ROBUST PRIDIT SCORING METHOD FOR CLASSIFICATION FRAUD CASES IN FINANCIAL DATA

NORBAITI BINTI TUKIMAN

UNIVERSITI TEKNOLOGI MALAYSIA

ROBUST PRIDIT SCORING METHOD FOR CLASSIFICATION FRAUD CASES IN FINANCIAL DATA

NORBAITI BINTI TUKIMAN

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Mathematics)

Faculty of Science
Universiti Teknologi Malaysia

MARCH 2022

DEDICATION

In the name of ALLAH, The Most Merciful and The Most Kind.
I dedicate this thesis to my late mother, Allahyarhamha Manisah Selak
and my late father, Allahyarham Tukiman bin Lehan
(Al Fatihah to them)
My beloved husband, Abd. Rahim bin Talib
My children, Nur Azizah, Muhammad Asyraf and Nur Ayuni
My supportive supervisor, Dr Norhaiza, as well as all the lecturers and friends who have helped me through this journey.

ACKNOWLEDGEMENT

Bismillahirahi ar-Rahman ar- Rahim. Alhamdulillahi Rabbil A'alamin. All praise for Allah, Lord of all Worlds, The Most Compassionate and The Most Merciful. Thank you, Allah, for giving me the strength and guiding me through many trials and tribulations during my Ph.D. journey. This thesis is the result of a lot of tears and hours of hard work, including reading articles and tons of books, a lot of computer programming and writing, more writing and then rewriting! This journey was extremely challenging, and not easy at all. My self-motivation had its ups and downs in the endeavour to achieve my goal.

Eventually, Alhamdulilla, Praise to Allah, who made me reset my focus on so many occasions until I finally finished writing this thesis. The work is done, the journey has come to a delightful end, I feel delighted, and it could not be complete without the tremendous support and motivation from so many people in my life. This acknowledgment is conveyed to my supervisor, Dr Norhaiza Ahmad. Thank you for your kindness, passion and hours of your precious time spent guiding, supervising, and supporting me through innumerable discussions and revisions to this thesis, even under strict deadlines. May Allah SWT bless you in kindness and keep you and your family in the best of health. Your invaluable guidance leading to the fruition of my thesis is indescribable. Undoubtedly, my deepest and sincere gratitude also goes to my husband, Abd. Rahim bin Talib, my children Nur Azizah, Muhammad Asyraf and Nur Ayuni, for understanding the pressures and seemingly unsurmountable workload I faced every other day while finishing my thesis. Thank you so much for being part of my Ph.D. journey - for always supporting and praying for my success in this challenging course of life. May Allah SWT bless all of us. Thank you.

ABSTRACT

Increasing number of fraud cases could jeopardize business solvency. Identification of fraud using effective statistical methods, such as classification, can protect organisations from this pitfall. However, identifying fraud cases can be a statistical challenge due to several characteristics of financial datasets. These data typically form large datasets that are highly dimensional, contain mixed data types and can involve an imbalanced number of fraud and non-fraud cases. This study employed the Principal Component Analysis (PCA) based on Relative to an Identified Distribution (RIDIT) scores, known as the PRIDIT method, to classify and identify data that could potentially be fraudulent cases. The classical PRIDIT method involves the transformation of each analysed dataset into a probability scale, RIDIT score. PCA is then employed to the RIDIT score data matrix to capture the highest variability in the dataset. However, the classical PRIDIT method framework has a limitation in the form of the PCA based Pearson correlation's measures being insensitive to the variability of the data. In addition, there are no specific measurements for assessing the PRIDIT method's performance under different data characteristics. Hence, this study proposed a robust PRIDIT methodology framework by incorporating several robust estimators (M-Huber, M-Tukey Bisquare, MM and LTS estimators) to improve the performance of classification tasks in identifying potentially fraudulent case data. The proposed method is applied on a German Credit Card Dataset. The analysis indicates that the highest accuracy rate of 48.5% was obtained by robust PRIDIT based on M-Tukey Bisquare estimator, followed by the results of robust PRIDIT based on MM and LTS estimators, which show similar accuracy scores of 48.1% with classical PRIDIT. The lowest accuracy score was obtained by robust PRIDIT based on M-Huber at 47.9%. A simulation study was also conducted to assess the performance of different PRIDIT methods. Behaviours of different PRIDIT methods were observed under different credibility percentage settings (Non-Fraud (NF); Fraud (F) cases, 95%NF;5%F, 90%NF;10%F, 80%NF;20%F and 70%NF;30%F) and variability levels (low, medium and high) in the datasets. The simulation results show that the accuracy rate obtained by classical PRIDIT, robust PRIDIT based M-Tukey Bisquare, MM, LTS and Huber are 64.3%, 65.3%, 65%, 63.7% and 61.7% respectively at credibility setting (70%NF;30%F) and medium variability. Thus, the findings indicate that the robust PRIDIT based on M-Tukey Bisquare outperform the other estimators by achieving the highest accuracy rate of 65.3%. In addition, the robust PRIDIT method also has a better rate of accuracy when data variability is medium or high compared to the classical PRIDIT method. Thus, this study has introduced a new method using robust PRIDIT to assess the credibility of financial data effectively.

ABSTRAK

Peningkatan bilangan kes penipuan boleh membahayakan kemampubayaran perniagaan. Pengenalpastian penipuan menggunakan kaedah statistik yang berkesan, seperti klasifikasi, boleh melindungi organisasi daripada masalah ini. Namun, mengenal pasti kes penipuan, adalah satu cabaran statistik disebabkan oleh beberapa ciri set data kewangan. Data ini biasanya memiliki set data besar yang berdimensi tinggi, mengandungi jenis data bercampur dan boleh melibatkan bilangan kes penipuan dan bukan penipuan yang tidak seimbang. Kajian ini menggunakan Analisis Komponen Utama (PCA) berdasarkan skor Relatif kepada skor Taburan Terpilih (RIDIT), yang dikenali sebagai kaedah PRIDIT, untuk mengklasifikasi dan mengenal pasti data yang berpotensi adalah kes penipuan. Kaedah PRIDIT klasik melibatkan transformasi setiap set data yang dianalisis kepada skala kebarangkalian, skor RIDIT. PCA kemudiannya digunakan pada matriks data skor RIDIT untuk mendapatkan kebolehubahan tertinggi dalam set data. Walau bagaimanapun, rangka kerja kaedah PRIDIT klasik mempunyai kelemahan di mana PCA menggunakan ukuran korelasi Pearson yang tidak sensitif terhadap kebolehubahan data. Selain itu, tidak ada ukuran khusus untuk menilai prestasi kaedah PRIDIT di bawah ciri data yang berbeza. Oleh itu, kajian ini mencadangkan rangka kerja metodologi PRIDIT teguh dengan menggabungkan beberapa penganggar teguh (penganggar M-Huber, M-Tukey Bisquare, MM dan LTS) untuk meningkatkan prestasi klasifikasi dalam mengenal pasti data yang berpotensi sebagai kes penipuan. Kaedah yang dicadangkan telah digunakan pada set data kad kredit Jerman. Analisis menunjukkan bahawa kadar ketepatan tertinggi pada 48.5% diperolehi oleh PRIDIT teguh berasaskan penganggar M-Tukey Bisquare, diikuti oleh keputusan PRIDIT teguh berasaskan penganggar MM dan LTS yang menunjukkan skor ketepatan sebanyak 48.1%, iaitu lebih kurang sama dengan nilai ketepatan PRIDIT klasik. Skor ketepatan terendah pula diperolehi oleh PRIDIT teguh berasaskan M-Huber pada 47.9%. Keputusan ini menunjukkan bahawa PRIDIT teguh berasaskan M-Tukey Bisquare adalah penganggar terbaik berbanding penganggar lain dengan kadar ketepatan 48.5%. Kajian simulasi juga telah dijalankan untuk mengukur prestasi kaedah PRIDIT yang berbeza. Tingkah laku kaedah PRIDIT yang berbeza diperhatikan di bawah tetapan peratusan kredibiliti yang berbeza (Kes Bukan Penipuan (NF); Penipuan (F), 95%NF;5%F, 90%NF;10%F, 80%NF;20%F dan 70%NF;30%F) dan tahap kebolehubahan (rendah, sederhana dan tinggi) dalam set data. Keputusan simulasi menunjukkan kadar ketepatan yang diperolehi oleh PRIDIT klasik, PRIDIT teguh berdasarkan M-Tukey Bisquare, MM, LTS dan Huber masing-masing ialah 64.3%, 65.3%, 65%, 63.7% dan 61.7% pada tetapan kredibiliti (70%NF; 30%F) dan kebolehubahan sederhana. Oleh itu, dapatan menunjukkan bahawa PRIDIT teguh berdasarkan M-Tukey Bisquare mengatasi penganggar lain dengan kadar ketepatan tertinggi iaitu 65.3%. Selain itu, kaedah PRIDIT teguh juga mempunyai kadar ketepatan yang lebih baik apabila kebolehubahan data adalah pada tahap sederhana atau tinggi berbanding kaedah PRIDIT klasik. Dengan demikian, kajian ini telah mengemukakan satu kaedah baharu menggunakan PRIDIT teguh untuk menilai kredibiliti data kewangan dengan lebih efektif.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

PCA	-	Principal Component Analysis
SVM	-	Support Vector Machine
RIDIT	-	Reference to Identified Distribution
PRIDIT	-	PCA on Reference to Identified Distribution (RIDIT)
UCI	-	University of California, Irvine
EMA	-	Europe, the Middle East and Africa
SVD	-	Singular Value Decomposition
LR	-	Logistic Regression
SVM	-	Support Vector Machine
KNN	-	K-nearest neighbors
NB	-	Naïve Bayes
DT	-	Decision Tree
FP	-	Frequent Pattern
SOM	-	Self Organizing Map
GA	-	Genetic algorithm
ANN	-	Artificial Neural Network
AIS	-	Artificial Immune System
BN	-	Bayesian Network
ILP	-	Inductive Logic Programming
CBR	-	Case Base Reasoning
ADR	-	Accuracy Detection Rate
TP	-	True positive
TN	-	True negative
FPR	-	False positive rate
ROC	-	Relative Operating Characteristics curve

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation of this Study

Finance data, such as business support, credit card, insurance, mortgage, and payment data, play a significant role in organizational controls from simplistic reconciliations to audits, forensics, and even governance. However, financial industries such as insurance companies, banks, and other financial institutions have recently encountered a significant number of fraud cases that has negatively impacted their businesses and the economy as a whole. According to KPMG's Global Banking Fraud report (2019), a survey on banking fraud involving 43 retail banks (13 in the Asia-Pacific region, 5 in the Americas, and 25 in Europe, the Middle East, and Africa or EMA) was conducted between November 2018 and February 2019. According to the survey, 61 percent of banks reported increased external fraud cases over the last three years, both in terms of value and volume. The survey also found that over half the respondents recovered less than 25 percent of losses incurred through fraud, which demonstrates that fraud prevention and related strategies are vital for protecting banks against fraudsters.

Subsequent expansion and increase in fraud cases will result in billions of dollars in losses for financial institutions, as well as a reduction in the company's ability to operate. Fraud in the workplace can have a negative effect on the economy and lead to businesses becoming insolvent. As a result, financial data monitoring and effective classification of potential fraud cases could protect businesses from unassuming pitfalls. Previous research on financial fraud, such as bank fraud, insurance fraud, and commodities fraud, has been hampered by several difficulties and obstacles. Scholars have made significant attempts to establish approaches or strategies for objectively predicting, classifying, clustering, or profiling possible fraud cases in the financial industry.

However, there are several issues in handling financial data and methods developed in previous studies. There is no exact figure or information regarding actual fraud cases in the datasets, making it difficult to identify and classify fraud cases through transactions. Most companies are very sensitive to the idea of publishing real fraud cases due to worries of declining clients' trust (Ai, 2008). In insurance cases, Brockett et al. (2002) stated that insurance investigators, adjusters, and insurance claim managers are often faced with situations where there is incomplete information for making decisions concerning the validity or possibility of a particular filed claim stamped with a fraudulent status.

In other cases, numerous commercial banks and insurance institutions still use the judgemental approach in classifying tasks, such as classifying fraud or non-fraud in a claim or whether to extend credit or not. The method involves high costs if the case is a high-profile fraud case and this makes the decision-making process ineffective. Therefore, a systematic and comprehensive statistical methodology or system is needed to discover, detect or classify fraud and improve decision-making in the claim process. Hence, the issue lies in classifying and identifying fraudsters in data transactions.

Understanding the characteristics of data variables is critical when dealing with financial data in order to examine and further analyse it. The issue in identifying financial fraud data is a crucial, complicated, and very challenging task since the database for such transactions is quite large, real-time and highly dimensional with more variables of interests and having a mixed type of data. Furthermore, these data are typically imbalanced in nature and fraudulent transactions make up a very small percentage compared to non-fraudulent cases. Therefore, it may be difficult to detect fraud. According to Wang, Zhao, & Li (2019), the banking industry faces problems as the amount of fraud transaction data is too small when using machine learning or other methods to construct fraud detection models, thus, it affects the training of anti-fraud models and the detection of fraud transactions.

Thus, an effective fraud detection technique should have the capability to address these difficulties to achieve an optimum performance. This study focuses on

improving the classification task involved in identifying potential fraud cases, specifically in credit card applications, based on their profile and past credit history.

One analytical method to counter the issue of credit financial datasets is the unsupervised scoring method, namely the classical Principal Component Analysis (PCA) Relative to an Identify Distribution (RIDIT) score or PRIDIT, as it suits the characteristics of a credit dataset. The classical PRIDIT method involves the transformation of each analysed dataset into a probability RIDIT scale (Mishra, Mohanty, and Mall (2018), Agostinho and Cherry, 2014, Brockett et al., 2002, Ai, Golden, and Brockett, 2009 and Bross, 1958). However, although there are limited references in the literature, unsupervised learning PRIDIT methods are still being employed in many applications, especially in financial data.

Principal Component Analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss by creating new uncorrelated variables that successively maximize variance (Jolliffe and Cadima, 2016). PCA deliberately reduces the dimensionality of a specific dataset and keeps the most significant variances in values in it (Andrić et al., 2016; Keng et al., 2015; Sriwijayanti, Raupong, and Sunusi, 2019). PCA is also intimately related to Singular Value Decomposition (SVD) since the mean arithmetic of the principal component is zero and equal to an eigenvector of the covariance matrix sorted by corresponding to the eigenvalue or equivalent to the variance of data. The principle coefficients are the linear coefficient used to construct the initial data set based on the principal component.

The classical PRIDIT employs PCA-based Pearson correlation in the RIDIT data matrix to capture the highest eigenvalue and eigenvector. The PCA technique, on the other hand, has its own flaws. According to Hubert and Engelen (2004), outliers are particularly susceptible to the standard PCA technique, which is based on the data's mean and sample covariance matrix. In the presence of outlying observations, classification algorithms based on this covariance matrix do not produce satisfactory results. Therefore, to overcome the flaws, a few scholars proposed robust estimators to classical PCA. For example, Engelen, Hubert, and Vanden Branden (2016)

compared three procedures for robust PCA in High Dimensions. Results revealed that the robust PCA (ROBPCA) is extremely robust and able to survive a wide range of contaminants.

As mention by Cheikh (2014), one way to 'robustify' a PCA is by replacing the empirical mean and covariance matrix with robust versions, such as M-estimators (Maronna, 1976), minimum covariance determinant (MCD) estimators, and reweighted versions of them (Croux and Haesbroeck, 1999; Rousseeuw, 1985), S-estimators (Davies, 1987), and other proposed versions, for example, by Ma and Genton (2001) and Kamiya (2001). Thus, in order to resist this variability and other effects using the Pearson coefficient of correlation, various robust correlation methods were employed by Ahad et al. (2018), Croux, Filzmoser, and Oliveira (2011) and Engelen et al. (2016a). Due to the issue of the Pearson correlation, a few robust estimators suggested incorporating the correlation matrix from the classical PRIDIT framework since many studies have found that robust estimators provide better results in terms of accuracy rate and classification task.

Current studies on credit application still lack references on the use of simulation data to assess the performance of the PRIDIT methods. Previous studies on the assessment of the classical PRIDIT approach have often compared competitive techniques, such as clustering, logistic regression, and SVM (Ai, J et al., 2012). No study until date has checked the performance under different dataset settings.

Hence, due to the above-mentioned issues concerning financial data and the classical PRIDIT method, this study aims to improve the classical PRIDIT framework and assess its performance by simulating data to address the issues. The main contribution of this research is the development of a robust PRIDIT methodology framework that can accommodate an approximate outlying measurement in the data matrix for the purpose of classifying financial data and identifying fraud cases.

Performances by PRIDIT and robust PRIDIT methods are assessed by comparing the ability of different PRIDIT methods to detect potential fraud cases in this unsupervised approach using simulated data. The goal of using simulated data is

to demonstrate and evaluate PRIDIT's performance in classifying fraud and non-fraud cases at numerous different credibility percentages and degrees of variability, as well as to identify the most effective type of robust estimator that can be incorporated into the robust PRIDIT framework.

1.2 Problem Statement

Handling financial data is very challenging due to the unique features of financial data; for example, dataset transactions are very large, and raw datasets are typically unbalanced or skewed. There is also no exact figure or information regarding actual fraud cases in the datasets, making it difficult to classify and identify fraud cases through these transactions.

The PRIDIT scoring method is an analytical method that can classify and identify fraud cases in financial credit datasets. The classical PRIDIT employs Principal Component Analysis (PCA) based Pearson correlation in the RIDIT data matrix to capture the highest eigenvalue and eigenvector. This method is still not robust since the PCA in RIDIT uses classical estimators which is insensitive to the variability present in high dimensional data. One method to counter the issue is to introduce robust estimators but the challenging task is to determine which is the best estimator and most suitable to be incorporated into the PRIDIT framework.

Current studies on credit application lack the focus on examining the use of simulation data to assess the performance of the PRIDIT method. Previous studies compare between competitive supervised methods such as Logistic Regression, Support Vector Machine and others when assessing the classical PRIDIT method. There is yet any study until now that has examined its performance under different simulated dataset settings.

1.3 Aim of the Research

The aim of this research is to develop a robust PRIDIT scoring method based on robust estimators correlation matrix to improve the performance of the classification task and identify potential fraud in financial datasets.

1.4 Research Objectives

The aim is expressed in a set of specific objectives that provide direction for this research. Therefore, the purpose of this research are listed as follows:

i) To determine the best family of robust estimators in a PRIDIT scoring method.

ii) To propose a modification of the PRIDIT scoring methodology framework for improving the classification task's performance.

iii) To analyse the performance of modified PRIDIT scoring on different levels of contaminated data (via credit applicant's variability and credibility setting).

1.5 Scope of the Research

The scope of this research deals with the following considerations:-

i) Type of data is the primary factor in any research; therefore, this research focuses on mixed datasets consisting of categorical and continuous data in the analysis. Data used are secondary data, (German Credit dataset) retrieved from the UCI Machine Learning Repositories (https://archive.ics.uci.edu/ml).

ii) Numerous classification tasks have applied supervised and unsupervised learning methods. However, this research focuses on the unsupervised method with no training data or predictor variable. The method employs unsupervised

Principle Component Analysis (PCA) based on RIDIT or PRIDIT to improve classification tasks and identifying potential fraud cases involving financial and credit datasets.

iii) This research used the R Statistical Programming method to develop and run the algorithm when assessing the PRIDIT scoring methodology's performance. The simulation datasets that mimic certain characteristics of original datasets also generated different levels of contaminated data (via credit applicant's variability and credibility settings), which will be then analyzed.

1.6 Significance of the Research

This research aims to develop the robust unsupervised scoring procedure based on the PRIDIT approach to tackle the problem of misclassification or inaccuracy of fraudulent cases in financial data. Therefore, this research contributes to the existing corpus of knowledge related to the financial industry through the robust PRIDIT scoring method. The significance of this research is listed as follows:

i) The robust PRIDIT scoring method can improve the classification task when detecting fraudulent cases.

ii) Results of the scores will assist the decision making process since this method can classify and indentify potential fraud cases in the financial dataset.

iii) Output from the statistical analysis of the robust PRIDIT scoring method will help the management or executives to protect the organisation and evaluate risks before proceeding with the credit loan or claim.

iv) Cost minimization of fraud investigations through a simple and more accurate unsupervised statistical method.

1.7 Contributions of the Research

Contributions by this research are wide and immense in scope. The improved methodology assesses the performance of the classification task and detects potential fraud cases in financial and credit datasets.

The main contribution of this research is the method that can determine the best estimators by comparing classical PRIDIT and robust PRIDIT methods using German Credit datasets. Thus, it would provide better results when estimating PRIDIT scores, leading to a more accurate classification task. The research has developed a robust PRIDIT methodology framework by incorporating robust estimators into the classical PRIDIT in order to improve the classification task and identify potential fraud cases in financial and credit datasets.

The research assesses the performance of methods through a simulation study to classify and identify potential fraud cases. This simulation enables the research to measure the ability of different PRIDIT methods to detect the implanted applicants labelled as 'fraudulent' through different levels of contaminated data (via credibility settings and level of variability). Finally, the research introduces an algorithm and an alternative method to rank the scores and determine the best estimators for the robust PRIDIT method.

1.8 Organization of the Thesis Structure

This thesis begins with Chapter 1, which introduces the motivation and background of the problems related to the issue of financial data. In addition, this chapter also provides the aim, objectives, scope, significance and contribution of the research. In essence, this chapter provides a general overview of the thesis.

Chapter 2 examines existing literature related to the area of research. This chapter presents a general overview of fraud detection using financial data, review of data mining techniques and scoring methodologies for fraud detection, reviews credit

scoring for data mining methodologies and its advantages, disadvantages as well as an evaluation of credit card fraud detection, reviews the development of the PRIDIT scoring method and the application of RIDIT and PRIDIT methods, discusses limitations of the classical PRIDIT method, and reviews the robust estimators. The comparison of the scoring method, RIDIT and PRIDIT are also discussed here.

Chapter 3 presents the PRIDIT methodology found in the research framework. The classical PRIDIT methodology is discussed, together with the robust PRIDIT framework employed in this research. The robust PRIDIT methodology framework developed by incorporating several robust estimators into the classical PRIDIT and the algorithm are discussed and this is one of the main contributions of this research. This chapter also provides the algorithm and an alternative method for ranking credit applicants.

Chapter 4 describes the selection of variables for further analysing the PRIDIT method and generating the simulated data. German credit data are then explored and this chapter provides a guide on how to select the variables. It also provides results of the analysis on classical PRIDIT and robust PRIDIT using the original German credit dataset. The result lead to the formulation of the study's first objective .

Chapter 5 presents the analysis and discussions regarding the use of classical PRIDIT and robust PRIDIT methodologies that employed simulated data. It discusses how to generate simulated data to acsess the performance of both, the classical and robust PRIDIT methods, to improve the classification task and identify fraud cases. The performance methods have been assessed by comparing the accuracy rate in classifying and detecting potential fraud cases using simulated data with different levels of contaminated data (via credit applicant's variability and credibility settings). The R programming method is used for simulating data, and analysing the output of financial and credit datasets.

Chapter 6 presents the achievement of all the research objectives in order to improve the classification task and identify potential fraud in credit datasets. This research contibutes to the existing corpus of knowledge on the robust PRIDIT

methodology framework and assessment of the simulation procedures. It also discusses the limitations of the research and suggests topics for future research to explore.

Therefore, it can be concluded that the core of financial and banking data analytics is to provide significant business intelligence in terms facilitating the classification task and identifying fraud as far as quantification of risks in financial data transactions are concerned. The Unsupervised PRIDIT scoring method is proposed since there is a lack of studies on this subject matter.

Previous studies have indicated that this method can solve the problem of classification task and identifying fraud in datasets. Therefore, this study should be conducted due to limitations in the classical PRIDIT method as well as the need to improve the classification task and identify potential fraud in financial datasets.

REFERRENCES

Ahmad, N (2007). Statistical Profiling of Proteomics Data. [PhD Thesis, The University of Kent]

Alberts M.J., Bertels C., & Dawson D.V. An analysis of time of presentation after stroke. JAMA. 1990 Jan 5;263(1):65-8. PMID: 2293690.

Agilandeeswari, L., & Ganesan, K. (2016). An efficient hilbert and integer wavelet transform based video watermarking. *Journal of Engineering Science and Technology*, 11(3), 327–345.

Agostinho, B. P. J. F., & Cherry, C. J. (2014). The significance of claims fraud in microinsurance and a statistical method to channel limited fraud identification resources. *Actuarial Society of South Africa's 2014 Convention*, October, 22–23.

Ahad, N. A., Zakaria, N. A., Abdullah, S., Yahaya, S. S. S., & Yusof, N. (2018). Robust correlation procedure via sn estimator. *Journal of Telecommunication, Electronic and Computer Engineering*, 10(1–10), 115–118.

Ai,J.(2008), *Supervised and Unsupervised PRIDIT for Active Insurance Fraud Detection,* [Doctoral Dissertation, University of Texas], https://repositories.lib.utexas.edu

Ai, J., Golden, L. L., & Brockett, P. L. (2009). Assessing consumer fraud risk in insurance claims: An unsupervised learning technique using discrete and continuous predictor variables. *North American Actuarial Journal*, 13(4), 438–458. https://doi.org/10.1080/10920277.2009.10597568

Ai, J., Brockett, P. L., Golden, L. L., & Guillén, M. (2013). A Robust Unsupervised Method for Fraud Rate Estimation. *Journal of Risk and Insurance*, 80(1), 121–143. https://doi.org/10.1111/j.1539-6975.2012.01467.x

Almetwally, E.M., & Almongy, H.M. (2018). Comparison Between M-Estimation, S-Estimation, And MM Estimation Methods of Robust Estimation with

Application and Simulation. International Journal of Mathematical, 9(11), 55-63.

Andrić, F., Bajusz, D., Rácz, A., Šegan, S., & Héberger, K. (2016). Multivariate assessment of lipophilicity scales—computational and reversed phase thin-layer chromatographic indices. *Journal of Pharmaceutical and Biomedical Analysis*, 127, 81–93. https://doi.org/10.1016/j.jpba.2016.04.001

Bandyopadhyay, U., & Biswas, A. (2016). Fixed-width confidence interval for two-stage response-adaptive designs in ridit analysis. *Statistics & Probability Letters*, 108, 45–51. https://doi.org/10.1016/j.spl.2015.09.020

Bashiri, M., & Moslemi, A. (2013). Simultaneous robust estimation of multi-response surfaces in the presence of outliers. *Journal of Industrial Engineering International*, 9(1), 1-12.

Bhattacharjee, B., Sridhar, A., & Shafi, M. (2017). An artificial neural network-based ensemble model for credit risk assessment & deployment as a graphical user interface. *International Journal of Data Mining, Modelling and Management*. https://doi.org/10.1504/IJDMMM.2017.085643

Bolton J.R. & Hand D.J (2002). Statistical Fraud Detection: A Review. *Statistical Science*. 17 (3), 235–255. https://doi: 10.1214/ss/1042727940

Brockett, P. L., Derrig, R. A., Golden, L. L., Levine, A., & Alpert, M. (2002). Fraud classification using principal component analysis of RIDITs. *Journal of Risk and Insurance*, 69(3), 341–371.

Bross, I. D. J. (1958). How to Use Ridit Analysis. *International Biometric Society* Stable URL : http://www.jstor.org/stable/2527727 . 14(1), 18–38.

Brause, R., Langsdorf, T., & Hepp, M. (1999). Neural data mining for credit card fraud detection. *Proceedings 11th International Conference on Tools with Artificial Intelligence*. doi:10.1109/tai.1999.809773

Chang, H.-L., & Yang, C.-H. (2008). Do airline self-service check-in kiosks meet the needs of passengers? *Tourism Management*, 29(5), 980–993. doi:10.1016/j.tourman.2007.12.002

Cheikh, M. (2014). Comparative study of robust estimators based on a sensitivity coefficient in principal component analysis. *Communications in Statistics: Simulation and Computation*, 43(10), 2639–2648. https://doi.org/10.1080/03610918.2012.762390

Chen, M.-C. and Huang, S.-H. (2003) 'Credit scoring and rejected instances reassigning through evolutionary computation techniques', *Expert Systems with Applications*, 24(4), pp. 433–441.

Chen, T. T., Lai, M. S., Lin, I. C., & Chung, K. P. (2012). Exploring and comparing the characteristics of nonlatent and latent composite scores: Implications for pay-for-performance incentive design. Medical Decision Making. https://doi.org/10.1177/0272989X10395596

Chen, W., Ma, C., & Ma, L. (2009). Mining the customer credit using hybrid support vector machine technique. *Expert Systems with Applications*, 36(4), 7611–7616. https://doi.org/10.1016/j.eswa.2008.09.054

Clerc, M. and Kennedy, J. (2002).The particle swarm - explosion, stability, and convergence in a multidimensional complex space. *IEEE Transactions on Evolutionary Computation*, 6(1), pp. 58–73.

Croux, C., Filzmoser, P., & Oliveira, M. R. (2011). Algorithms for Projection-Pursuit Robust Principal Component Analysis. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.968376

Charles D. Craig, G. David Faulkenberry (1979), The application of ridit analysis to detect trends in visibility, *Atmospheric Environment* (1967), 13(12), 1617-1622, https://doi.org/10.1016/0004-6981(79)90319-6.

Donaldson, G.W. (1998). Ridit scores for analysis and interpretation of ordinal pain data. *European Journal of Pain*, 2(3), 221-227. doi: 10.1016/s1090-3801(98)90018-0

De Menezes, D. Q. F., Prata, D. M., Secchi, A. R., & Pinto, J. C. (2021). A review on robust M-estimators for regression analysis. *Computers & Chemical Engineering*, 147, 107254. doi:10.1016/j.compchemeng.2021.10

El-rouby MG. (1994). Analysing attitude data through ridit schemes. *The Egyptian population and family planning review*, 28, 183-203.

Engelen, S., Hubert, M., & Vanden Branden, K. (2016). A Comparison of Three Procedures for Robust PCA in High Dimensions. *Austrian Journal of Statistics*, 34(2), 117. https://doi.org/10.17713/ajs.v34i2.405

Filzmoser, P., Höppner, S., Ortner, I., Serneels, S., & Verdonck, T. (2020). Cellwise robust M regression. *Computational Statistics & Data Analysis*, 106944, 1-14. doi:10.1016/j.csda.2020.106944

Flora, D.J. (2014). Ridit Analysis. Wiley StatsRef: Statistics Reference Online. https://doi.org/10.1002/9781118445112.stat00378

Francis, L. A. (2016). Application of Two Unsupervised Learning Techniques to Questionable Claims: PRIDIT and Random Forest. Predictive Modeling Applications in Actuarial Science, 180–207. https://doi.org/10.1017/cbo9781139342681.008

Friedman, S., & Weisberg, H. F. (1981). Interpreting the First Eigenvalue of a Correlation Matrix. *Educational and Psychological Measurement*, 41(1), 11–21. doi.org/10.1177/001316448104100102

Fleiss, J.L, Chilton N.W. and Wallenstein S. (1979). Ridit Analysis in Dental Studies *Journal of Dental Research*. 58(11). 2080-2084. doi: 10.1177/00220345790580110701

Fielding, A. (1993). Scoring functions for ordered classifications in statistical analysis. *Quality and Quantity*, 27(1), 1-17

Gibson, D., & de Freitas, S. (2016). Exploratory Analysis in Learning Analytics. *Technology, Knowledge and Learning*. https://doi.org/10.1007/s10758-015-9249-5

Golden L.L., Brockett P.L., Guillén M. & Manika D. (2019). aPRIDIT Unsupervised Classification with Asymmetric Valuation of Variable Discriminatory Worth. *Multivariate Behavioral Research*. doi: 10.1080/00273171.2019.1665979

Goossen, W. T. ., Epping, P. J. M. ., Feuth, T., van den Heuvel, W. J. ., Hasman, A., & Dassen, T. W. . (2001). Using the nursing minimum data set for the Netherlands (NMDSN) to illustrate differences in patient populations and variations in nursing activities. *International Journal of Nursing Studies*, 38(3), 243–257. doi:10.1016/s0020-7489(00)00075-4

Hassibi, K. (2000). Detecting Payment Card Fraud with Neural Networks. Business Applications of Neural Networks, 141–157. doi: 10.1142/9789812813312_0009

Ha, V. S., & Nguyen, H. N. (2016). Credit scoring with a feature selection approach based deep learning. *MATEC Web of Conferences.* https://doi.org/10.1051/matecconf/20165405004

Howard, P. J. A. & Howard, D. M. (1985). The application of ridit analysis to phenological observations, *Journal of Applied Statistics*, 12(1), 29-35, DOI: 10.1080/02664768500000004

Huber, P.J. , 1964. Robust estimation of a location parameter. Ann. Math. Stat. 35 (1), 73–101 .

Huber, P.J. , 1981. Robust Statistics. John Wiley & Sons . Huber, P.J. , 1984. Finite sample breakdown of m-and p-estimators. Ann. Stat. 119–126 .

Huber, P.J. , Ronchetti, E.M. , 2009. Robust Statistics, Vol. 10. A John Wiley & Sons, Inc., Publication .

Huber, P.J. , et al. , 1973. Robust regression: asymptotics, conjectures and Monte Carlo. Ann. Stat. 1 (5), 799–821 .

Hu, Y. R., & Hu, A. R. (2009). Analysis of risk factors of patients with chronic liver failure complicated invasive fungal infections. Chinese journal of experimental and clinical virology, 23(3), 214-217

Hu, M.-L. M., Horng, J.-S., Teng, C.-C. C., & Yen, C.-D. (2013). Assessing students' low carbon literacy by Ridit IPA approach. Journal of Hospitality, Leisure, Sport & Tourism Education, 13, 202–212. doi:10.1016/j.jhlste.2013.09.006

Hubert, M., Rousseeuw, P. J., & Vanden Branden, K. (2005). ROBPCA: A new approach to robust principal component analysis. *Technometrics.* https://doi.org/10.1198/004017004000000563

Hubert, M., Rousseeuw, P. J., & Verdonck, T. (2012). A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics*, 21(3), 618–637. https://doi.org/10.1080/10618600.2012.672100

Huang, L., & Tsai, H. T. (2003). The study of senior traveler behavior in Taiwan. *Tourism management*, 24(5), 561-574.

Jiang, Y., Wang, Y-G, Fu, L.& Wang,X.,(2019). Robust Estimation Using Modified Huber's Functions With New Tails, *Technometrics*, 61:1, 111-122, https://doi: 10.1080/00401706.2018.1470037

Jolliffe IT, Cadima J. (2016). Principal component analysis: a review and recent developments.Phil. *Trans. R. Soc*. A 374:20150202. http://dx.doi.org/10.1098/rsta.2015.0202

Lieberthal, R. D., & Comer, D. M. (2013). What Are the Characteristics That Explain Hospital Quality? A Longitudinal Pridit Approach. *Risk Management and Insurance Review*, 17(1), 17–35. doi:10.1111/rmir.12017

Kafadar, K. (1983). Efficiency of the Biweight As a Robust Estimator of Location. *Journal of Research of the National Bureau of Standards (United States*), 88(2), 105–116. https://doi.org/10.6028/jres.088.006

Kang, Z., Peng, C., & Cheng, Q. (2016). Robust PCA via nonconvex rank approximation. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2016-January, 211–220. https://doi.org/10.1109/ICDM.2015.15

Keng, S. E., Easa, A. M., Hoong, C. L., Al-Karkhi, A. F. M., & Mohd Talib, M. K. (2015). An investigation of potential fraud in commercial orange juice products

in malaysian market by cluster analysis and principal component analysis. *Malaysian Journal of Analytical Sciences*, 19(2), 377–387.

Kinser, J. M., & Kinser, J. M. (2018). Principle Component Analysis. In Image Operators. https://doi.org/10.1201/9780429451188-8

KPMG. (2019). Global Banking Fraud Suvey. https://assets.kpmg/content/dam/kpmg/au/pdf/2019/global-banking-fraud-survey-2019-au.pdf

Liu, C., White, M., & Newell, G. (2018). Detecting outliers in species distribution data. *Journal of Biogeography*, 45(1), 164–176. https://doi.org/10.1111/jbi.13122

Lin, W.-T., & Chen, C.-Y. (2013). Shopping Satisfaction at Airport Duty-Free Stores: A Cross-Cultural Comparison. *Journal of Hospitality Marketing & Management*, 22(1), 47–66. doi:10.1080/19368623.2012.680242

De Menezes, D. Q. F., Prata, D. M., Secchi, A. R., & Pinto, J. C. (2021). A review on robust M-estimators for regression analysis. Computers & Chemical Engineering, 147, 107254. doi:10.1016/j.compchemeng.2021.10

Mishra, P. C., Mohanty, M. K., & Mall, M. (2018). Operational attributes influencing productivity in Indian opencast mines. *International Journal of Productivity and Quality Management*. https://doi.org/10.1504/IJPQM.2018.091790

Miao, E. Y., Miao, M. Y. M., Kildea, D. G., & Lao, Y. W. (2014). Effects of electroacupuncture and electroacupuncture plus Tao Hong Si Wu Wan in treating primary dysmenorrhea. *Journal of acupuncture and meridian studies*, 7(1), 6-14.

Mielke Jr, P. W., Long, M. A., Berry, K. J., & Johnston, J. E. (2009). g-Treatment ridit analyses: Resampling permutation methods. *Statistical Methodology*, 6(3), 223-229.

Muehrer, R. J., Brown, R. L., & Lanuza, D. M. (2014). Depicting Changes in Multiple Symptoms Over Time. *Western Journal of Nursing Research*, 37(9), 1214–1228. doi:10.1177/0193945914542163

Munzel, U. and Langer, F. (2004), A Global View on Parametric and Nonparametric Approaches to the Analysis of Ordered Categorical Data. *Biometrical Journal*, 46: 7-18. https://doi.org/10.1002/bimj.200210001

Nataraja, P., & Raju, G. T. (2013). Quantitative influence of HCI characteristics in a blended learning system. Education and Information Technologies, 18(4), 687-699.

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, 50(3), 559–569. https://doi.org/10.1016/j.dss.2010.08.006

Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., & Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. Applied Soft Computing Journal. https://doi.org/10.1016/j.asoc.2018.10.004

Owen, M. (2010). Tukey's Biweight Correlation and the Breakdown. Master's Thesis, Pomona College. http://www.pages.pomona.edu/~jsh04747/Student Theses/MaryOwen10.pdf

Patil, S., Nemade, V., & Soni, P. K. (2018). Predictive Modelling for Credit Card Fraud Detection Using Data Analytics. Procedia Computer Science, 132, 385–395. https://doi.org/10.1016/j.procs.2018.05.199

Pouplard, N., Qannari, E. M., & Simon, S. (1997). Use of ridits to analyse categorical data in preference studies. *Food quality and preference*, 8(5-6), 419-422.

Pejic-Bach, M. (2010). Profiling intelligent systems applications in fraud detection and prevention: Survey of research articles. ISMS 2010 - UKSim/AMSS 1st International Conference on Intelligent Systems, Modelling and Simulation. https://doi.org/10.1109/ISMS.2010.26

Randhawa, K., Loo, C. K., Seera, M., Lim, C. P., & Nandi, A. K. (2018). Credit Card Fraud Detection Using AdaBoost and Majority Voting. IEEE Access. https://doi.org/10.1109/ACCESS.2018.2806420

Ren, H., Ye, Z., & Li, Z. (2017). Anomaly detection based on a dynamic Markov model. Information Sciences. https://doi.org/10.1016/j.ins.2017.05.021

Sinova, B., & Van Aelst, S. (2018). Advantages of M-estimators of location for fuzzy numbers based on Tukey's biweight loss function. *International Journal of Approximate Reasoning*, 93(November), 219–237. https://doi.org/10.1016/j.ijar.2017.10.032

Shaharudin, M. S., Ahmad, N., & Yusof, F. (2013). Improved Cluster Partition in Principal Component Analysis Guided Clustering. *International Journal of Computer Applications*, 75(11), 22–25. https://doi.org/10.5120/13156-0839

Shaharudin, M. S., Ahmad, N., & C. M. Nor, S. M. (2020). A modified correlation in principal component analysis for torrential rainfall patterns identification. *International Journal of Artificial Intelligence*, 9(4), 655-661. https://doi: 10.11591/ijai.v9.i4.pp655-661

Sriwijayanti, Raupong, & Sunusi, N. (2019). Robust Principal Component Analysis with Modified One-Step M-Estimator Method. *Journal of Physics: Conference Series*. https://doi.org/10.1088/1742-6596/1341/9/092008

Thalor, M. A., & Patil, S. T. (2018). Propagation misclassified instances to handle nonstationary imbalanced data stream. *Journal of Engineering Science and Technology*, 13(4), 1134–1142.

Uwawunkonye, E. G., & Anaene, O. I. C. (2013). A comparative study between ridit and modified ridit analysis. *American Journal of Theoretical and Applied Statistics*, 2(6), 248-254. https://doi.org/10.11648/j.ajtas.20130206.23

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, 19(1). doi:10.1186/s12911-019-1004-8

Vandewalle, K. S., Ferracane, J. L., Hilton, T. J., Erickson, R. L., & Sakaguchi, R. L. (2004). Effect of energy density on properties and marginal integrity of posterior resin composite restorations. Dental Materials, 20(1), 96–106.

Vaughan, G. (2018). Efficient big data model selection with applications to fraud detection. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2018.03.002

Wang, L., Xia, J. L., Yu, L. L., Li, C. J., & Wang, S. Z. (2008). The relationship between Ridit analysis and rank sum test for one-way ordinal contingency table in medical research. Zhonghua yu Fang yi xue za zhi [Chinese Journal of Preventive Medicine], 42(6), 427-430.

Wang, X., Bai, M., Shen, D., Nie, T., Kou, Y., & Yu, G. (2017). A distributed algorithm for the cluster-based outlier detection using unsupervised extreme learning machines. *Mathematical Problems in Engineering*. https://doi.org/10.1155/2017/2649535

Wang, X., Zhao, R., & Li, Y. (2019). A Fraudulent Data Simulation Method Based on Generative Adversarial Networks. *Journal of Physics: Conference Series*. https://doi.org/10.1088/1742-6596/1302/2/022089

Wang, B., Kong, Y., Zhang, Y., Liu, D., & Ning, L. (2019). Integration of Unsupervised and Supervised Machine Learning Algorithms for Credit Risk Assessment. *Expert Systems with Applications*. doi:10.1016/j.eswa.2019.02.033

Yao, J., Zhang, J., & Wang, L. (2018). A financial statement fraud detection model based on hybrid data mining methods. *2018 International Conference on Artificial Intelligence and Big Data* (ICAIBD). doi:10.1109/icaibd.2018.8396167

Yohai, V. J. (1987). High Breakdown-Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics*, 15(2), 642–656. doi:10.1214/aos/1176350366

Yu, C. & Yao, W. (2016). Robust Linear Regression: A Review and Comparison, *Communications in Statistics - Simulation and Computation*, DOI:10.1080/03610918.2016.1202271

LIST OF PUBLICATIONS

Index Jurnal

1. **Tukiman, N**., Ahmad, N., Mohamed, S., Othman, Z. S., Shafee, C. T. M. N. M., & Rizman, Z. I. (2017). Credit card detection system based on RIDIT approach. *International Journal on Advanced Science, Engineering and Information Technology*, 7(6), 2071–2077. https://doi.org/10.18517/ijaseit.7.6.1316