UNIVERSITI TEKNOLOGI MARA

IMBALANCED MULTI-CLASS POWER TRANSFORMER FAULT DATA CLASSIFICATION THROUGH EDITED NEAREST NEIGHBOUR-MANHATTAN-RANDOM FOREST

PUTRI AZMIRA BINTI R AZMI

MSc

October 2025

ABSTRACT

This study highlights the global significance of the O&G industry, emphasizing the need for efficient operational management to ensure energy reliability, economic optimization, and environmental protection. Transformers are critical in power systems and require constant monitoring to maintain stability. Dissolved gas analysis uses gas chromatography to detect combustible gases generated during abnormal operations, playing a key role in transformer fault diagnosis. Artificial intelligence techniques, such as Support Vector Machines and Artificial Neural Networks, have been widely applied, enhancing diagnostic accuracy through predictive analytics. However, imbalanced datasets, particularly in dissolved gas analysis, severely affect classification performance by causing machine learning models to favour majority fault types while overlooking minority classes. This misclassification and data loss can lead to the failure to detect rare but critical transformer faults, jeopardizing system reliability and early fault mitigation. To address this challenge, the study focuses on improving Edited Nearest Neighbour techniques using alternative distance measures to enhance classification accuracy in imbalanced dissolved gas analysis datasets. The Edited Nearest Neighbour technique, shown to be effective in other O&G subdomains, is evaluated using the Random Forest algorithm, which is widely used for its precision and ability to handle non-linear datasets. To validate the effectiveness of Edited Nearest Neighbour-Random Forest, it is compared to four data-level techniques including Random Under-Sampling, NearMiss, Random Oversampling, and Adaptive Synthetic Sampling. Furthermore, Random Forest is compared to four machine learning algorithms including Support Vector Machine, XGBoost, Convolutional Neural Networks, and Decision Trees. Edited Nearest Neighbour with Manhattan distance measure, which demonstrated over 85.00% accuracy in previous studies, is assessed alongside Minkowski and Mahalanobis distances to achieve the best model. After parameter tuning of Random Forest, the Edited Nearest Neighbour-Manhattan-Random Forest model outperformed, achieving 90.77% accuracy and reducing data loss from 70.00% to 17.50%. These findings show that Edited Nearest Neighbour-Manhattan-Random Forest effectively balances the dissolved gas analysis dataset while enhancing classification accuracy. Further research is required to explore the broader applicability of this technique in other domains with imbalanced multi-class datasets, as real-world datasets are often scattered and imbalanced.

ACKNOWLEDGEMENT

First and foremost, I want to thank Allah S.W.T. for allowing me to embark on my master's degree and for completing this challenging yet beautiful journey successfully. I also want to express my gratitude and big thanks to my main supervisor, Assoc. Prof. Dr Marina Yusoff, and my co-supervisor, Dr Yuzi Mahmud, for their guidance and support throughout the study. Their insights and expertise greatly contributed to the successful completion of this research.

I am deeply grateful to my family for their unconditional love, support, and patience during this journey. I would also like to thank my friends for their encouragement and helpful discussions.

Lastly, I would like to take a moment to acknowledge my own perseverance and dedication throughout this research journey. The countless hours of hard work, self-discipline, and determination have played a crucial role in bringing this research to completion. This experience has been both challenging and rewarding, and I am proud of the growth and resilience I have demonstrated along the way. thank you for myself that always believe.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

This chapter provides a concise summary of the research, encompassing six sections. Section 1.1 offers an overview of the research background, while Section 1.2 outlines the current research problem. Section 1.3 defines the research questions, followed by the research objectives in Section 1.4. The significance of the research is expounded upon in Section 1.5, and the entire research paper is succinctly summarized in Section 1.6.

1.1 Research Background

Insights from the International Energy Agency in 2020 show that the oil and gas (O&G) industry plays an essential role in the global economy, providing a substantial portion of the world's energy needs. Effective management and optimization of operations in this business are critical for providing a reliable energy supply, reducing environmental effects, and optimizing economic returns (Liang et al., 2022; Xiang et al., 2021). The power transformer, serving as the primary component of the power system, plays a role in converting alternating current (AC) voltage and current for the transmission of AC power. An oil-immersed power transformer can generate combustible gases during abnormal operation due to thermal, electrical, and mechanical stresses. This gas production directly impacts the overall stability and safety of the power system. The condition of the dissolved gas in the oil undergoes gradual changes when latent faults exist in the transformer (Liu et al. 2022).

The critical role of a transformer in the power grid in monitoring and diagnosing its operational status is of most importance (Wang et al. 2022). Dissolved gas analysis is a preferred method for predicting faults in power transformers, given its simplicity and suitability for online diagnosis (Taha & Mansour 2021). The dissolved gas analysis process involves several steps. Initially, samples of oil are extracted from the transformer. These oil samples are subsequently introduced into a gas chromatograph designed to separate and analyze the various gas components present in the oil sample. It functions by detecting the chemical composition of the mixture and quantifying the concentration of each component (Flanagan et al., 2008). Gas chromatography utilizes