produces more damage, while arcing is most destructive. Table 4.4 shows the criteria for four methods used with DGA. A problem common to all the ratio methods is that the ratios do not cover all ranges of data, so sometimes ratios reflect an unknown cause because it falls outside the range of tables (Akbari et al. 2005). Even when the different methods are combined, the condition *unknown cause* still occurred in 63% of cases in the small sample and the expert criteria diagnosed 48% as unknown for the same sample. In the case of the larger sample 55% were diagnosed as unknown causes by the expert. The ratios are as follows (Doernenberg & Gerber 1967), (Rogers 1975), (IEC60599 1999):

$$\frac{CH_4}{H_2} = Ratio_1 = (R_1) \tag{4.1}$$

$$\frac{C_2H_2}{C_2H_4} = Ratio_2 = (R_2) \tag{4.2}$$

$$\frac{C_2H_2}{CH_4} = Ratio_3 = (R_3) \tag{4.3}$$

$$\frac{C_2H_6}{C_2H_2} = Ratio_4 = (R_4) \tag{4.4}$$

$$\frac{C_2H_4}{C_2H_6} = Ratio_5 = (R_5) \tag{4.5}$$

$$\frac{CO_2}{CO} = Ratio_6 = (R_6) \tag{4.6}$$

4.4 EXPERT DIAGNOSIS CRITERIA FOR BUSHINGS

Diagnosing the condition of the bushing means classifying it as having dissolved gases above the safe limit, and then deciding on the cause of degradation. The expert diagnosis criteria must consider that if the alarm and rejection criteria are too strict, then there can be many false alarms and the cost to act on them does not make the methods viable. If the evaluation criteria is not strict enough, i.e. above the safe operating limits, then the bushings will fail before detection.

The combination of the gases is a more important indication of the extent of degradation, than the individual gases alone. This is because the cause of initial degradation is not always the reason for failure of the bushing. For example, PD can initiate breakdown, develop into a combination of arcing and thermally induced degradation, which can result in failure due to thermal runaway.

Total dissolved combustible gas limits for the different methods are shown in Table 4.5. The concentrations of the gases are in parts per million (ppm). TDCG is calculated as shown in Equation 3.33.

4.5 EXPERT DIAGNOSIS ACCORDING TO CAUSE OF DEGRADATION

The expert criteria is the expert limits + relevant ratio, where relevant ratio is for corona or thermal or arcing criteria as listed below.

Corona: $R_1 < 0.07 or R_2 < 0.1 or R_3 < 0.3 or R_4 > 0.4 or R_5 < 1$ Arcing: $0.1 \leq R_1 \leq 1, 1 < R_2 < 3, R_3 > 0.3, R_4 < 0.4, R_5 > 3$ Thermal: $R_1 > 1, R_2 < 0.1, R_3 < 0.3, R_4 > 0.4, R_5 > 1, R_6 > 20, R_6 < 1$

Table 4.4: Interpretation criteria for DGA using CSUS,IEC60599, Rogers and Doernenburg

Criterion	CSUS	Doernenburg	Rogers	IEC60599
Minimum condition to raise alarm	$H_2 > 1000$ $CH_4 > 80$ $C_2H_6 > 35$ $C_2H_4 > 150$, $C_2H_2 > 70$ $CO > 1000$ $CO_2 > 15000$ TDCG> 5000	If $> 2^*$ anyone: $H_2 = 100$ $CH_4 = 120$ $C_2H_2 = 35$ $C_2H_4 = 50$ and > other gases: $C_2H_6 = 65$ $CO = 350$	$R_3 < 0.1$ $0.1 \leq R_1 \leq 1$, $R_5 < 1$	Concentration: $H_2 > 140ppm, CH_4 > 40$, $C_2H_6 > 70, C_2H_4 > 30$, $C_2H_2 > 2, CO > 1000$, $CO_2 > 3400ppm$ Rate of increase: $H_2 > 5ppm/day, CH_4 > 2$ $C_2H_6 > 2, C_2H_4 > 2$, $C_2H_2 > 0.1, CO > 50$, $CO_2 > 200ppm/day$
Partial discharge/corona	$H_2 > 1000$	$R_1 < 0.1, R_3 < 0.3$, $R_4 > 0.4$	$R_1 < 0.1, R_2 < 0.$ 1, $R_5 < 1$	$R_1 < 0.1, R_5 < 0.2$ Added for bushings $R_1 < 0.07$
Discharge of low energy/sparking	$CH_4 > 80$			$0.1 \leq R_1 \leq 0.5$, $R_2 > 1, R_5 > 1$
Discharge of high energy/ arcing	$H_2 > 1000$, $C_2H_2 > 70$	$0.1 < R_1 < 1$, $R_2 > 0.75, R_3 > 0.3$, $R_4 < 0.4$	$0.1 \leq R_1 \leq 1$, $0.1 \leq R_2 \leq 3$ $R_5 > 3$	$0.1 \leq R_1 \leq 1$, $0.6 \leq R_2 \leq 2.5$ $R_5 > 2$ Added for bushings $R_2 > 1$
Thermal< 300°C	$C_2H_6 > 35$ $C_2H_4 > 150$, $CO > 1000$ $CO_2 > 15000$	$R_1 > 1.0, R_2 < 0.75$ $R_3 < 0.3, R_4 > 0.4$	$R_1 > 1, R_2 < 0.1$, $1 \leq R_5 \leq 3$	$R_5 < 1$
$300° < Thermal$ < 700°C			$R_1 > 1, R_2 < 0.1$ $1 \leq R_5 \leq 3$	$R_1 > 1, R_2 < 0.2$ $1 \leq R_5 \leq 4$
Thermal> 700°C			$R_1 > 1, R_2 < 0.1$ $R_5 > 3$	$R_1 > 1, R_2 < 0.2$ $R_5 > 4$ Added for bushings $R_5 > 1$ $R_6 < 1 or R_6 > 20$

Table 4.5: Various TDCG limits
(IEEEC57-104 1991) and (IEC60599 1999)

	TDCG	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	CO
IEC60599	1282	140	40	70	30	2	1000
CSUS	5000	1000	80	35	150	70	1000
IEEEc57.104	4630	1800	1000	80	200	150	1400
Doernenburg	1025	200	240	65	100	70	350

4.6 COMBINED DIAGNOSING CRITERION

The combined criteria are implemented as follows: Sum all results from all methods for each category. If more than two methods identify the same cause, accept that as the cause. Otherwise conclude unknown result and run the expert criterion network and use the result from the expert criteria network. If the combined result does produce a majority decision, then use the combined result from the individual networks. Figure 4.2 shows the flow chart of how the artificial intelligence expert system was implemented to perform the human task of supervising and deciding on the correct decision when the four methods disagreed.

When the cause is indeterminate then it means the bushing has some degree of degradation but has not reached a severe enough state of degradation to conclusively isolate the cause as being thermal or arcing or corona based on the models. This means the bushing is still healthy, with abnormal gassing. Another condition which is not a cause for alarm is when the ratios criteria are satisfied but the alarm threshold for all the gases is not met. This condition is ignored. Sometimes a diagnosis appears in two categories simultaneously. The most damaging class was selected in this case. In practice three conditions are considered when evaluating a bushing, (1) individual gas limits, (2) rates of increase of gases, (3) gas ratios. To eliminate false alarms three conditions must be met for a bushing to be considered defective based on any model. Because only static data was used for this project only conditions 1 and 3 were considered. Table 4.6 shows the diagnosis decision using different criteria.

4.7 CONCLUSIONS

The expert criterion proposed in this work when compared to CSUS produced the same results 98.33% of the time, and produced the same result as the combined

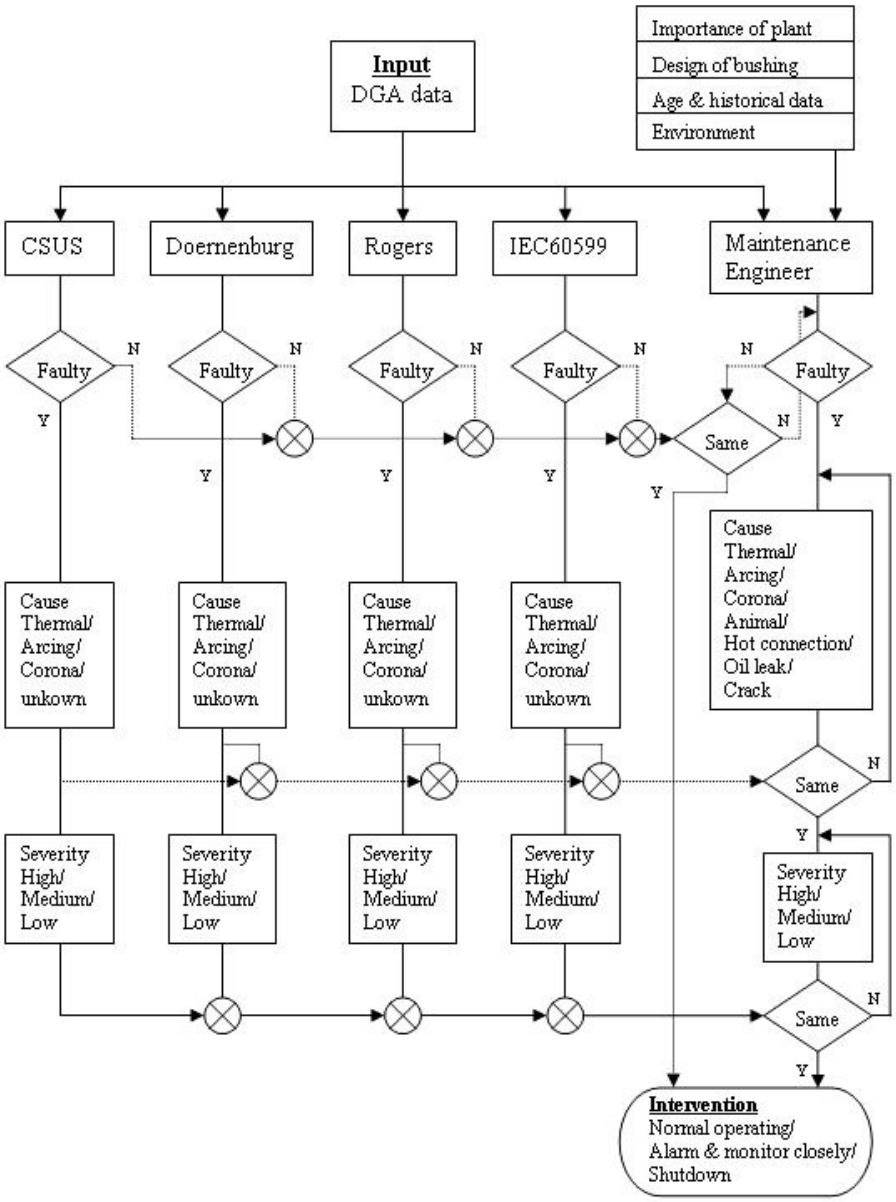

Figure 4.2: Expert intervention flow process for DGA interpretation

Table 4.6: Results of diagnosis using different criteria

Criterion	Training Data				Actual			
	Thermal	Corona	Arcing	Unknown	Thermal	Corona	Arcing	Unknown
CSUS	0	1	1	46/47	1	159	26	2655/2841
Doernenburg	7	2	5	18/32	219	48	131	556/954
Rogers	32	0	7	57/96	710	0	75	1885/2670
IEC60599	20	1	12	10/33	710	46	342	15/1113
4 same	0	0	0					
3 same	0	1	0					
2 same	13	1	5					
All different	4		15					
Combined	10	2	19	41/65	223	99	140	2302/2764
Expert	15	2	18	33/68	697	217	243	1590/2866

methods 96.4% of the time. The expert diagnosis criteria produce fewer cases of unknown results than other methods, making investigations of probable causes of degradation easier, because it gives a starting point for the investigation.

Chapter 5

IMPACT OF SENSOR FAILURE ON CLASSIFICATION RESULTS

5.1 INTRODUCTION

The work proposes the application of neural networks with particle swarm optimisation (PSO) and genetic algorithms (GA) to compensate for missing data in classifying high voltage bushings. The classification is done using DGA data from 60699 bushings based on IEEEC57-104 (1991), IEC60599 (1999) and IEEE production rates methods for oil impregnated paper (OIP) bushings. PSO and GA were compared in terms of accuracy and computational efficiency.

5.2 BACKGROUND TO MISSING DATA

This work investigates tools that compensate for sensor failure in systems that are used for condition monitoring of high voltage bushings. Sensor failure is a concern that needs to be taken into account when designing an automated system of bushing condition monitoring. If a sensor fails on an online bushing monitoring system and the system trips a transformer, the financial and legal consequences can be serious. A failed sensor may take a few hours or months to repair so one will need to answer the following questions before implementing an online diagnostics system. What happens if one or more of the sensors fail? How many sensors can fail before the

online system is rendered ineffective? Apart from the question of reduced reliability due to additional series linked components cost is also a concern. Sensors cost money to purchase, install, and maintain. The cost of transmitting volumes of data continuously, consumes bandwidth and costs the utility money. Being able to estimate missing data accurately allows one to extract important information about the system from the remaining variables and reduce monitoring costs.

In a practical application for monitoring all the components on a plant one would use multiple sensors. Zhang, Zhao & Yan (2006) in their study used leakage current to monitor the bushings, DGA, water content and temperature to monitor transformer oil, and other variables, as shown in Table 5.1, to monitor the other parts of a transformer. They described the machine learning system in terms of five layers, each of which generates and uses data. If any one of the data layers goes missing the reliability of the whole system is jeopardized. The layers are (1) Plant and Environment layer. (2) Sensors layer, i.e. what to measure. (3) Data acquisition layer, i.e. how does one transmit the data, i.e. what telecommunications options are available. (4) Diagnostics layer, i.e. what tools are useful for data processing, criteria for condition evaluation, data fusion and importance identification. (5) Control layer i.e. level of degradation, cause of degradation, remedial action, time taken to decide and respond, and graphic user interface.

Table 5.1: Example of sensors used on a bushing system (Zhang et al. 2006)

Monitoring item	Quantity	Type of sensor
Partial Discharge	1	UHF
Dissolved gas in oil	1	Single and multiple gas
Moisture in oil	1	Capacitive
Top oil	1	Resistive sensor
Current of iron core	2	Active Rogowski coil
Current of bushing	6	Active Rogowski coil
Operating voltage	1	Active Rogowski coil
Load current	2	Active Rogowski coil
Position of Tap Changer	1	BCD code relay
On/Off cooler	1	Relay signal
Ambient temperature	2	Integrated circuits

Recognising the value of artificial intelligence and the risks associated with its usage in power systems, Cigre formed Study Committee 15 WG 11 in 2002 with the task of standardising and improving applications of data mining techniques within power

systems (Mcgrail, Gulski, Groot, Allan, Birtwhistle & Blackburn 2003). Mcgrail et al. (2003) identified several areas in reliability centered maintenance that can be improved by using artificial intelligence online and offline. Artificial intelligence (AI) can add significant value in power systems operations, maintenance and control, by providing a single engine to execute the critical role of data-fusion (Hall & Llinas 2001). However, it is also clear that AI systems need to be robust and reliable.

Some of the methods used previously to account for missing data include regression techniques by Jackson (1991) as well as Jolliffe (1986) and Madow et al. (1983) who used principal component analysis. Abdella & Marwala (2005) proposed a method of accurately compensating for missing data using an autoencoder. Markey & Stevens (2003) chose to use zeros where there was missing data. Other work by Ghahramani & Jordan (1997) and Tresp et al. (1994) also used regression to address the problem of missing data. The accuracy of the solution derived by replacing missing data with averages or zeros or an iterated number depends on whether the final diagnosis decision is based on the missing variable alone, or if that decision depends collectively on all the variables. The accuracy of the approximated variable will be considered in this paper.

What is important at all stages in evaluating data is to note that data or measurements or numbers by themselves have little meaning, and historically human intervention is required before decisions can be made and executed. Artificial Intelligence gives the option to autonomously transform, correlate and interpret data, so that it becomes valuable knowledge.

5.3 WHY MISSING DATA IS A PROBLEM

The first reason for concern when a sensor fails is that no information is available for a particular measured parameter, and that missing variable might be important. The second reason to be concerned is that the processor of the online diagnostics tool will identify an undefined value where there is missing data due to sensor failure. The third reason for concern is that data from a sensor can become corrupted due to the sensor loosing calibration or excessive noise. With online learning tools being applied to condition monitoring missing data approximation is very important Vilakazi et al. (2006). The problem is that analytical methods such as neural network, fuzzy set theory, principal component analysis, etc., cannot process undefined values. Currently, there is no method for specifying missing data within a neural network

or fuzzy structure, so missing data must be approximated prior to processing. For multivariate data, six methods are available to address missing data, as detailed in section 3.4.4. The methods are as follows: (1) Ignore the missing value; (2) Fill in the missing value manually; (3) Use a global constant to fill missing values; (4) Use the attribute mean value to fill the missing value; (5) Use the attribute mean value for all samples belonging to the same class; (6) Use the most probable value to fill the missing value

This work approximates data which is missing at random (MAR), missing completely at random (MCAR) and non-ignorable data as described in Little and Rubin (1987). MCAR arises if the probability of missing value for variable X is unrelated to the value X itself or to any other variable in the data set (Abdella & Marwala 2005). MAR arises if the probability of missing data on a particular variable X depends on other variables, but not on X itself (Abdella & Marwala 2005).

5.4 TECHNIQUES TO COMPENSATE FOR MISSING DATA

There are many methods of data fusion for extracting features among variables of highly dimensional data, among these are dimension reduction techniques such as Principal Component Analysis (PCA), Fisher Linear Discriminant (FLD), Multidimensional Scaling (MDS), Independent Component Analysis (ICA), Factor Analysis (FA), and Auto associative neural network encoder (autoencoder).

All the above methods, except autoencoders, are dimension-reduction techniques, in the sense that they can be used to replace a large set of observed variables with a smaller set of new variables, whilst preserving all of the original information. PCA takes advantage of redundancy of information and simplifies data by replacing a group of variables with fewer new variables, called principal components. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other, so there is no redundant information. Everitt & Dunn (1991) found that FA works well for finding correlations among data, in contrast to PCA which only summarises data using fewer dimensions. MDS is used to detect underlying dimensions that allows one to explain observed similarities or dissimilarities between the investigated objects. MDS is comparable with FA, the only difference being that FA tends to extract more factors than MDS, forcing the user manually to interpret the results; as a result, MDS often yields more useful

solutions. Welling & Weber (1999) established that ICA searches for directions in data-space, which are independent across all statistical orders. ICA is related to principal component analysis and factor analysis, yet ICA is a much more powerful technique, because ICA is capable of finding the underlying factors or sources when the classic methods such as PCA and FA fail completely. Although PCA finds the minimum number of components that best represents the data, this best representation is in the least square sense and it does not guarantee any usefulness for discrimination. One needs to reduce the dimensionality, under some constraint of maximizing the class discrimination. Maximizing the discrimination can be achieved by increasing the inter-cluster distances and reducing the intra-cluster distances.

These distances are obtained using between and within class scatter matrices through the FLD method. FLD is referred to as multiple discriminant analysis (MDA) when the number of classes of data exceeds two.

Dimension reduction of original complete data can be used to generate a smaller data set, which can then be enlarged into an approximation of the original data set using a prediction neural network (NN). If there is missing data in the large data set, one can ignore the missing variable and use a dimension reduction technique with the available variables and then run the output of the PCA or FLD, etc., through a neural network to regenerate the original complete data set with the approximated variable. The above approximation process has many steps where errors can be introduced. For this reason this chapter will evaluate the autoencoders only.

5.5 AUTOENCODER

Autoencoders use the principle of contractive mapping to locate a point of convergence, given a set of known sensors' data and some unknown sensor data. Contractive mapping occurs when the output distance between two points is less than the input distance between the same points. Mathematically, it is a mapping $O : X \rightarrow X$ on a complete metric space (X, d) in which, for any x and y in that space (Thompson et al. 2003):

$$d(O_x, O_y) \leq k.d(x, y) \tag{5.1}$$

where, $0 \leq k \leq 1$

Autoencoders are an application of the Banach Fixed-Point Theorem, which states that if f is a contractive mapping, then there exists a unique fixed point x_0 for which $f(x_0) = x_0$. Moreover, there exists a sequence x_n, for which any element $x_{n+1} = f(x_n)$, converges, and that convergent point is x_0. Figure 5.1 shows the structure of an autoencoder. Thompson et al. (2003) successfully applied autoncoders to restore missing sensors by minimising the error between the missing sensor inputs and outputs and also minimising the error between the entire input pattern and output pattern using both missing and known sensors to achieve a final answer.

The input to the network is two vectors whose total dimension is n. By definition the input dimension and the output dimension are the same in an autoencoder, while the hidden layer has a lower dimensionality than the input/output. As a rule, the ratio of input to hidden layer neurons is 2:1, as in Thompson et al. (2002). The input into the neural network is given by x. The first vector, x_k, is the set of known sensor values for a given input pattern. The second vector is x_m, the set of missing sensor values. The autoencoder that is used in this chapter is a feed-forward multilayer perceptron (MLP). The parameter w_{ijk} is the matrix of weights whose $(i,j)th$ element is the weight connecting the ith known sensor value to the jth neuron in the hidden layer; w_{ijm} is the matrix for the missing sensors, b_1 is the vector of bias weights for the first layer; w_{kjk} and w_{kjm} are the weights on the output. Within the input layer the input is transformed from x into a using the weights (w_{ji*}) and biases (b_j) (Bishop 1995) as in:

$$a_j = \sum_{i=1}^{d} w_{ji*}.x_i + b_j \tag{5.2}$$

Within the hidden layer the input is further transformed using an activation function, such as (Nabney 2003):

$$z_j = tanh(a_j) = \left[\frac{e^{ca_j} - e^{-ca_j}}{e^{ca_j} + e^{-ca_j}}\right] \tag{5.3}$$

The variable c is a constant term. In the output layer the input is further transformed into a variable which can be optimised using weights (Hines 1997).

$$a_k = \sum_{j=1}^{m} w_{jk*}.z_j + b_k \tag{5.4}$$

At the output stage the variable is passed through another activation function.

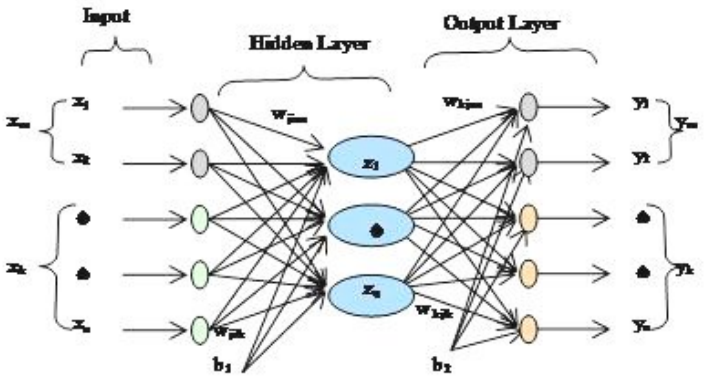

Figure 5.1: Autoencoder architecture

In the case of this work, the activation function that produced the most accurate results was the sigmoid function, shown in Equation 5.5.

$$y_k = \frac{1}{1 + e^{(-a_k)}} \tag{5.5}$$

The error at the output of the neural network is given as (Hines 1997):

$$e_1^2(k) = \left[y_j(k) - t_j(k)\right]^2 \tag{5.6}$$

where t_j is the desired output, y_j is the neuron output, and k is the k^{th} output. Within the neural network several methods of optimisation were used to minimise errors in the weights. The optimisation methods tested within the neural networks included gradient methods such as conjugate gradient (CG), scaled conjugate gradient (SCG), quasi-Newton (QN), batch gradient descent (GD). Each stage of the optimisation introduces some error as highlighted in Dhlamini & Marwala (2005). The sum of the squared output error is used to prevent the error from being a negative number, i.e. when target is greater than ANN output. The sum squared error is given by (Hines 1997):

$$\varepsilon^2(k) = \frac{1}{n}\sum_{j=1}^{n} e_j^2(k) \tag{5.7}$$

where n is the number of neurons in the output layer, the average squared error is calculated by summing the squared error (ε_i) of all the outputs and dividing by the size of the output set (N), giving an average error (ε_{av}) (Hines 1997).

$$\varepsilon_{av} = \frac{1}{N}\sum_{j=1}^{n}\varepsilon(k) \tag{5.8}$$

If there is missing data, then the error function becomes (e_m) (Abdella & Marwala 2005).

$$e_m = \left\{ \begin{pmatrix} x_k \\ x_m \end{pmatrix} - f\left\{ \begin{pmatrix} x_k \\ x_m \end{pmatrix}, \begin{pmatrix} w_k \\ w_m \end{pmatrix}, b \right\} \right\}^2 \tag{5.9}$$

where the subscript m stands for missing, and k stands for known. Two evolutionary algorithms were used to optimise the error function and these are particle swarm optimisation and genetic algorithms (GA). Because the GA used in this work seeks to maximise the error function a negative was inserted in the error equation to obtain a minimum value. So the GA error function is given in Equation 5.10 (Abdella & Marwala 2005).

$$e_m = -\left\{ \begin{pmatrix} x_k \\ x_m \end{pmatrix} - f\left\{ \begin{pmatrix} x_k \\ x_m \end{pmatrix}, \begin{pmatrix} w_k \\ w_m \end{pmatrix}, b \right\} \right\}^2 \tag{5.10}$$

The process diagram of the evolutionary neural network is as shown in Figure 5.2. Approximating missing data using iterative optimisation techniques such as PSO and GA is an expectation maximisation (EM) type of approach, because the first step, called the expectation (E) step computes the expected value of the missing variable. And the second step, called the maximization (M) step, substitutes the expected values for the missing data obtained from the E step and then maximizes the likelihood function as if no data were missing to obtain new parameter estimates. The cycle of expectation and maximisation is repeated until the error is within the tolerance or until the number of cycles has been exceeded.

In this work the autoencoder had 10 inputs and outputs with 7 hidden neurons. The number of neurons in the hidden layer of the classifying MLP was optimised to 31, with 10 inputs and 1 output.

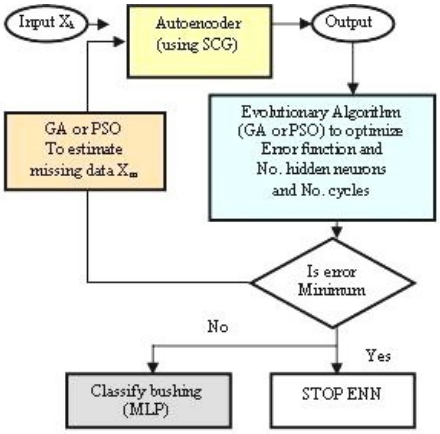

Figure 5.2: Missing data classification flow chart

5.6 GENETIC ALGORITHMS

Genetic algorithms (GA) search the solution space of a function by simulating the survival of the fittest strategy, similar to evolution (Holland 1975). The evolutionary strategy used in GA is similar to that used in evolutionary programming (Fogel et al. 1966), scatter search (Glover 1977), genetic programming (Koza 1992). Collectively all the methods are called evolution programs. Evolutionary programming is used for searching a space of finite state machines. Scatter search keeps a population of reference points and produces a new population by combining the individuals in the population linearly using weights. Genetic programming searches for the most suitable computer program to solve a particular problem.

All evolution programs are probabilistic methods that maintain a population of individuals, $P(t) = \left\{x_1^t, ..., x_n^t\right\}$, where $P(t)$ is the population at time t and represents the individuals in the population. A selection process is applied to the population based on the fitness of the individuals and some genetic operators to isolate the best individual. The probability that an individual will be selected depends on that individual's fitness relative to the fitness of other individuals as shown below (Michalewicz 1996).

$$P_s(x_i) = \frac{f(x)}{\sum_{k=1}^{n} f(x_k)} \tag{5.11}$$

where $P_s(x_i)$ is the probability that the individual x_is will be selected, $f(x_i)$is the fitness of the individual and $\sum_{k=1}^{n} f(x_k)$ represents the fitness of the entire population. Each individual is a potential solution which is assigned a fitness value. At a

new time interval $t+1$, a new population is formed by selecting the fitter individuals and changing them by applying genetic operators to them. Mutation is a kind of genetic operation which creates new individuals by a small change in a single individual. Crossover is another type of genetic operation that creates new individuals by combining parts from two or more individuals (Michalewicz 1996). After several generations the program converges and the solution identifies the best individual.

GA uses components and stages which include chromosomes, selection functions, genetic functions, reproduction functions, a random initial population, terminating criteria, and an evaluation function (Houck, Joines & Kay 1995). The five stages in a GA optimisation cycle are to create a random initial state, evaluate fitness, select the fittest population, undergo crossover, undergo mutation and repeat until successful. Selection methods include roulette wheel and its variations, scaling techniques, tournament, elitist, and ranking methods.

For this work 25 generations and a population size of 20 was used. The roulette wheel was the selection method, and arithmetic crossover and non-uniform mutation were used. With roulette wheel selection the probability of the section in the roulette wheel is proportional to the value of the fitness function of every chromosome (Michalewicz 1996). The bigger the fitness value is the larger is the allocated section in the roulette.

Arithmetic crossover results in a linear combination of two vectors. If x_1 and x_2 are the two parents that need to be crossed, the offspring are determined as shown below (Michalewicz 1996).

$$\acute{x_1} = a \cdot x_2 + (1 - a) \cdot x_2 \tag{5.12}$$

$$\acute{x_2} = a \cdot x_2 + (1 - a) \cdot x_1 \tag{5.13}$$

where $\acute{x_1}$ and $\acute{x_2}$ are the next generation and a is a random value, $a \in [0...1]$. Variations of arithmetic crossover include intermediate crossover, guaranteed average crossover and linear crossover (Michalewicz 1996).

In non-uniform mutation an individual parent x, mutates to form a new individual $x_k^{'}$ such that (Michalewicz 1996):

$$x_k^{'} = \begin{cases} x_k + \Delta(t, right(k) - x_k) & if\ a\ random\ binary\ digit\ is\ zero \\ x_k - \Delta(t, x_k - left(k)) & if\ a\ random\ binary\ digit\ is\ 1 \end{cases} \tag{5.14}$$

where $\Delta(t, y)$ is a value in the range $[0, y]$ such that the probability of $\Delta(t, y)$ being close to 0 increases as number of generations (t) increases. Non-uniformity of this approach is due to the fact that the $\Delta(t, y)$ operator searches the entire space uniformly while t is small and then searches locally as t becomes larger. Table 5.2 compares the results of the simulations done using PSO and GA.

5.7 PARTICLE SWARM OPTIMISATION

Particle swarm optimisation (PSO) takes its origins from the social behaviour of bird flocking. PSO is a random search technique using global optimisation method. There are many similarities between GA and PSO, yet unlike GA, PSO has no evolution operators such as crossover and mutation. In PSO, the potential solutions, called particles, fly through the problem space by following the current optimum particle. Each particle is defined by two variables, a velocity and position, as shown in Equation 5.15 and 5.16 respectively (Shi & Eberhart 1998).

$$v[\ ] = v[\ ] + c_1.rand() \cdot (pbest[\ ])c_2 \cdot rand() \cdot (gbest[\ ] - present[\ ]) \tag{5.15}$$

$$present[\ ] = present[\ ] + v[\ ] \tag{5.16}$$

where, $v[\ ]$ is the particle velocity, $present[\ ]$ is the current particle or current solution, $pbest[\ ]$ is the best solution or fitness achieved so far, $gbest[\ ]$ is the best value in the global set of particles, $rand()$ is a random number between $(0, 1)$ as well as c_1 and c_2 which are learning factors as defined by Shi & Eberhart (1998).

Each particle keeps track of its coordinates in the problem space which are associated with the best solution or fitness it has achieved so far. The fitness value is also stored as *pbest*. The best value, obtained so far by any particle in the neighbourhood of the particle is also tracked. This location is called *lbest*. When a particle

takes all the population as its topological neighbours, the best value is a global best and is called *gbest*. At each time step, the velocity changes, thus accelerating each particle toward its *pbest* and *lbest* locations. Acceleration is weighted by a random term, with separate random numbers being generated for acceleration toward *pbest* and *lbest* locations. Clearly the randomisation of the position of the particles is at the start of the PSO optimisation, i.e. only the first set of particle values is truly random. Subsequent particles all move towards the best particle, *pbest* located at *lbest*. If the initial *pbest* is at a local minimum *lbest* then the entire swarm will search and converge within that local minimum. This is in contrast with GA where the entire search process is random. The starting positions of the chromosomes are random, the mutation of each chromosome is random and further randomised by crossover of mutated chromosomes. When the fittest chromosome is selected, mutated, and reproduced then the search surface is less likely to be a local minimum, because the offspring can end up in a completely unexplored surface, or search the same surface more than once. By remembering the best result over the total number of generations, the GA is able to obtain a global optimum. GA performs a more exhaustive search for an optimum than PSO.

The PSO simulation used a swarm size of 20, with 50 iterations to produce an accuracy of 95% with one missing data point. Table 5.2 shows more results. What is common in the implementation of evolutionary techniques such as PSO and GA are the following procedures: (1) random generation of an initial population; (2) calculating a fitness value for each subject which is directly dependent on the distance to the optimum; and (3) reproduction of the population based on fitness values.

5.8 RESULTS

Using evolutionary algorithms together with neural networks, the work was able to minimise the error function and number of hidden neurons as well as the number of cycles in the iteration. The missing data approximation simulations were done with 500 bushings each with 10 variables. The criterion for a correct approximation of missing value is that it should lie within the standard deviation for each variable, e.g. CH_2, and be positive. Standard deviations calculated gave values ranging from 0.96 to 7226.

If the standard deviation is small, the approximated value is closer to the actual missing value. Large standard deviations result in approximated values which are

further from the target value even if the approximation is within the standard deviation. Based on the approximated missing variables, 500 bushings were then evaluated according to IEEEC57-104 (1991) and IEC60599 (1999) criteria, and classified as acceptable or unusable. Table 5.2 shows the average values of three simulations of the accuracy of approximated values that were calculated using PSO and GA.

Table 5.2: Accuracy of predictions for missing variables using PSO and GA

Criterion	Number if missing data points				Method
	1	2	3	4	
Accuracy	95%	84%	76%	54%	GA
Time	4608	4799.5	5006.2	4999	GA
Accuracy	95%	66%	66%	51%	PSO
Time	1050.1	1057	1071.25	1060	PSO

The difference in the results is due to using fewer particles in the PSO swarm. PSO was found to be 4 times faster than GA to achieve the same level of accuracy for one missing data point. But it can be argued that it depends on the number of iteration and swarm size. By setting both at very high values, e.g. swarm size of 100 and iterations of 500, the average accuracy of the results can be improved by 1%.

To approximate a missing value the known data and the guessed unknown/missing value was put into a trained autoencoder. The average error between the input and output values of the known values was calculated ignoring the approximated missing variable. If the error in the known variables was greater than 1×10^{-3}, then the approximated missing value was recalculated by the GA or PSO. The process continued until the number of iterations of the swarm or generations for the GA was exceeded or until the error was less than 1×10^{-3}. The results obtained show that an autoencoder can reliably trace correlations between the missing data and known data. The results further show that where 50% of data was not available, the network could only approximate 30% of the missing variables to within the standard deviation.

If PSO and GA performance is compared based on time alone, then PSO is better than GA. If accuracy is the criteria for evaluation, and time is not limited then both methods perform the same, because the number of iteration or generations as well as the swarm size or population size can be increased to achieve the desired

accuracy. PSO has few parameters to adjust during the optimisation.

5.9 CONCLUSIONS

The work finds that an autoencoder can trace correlation between the missing data and known data if 10% of the data is missing. Both PSO and GA could produce an average accuracy of 95%. With 10% of the data missing all approximated data was within standard deviation, and therefore accepted as correct, i.e. 100% accuracy. If the percentage of missing data is less than 30% of total number variables, the autoencoder approximated the missing data with an average accuracy of 68% for PSO and 76% for GA. PSO was found to be 4 times faster than GA to achieve the same level of accuracy for one missing data point. But it can be argued that it depends on the number of iterations and the swarm size. By setting both at very high values, e.g. swarm size of 100 and iterations of 500, the accuracy of the results can improve by 1%, but the time to simulate will be increased by 400%. GA performs a more exhaustive search for an optimum than PSO.

Chapter 6

FUZZY-NEURO CLASSIFICATION

6.1 INTRODUCTION

This section presents fuzzy set theory (FST) used in condition monitoring for high voltage bushings. Fuzzy set theory has been used in diverse applications in the last decade. Majozi & Zhu (2005) used fuzzy inference system (FIS) to match operators and chemical plants based on their skill, availability, health and age; Kubica, Wang & Winter (1995) used FST in control systems; Flaig, Barner & Arce (2000) applied FST in pattern recognition. Ammar & Wright (2000) applied FST in the evaluation of state government performance, client satisfaction surveys, and economic impact of state-funded agencies. Its main strength is the ability to model imprecise or uncertain data that characterise many systems and environments. Fuzzy theory allows one to explore the interaction of variables which define a system, and how the variables affect the system's output. Majozi & Zhu (2005) emphasise that attempting to linearly combine these inputs would not be able to explore these interactions, hence would lack robustness. Neural Networks have been tested by Dhlamini & Marwala (2005) for condition monitoring of bushings, and by Wang (2000) for transformer condition monitoring. In the case of bushings that are evaluated using IEC60599 (1999) there is a large range of values associated with normal, elevated and abnormal amounts of gas. FST will help in objectively answering the question: How high is too high or too low for an elevated condition to be classified as dangerous and require the bushing to be maintained?

6.2 FUZZY SET THEORY BACKGROUND

There are 4 steps involved in fuzzy logic implementation, i.e. (1) Fuzzify inputs; (2) Select membership functions; (3) Apply fuzzy operators; and finally (4) Defuzzify (Majozi & Zhu 2005).

6.2.1 Fuzzify inputs

Fuzzify inputs means: to identify the inputs or attributes which describe the system.

6.2.2 Select membership functions

Select membership functions means: to resolve all fuzzy statements (inputs) into a degree of membership between 0 and 1 for each attribute.

6.2.3 Apply fuzzy operators

Apply fuzzy operators means: AND or OR or NOT or ANY the inputs similarly to Boolean algebra. AND is the *minimum* fuzzy operator, which chooses the least of all values input in the same rule. OR is the *maximum* fuzzy operator, which chooses the greatest of all values input in the same rule. The NOT operator makes the value the opposite, i.e. (1-value). The ANY operator sums the values in the rule. The result after applying the fuzzy operator is called the degree of support for the rule, e.g. max (0.0, 0.7) = 0.7 means if inputs are 0.0 or 0.7 then choose 0.7. Fuzzy sets need more than one rule, Majozi & Zhu (2005) generally used three rules.

6.2.4 Defuzzify

Defuzzify means: to apply the implication or consequence. This is done by using the degree of support for the entire rule to shape the fuzzy set output. Mamdani & Assilian (1975) and Sugeno (1985) proposed two types of fuzzy inference systems (FIS) that are commonly used. The more popular of the two is Mamdani fuzzy inference (MFI), first proposed by Ebrahim Mamdani in 1975. He used the method to control a steam engine boiler by using linguistic control rules from experienced human operators in a machine controlled system. Mamdani based that work on Lofti

Zadeh's work which was published in 1973 (Zadeh 1973). Mamdani fuzzy inference expects the membership function to be fuzzy sets. After summation the output must be defuzzified. MFI finds the centroid of a 2D function. It uses a single output membership function because it greatly simplifies computation of MFI. Rather than integrate across the entire 2D function to find the centroid, MFI uses the weighted average of a few data points. Sugeno fuzzy inference is normally used to model systems where the output is linear or constant.

6.3 FUZZY SET THEORY FOR BUSHING EVALUATION

Fuzzy set theory is used to explore the interrelation between each bushing's identifying attributes, i.e. the dissolved gases in oil. In dissolved gas analysis (DGA) there is a relation between consequent failure and the simultaneous presence of oxygen with a secondary gas such as hydrogen, methane, ethane, ethylene, acetylene, and carbon monoxide in a bushing. The presence of combustible gases in the absence of oxygen is itself not an indication of eminent failure. Applying fuzzy sets on bushing data is necessary because the extent to which the evaluation criterion is below the threshold for a safe and acceptable, or rejected due risk of explosion, is not uniform for each bushing. This discrepancy can be accounted for in the evaluation process by applying fuzzy set theory. Temperature is an important criterion in the evaluation. Temperature refers both to the operating temperature of the oil and the difference between the ambient and the oil temperature. Bushings that continuously operate at temperatures near or above the auto-ignition temperature of any of the gases or oil have a significantly higher probability of explosion than those that operate at lower temperatures with the same ratio of gases. Auto-ignition temperature of a substance is the temperature at or above which a material will spontaneously ignite or catch fire without an external spark or flame.

Auto-ignition temperature should not to be confused with flash or fire points, which are generally a few hundred degrees lower. The flash point is the lowest temperature at which a liquid can form an ignitable mixture with air near the surface of the liquid. The lower the flash point, the easier it is to ignite the material. Fire point is the minimum sample temperature at which vapour is produced at a sufficient rate to sustain combustion. It is the lowest temperature at which the ignited vapour persists in burning for at least 5 seconds.

Flash point may be determined by the ASTM D 93 Method called "Flash Point by Pensky-Martens Closed Tester" for fuel oils. Alternatively, ASTM D 92 Method called "Flash and Fire Points by Cleveland Open Cup" can determine flash points of lubricating oils. At the fire point, the temperature of the flame becomes self-sustained so as to continue burning the liquid, while at the flash point the flame does not need to be sustained. The fire point is usually a few degrees above the flash point. Transformer oil which is used for both cooling and electrical insulation has characteristics as shown in Table 6.1.

Table 6.1: Properties of bushing oil
(IEC60867 1986)

Property	**Magnitude**
Boiling point	140°C
Vapour pressure	$0.1mbar$ *or* $10Pa$ *at* 20°C
Density	$840kg/m^3$ at 15°C
Specific gravity	0.8890
Solubility in H2O	Insoluble
Viscosity	7.7mm at 40°C
Flashpoint	156°C
Auto-ignition temperature	250°C

6.3.1 Identifying attributes

In this study, ten identifying attributes were selected to develop membership functions. These are concentrations of hydrogen, oxygen, nitrogen, methane, carbon monoxide, carbon dioxide, ethylene, ethane, acetylene and total dissolved combustibles gases. The concentrations are in parts per million (ppm). IEC60599 (1999) and IEEEC57-104 (1991) criteria were used in decision making.

6.3.2 Membership functions

Defining the membership functions (MF) is the most important step in fuzzy set theory application. This step takes the most time and must be accurate. One can use other MF curves such as a straight-line, Gaussian-bell, sigmoid, polynomial or a combination, if one can justify the decision after analysis of the data. Bojadziev & Bojadziev (1995) discussed that triangular functions accurately represent most

memberships. In general, triangular and trapezoidal membership functions are representative of most cases (Majozi & Zhu 2005), (Zadeh 1973). In this application the trapezoidal and triangular shapes of membership functions were selected to coincide with the safe operating limits for gas contaminants inside the bushing's oil. Each of the attributes is rated in terms of high, medium or low. The rating depends on the measured magnitude of the attribute compared to the reject threshold obtained in IEC60599 (1999) criteria. The membership functions (MF) are as given in Equations 6.1 to 6.30. The membership function for hydrogen described by equations 6.1 to 6.3 is shown in Figure 6.1. Equations 6.4 to 6.30 have similar membership functions.

Hydrogen

$$\mu_{Normal}(x) = \begin{cases} 1 & for\ 0 \leq x \leq 135 \\ -0.067x + 10 & for\ 135 \leq x \leq 150 \end{cases} \tag{6.1}$$

$$\mu_{Elevated}(x) = \begin{cases} 0.067x - 9 & for\ 135 \leq x \leq 150 \\ 1 & for\ 150 \leq x \leq 900 \\ -0.067x + 10 & for\ 900 \leq x \leq 1000 \end{cases} \tag{6.2}$$

$$\mu_{Dangerous}(x) = \begin{cases} 0.01x - 9 & for\ 900 \leq x \leq 1000 \\ 1 & for\ x \geq 1000 \end{cases} \tag{6.3}$$

Methane

$$\mu_{Normal}(x) = \begin{cases} 1 & for\ 0 \leq x \leq 23 \\ -0.5x + 12.5 & for\ 23 \leq x \leq 25 \end{cases} \tag{6.4}$$

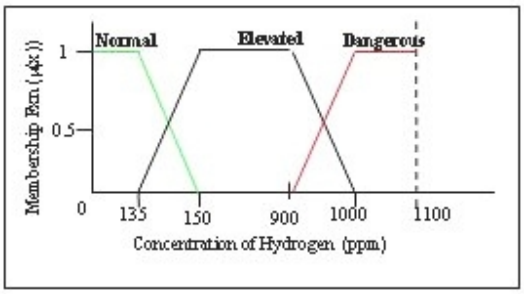

Figure 6.1: Membership functions of Hydrogen

$$\mu_{Elevated}(x) = \begin{cases} 0.05x - 11.5 & for\ 23 \leq x \leq 25 \\ 1 & for\ 25 \leq x \leq 72 \\ -0.125x + 10 & for\ 72 \leq x \leq 80 \end{cases} \tag{6.5}$$

$$\mu_{Dangerous}(x) = \begin{cases} 0.125x - 9 & for\ 72 \leq x \leq 80 \\ 1 & for\ x \geq 80 \end{cases} \tag{6.6}$$

Ethane

$$\mu_{Normal}(x) = \begin{cases} 1 & for\ 0 \leq x \leq 9 \\ -x + 10 & for\ 9 \leq x \leq 10 \end{cases} \tag{6.7}$$

$$\mu_{Elevated}(x) = \begin{cases} x - 9 & for\ 9 \leq x \leq 10 \\ 1 & for\ 10 \leq x \leq 32 \\ -0.333x + 11.66 & for\ 32 \leq x \leq 35 \end{cases} \tag{6.8}$$

$$\mu_{Dangerous}(x) = \begin{cases} 0.333x - 10.66 & for\ 32 \leq x \leq 35 \\ 1 & for\ x \geq 35 \end{cases} \tag{6.9}$$

Ethylene

$$\mu_{Normal}(x) = \begin{cases} 1 & for\ 0 \leq x \leq 18 \\ -0.5x + 10 & for\ 18 \leq x \leq 20 \end{cases} \tag{6.10}$$

$$\mu_{Elevated}(x) = \begin{cases} 0.5x - 9 & for\ 18 \leq x \leq 20 \\ 1 & for\ 20 \leq x \leq 90 \\ -0.1x + 10 & for\ 90 \leq x \leq 100 \end{cases} \tag{6.11}$$

$$\mu_{Dangerous}(x) = \begin{cases} 0.1x - 9 & for\ 90 \leq x \leq 100 \\ 1 & for\ x \geq 100 \end{cases} \tag{6.12}$$

Acetylene

$$\mu_{Normal}(x) = \begin{cases} 1 & for\ 0 \leq x \leq 14 \\ -x + 15 & for\ 14 \leq x \leq 15 \end{cases} \tag{6.13}$$

$$\mu_{Elevated}(x) = \begin{cases} x - 14 & for\ 14 \leq x \leq 15 \\ 1 & for\ 15 \leq x \leq 63 \\ -0.142857x + 10 & for\ 63 \leq x \leq 70 \end{cases} \tag{6.14}$$

$$\mu_{Dangerous}(x) = \begin{cases} 0.142857x - 9 & for\ 63 \leq x \leq 70 \\ 1 & for\ x \geq 70 \end{cases} \tag{6.15}$$

Carbon Monoxide

$$\mu_{Normal}(x) = \begin{cases} 1 & for\ 0 \leq x \leq 450 \\ -0.02x + 10 & for\ 450 \leq x \leq 500 \end{cases} \tag{6.16}$$

$$\mu_{Elevated}(x) = \begin{cases} 0.02x - 9 & for\ 450 \leq x \leq 500 \\ 1 & for\ 500 \leq x \leq 900 \\ -0.01x + 10 & for\ 900 \leq x \leq 1000 \end{cases} \tag{6.17}$$

$$\mu_{Dangerous}(x) = \begin{cases} 0.01x - 9 & for\ 900 \leq x \leq 1000 \\ 1 & for\ x \geq 1000 \end{cases} \tag{6.18}$$

Nitrogen

$$\mu_{Normal}(x) = \begin{cases} 1 & for\ 0 \leq x \leq 0.9 \\ -10x + 10 & for\ 0.9 \leq x \leq 1 \end{cases} \tag{6.19}$$

$$\mu_{Elevated}(x) = \begin{cases} 10x - 9 & for\ 0.9 \leq x \leq 1 \\ 1 & for\ 1 \leq x \leq 9 \\ -x + 10 & for\ 9 \leq x \leq 10 \end{cases} \tag{6.20}$$

$$\mu_{Dangerous}(x) = \begin{cases} x - 9 & for\ 9 \leq x \leq 10 \\ 1 & for\ x \geq 10 \end{cases} \tag{6.21}$$

Oxygen

$$\mu_{Normal}(x) = \begin{cases} 1 & for\ 0 \leq x \leq 0.09 \\ -100x + 10 & for\ 0.09 \leq x \leq 0.1 \end{cases} \tag{6.22}$$

$$\mu_{Elevated}(x) = \begin{cases} 100x - 9 & for\ 0.09 \leq x \leq 0.10 \\ 1 & for\ 0.10 \leq x \leq 0.18 \\ -50x + 10 & for\ 0.18 \leq x \leq 0.20 \end{cases} \tag{6.23}$$

$$\mu_{Dangerous}(x) = \begin{cases} 50x - 9 & for\ 0.18 \leq x \leq 0.20 \\ 1 & for\ x \geq 0.20 \end{cases} \tag{6.24}$$

Carbon Dioxide

$$\mu_{Normal}(x) = \begin{cases} 1 & for\ 0 \leq x \leq 9000 \\ -0.001x + 10 & for\ 9000 \leq x \leq 10000 \end{cases} \tag{6.25}$$

$$\mu_{Elevated}(x) = \begin{cases} 0.001x - 9 & for\ 9000 \leq x \leq 10000 \\ 1 & for\ 10000 \leq x \leq 13500 \\ -0.00067x + 10 & for\ 13500 \leq x \leq 15000 \end{cases} \tag{6.26}$$

$$\mu_{Dangerous}(x) = \begin{cases} 0.00067x - 9 & for\ 13500 \leq x \leq 15000 \\ 1 & for\ x \geq 15000 \end{cases} \tag{6.27}$$

Total Combustible Gases

$$\mu_{Normal}(x) = \begin{cases} 1 & for\ 0 \leq x \leq 648 \\ -0.01389x + 10 & for\ 648 \leq x \leq 720 \end{cases} \tag{6.28}$$

$$\mu_{Elevated}(x) = \begin{cases} 0.01389x - 9 & for\ 648 \leq x \leq 720 \\ 1 & for\ 720 \leq x \leq 4500 \\ -0.002x + 10 & for\ 4500 \leq x \leq 5000 \end{cases} \tag{6.29}$$

$$\mu_{Dangerous}(x) = \begin{cases} 0.002x - 9 & for\ 4500 \leq x \leq 5000 \\ 1 & for\ x \geq 5000 \end{cases} \tag{6.30}$$

6.3.3 Fuzzy rules

Fuzzy rules represent interrelation between all the inputs. The number of rules is theoretically equal to the number of fuzzy categories raised to the power of the number of fuzzy criteria. Fuzzy categories (FC) used in this case were, the membership functions *dangerous*, *elevated* or *normal*. Fuzzy criteria (NC) used in this case were the different gases that are present, i.e. hydrogen, methane, ethane, ethylene, acetylene, carbon monoxide, nitrogen, oxygen, carbon dioxide and total combustible gases. The rates of change of the gases were not used because the available data is taken on one day only. The required number of fuzzy rules is calculated according to Equation 6.31. Rules have an antecedent and a consequence. Rules can be expressed in the form:

IF $Attribute_1$ is $A1$ AND $Attribute_2$ is A_2 AND ... AND $Attribute_N$ is A_N, THEN Consequent is C_i,

In the expression, $Attribute_1$, $Attribute_2$, ..., $Attribute_N$ collectively form an Antecedent. Antecedents and Consequents are variables or concepts and A_1, A_2; ..., C_i are linguistic terms or fuzzy sets of these variables, such as, *low*, *dangerous* or *high* (Bandemer & Gottwald 1995).

$$N_{rules} = (FC)^{NC} \tag{6.31}$$

$$N_{rules} = (3)^{10} = 59049 \tag{6.32}$$

6.3.4 Simplification of fuzzy rules

In the case of bushing diagnosis the combinations of the combustible gases in the absence of oxygen does not create a failure. With transformer oil, failure occurs when oxygen is present in quantities above 0.2% at temperatures above 250°C without any spark present (auto-ignition) or at 156°C if a spark is present (flash point). This condition reduces the number of fuzzy rules significantly, to only 81 fuzzy rules. The combinations are modelled in 24 compartments shown in Table 6.3 and Table 6.4. Two examples of fuzzy rules in spoken language (as opposed to machine language) are (1) If hydrogen is High only then Low Risk and (2) If hydrogen is High and Oxygen is High then High Risk.

6.3.5 Consequence table or decision table

Based on the rules the bushing is given a risk rating for which certain maintenance actions must be taken on the plant. For safe operation of bushings it is recommended that for all HR cases, trip the transformer and remove the bushing from the transformer. For all MR cases monitor the bushings more frequently, i.e. reduce the sampling interval by half. All LR cases operate as normal. From the decision table an aggregated membership is developed, shown in Equations 6.33 and 6.34 (Bandemer & Gottwald 1995).

$$\mu_{agg} = \mu_{LR} \cup \mu_{MR} \cup \mu_{HR} \tag{6.33}$$

where,

$$\mu_{LR}(x) = \begin{cases} 1 & for\ x \leq 10 \\ -0.01667x + 1 & for\ 10 \leq x \leq 60 \end{cases} \tag{6.34}$$

$$\mu_{MR}(x) = \begin{cases} 0.01667x - 1 & for\ 10 \leq x \leq 60 \\ -0.05x + 4 & for\ 60 \leq x \leq 80 \end{cases} \tag{6.35}$$

Table 6.2: Conclusion table

x	LR	MR	HR
0	1		
10	1	0	
60	0	1	0
80		0	1
100			
	Group A	**Group B**	**Group C**

$$\mu_{HR}(x) = \begin{cases} 0.05x - 3 & for\ 60 \leq x \leq 80 \\ 1 & for\ x \leq 80 \end{cases} \tag{6.36}$$

The decision table and a graph of the membership functions are shown in Figure 6.2 and Figure 6.3 respectively. Table 6.2 is the conclusion table, which shows values and classes. The values of 10, 60 and 80 were selected to represent the levels of risk of failure of a bushing. These levels were then taken as the limits of each of the groups in the conclusion membership function.

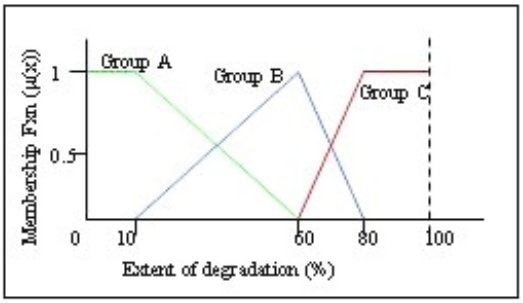

Figure 6.2: Membership functions of decision

The membership function is asymmetrical so that a decision to exclude damaged bushings is more stringent than that of marginally safe bushings. In other words, small changes in a condition that is becoming dangerous are highlighted by the

membership function. A steeper gradient on the graph allows the user to identify those components which have small differences in dangerous levels of concentrations of dangerous gases.

6.4 RESULTS OF FUZZY SET APPLICATION

Fuzzy set theory was applied to ten bushings. The fuzzy rules were applied to each bushing. For each rule, the truth value of the consequence is the minimum membership value of the antecedent. The degrees of membership of the other gases are shown in Table 6.5. The aggregated result is shown in Table 6.3 and Table 6.4.

6.4.1 Aggregated Rules

The table of fuzzy rules can be simplified further by finding cells with common features within compartments. This process is called aggregating. One can develop the following aggregated rules (AR) based on the highlighted compartment in Table 6.3 and Table 6.4:

(AR4) IF bushing has *Dangerous level of TDCG* AND *NOT Normal Oxygen* AND *Not Normal Methane*, THEN the bushing belongs to Group A (*high risk* or *dangerous*). (AR5) IF bushing has *Dangerous TDCG* AND *NOT Normal Oxygen* AND *Normal Methane* THEN the bushing belongs to Group B (*medium risk* or *elevated*). (AR6) IF bushing has *Dangerous TDCG* AND *Normal Oxygen* AND *Not Normal Methane*, THEN the bushing belongs to Group B (*medium risk* or *elevated*). (AR7) IF bushing has *Dangerous TDCG* AND *Normal Oxygen* AND *Normal Methane*, THEN the bushing belongs to Group C (*low risk* or *normal*).

In rule AR1, the consequence is 'the bushing belongs to Group A'. The truth value of this consequence (CAR4) is shown in Equation 6.37.

$$CAR_4 = min(1, 1, 1) = 1 \tag{6.37}$$

where, the values in the *min* function are obtained as follows: The first 1 is the degree of membership of TDCG for bushing #200323106 in the set 'Dangerous'. The second 1 is the degree of membership of NOT Normal Oxygen for bushing #200323106 is in the set 'NOT normal' is 1, which is obtained by subtracting the

Table 6.3: Fuzzy decision table - Part I

Oxygen	Total Dissolved Combustible Gases (TDCG)											
	Dangerous											
	Hydrogen			Methane			Ethane			Acetylene		
	Dangerous	Elevated	Normal	Dangerous	Elevated	Normal	Dangerous	Elevated	Normal	Dangerous	Elevated	Normal
Dangerous	HR	HR	MR	HR	HR	MR	HR	HR	MR	HR	HR	MR
Elevated	HR	HR	MR	HR	HR	MR	HR	HR	MR	HR	HR	MR
Normal	MR	MR	LR	MR	MR	LR	MR	MR	LR	MR	MR	LR

Oxygen	Total Dissolved Combustible Gases (TDCG)											
	Dangerous											
	Carbon Monoxide			Nitrogen			Carbon Dioxide			Ethylene		
	Dangerous	Elevated	Normal	Dangerous	Elevated	Normal	Dangerous	Elevated	Normal	Dangerous	Elevated	Normal
Dangerous	HR	HR	MR	MR	HR	HR	MR	MR	MR	HR	HR	MR
Elevated	HR	HR	MR	MR	HR	HR	MR	MR	LR	HR	HR	MR
Normal	MR	MR	LR	LR	LR	LR	LR	LR	LR	MR	MR	LR

Oxygen	Total Dissolved Combustible Gases (TDCG)											
	Elevated											
	Hydrogen			Methane			Ethane			Acetylene		
	Dangerous	Elevated	Normal	Dangerous	Elevated	Normal	Dangerous	Elevated	Normal	Dangerous	Elevated	Normal
Dangerous	HR	HR	MR	HR	HR	MR	HR	HR	MR	HR	HR	MR
Elevated	HR	HR	MR	HR	HR	MR	HR	HR	MR	HR	HR	MR
Normal	MR	LR	LR	MR	LR	LR	MR	LR	LR	MR	LR	LR

Table 6.4: Fuzzy decision table - Part II

Oxygen	Total Dissolved Combustible Gases (TDCG)											
	Elevated											
	Carbon Monoxide			Nitrogen			Carbon Dioxide			Ethylene		
	Dangerous	Elevated	Normal	Dangerous	Elevated	Normal	Dangerous	Elevated	Normal	Dangerous	Elevated	Normal
Dangerous	HR	HR	MR	MR	HR	HR	MR	MR	MR	HR	HR	MR
Elevated	HR	HR	MR	MR	HR	HR	MR	MR	LR	HR	HR	MR
Normal	MR	LR	LR	LR	LR	LR	LR	LR	LR	MR	LR	LR

Oxygen	Total Dissolved Combustible Gases (TDCG)											
	Normal											
	Hydrogen			Methane			Ethane			Acetylene		
	Dangerous	Elevated	Normal	Dangerous	Elevated	Normal	Dangerous	Elevated	Normal	Dangerous	Elevated	Normal
Dangerous	HR	HR	MR	HR	HR	MR	HR	HR	MR	HR	HR	MR
Elevated	HR	HR	MR	HR	HR	MR	HR	HR	MR	HR	HR	MR
Normal	LR	LR	LR	LR	LR	LR	LR	LR	LR	LR	LR	LR

Oxygen	Total Dissolved Combustible Gases (TDCG)											
	Normal											
	Carbon Monoxide			Nitrogen			Carbon Dioxide			Ethylene		
	Dangerous	Elevated	Normal	Dangerous	Elevated	Normal	Dangerous	Elevated	Normal	Dangerous	Elevated	Normal
Dangerous	HR	HR	MR	LR	MR	MR	MR	MR	MR	HR	HR	MR
Elevated	HR	HR	MR	LR	MR	MR	MR	LR	LR	HR	HR	MR
Normal	LR	LR	LR	LR	LR	LR	LR	LR	LR	LR	LR	LR

degree of membership of NOT Normal Oxygen in the set 'Normal', i.e. 0, from 1. The third 1 is the degree of membership of 'NOT Normal Methane' for bushing #200323106 in the set 'Normal' i.e. 0, from 1. Note that the 'NOT' operator requires that the corresponding degree of membership be subtracted from 1. An 'ANY' term entails summing of all the degrees of membership of a particular quality, e.g. acetylene, or TDCG in different corresponding sets. For example, the condition 'ANY Level of TDCG' has a degree of membership of 1. This is obtained by summing the degrees of membership of TDCG for bushing #200323106 in the sets 'Dangerous' (1), 'Elevated' (0) and 'Normal' (0) as shown in Table 6.3.

Table 6.3 is only a subset of the fuzzy decision table. It is presented in this document in order to highlight how the fuzzy rules are implemented. The application of the rules in all the other compartments follows the same pattern. Since different rules can result in the same conclusion or consequence, the truth values of a particular consequence will vary according to the rule applied to the bushing.

Once all the rules have been applied to a particular bushing, and different truth values of each consequence obtained, the maximum value of each consequence among all the rules that result in that consequence, is taken as the degree to which that consequence applies to a given bushing. This eventually gives rise to an aggregated fuzzy output as shown in Table 6.6 and Equation 6.38 (Bandemer & Gottwald 1995).

$$AGD_i = max(CAR1_i \cap CAR2_i \cap \cdots \cap CARn_i) \tag{6.38}$$

where, AGD_i is the aggregated decision for category i, e.g. group HR, CAR_i is the consequence of aggregated rules in a particular category i, in a certain compartment. i is the number of categories, in this case the categories are HR, MR and LR.

According to Table 6.6, bushing #200323106 belongs to Group C (low risk or normal), it also belongs to Group B (elevated or medium risk) as well as Group A (high risk or dangerous) to degrees 1, 1 and 1, respectively. This step is the end of the fuzzification steps. The membership value 1 in the HR category indicates that the degradation is severe but does not indicate how bad it is. It can be anywhere on the highlighted line in the MF curve. In the case of bushing #200323106, the position on the graph is insignificant because the value is already in the saturation region. But if the bushing had lower degrees of membership, (i.e. less than 1) in Group

Table 6.5: Memberships of gases in bushing #200323106

Gas	Quantity	Membership function	Degree of Membership
Acetylene	0	Normal	1
		Elevated	0
		Dangerous	0
Carbon Dioxide	72	Normal	1
		Elevated	0
		Dangerous	0
Carbon Monoxide	44	Normal	1
		Elevated	0
		Dangerous	0
Ethane	22	Normal	0
		Elevated	1
		Dangerous	0
Ethylene	2	Normal	1
		Elevated	0
		Dangerous	0
Hydrogen	5782	Normal	0
		Elevated	0
		Dangerous	1
Methane	240	Normal	0
		Elevated	0
		Dangerous	1
Nitrogen	4.58	Normal	1
		Elevated	1
		Dangerous	0
Oxygen	0.2535	Normal	0
		Elevated	0
		Dangerous	1
TDCG	6090	Normal	0
		Elevated	0
		Dangerous	1

Table 6.6: Aggregated output for bushing #200323106

Category	Degree of membership
Group C (HR)	1
Group B (MR)	1
Group A (LR)	1

B or C, then one would be able to determine the extent of degradation and thus make an informed maintenance decision of a likely time to replace the bushing. To determine the extent of degradation one needs to defuzzify the membership function into crisp data. The crisp result is useful for determining the degree of degradation, and not only for informing of rejection or acceptance of a component. The crisp result takes into account the degrees of membership in all the groups in the decision truth table. In other words, the crisp result of the fuzzy analysis is useful because it can be used by the maintenance department to make informed decisions about spares level at any given time and deciding the maintenance action and the timing of the maintenance action, based on the current state of the bushing and the rate of change in that state. To quantify the extent of damage, the fuzzy information needs to be ranked to give crisp data.

6.5 DEFUZZIFICATION

Defuzzification is aimed at converting fuzzy information into crisp data. The method used for defuzzification in this case is called the weighted average of maximum values of membership functions method used by Flaig et al. (2000) and Majozi & Zhu (2005). The method was selected because it is effective and computationally inexpensive. The result from the application of this method gives the rank or level of risk of each bushing. For bushing #200323106 with an aggregated output is shown in Table 6.3, the rank is obtained using Equation 6.39. Figure 6.3 shows the aggregated membership function from which the values for Equation 6.39 are taken (Majozi & Zhu 2005).

$$Rank(x) = \frac{\left(\frac{0+10}{2}\right) \cdot \mu(Group_A(x)) + 60 \cdot \mu(Group_B(x)) + \left(\frac{80+100}{2}\right) \cdot \mu(Group_C(x))}{\mu(Group_A(x)) + \mu(Group_B(x)) + \mu(Group_c(x))} \tag{6.39}$$

where x=200323106

The coefficients appearing in Equation 6.39 are the levels of risk of failure corresponding to the maximum values, i.e. 1, of the respective sets as shown in the conclusion table, for example, a risk of rating of 60 corresponds with the maximum value of the membership function of set B. In case there is a flat, as in the set A membership function as well as set C membership function, an average value of the extreme values at the maximum is used as a coefficient, e.g. (80+100). Thus the

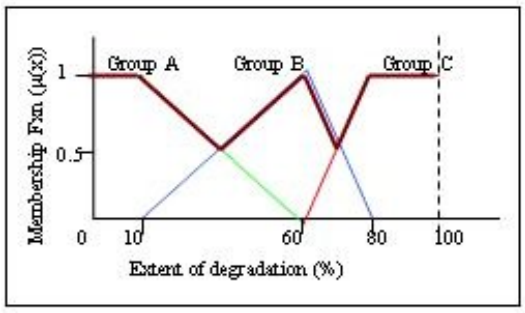

Figure 6.3: Aggregated output for bushing #200323106

solution for the rank of bushing #200323106 is shown to be 51.66, calculated as shown below.

$$Rank_{(200323106)} = \frac{\left(\frac{0+10}{2}\right) \cdot (1) + 60 \cdot (1) + \left(\frac{80+100}{2} \cdot (1)\right)}{(1) + (1) + (1)} = 51.66$$

Action is taken according to the crisp result. Bushings with a value of more than 30 are removed from service. Between 10 and 30 the interval of monitoring is halved or the frequency is doubled. Below 10 the bushing is left to operate as normal. Clearly bushing #200223106 with a crisp output value of 51.66 should be removed from service. Table 6.7 shows the results of all ten bushings.

Table 6.7 shows how a multilayer perceptron (MLP) classified the same bushings that were used to demonstrate the application of fuzzy inference. The two methods show similar levels of accuracy. Evaluation of the crisp result from the fuzzy analysis showed no false acceptance rate, i.e. 100% accuracy. A neural network was able to classify the crisp data, using the criteria of $x > 30$ for reject, $x < 30$ for accept. Because the neural network in the second case used results from a fuzzy analysis it is called a neuro-fuzzy system. The neural network (ANN) classified data directly from the DGA gas chromatography sheet using a multilayer perceptron with 7 hidden neurons, as done previously by Dhlamini, Marwala & Majozi (2006). The manual method used an experienced maintenance operator, who is supposed to be 100% accurate. The results prove that ANN and neuro-fuzzy have similar levels of accuracy (90%). While the purely fuzzy method showed 100% accuracy, ANNs are fast and efficient, taking 1.35s to train and classify the data compared to 30 minutes

for the fuzzy set system and the neuro-fuzzy system, compared to 5 minutes for the manual method of classification of 10 bushings.

Table 6.7: Classification of bushings using neuro-fuzzy network

Bushing	**Rank**	**Fuzzy**	**Neurofuzzy**	**NN**	**Human**
200323106	51.666667	Reject	Reject	Reject	Reject
200373387	60	Reject	Reject	Reject	Reject
200323104	32.5	Reject	Reject	Accept	Reject
200323105	32.5	Reject	Accept	Reject	Reject
200302381	60	Reject	Reject	Reject	Reject
200355292	5	Accept	Accept	Accept	Accept
200367794	5	Accept	Accept	Accept	Accept
200378937	5	Accept	Accept	Accept	Accept
200328202	5	Accept	Accept	Accept	Accept
	Accuracy	100%	90%	90%	100%

6.6 CONCLUSIONS

The method of using fuzzy set theory (FST) compares well with the method of diagnosing using neural networks. Fuzzy inference system (FIS) tells maintenance personnel whether or not there is damage and also how severe the damage is, thus helping to make operational decisions of whether the bushings should be replaced or remain in service. The benefit of using FIS over neural networks is that it allows the user to evaluate the extent of damage more objectively and comprehensively. The crisp result from fuzzy analysis is useful for determining equipment spares levels at any given time and deciding the maintenance action. The neural network (NN) classified data directly from the DGA gas chromatography sheet using a multilayer perceptron with 7 hidden neurons, as done previously by Dhlamini & Marwala (2005). The manual method used an experienced maintenance operator, who was 100% accurate. The results prove that NN and neuro-fuzzy have similar levels of accuracy (90%). While the purely fuzzy method showed 100% accuracy, NN are fast and efficient, taking 1.35s to train and classify the data compared to 30 minutes for the fuzzy set system and the neuro-fuzzy system, compared to 5 minutes for the manual method of classification of 10 bushings.

Chapter 7

DIMENSION REDUCTION AND REDUNDANCY AND RELEVANCE DETERMINATION

7.1 INTRODUCTION

Most statistical methods of data analysis require that observations conform to a particular distribution, generally this is *normal distribution* (Reyment & Joreskog 1993). If that were the case, then it would greatly simplify extraction of latent variables to reduce data dimensionality. Latent variables are probabilistic models that try to explain high dimensional data in a summarised form with fewer degrees of freedom (Bartholomew 1987). Mathematical methods for data reduction must be used because often data has a skewed distribution, and often the system is not well known to the person performing the analysis. The data analyst often does not have experience and theoretical knowledge particular to that system to be able isolate the important variables of the system. To reduce complexity of raw data, mathematical methods of summarising, delineating and investigating interrelations between the data points are available. To simplify data Reyment & Joreskog (1993) and Rummel (1970) used Principal Component Analysis (PCA), Multi-Dimensional Scaling (MDS) and Factor Analysis (FA). Carreira-Perpin (2001) used PCA, FA, Independent component analysis (ICA), an autoassociative network (autoencoder) and generative topographic mapping to reduce dimensionality of electropalatographic data. Han & Kamber (2001) recommended PCA as an effective method of dimension reduction of customer data in a retail database. Mdlazi et al. (2003) used ARD and PCA to reduce the dimension of vibration data. Bishop (1995) used PCA as a

linear dimension reduction transformation of data, prior to classification by neural networks. Reyment & Joreskog (1993) used FA and PCA to identify important variables in the analysis of mineral chemistry, maturation of crude oil and sedimentology.

This section investigated methods with applications across a wide spectrum of diverse data sets to reduce the dimensionality of bushing DGA data. The methods are PCA, MDS, FA and an autoencoder (AE) with a genetic algorithm (GA). Automatic Relevance Determination (ARD) was used to detect relevance of variables used to diagnose high voltage bushings. The classification is done using Dissolved Gas Analysis (DGA) data based mainly on IEEEC57-104 (1991) and IEC60599 (1999). Figure 7.1 shows the process that was followed. After the data was preprocessed, the bushings were classified using a trained multilayer perceptron (MLP) neural network. The results of classification were compared. Other methods of data reduction such as data compression, numerosity reduction, discretization and concept hierarchy generation were not applied (Han & Kamber 2001). Data reduction by data cube aggregation, was done to a limited extent through usage of the TDCG variable which combined several of the gas variables into one.

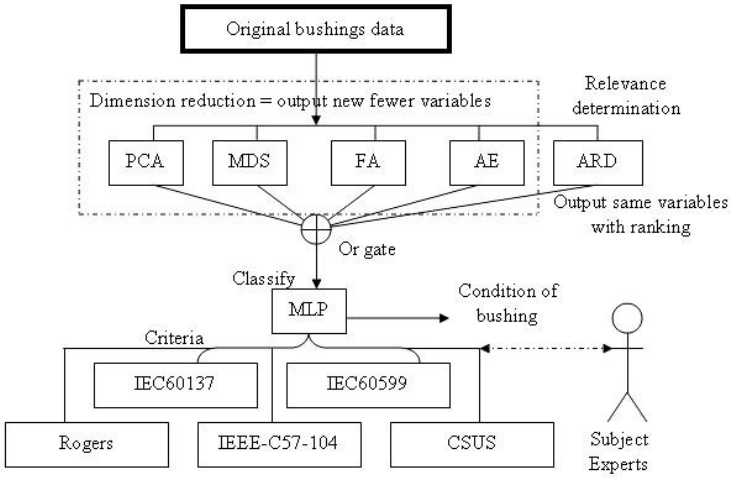

Figure 7.1: Data reduction and classification

Reducing the size of the data while maintaining the important decision-making information means that fewer sensors may be used to monitor the health of a system, possibly reducing the cost of transducers. Fewer sensors also mean faster computational time or less computation power, thus allowing the system to be simpler with fewer components and higher reliability.

7.2 METHODS FOR DIMENSION REDUCTION AND REDUNDANCY DETECTION

7.2.1 Principal component analysis

PCA is also called the Karhunen-Loeve transform (Jolliffe 1986). It is the most commonly used method for dimension reduction. For any data set $x_i, ...x_n$, the principal components are calculated by determining the mean (μ) and covariance (K) as in Equation 7.1 and 7.2 (Han & Kamber 2001).

$$\mu = \left(\frac{1}{n}\right)\sum_{i=1}^{n} x_i \tag{7.1}$$

$$K = \left(\frac{1}{n}\right)\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T \tag{7.2}$$

Next the eigenvalues (λ) and eigenvectors (e) for the covariance matrix K are calculated by solving the equality $Ke = \lambda e$. Dimension reduction occurs by ignoring the directions in which the covariance (K) is small. The eigenvalue (λ) measures the variation in the direction of the eigenvector (e). The eigenvalues (λ) are then ranked in order of magnitude $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \lambda_n$. For many datasets, most of the eigenvalues (λ) are negligible and can be discarded. PCA keeps the eigenvectors associated with the largest eigenvalues. The threshold for magnitudes of the acceptable eigenvalues is determined by the user. Dimensionally reduced data consists of all the data $x_i \cdots x_n$ projected onto the selected eigenvectors (e_i),

$$x \propto \mu + \sum_{i=1}^{m} a_i e_i,$$

where

$$a_i = (x_i - \mu) \cdot e_i^T$$

where a_i are the projection coefficients of the data vectors (x_k) onto the eigenvectors (e_i) and m is the number of accepted eigenvalues. Each principal component is a linear combination of the original variables. All the principal components are

orthogonal to each other so there is no redundant information. The cost function that is optimized in PCA is shown in Equation 7.3 (Han & Kamber 2001).

$$\jmath(\mu, a, e,) = \sum_{k=1}^{n} \left\| \left(\mu + \sum_{i=1}^{m} a_{ki} e_i \right) - x_k \right\| \tag{7.3}$$

(Han & Kamber 2001) highlighted two limitations of PCA. (1) PCA may not find the best directions for discriminating between two classes if the data is in parallel separated clusters, because only one principal component will represent all the data, e.g. if two classes have 2D Gaussian densities as ellipsoids; (2) PCA is not effective for some datasets, e.g. if the data is a set of strings (1,0,0,0,$\cdots$),(0,1,0,0,$\cdots$) ,$\cdots$,(0,0,0,$\cdots$,1) then the eigenvalues do not fall off as PCA requires. PCA perform linear dimension reduction using the sum-of-squares error criterion, the same as MLP neural networks (Bishop 1995). Dimensionality reduction based on covariance measures and population entropy give the same results as PCA (Fukunaga 1990). PCA is not able to summarize non-linear data or data with intrinsic dimensionality (Bishop 1995).

7.2.2 Factor analysis

Factor analysis is a global term used to describe several methods that were developed to analyse the interrelation between variables in a data set. These methods include PCA, true factor analysis (FA), canonical correlation analysis, principal coordinate analysis, R-mode analysis, Q-mode analysis and QR-mode analysis (Reyment & Joreskog 1993). The common feature of all the methods is that they can construct a set of variables or factors into a smaller number of variables or factors. The factors are constructed such that the overall complexity of the original data is reduced by using the interdependencies found in the original data set. Factor analysis is a form of multivariate data reduction. In PCA factors are determined so as to allow for a maximum variance between all variables. Everitt & Dunn (1991) observed that factor analysis factors are defined to allow for a maximum correlation of the variables, in contrast to PCA which only summarises data using fewer dimensions.

Factor analysis uses a Gaussian distribution prior and noise model, as well as linear mapping from data space to transform to latent space (Everitt & Dunn 1991). The latent space prior $p(x)$ is unit normal (Bartholomew 1987):

$$x \sim \aleph(0, I) \tag{7.4}$$

where x represents the input variables. $\aleph(0, I)$ means normal distribution of the input. The latent variables of x are referred to as factors. x is mapped onto the latent space by the function in Equation 7.5.

$$f(x) = \Lambda x + \mu \tag{7.5}$$

where Λ is a $D \times L$ matrix whose columns are called $factor\ loadings$, by assuming $rank\ (\Lambda) = L$ i.e. linearly independent factors.

The data space noise model is normal centered at $f(x)$ with diagonal covariance matrix (ψ):

$$t|x \sim \aleph\left(f(x), \psi\right) \tag{7.6}$$

where the diagonal elements of ψ are referred to as the *uniquenesses*.

The marginal distribution in data space can be computed analytically to be normal with constrained covariance matrix given by:

$$t \sim \aleph\left(\mu, \Lambda\Lambda^T + \psi\right) \tag{7.7}$$

The posterior in the latent space is also normal:

$$x|t \sim \aleph\left(A(t-\mu), (I + \Lambda^T\psi^{-1}\Lambda)^{-1}\right) \tag{7.8}$$

$$A = \Lambda^T\left(\Lambda\Lambda^T + \psi\right)^{-1} = \left(I + \Lambda^T\psi^{-1}\Lambda\right)^{-1}\Lambda^T\psi^{-1} \tag{7.9}$$

The reduced dimension representative is taken as the posterior mean, coinciding with the mode, commonly referred to the *Thomson scores*.

$$F(t) = E(x|t) = A(t-\mu) \tag{7.10}$$

The dimensionality reduction mapping F is linear and smooth (Carreira-Perpin 2001). If an invertible linear transformation g with matrix R is applied to the factors x, then a new set of factors is obtained $y = Rx$. The prior distribution $p(y)$ is still normal, and the new mapping becomes $t = \dot{f}(y) = f(g^{-1}(y)) = \Lambda\ R^{-1}\ \ y+\mu$. The new *factor loadings* become $\Lambda\ = \Lambda\ R^{-1}$. If R is an orthogonal matrix, i.e. $R^{-}1{=}R^{T}$, then the new factors (y) will still be independent and $\dot{\psi}{=}\psi$. This is called *orthogonal rotation* of factors. If R is an arbitrary nonsingular matrix, the new factors (y) will not be *independent*. This is called *oblique rotation* of factors.

With FA the principal factors are rotated to achieve maximum variance in the factor loading. In other words, the principal components need not always be orthogonal. Using the orthomax rotation function gives orthogonal principal components. Non-orthogonal rotation functions include varimax, procrustes, and promaxpm. For this work, a varimax rotation function was used, because it is the most commonly used rotation function when applying FA (Lawley & Maxwell 1971).

7.2.3 Independent component analysis

Welling & Weber (1999) established that ICA searches for directions in data-space, which are independent across all statistical orders. Independent component analysis (ICA) has the following formulation (Carreira-Perpin 2001):

1. Call $x(\tau)$ in R^{L}, the L time-varying input or $x(x, y)$ space-varying input. τ may be continuous or discreet time variable. Assume that the input has a zero mean, and all inputs are independent.

2. Call $t(\tau)$ *in* R^{D} the D data input measured without noise and with instantaneous mixing. So that there is no time delay between source l into channel d.

3. Then $t(\tau) = \Lambda\ x(\tau)$,
   where $\Lambda_{D\times L}$ is the mixing matrix, and $D \geq L$

4. The goal of ICA is to find a linear transformation $A_{L\times D}$ *separation matrix* of the sensor signals t that makes the outputs $u(\tau) = At(\tau) = A\Lambda x(\tau)$ as independent as possible (Bell & Sejnowski 1995).

ICA is related to principal component analysis and factor analysis, yet ICA is a much more powerful technique, because ICA is capable of finding the underlying factors or sources when the classic methods such as PCA and FA fail completely (Welling & Weber 1999).

7.2.4 Autoencoder

Autoencoders are networks which are trained to recall their inputs (Bourlard & Kamp 1988). A multilayer perceptron configured as an autoassociative neural network was used for dimension reduction. Autoencoders overcome the limitation of PCA, by performing dimension reduction on all types of data, including non-linear data (Bishop 1995). The inputs are used as the targets to train the neural network. The hidden layer neurons are fewer than the input vectors so it is not possible to reconstruct all the outputs from the inputs (Rumelhart et al. 1986). In this way the less relevant inputs are eliminated. Figure 7.2 shows the structure of the network.

The network is trained by minimising the sum-of-square errors shown in Equation 7.11 (Rumelhart et al. 1986).

$$\varepsilon = \frac{1}{2}\sum_{n=1}^{m}\sum_{k=1}^{d}[y_k(x^n) - x_k^n]^2 \tag{7.11}$$

where x_k is the actual target, $y_k(x^n)$ is the target predicted by the neural network and d is the total number of inputs, which is the same as the targets. m is the number of hidden neurons ($m < d$).

The error minimisation used by the autoenconder is classified as unsupervised training because, there are no independent targets. Bourlard & Kamp (1988) showed that even if non-linear activation functions such as sigmoidal functions are used inside the hidden neurons, there is no advantage in using neural networks over PCA for dimension reduction. However if more than one hidden layer is used with non-linear activation functions, then non-linear principal component analysis can be performed by the network (Kramer 1991).

7.2.5 Multi-dimensional scaling

Multi-Dimensional Scaling (MDS) or Principal Coordinates Analysis is used to detect underlying dimensions that allow one to explain observed similarities or dissim-

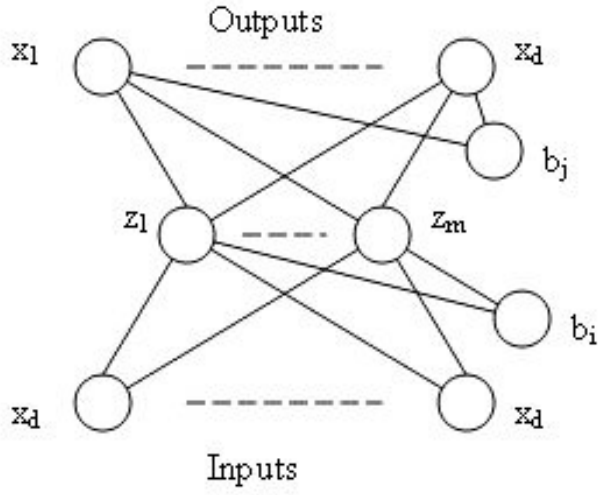

Figure 7.2: Autoassociative multilayer perceptron

ilarities between the data points. MDS is comparable with FA, the only difference being that FA tends to extract more factors than MDS, forcing the user manually to interpret the results; as a result, MDS often yields more useful solutions (Seber 1984). MDS produces a representation of the original data in a smaller number of dimensions.

7.2.6 Redundancy detection

Recall from Chapter 3 that a variable is redundant if it can be derived from another variable in the same database (Han & Kamber 2001). Redundancy determination is commonly conducted using correlation analysis shown in Equation 7.12 (Reyment & Joreskog 1993). The correlation coefficient (r) is the measure of correlation. This measure is defined as the ratio of the covariances of the two variables to the product of their standard deviations (Reyment & Joreskog 1993):

$$r_{ij} = \frac{s_{ij}}{s_i s_j} \tag{7.12}$$

where r_{ij} is the correlation coefficient, s_i and s_j are the standard deviation of variables x_i and x_j, N is the sample size and s_{ij} is the covariance.

The correlation coefficient is more generally written as in Equation 7.13:

$$r_{ij} = \frac{\sum_{n=1}^{N} y_{ni} \cdot y_{nj}}{\left(\sum_{n=1}^{N} y_{ni}^2 \cdot \sum_{n=1}^{N} y_{nj}^2 \right)^{1/2}} \tag{7.13}$$

where y_{nj} is the difference between a variable x_{nj} and the mean of that variable x_j, y_{ni} is the difference between a variable x_{ni} and the mean of that variable x_i.

If the correlation coefficient is greater than 0, then there is positive correlation between variables x_i and x_j (Han & Kamber 2001). Meaning that the value of x_i increases as x_j increases. A larger correlation coefficient means close relation between the variables. As a result x_i or x_j should be removed because it is redundant. If r_{ij} is 0 then the variables are independent. If r_{ij} is less than 0 the variables are negatively correlated, i.e. as one increases the other decreases.

7.2.7 Automatic relevance determination

Automatic relevance determination (ARD) finds the variables which represent the system. Mdlazi et al. (2003) explain that the basis of ARD is that each input has a prior variance. Large variances suggest that all weights associated with that input are large, and therefore that input is important. MacKay (1994) indicates that irrelevant inputs have small variances and are not important for explaining the data. A multilayer perceptron was used, with re-estimation of the hyper parameters. The final values for the hyperparameters reflect the relative importance of the ten inputs.

7.3 RESULTS OF DIMENSIONALITY REDUCTION AND CLASSIFICATION

All the methods were able to prune the redundant measured variables. With PCA, FA and MDS transformation from the data space to latent space resulted in a reduction from ten to as few as four variables. These methods, however were not able to indicate which of the inputs is more important than others. The variable TDCG which is referred to in most of the standards are an aggregation dimension reduction because it includes most of the individual gases which are also used in determining

a bushings condition. As a result it is expected that with fewer variables used in the classification the results should be sufficiently closely aligned with a decision made using complete input data. ARD was able to identify six important variables to be TDCG, Nitrogen, Carbon Monoxide, Ethylene, Hydrogen, and Ethane. Table 7.1 gives the results. The accuracy of classification using only 6 variables is 76% compared to 97% when using all 10 variables.

Table 7.1: Dimension reduction results

Method	**No. latent variables**	**Classification**
Original	10	97%
PCA	4	52%
FA	4	47%
ARD	6	76%
MDS	6	61%
AE	5	54%

7.4 CONCLUSIONS

All the methods were able to prune the redundant measured variables. The difference in rankings or levels of importance of the different gases as highlighted by the different methods is due to the fact that some methods use weights to correlate the data and then rank the importance depending on the magnitude of the weights, while other methods use variance in the actual data to decide on the importance of the variables. Ten input data points were reduced to six using automatic relevance determination. The accuracy of classification using only 6 variables is 76% compared to 97% when using all 10 variables. The six relevant variables are carbon dioxide, oxygen, total dissolved combustible gases, nitrogen, hydrogen and carbon monoxide. The work finds that using automatic relevance determination and autoencoder traces correlations between variables. Principal Component Analysis (PCA), Multi-Dimensional Scaling (MDS), and Factor Analysis (FA) focus on the variance in the variables to allocate importance to the data. The variable TDCG which is referred to in most of the standards are an aggregation dimension reduction because it includes most of the individual gases which are also used in determining a bushings condition. As a result it is expected that with fewer variables used in the classification the results should be sufficiently closely aligned with a decision made using complete input data.

Chapter 8

CONCLUSIONS AND RECOMMENDATIONS

This thesis concludes by making comments on the approach that was followed in conducting the research, the contributions of the thesis and suggestions of future work.

8.1 General approach of the thesis

Several methods of pattern recognition were reviewed and proposed for application in condition monitoring of bushings by analysing dissolved gases in oil. Before a solution could be proposed, the failure mechanism of bushings was investigated and understood. Investigation of failures was by analysis of raw data from failed bushings obtained from the maintenance workshop; review of manufactures specifications for oil transformers and bushings; analysis of failure reports; review of relevant standards documents; and by inspection of failed samples as well as healthy bushings during the test process.

Once the limiting conditions for the safe operation of the bushing had been established a study was done to determine the effect on a bushing of common field conditions such as misaligned bushing sheaths, poor quality insulation oil and unearthed test tap. Simulations were done using boundary element and finite element software to determine electric field strength at important points inside a bushing, to get an understanding of the level of stress and the expected stress indicators from a bushing. The recommendations Duval (1974), Doernenberg & Gerber (1967), Hubacher (1976) and Rogers (1975) as well as of standards such as IEC60599 (1999), IEEEC57-

104 (1991) and IEC60137 (2003) were considered in order to develop a criterion that could be applied specifically to bushings oil, and not transformer oil, when using a machine learning tool.

Different methods used for determining the condition of bushings were evaluated and compared. Such methods include partial discharge monitoring, visual inspection, noise monitoring, dissolved gas analysis (DGA) assessment, dielectric dissipation factor monitoring, temperature monitoring and mechanical assessment. Once a suitable method had been selected, data from healthy and failed bushings were analysed. Taking into account the availability of test equipment and reproducibility of the test results, a decision was made on which manual assessment method to use jointly with machine learning methods. The dissolved gas analysis (DGA) method was determined to be the most widely accepted tool within the electricity industry in South Africa and most reliable diagnostics tool for monitoring the condition of bushings and transformers, and so it was decided to apply machine learning tools to DGA data.

Neural networks, support vector machines, fuzzy set theory and data reduction methods such as principal component analysis, multidimensional scaling and factor analysis were applied to a set of data from 60699 oil impregnated bushings. Evolutionary computing methods in the form of genetic algorithms and particle swarm optimization within neural networks were compared to gradient methods such as scaled conjugate gradient and quasi-Newton and non-gradient methods such as the simplex method. The work is proposing a solution to an industry problem was identified to be the failure of bushing, especially those that are mounted on transformers. Applying artificial intelligence tools that are proposed in this thesis will introduce a reliable way of monitoring bushings using the data that is routinely taken by field service staff, by alarming in time when a dangerous condition arises.

Assessment of the database after collection of data from the field and interaction with staff revealed that the size of the database of historical information as well as the loss of skills due to migration and retirement of experienced technical staff highlights the need for an automated diagnosis tool that integrates information from the people, sensors and the database. Users of the methods proposed in the work need not be subject experts in dissolved gas analysis, in order to translate the results correctly. Data reduction methods were implemented to address the problem of large number of bushings, together with the number of variables that are used to decide on the condition of each bushing using DGA, together with the fact that the variable is sometimes not available. During the work, it was found that many of the variables

did not match the threshold criteria exactly to be classified as usable or unusable. There is a degree of tolerance within the DGA criteria that can be exploited using fuzzy set theory (FST). FST was integrated with ANN's to classify bushings and indicate the extent of degradation of each diagnosed bushing. The research ended off by comparing different machine learning techniques and highlighting the deficiencies that were identified.

8.2 Contribution of this thesis

Machine learning tools which included two neural networks and a kernel machine were implemented on a DGA data set for bushings. The methods were compared based on speed, stability and accuracy in classifying. The networks were radial basis function (RBF) and multilayer perceptron (MLP). The kernel machine used is a support vector machine (SVM). In this work 60699 bushings were classified based on ten criteria recommended in IEEEC57-104 (1991), IEC60599 (1999), California State University Sacramento (CSUS) criteria (Hubacher 1976), Rogers (1975) and Doernenberg & Gerber (1967) criteria. This work developed an *expert* machine learning method and criterion to allow a decision to be made when the above methods reach deadlock or meet a condition that is unaccounted for within their criteria. The ANNs and SVM were configured to form a committee of classifiers. Classification by the committee was done based on a majority vote. The work proposed the application of neural networks with particle swarm optimisation (PSO) and genetic algorithms (GA) to compensate for missing data in classifying high voltage bushings. The work applied fuzzy set theory (FST) to diagnose the condition of high voltage bushings.

Relevance and redundancy detection methods were used successfully to prune the redundant measured variables and accurately diagnose the condition of bushings with fewer variables. The work found that a bushing could be accurately diagnosed using only six of the ten criteria, by applying dimension reduction techniques. Experimental results from bushings that were evaluated in the field were used to verify the simulations. The results of this work can be used to develop commercial real-time monitoring and decision making tools that combine information from chemical, electrical, optical, thermal and mechanical measurements that are taken from bushings.

8.3 Future work on this topic

Oil impregnated bushings account for more than 99% of bushings currently used on transformers and other power equipment. New bushings made of resin are beginning to gain acceptance in industry. A useful project would be to determine which attributes of this new design of bushing can be monitored and used by machine intelligence methods for diagnostics. Online learning methods for condition monitoring need further investigation.

Bibliography

Abdella, M. & Marwala, T. (2005), The use of genetic algorithms and neural networks to approximate missing data in database, *in* 'Computing and Informatics', pp. 577–589.

Ammar, S. & Wright, R. (2000), Applying fuzzy-set theory to performance evaluation, *in* 'Socio-Economic Planning Sciences', Vol. 34, pp. 285–302.

Bandemer, H. & Gottwald, S. (1990), *Neural Computation: An Introduction*, Institute of Physics Publishing, Bristol.

Bandemer, H. & Gottwald, S. (1995), *Fuzzy Sets, Fuzzy Logic, Fuzzy Methods with Applications*, John Wiley and Sons,New York.

Bartholomew, G. J. (1987), *Latent Variable Models and Factor Analysis*, Charles Griffin and Company Ltd, London.

Bartnikas, R. & Novak, J. P. (1993), 'On the character of different forms of pd and their related terminologies', *IEEE Transactions,Electric Insulations* **28**, 956–968.

Behnke, S. & Karayiannis, N. (1998), 'Competitive neural trees for pattern classification', *IEEE Transactions on Neural Networks* **9**(6), 1352–1370.

Bell, A. J. & Sejnowski, T. J. (1995), 'An information maximisation approach to blind separation and blind deconvolution', *Neural Computation* **7**(6), 1129–1159.

Bishop, C. N. (1995), *Neural Networks for Pattern Recognition*.

Bojadziev, G. & Bojadziev, M. (1995), *Fuzzy sets, fuzzy logic, applications,*, World Scientific Publishing Co, Pte, Ltd.

Bourlard, H. & Kamp, Y. (1988), *Auto-association by multilayer perceptrons and singular value decomposition*, Vol. 2.

Carreira-Perpin, M. . (2001), Continuous Latent Variable Models for Dimensionality Reduction and Sequential Data Reconstruction, PhD thesis, University of Sheffield, Deparment of Computer Science.

Cristianini, N. & Shawe-Taylor, J. (2000), 'An introduction to support vector machines and other kernel based learning methods', *Cambridge University Press,Cambridge* .

Cybenko, G. (1989), 'Approximating by superpositions of a sigmoid function', *Mathematics of Control, Signals and Systems* **2**, 303–314.

de Klerk, P. (2005), 'Transformer condition assessment', *Eskom Report Cleveland* .

Dhlamini, S. M., Crowdy, P. & van Coller, J. (2000), A study of the issues involved in the design of a 132kv polymer bushing, *in* 'South African Universities Power Engineering Conference (SAUPEC)', Durban.

Dhlamini, S. M., Marwala, T. & Majozi, T. (2006), Fuzzy and multilayer perceptron for evaluation of hv bushings, *in* 'Proceedings of the IEEE International Conference on System Man and Cybernetics, Taipei (Taiwan)', pp. 1331–1336.

Dhlamini, S. M., Marwala, T. & van Coller, J. (2005), Modeling inaccuracies from simulators for hv polymer bushings, *in* 'International Symposium on High Voltage', p. Paper A18.

Dhlamini, S. M. & Marwala, T. (2004), An application of svm, rbm and mlp with ard on bushings, *in* 'IEEE-CIS Conference Proceedings', pp. 1253–1258.

Dhlamini, S. M. & Marwala, T. (2005), Bushing diagnostics using an ensemble of parallel neural networks, *in* 'Proceedings of the IEEJ-IEEE Symposium on Electrical Insulating Materials (ISEIM05)', pp. 289–292.

Dhlamini, S. M., Nelwamondo, F. V. & Marwala, T. (2006), 'Sensor failure compensation techniques for hv bushing monitoring using evolutionary computing', *WSEAS Transactions on Power Systems* pp. 280–287.

Dhlamini, S. M. (2000), A study of the issues involved in the design of a 132kv polymer bushing, Master's thesis, School of Electrical and Information Engineering, University of Witwatersrand, Johannesburg.

DiGiorgio, J. (2005), 'Dissolved gas analysis of mineral oil insulating fluid, northern technology and testing report'.

Doernenberg, E. & Gerber, O. E. (1967), 'Analysis of dissolved and free gases for monitoring performance of oil- filled transformers', **54**(213), 104.

Dominelli, N. & Lau, M. (2004), A new approach to dga interpretation, *in* '71st Doble Conference for Clients', Boston.

Duval, M. (1974), Fault gases formed in oil-filled breathing e.h.v. transformers, *in* 'Interpretation of Gas Analysis Data', number C 74, pp. 476–478.

Everitt, B. S. & Dunn, G. (1991), *Applied Multivariate Data Analysis*, Edward Arnold, london.

Fiesler, E. & Beale, R. (1997), *The Handbook of Neural Computation*, Oxford University Press, Oxford.

Flaig, A., Barner, K. E. & Arce, G. R. (2000), 'Fuzzy ranking: Theory and applications, signal processing', **80**, 1017–1036.

Friedman, J. (1977), 'A recursive partitioning decision rule for nonparametric classifiers', *IEEE Transactions on Computers* **26**, 404–408.

Fukunaga, K. (1990), *Introducation to Statistical Pattern Recognition*, San Diego Academic Press.

Ghahramani, Z. & Jordan, M. (1997), 'Mixture models for learning from incomplete data', *MIT Press* **4**, 67–85.

Glover, F. (1977), 'Heuristics for integer programming using surrogate constraints, decision sciences', **8**(1), 156–166.

Gunn, S. R. (1998), 'Support vector machines for classification and regression'.

Halliday, D., Resnick, R. & Merrill, J. (1988), *Fundamentals of Physics*, Wiley and Sons, New York.

Hall, D. L. & Llinas, J. (2001), *Handbook of Multisensor Data Fusion*, CRC Press.

Han, J. & Kamber, M. (2001), *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco.

Harpe, S. & Kranz, H. (1999), 'Practical application of machine intelligent partial discharge pulse suppression system neurotek ii', *CIGRE-paper TF 15.03.09. IWD 16* .

Haykin, S. (1994), *Neural Networks,*, Macmillan College Publishing Company Inc, New York.

Hebb, D. O. (1949), *The Organization of Behavior*, John Wiley and Sons,New York.

Hines, J. W. (1997), *Fuzzy and Neural Approaches in Engineering,Matlab Supplement*, John Wiley and Sons,New York.

Holland, J. H. (1975), *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor.

Holman, J. P. (1989), *Experimental Methods for Engineers*, McGraw-Hill, New York.

Hornick, K., Maxwell, S. & Halbert, W. (1989), 'Multilayer feedforward networks are universal approximators', *Neural Networks* **2**, 359–366.

Houck, C. R., Joines, J. A. & Kay, M. G. (1995), A genetic algorithm for function optimisation: A matlab implementation, Technical report, North Carolina State University. Technical Report NCSU-IE Technical Report 95-09.

Hubacher, E. J. (1976), Analysis of dissolved gas in transformer oil to evaluate equipment condition, *in* 'P.C.E.A. Engineering and Operation Section Spring Conference'.

IEC60137 (2003), *Insulated Bushings for Alternating Voltages Above 1 000 V*, fifth edn, IEC Press, Geneva.

IEC60247 (1978), *Measurement of Relative Permittivity, Dielectric Dissipation Factor and D.C. Resistivity of Insulating Liquids*, IEC, Press, Geneva.

IEC60475 (1974), *Method of Sampling Liquid Dielectrics*, IEC Press, Geneva.

IEC60567 (1992), *Guide for the Sampling of Gases and of Oil from Oil-Filled Electrical Equipment and for the Analysis of Free and Dissolved Gases*, IEC Press, Geneva.

IEC60599 (1999), *Mineral Oil-Impregnated Electrical Equipment in Service Guide to the Interpretation of Dissolved and free Gases Analysis*, second edn, IEC Press, Geneva.

IEC60628 (1985), *Gassing of Liquids under Electrical Stress and Ionisation*, IEC Press, Geneva.

IEC60813 (1985), *Test Method for Evaluating the Oxidation Stability of Hydro-Carbon Insulating Liquids*, IEC Press, Geneva.

IEC60814 (1997), *Insulating liquids: Oil-impregnated Paper and Pressboard, Determination of Water by Automatic Coulometric Karl Fischer Titration*, IEC Press, Geneva.

IEC60833 (1987), *Measurement of Power-Frequency Electric Fields*, IEC Press, Geneva.

IEC60867 (1986), *Specifications for Unused Insulating Liquids Based on Synthetic Aromatic Hydrocarbons*, IEC Press, Geneva.

IEC60897 (1987), *Methods for the Determination of the Lightning Impulse Breakdown Voltage of Insulating Liquids*, IEC Press, Geneva.

IEC61294 (1993), *Insulating Liquids - Determination of the Partial Discharge Inception Voltage (PDIV) - Test Procedure*, IEC Press, Geneva.

IEC61463 (1996), *Bushings - Seismic Qualification*, IEC Press, Geneva.

IEC61620 (1998), *Insulating Liquids - Determination of the Dielectric Dissipation Factor by Measurement of Conductance and Capacitance- Test Method*, IEC Press, Geneva.

IEEEC57-104 (1991), *Guide for the Interpretation of Gases Generated in Oil-Immersed Transformers*, IEEE Press, New York.

Jackson, J. E. (1991), *Users Guide to Principal Components*, John Wiley and Sons, New York.

Janick, M. (2001), 'Operations-extreme loading'.

Jolliffe, I. T. (1986), *Principle Component Analysis*, Springer-Verlag, Berlin.

Kecman, V. (2001), *Learning and Soft Computing: Support Vector Machines, Neural Networks and Fuzzy Logic Models*, MIT Press,, Cambridge, MA.

Koza, J. R. (1992), *Genetic Programming*, MIT Press,, Cambridge.

Kramer, M. A. (1991), 'Nonlinear principal component analysis using autoassociative neural networks', *AIChe Journal* **37**(2), 233–243.

Kraus, J. D. (1991), *Electromagnetics*, McGraw-Hill Inc, New York.

Kubica, E. G., Wang, D. & Winter, A. D. (1995), 'Modelling balance and posture control mechanisms of the upper body using conventional and fuzzy techniques', *Gait and Posture* **3**(2), 111–119.

Kuffel, E. & Zaengl, W. (1984), *High Voltage Engineering Fundamentals*, Pergamon Press Ltd, Oxford.

Lau, M. (2003), 500kv bushings failures and bushing oil sampling program monitors, *in* 'Avo Conference', New Zealand.

Lawley, D. N. & Maxwell, A. E. (1971), *Factor Analysis as a Statistical Method*, second edn, American Elsevier Pub. Co, New York.

Leke-Betechuoh, B. (2004), Optimal selection of stocks using computational intelligence methods, Master's thesis, School of Electrical and Information Engineering, University of Witwatersrand, Johannesburg.

Lennon, S. J. (2006), 'Resources and strategy division', *Eskom critical issues report* .

Levenberg, K. (1944), 'A method for solution of certain non-linear problems in least squares', *Quarterly Journal of Applied Mathematics* **2**, 164–168.

Looney, C. G. (1997), *Pattern Recognition Using Neural Networks, Theory and Algorithms for Engineers and Scientists*, Oxford University Press, New York.

Lord, T. & Hodge, G. (2003), On-line monitoring technology applied to hv bushings, *in* 'Avo Conference', New Zealand.

MacKay, D. J. C. (1994), 'Bayesian non-linear modelling for the energy prediction competition', *ASHRAE Transactions* **100**(2), 1053–1062.

Majozi, T. & Zhu, X. X. (2005), 'A combined fuzzy set theory and milp approach in integration of planning and scheduling of batch plants - personnel evaluation and allocation', *Computers and Chemical Engineering 29, Elsevier Science Publishing Company* pp. 2029–2047.

Mak, M. W., Allen, W. G. & Sexton, G. (1994), 'Speaker identification using multilayer perceptrons and radial basis function networks', *Neurocomputing* **6**, 99–117.

Mamdani, E. H. & Assilian, S. (1975), 'An experiment in linguistic synthesis with a fuzzy controller', *International Journal of Man Machine Studies* **7**(1), 1–13.

Markey, L. & Stevens, M. (2003), 'Microstructural characterisation of xlpe electrical insulation in power cables: determination of void size distribution using tem', *Applied Physics D: Applied Physics* **36**, 2569–2583.

Marquardt, D. W. (1963), 'An algorithm for least-squares estimation of non-linear parameters', *Journal of the Society of Industrial and Applied Mathematics* **11**, 431–441.

Marwala, T. (2001), Fault Identification Using Neural Networks and Vibration Data, PhD thesis, University of Cambridge.

Mcgrail, A. J., Gulski, E., Groot, E. R. S., Allan, D., Birtwhistle, D. & Blackburn, T. R. (2003), Data mining techniques to assess the condition of high voltage electrical plant, *in* 'Cigre Conference', Paris.

Mewalall, R., de Klerk, P. & Moroaswi, J. (2006), Transformer condition monitoring from the system operator perspective, *in* 'IEEE Conference on Condition Monitoring and Diagnosis', Korea.

Michalewicz, Z. (1996), *Genetic Algorithms + Data Structures = Evolution Programs*, Springer-Verlag, New York.

Moller, M. F. (1993), 'A scaled conjugate gradient algorithm for fast supervised learning', *Neural Networks* **6**, 525–533.

Morhuis, P. H. & Kreuger, F. H. (1990), 'Transition from streamer to townsend mechanisms in dielectric voids', *Journal of Applied Physics* **23**, 562–1568.

Nabney, I. T. (2003), *NETLAB Algorithms for Pattern Recognition*, Springer, London.

Nasser, E. (1991), *Fundamentals of Gaseous Ionisation and Plasma Electronics*, John Wiley and Sons Inc, New York.

Nelson, M. M. & Illingworth, W. T. (1991), *A Practical Guide to Neural Nets*, Addison-Wesley Publishing Company Inc, New York.

Nyamupangedengu, C. & Reynders, J. P. (2003), An investigation into the characteristics of high frequency current sensor with respect to partial discharge detection in power cables, *in* 'South African Universities Power Engineering Conference', Pretoria.

Orton, H. E. (2002), 'Diagnosing the health of your underground cable', *IEEE Transmission and Distribution Magazine* .

Pacheco, P. (1997), *Parallel Programming with MPI*, Morgan Kaufmann Publishers Inc.

Palgrave, P. (2000), *Neural Networks*, Picton, Hampshire.

Pukel, G. J., Muhr, H. M. & Lick, W. (2006), Transformer diagnostics: Common used and new methods, *in* 'IEEE Conference on Condition Monitoring and Diagnostics', Korea.

Purkait, P. & Chakravorti, S. (2003), 'Investigations on the usefulness of an expert system for impulse fault analysis in distribution transformers', *Electric Power System Research* **65**(2), 149–157.

Pyle, D. (1999), *Data Preparation for Dat Mining*, Morgan Kauffmann Publishers, San Francisco.

Quinlan, J. R. (1989), Unknown attribute values in induction, *in* 'Sixth International Workshop on Machine Learning', New York, pp. 164–168.

Reyment, R. & Joreskog, K. G. (1993), *Applied Factor Analysis for Natural Sciences*, 2nd edn, Cambridge University Press, Cambridge.

Rogers, R. R. (1975), 'U.k. experience in the interpretation of incipient faults in power transformers by dissolved gas-in-oil chromatographic analysis'. Doble Conference Index of Minutes, Sec. 10-201.

Rummel, E. J. (1970), *Applied Factor Analysis*, Northwestern University Press, Evanston, Illinois.

Sears, F. W., Zemansky, M. W. & Young, H. D. (1987), *University Physics*, 7th edn, Addison-Wesley Publishing Co, Reading Massachusetts.

Seber, G. A. F. (1984), *Multivariate Observations*, John Wiley and Sons, New York.

Shi, Y. & Eberhart, R. C. (1998), 'Parameter selection in swarm optimisation', *Evolutionary Programing* pp. 591–600. VII Proc. EP98, Springer-Verlag, New York.

Sokolov, V. (2001), Transformer life management, *in* 'II Workshop on Power Transformers-Deregulation and Transformers Technical, Economic, and Strategical Issues', Salvador, Brazil.

Sugeno, M. (1985), *Industrial Applications of Fuzzy Control*, Elsevier Science Publishing Company.

Urban, R. & de Villiers, W. (2006), 'Radar alarm system', *University of Stellenbosch* .

van Wyk, S. (1997), 'Transformer field performance in sub-sahara africa 1996', *Energize* pp. 44–47.

Vapnik, N. V. (1999), *Statistics for Engineering and Information Science, The Nature of Statistical Learning Theory*, 2nd edn, Springer-Verlag, New York.

Wald, A. (1947), *Sequential Analysis*, Dover Phoenix Publications.

Wang, Z. (2000), Artificial Intelligence Applications in the Diagnosis of Power Transformer Incipient Faults, PhD thesis, Virginia Polytechnic Institute and State University.

Weiss, S. M. & Indurkhya, N. (1998), *Predictive Data Mining*, Morgan and Kaufmann Publishers, San francisco.

Welling, M. & Weber, M. (1999), Independent component analysis of incomplete data, *in* 'Proceedings of the 6th Joint Symposium on Neural Computation', UCSD.

Willshaw, D. J. & von der Malsberg, C. (1976), How patterned neural connectionscan be set up by self-organisation, *in* 'Proceedings of the Royal Society, Series B vol 194', pp. 431–445.

Workgroup, A. T. (1996), Australasia transformer reliability survey 1996 report, Technical report, Western Power Transmission Projects Branch.

Zadeh, L. (1973), 'Outline of a new approach to the analysis of complex systems and decision processes', *IEEE Transactions on Systems, Man and Cybernetics* **3**(1), 28–44.

Zhang, G. J., Zhao, W. B. & Yan, Z. (2006), Monitoring and diagnostic strategy for insulation condition of oil-immersed power transformer, *in* 'IEEE Conference on Condition Monitoring and Diagnostics', Korea.

Zheng, Z. & Boggs, S. (2005), 'Defect tolerance of solid dielectric transmission class cable', *IEEE Electrical Insulation Magazine* pp. 35–41.

Appendix A

LIST OF PUBLISHED WORK

A.1 Journals

1. S.M. Dhlamini, F.V. Nelwamondo and T. Marwala. (2006) Condition monitoring of HV bushings in the presence of missing data using evolutionary computing. WSEAS Transactions on Power Systems, pp. 280-287.

2. S.M. Dhlamini, T Marwala and T Majozi, (2006) Neuro-Fuzzy Networks for Bushing Evaluation, International Journal of Fuzzy Systems, Taiwan, Submitted, October.

A.2 Conference Presentations

1. S.M. Dhlamini, M. Kachienga, and T. Marwala. (2007) Artificial Intelligence as an Aide for Security Technology Management. Proceedings of the IEEE Africon International Conference, Windhoek (Namibia), September 7- 11, Paper No 194.

2. S.M. Dhlamini, T. Marwala, and T. Majozi. (2006) Fuzzy and Multilayer Perceptron for Evaluation of HV Bushings. Proceedings of the IEEE International Conference on System Man and Cybernetics, Taipei (Taiwan), October 7- 11, pp 1331-1336.

3. S. Dhlamini, B. Duma, F. Nelwamondo, L. Mdlazi, T. Marwala, (2006) Redundancy Reduction Techniques for HV Bushing, International Conference on Condition Monitoring and Diagnosis, Changwon (Korea), 2-5 April. Paper No.700.

4. Dhlamini, S.M., Marwala, T., (2005) Bushing diagnostics using an ensemble of parallel neural networks, Proceedings of the IEEJ-IEEE Symposium on Electrical Insulating Materials (ISEIM05), Fukuoka (Japan), 5-9 June, pp. 289-292.

5. Dhlamini, S.M., Marwala, T., (2005) Cost benefit of using a committee of parallel neural networks for bushing diagnostics, IEEE Power Engineering Society Conference (PES05), Durban, July 11-15, pp. 485-488.

6. Dhlamini, S.M., Marwala, T., and van Coller, J., (2005), Modelling inaccuracies from simulators for HV polymer bushings, Proceedings of the XIVth International Symposium on High Voltage Engineering, Tsinghua University, Beijing, China, Paper A18.

7. Dhlamini, S.M., Nelwamondo F.V. and Marwala, T., (2005) Sensor failure compensation techniques for HV bushing monitoring using evolutionary computing. 5th WSEAS / IASME International Conference on Electric Power Systems, High Voltages, Electric Machines, Tenerife (Spain), December 16-18, pp. 430-435.

8. Dhlamini, S.M., Marwala, T., (2004) Bushing monitoring using MLP and RBF, IEEE Africon 2004, Gabarone (Botswana), 14-17 September, pp. 613-617.

9. Dhlamini, S.M., Marwala, T., (2004) An application of SVM, RBF and MLP with ARD on Bushings, IEEE Conference on Cybernetics and Intelligent Systems, Singapore, 1-3 December, pp. 1245-1258.

Appendix B

CLEARANCES AND DIELECTRIC STRENGTH

B.1 Clearance in Air for 50Hz

Dielectric strength of air $= 13.5kV/cm = 1.35kV/mmat0m.$

Phase- phase clearance:

$$D_{\phi-\phi} = \frac{145 \times \sqrt{2}}{1.35} = 151.89mm$$

Phase- ground clearance:

$$D_{\phi-g} = \frac{145 \times \sqrt{2/3}}{1.35} = 87.7mm$$

At 1800m:

$$D_{(\phi-\phi)1800} = D\phi - \phi.\left(1 + \frac{1.25}{100}\left(\frac{1800-1000}{100}\right)\right) = 167.1mm$$

$$D_{(\phi-g)1800} = D\phi - g.\left(1 + \frac{1.25}{100}\left(\frac{1800-1000}{100}\right)\right) = 96.5mm$$

B.2 Clearance in oil for 50Hz

Dielectric strength of oil = 28kV/mm (IEC60137 2003).
Phase- phase clearance

$$D_{\phi-\phi} = \frac{145 \times \sqrt{2}}{28} = 7.32mm$$

Phase-ground clearance

$$D_{\phi-g} = \frac{145 \times \sqrt{2/3}}{28} = 4.23mm$$

B.3 Clearance in air for lightning

Dielectric strength of air = 500kV/m = 0.5kV/mm at 0m
BIL at 132kV = 650kV (IEC60137 2003).
Phase- phase and phase- ground clearance are the same (IEC60137 2003).
At 0m:

$$D_{\phi-\phi} = \frac{650}{0.5} = 1300mm$$

At 1800m:

$$D_{(\phi-\phi)1800} = D\phi - \phi.\left(1 + \frac{1.25}{100}\left(\frac{1800-1000}{100}\right)\right) = 1430mm$$

B.4 Clearance in oil for lightning

Dielectric strength of air = 145kV/mm
BIL at 132kV = 650kV (IEC60137 2003), so

$$D_{\phi-\phi} = \frac{650}{145} = 4.48mm$$

B.5 Clearance for switching impulses

Dielectric strength of air = 500kV/m = 0.5kV/mm at 0m.
Switching impulse at 132kV = 365kV (IEC60137 2003)
Phase- phase and phase- ground clearance are the same (IEC60137 2003).
At 0m:

$$D_{\phi-\phi} = \frac{365}{0.5} = 730mm$$

At 1800m:

$$D_{(\phi-\phi)1800} = D\phi - \phi.\left(1 + \frac{1.25}{100}\left(\frac{1800-1000}{100}\right)\right) = 803mm$$

Appendix C

DISSOLVED GAS ANALYSIS RATIOS

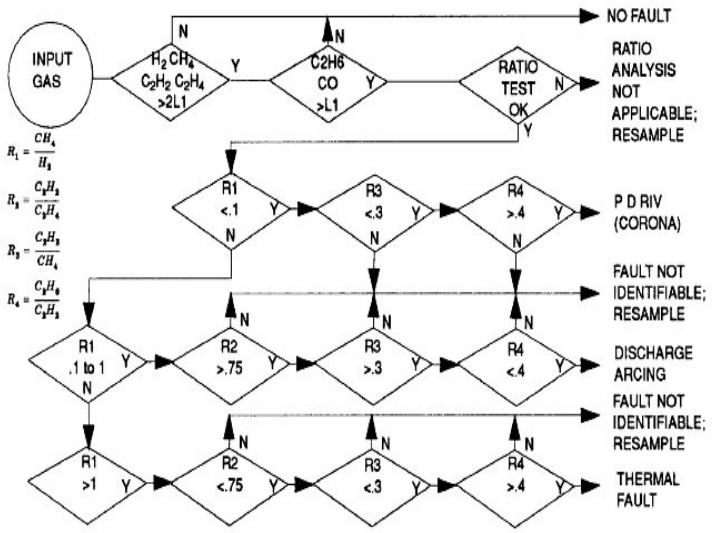

Figure C.1: Flow chart of Doernenberg ration method (IEEEC57-104 1991)

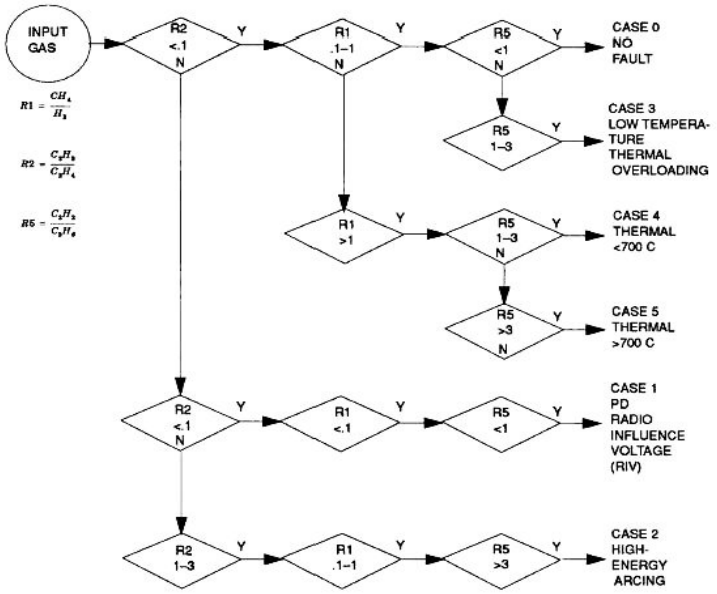

Figure C.2: Flow chart of Rogers ration method (IEEEC57-104 1991)

Appendix D

COMPUTER SPECIFICATIONS

All software was run on a Windows 2000 personal computer with 1.7GHz processor, 1GB RAM. Matlab version 6.5.0.1 was used.

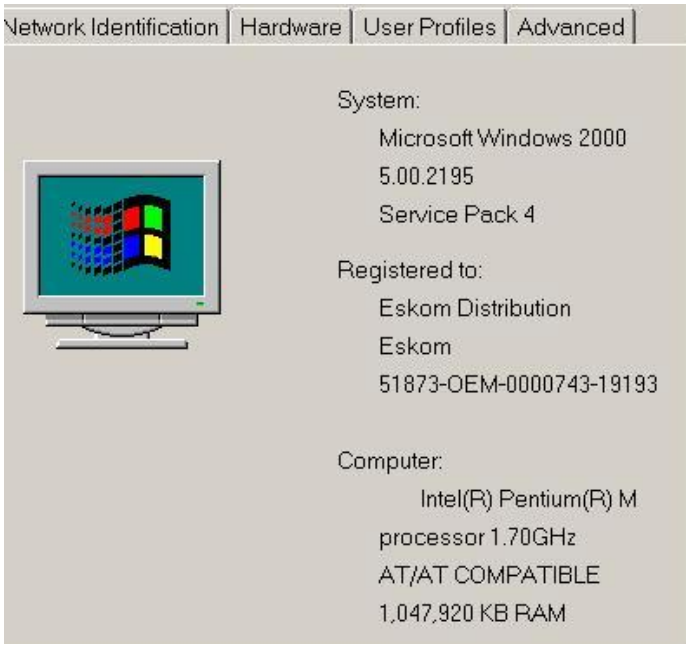

Figure D.1: PC hardware specifications

Appendix E

ELECTRIC FIELD SENSORS

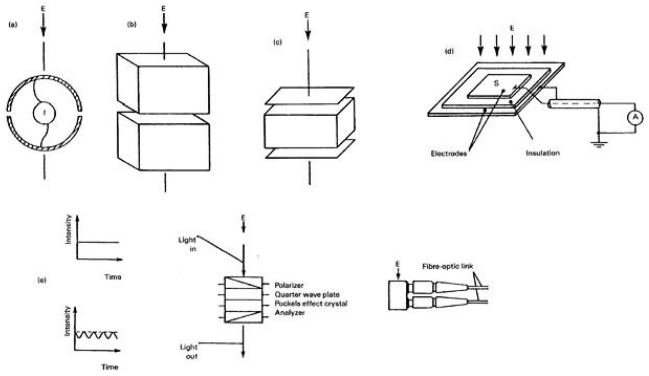

Figure E.1: Main Types of Electric Field Sensors
(IEC60833 1987)

Type (a) is the spherical probe, (b) is the box type probe, (c) is the parallel plate probe, (d) is the ground-reference probe, (d) is the electro-optic probe