AI IN ACCOUNTING: COMBINED AUTOENCODER AND BENFORD'S LAW FIRST DIGIT ANALYSIS APPROACH FOR ANOMALY DETECTION IN ACCOUNTING JOURNAL ENTRY DATA

BACHELOR THESIS

Submitted as part of the requirements to complete a Bachelor's Degree at the Undergraduate Program (S1) of the Faculty of Economics and Business, Diponegoro University

Written by:
Muhammad Dzaki Al Aqsha
NIM. 12030121120034

FACULTY OF ECONOMICS AND BUSINESS
UNIVERSITAS DIPONEGORO
SEMARANG
2025

APPROVAL OF PROPOSED RESEARCH

Name	:	Muhammad Dzaki Al Aqsha
Student ID Number	:	12030121120034
Faculty/Department	:	Economics and Business/Accounting
Title of Thesis	:	AI IN ACCOUNTING: COMBINED AUTOENCODER AND BENFORD'S LAW FIRST DIGIT ANALYSIS APPROACH FOR ANOMALY DETECTION IN ACCOUNTING JOURNAL ENTRY DATA
Supervisor	:	Adi Firman Ramadhan, S.E., M.Ak.

Semarang, November 12, 2025

Supervisor

Adi Firman Ramadhan, S.E., M.Ak.

NIP. 198406202015041001

APPROVAL OF THESIS

Name	:	Muhammad Dzaki Al Aqsha
Student ID Number	:	12030121120034
Faculty/Department	:	Economics and Business/Accounting
Title of Thesis	:	AI IN ACCOUNTING: COMBINED AUTOENCODER AND BENFORD'S LAW FIRST DIGIT ANALYSIS APPROACH FOR ANOMALY DETECTION IN ACCOUNTING JOURNAL ENTRY DATA
Supervisor	:	Adi Firman Ramadhan, S.E., M.Ak.

Semarang, November 12, 2025

Supervisor

Adi Firman Ramadhan, S.E., M.Ak.

NIP. 198406202015041001

APPROVAL OF EXAM PASSAGE

Name	:	Muhammad Dzaki Al Aqsha
Student ID Number	:	12030121120034
Faculty/Department	:	Economics and Business/Accounting
Title of Thesis	:	AI IN ACCOUNTING: COMBINED AUTOENCODER AND BENFORD'S LAW FIRST DIGIT ANALYSIS APPROACH FOR ANOMALY DETECTION IN ACCOUNTING JOURNAL ENTRY DATA

It is hereby certified that the aforementioned undergraduate thesis has been successfully defended in front of a Board of Examiners on November 20, 2025. Consequently, the student is therefore declared to have **PASSED** the oral defense protocol.

Chairman of the Board of Examiners	:	Adi Firman Ramadhan, S.E., M.Ak.
Team of Examiners	:	1. Adi Firman Ramadhan, S.E., M.Ak. 2. Dr. Totok Dewayanto, S.E., M.Si., Akt. 3. Dr. Shiddiq Nur Rahardjo, S.E., M.Si., Akt.

Semarang, November 20, 2025

Head of the Accounting Department	Chairman of the Board of Examiners
Agung Juliarto S.E., M.Si., Akt., Ph.D.	Adi Firman Ramadhan, S.E., M.Ak.
NIP 197307222002121002	NIP 198406202015041001

STATEMENT OF THESIS ORIGINALITY

The undersigned party, Muhammad Dzaki Al Aqsha, hereby declares that this thesis titled **AI IN ACCOUNTING: COMBINED AUTOENCODER AND BENFORD'S LAW FIRST DIGIT ANALYSIS APPROACH FOR ANOMALY DETECTION IN ACCOUNTING JOURNAL ENTRY DATA** contains original research work as conducted by the undersigned party unless otherwise properly cited. This declaration entails a full, legally-binding, statement testifying the self-proclaimed lack thereof contents, whether in its entirety and/or partially, attributable to other authors without stating, citing, and/or crediting the original author(s).

In cases where circumstances arise whereby nonperformance and/or material breach in non-compliance to the covenants of this declaration is substantiated and conclusively proven through competent evidence, the undersigned party is therefore contractually obliged to withdraw the submission of the relevant work. In relevance to the subsequent turn-of-events in cases whereby evidence has been competently substantiated with regard to breaches to this non-compliance, the degrees and/or diplomas that have been rewarded in association to this work are therefore liable for cancellations.

Semarang, November 12, 2024

(Muhammad Dzaki Al Aqsha)

NIM. 12030121120034

ABSTRACT

The efficacy vis-à-vis the generalizability of conventional anomaly detection tools in large-scale accounting data has not presented itself devoid of skepticisms. In general, the current status quo with regard to the performance of the commonly used methodologies can be justifiably characterized to be a factor of dissatisfaction. The current repertoire of tools is largely based on known fraud scenarios whereby generalizability for new data tends to be compromised. This situation is further exacerbated given the costly business implications of failures in such systems, as compromised detection leads to wasteful follow-up procedures by way of operations under a pretense of false flags. At the same time, the growth in artificial intelligence and/or machine learning in other disciplines has demonstrated a sense of urgency for technological realignments in leveraging the new technologies. Within this research, we tested the efficacy of an artificial intelligence/machine learning based autoencoder neural network approach in conducting anomaly detection, benchmarked against a popular anomaly detection tool namely the Benford's law first digit analysis, alongside proposing a new method that combines the two approaches through a heuristical multiplier mechanism. By conducting an experiment on a real world dataset containing 500,000+ rows of accounting journal entry data obtained from an anonymized entity's SAP ERP BKPF and BSEG table that has been injected with synthetic anomalies, we found that the autoencoder based approach to be the best performing method both in terms of recall (sensitivity) and in balancing the precision-recall tradeoffs measured in terms of F1-Score, highlighting its great potential within the context of internal auditing. The novel proposed method was found to be inconsequential in alleviating the identified recall problem of the baseline autoencoder, although the discrepancy in performance is not conclusive enough to derive a generalizable conclusion. Collectively, these results suggest that the autoencoder neural network approach represents a promising framework in conducting anomaly detection especially in cases whereby recall is more preferable than precision (such as in internal auditing), as the opposite is true in the case for Benford's law first digit analysis (such as in external auditing).

Keywords: Anomaly Detection . Accounting . Auditing . Autoencoder . Benford's Law . Artificial Neural Network . Machine Learning . Deep learning . Artificial Intelligence

ABSTRAK

Efektivitas serta cakupan generalisasi metode deteksi anomali yang umum digunakan untuk data akuntansi berskala besar telah lama menjadi permasalahan yang diperdebatkan. Umumnya, performa pendekatan-pendekatan yang umum digunakan saat ini seringkali dinilai kurang memuaskan. Pendekatan-pendekatan yang lazim digunakan umumnya didasarkan pada kasus-kasus fraud lampau yang tidak dapat digeneralisasi dengan baik pada data-data baru. Terlebih lagi, permasalahan ini juga memiliki konsekuensi bisnis yang bersifat negatif apabila indikator temuan yang tidak tepat menyetir arah prosedur-prosedur lanjutan. Di saat yang bersamaan, perkembangan konsep serta penerapan teknologi kecerdasan buatan (AI) dan pembelajaran mesin (ML) sebagaimana telah didemonstrasikan pada bidang-bidang keilmuan lainnya sebaliknya merepresentasikan potensi solutif menjanjikan dalam menyelesaikan polemik tersebut. Penelitian ini dilakukan untuk menguji efektivitas pendekatan berbasis kecerdasan buatan/pembelajaran mesin yakni jaringan syaraf tiruan autoencoder dalam konteks deteksi anomali, yang juga dibandingkan dengan salah satu metode deteksi anomali akuntansi yang populer digunakan yaitu analisis digit hukum Benford, serta mengajukan suatu pendekatan baru yang menggabungkan kedua metode tersebut melalui suatu mekanisme multiplier yang didapatkan secara heuristik melalui eksperimen berulang. Dengan melakukan pengujian penerapan ketiga metode pada data nyata yang terdiri dari 500,000+ baris data jurnal akuntansi yang diperoleh dari tabel BKPF dan BSEG dari sistem ERP SAP suatu entitas nyata yang telah disamarkan dan disuntik dengan anomali sintetis, ditemukan bahwa pendekatan berbasis jaringan syaraf tiruan autoencoder sebagai metode yang paling efektif diukur dari sensitivitas (recall) serta dalam menyeimbangkan tradeoff precision dan recall melalui metrik F1-score. Temuan ini menggarisbawahi potensi menjanjikan penerapan pendekatan berbasis kecerdasan buatan/pembelajaran mesin terutama sekali pada konteks pengauditan internal. Pendekatan baru yang diajukan mendemonstrasikan wanprestasi dalam tujuannya untuk memperbaiki masalah recall dari model baseline autoencoder, kendatipun demikian perbedaan kinerja di antara keduanya tidak cukup signifikan untuk memberikan kesimpulan yang dapat digeneralisasi secara luas. Temuan-temuan ini menunjukkan tingginya potensi kinerja model autoencoder dalam implementasi yang mengutamakan performa recall dibandingkan dengan precision (misalnya dalam audit internal), sedangkan kondisi sebaliknya dapat dengan lebih baik diakomodasi menggunakan analisis digit pertama Hukum Benford (misalnya dalam audit eksternal).

Keywords: Deteksi Anomali . Akuntansi . Pengauditan . Autoencoder . Hukum Benford . Jaringan Syaraf Tiruan . Pembelajaran Mesin . Deep learning . Kecerdasan Buatan

FOREWORD

This thesis is presented as part of the graduation requirements in order to complete an Undergraduate Program in Accounting at the Faculty of Economics and Business, Universitas Diponegoro. In acknowledgement of the divine role involved in the unfolding of fate, prayers of gratitude are therefore extended in veneration to Allah SWT. In relation to the non-discretionary expected level of disclosures with regard to contributory academic involvements in the writing process, several acknowledgements are characterized to be imperative. The relevant parties of which contributory acknowledgements can be attributed to are as follows:

1. Adi Firman Ramadhan, S.E., M.Ak., as the Secretary of the Accounting Department, Universitas Diponegoro and the main supervisory lecturer in addition to second author of the project;
2. Agung Juliarto, S.E., M.Si., Akt., Ph.D., as the Head of the Accounting Department, Universitas Diponegoro;
3. Andrian Budi Prasetyo, S.E., Akt., M.Si., as the author's main Academic Advisory lecturer;
4. Prof. Faisal, S.E., M.Si., Ph.D., as the Dean of Faculty of Economics and Business, Universitas Diponegoro;
5. Dr. Amir Husin, S.T., M.T. and Dra. Deny Supriharti, M.Sc. as the author's biological parental figures that have provided the necessary operational funds in conducting the research, alongside substantial academic advice in scientific writing as well as in proofreading the thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I
INTRODUCTION

1.1. Research Background

The sub-discipline of accounting information system pertains to discourses on efficiency to a degree that its architecture helps in alleviating inefficiencies in its sequential and parallelized procedures. In the context of the modern environment contemporary accounting operates on, automation has been a mainstay topic given the repetitive nature of some of its procedures. Even in areas that necessitate dynamic processes, partial implementation of automation has presented itself as inseparable from the modern business process. Automated solutions in accounting in and of themselves are made possible with the advent of digitization. An example of this relationship is demonstrated by the rapid adoption of digitization in the form of spreadsheet services that automate processing and calculations, even if the majority of the procedures entail human involvement. Highlighting the dichotomous divide between automation and digitization notwithstanding, it is clear that efficiency-improving solutions have been and will continue to be the centrepiece of any, if not all, process design and organizational transformation, as is the case with automation and digitization.

In lieu of the traditional automation paradigm in the interest of a more granular context within the accounting discipline, Ning (2022) indicated non-verbatim that artificial intelligence in particular has been a factor that drives

the acceleration rate in accounting transformation in general. In which emphasis was placed on its ability to process larger, more-complex data, with a relatively simpler setup in a shorter time frame. By way of interpretation this can also extend to the underlying relative ease-of-use and setup of established AI tools such as LLM APIs in performing tasks such as discriminative modelling especially involving natural language processing procedures. In contrast, traditional procedural automation might require a more extensive development, and in the case of machine learning models, a development phase in model training.

Furthermore, conceptual and implementative overviews have also brought forward the need for a paradigm shift in accounting process automation, highlighted in a study by Nwaimo et al. (2024), accentuating the use of accounting predictive analysis using machine learning techniques in response to the fast-paced and rapidly growing contemporary economy, placing importance especially on the real-time nature of such implementation in decision making processes. In a manner similar to the previous citation, Theodorakopoulos et al. (2025) cited real-time capability in conjunction with predictive accuracy as the primary justification for the necessity in introducing, in this case, deep-learning based (a subset of machine learning) accounting analytics. The proposition was similarly postulated and subsequently demonstrated in response to the fast-paced nature of contemporary processes, oftentimes in fractions of a minute.

In discussion of a broader overview on automation and digitization, an *a priori* postulate can be justifiably formulated that the growing trend in automation

and digitization has both contributed to the intensification of and be subjected to the complexity of data generation by accounting for sheer volume alone. This presumptive characterization is not entirely unfounded for. The adoption of new mediums that facilitates the preservation and transfer of information has historically resulted in the intensification in volume and complexity for both production and consumption of said medium. An argument for which is self-evidently conspicuous by the invention of the printing press, and further efficiency in the printing process as the advent of the newspaper has demonstrated. Likewise, digitization of the anthroposphere in general, and the economy in particular is of no difference. In a way, the *ad nauseam* discourse to decide a binary decision of whether to automate or not to automate, as a consequence and in response to the rate of digitization, is of no concern under the implication of such matter. That the discourse should and in a way has largely shifted to the more granular implementations, given the implication of inevitability.

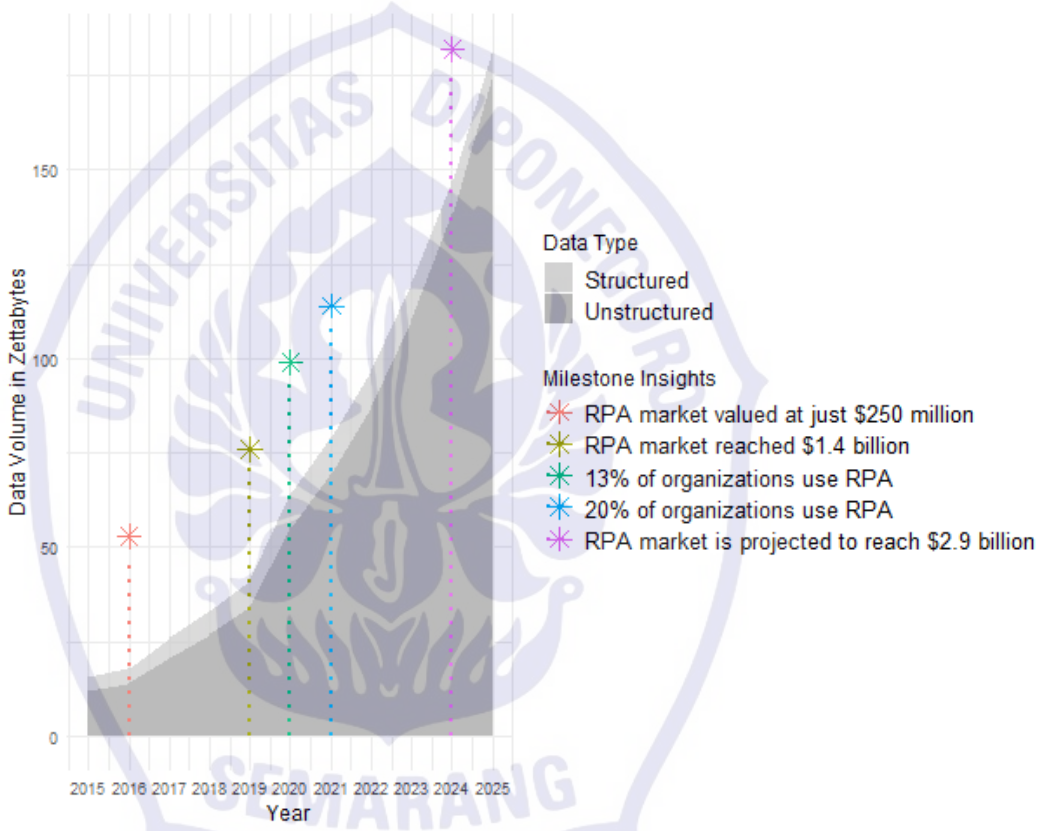
In contemporary terms, complexity can be heuristically characterized as both the main driver to and the consequence of automation and digitization (in the case of accounting, automation that arises from digitization). While data complexity can be specifically measured, the complexity and the nature of its calculation pertain to constraints of applicability to a known dataset. In terms of estimating complexity for the entire data generated in the contemporary world, the use of proxy variables is therefore reasonably justifiable. By using the estimated proportion of all data in the world, bifurcated between structured and unstructured

data, the smaller the ratio indicates higher share of unstructured data, thus serving as a proxy for higher worldwide data complexity. The stacked area plot below in Figure 1.1. visualizes the increasing complexity of worldwide data, as the interpolation (using linear method) of two conveniently accessible observations both in 2011, in Villars et al. (2011) as well as for 2022 in a whitepaper from IDC (2023), where the proportion of unstructured data in 2011 was estimated to be 61.8%, and 90% in 2022. The state of impairment as accessibility to other data points is otherwise impaired or subjected to resource constraint notwithstanding, it is imperative to note by way of this proxy the notion of increasing complexity, as the gap between the stacked area plot gets narrower over time.

Likewise, the use of proxy is once again demonstrated in the presence of the total annual worldwide data volume in zettabytes, in Petroc (2024), as the volume of data worldwide is approximated to have increased more than seven times the number it was in the last decade. This increase in data volume functions as a proxy to estimate the rate of digitization in the modern economy, and moves in a concurrent manner to the rise in complexity as is proxied by the share of unstructured data. Similarly, the rise of automation, particularly in the business sphere in which case the use of important RPA (Robotic Process Automation) milestone years demonstrate the concurrent nature of the rise of data complexity, digitization, and automation. In Le Clair, et al. (2017), the market share of RPA in 2016 was valued to be just around $250 million. In GVR (2025), this valuation has reached an approximated $1.4 billion by the year 2019, and is forecasted to reach $2.9 billion by the year 2024. Similarly, the use of RPA was also indicated

to have almost doubled the rate of usage between 2019 to 2020 alone, as indicated in Computer Economics (2020).

Figure 1.1. Diagram Illustrating Trends in Data Complexity, Rate of Automation, and Rate of Digitization



Source: Author's analysis based on data from Computer Economics, (2020); GVR, (2025); IDC, (2023); Le Clair et al., (2017); Petroc, (2024); Villars et al., (2011).

Simultaneously, complexity by means of dynamic and demanding cognition in higher-level decision making processes has also presented itself to be a hindrance to computer automation, as is indicated by Bresnahan et al. (1999). The same work also highlighted previous works describing the lack thereof compatibility for automation in tasks that involve exception processing (in cases

where unforeseeable circumstances where the automation procedure has not been specifically programmed for), visual and/or spatial processing, as well as non-algorithmic reasoning as indicated in Autor, Levy, and Murnane (2000). Dissuading conclusion notwithstanding, the study was based on firm-level observation of the status quo of its time that whilst adaptive processing techniques such as machine learning and artificial intelligence have existed at the time, the firm-level viability of implementing such systems was seen as impractical. It is imperative to note that the advent of these new paradigms and methodologies have and will continue to open up new possibilities for both more intensive and more extensive implementations.

As the digitization (and subsequent automation) of informational infrastructures in general and economic activities in particular intensified, a reasonable argument by way of observation can be made that a theme of disparity is prevalent in juxtaposition between conceptual/theoretical researches on the use of artificial intelligence and machine learning in accounting, and actual empirical case studies that explore the technicality of such implementation. At the very least as of the time of writing, an identifiable pattern emerges that the *zeitgeist* of accounting in Indonesia by means of preliminary surface-level attempt in discovering literatures that meet the specified criteria (within the constraint of the geographical location) has concluded with an apparent elusiveness. In inquiry of research conclusions of similar manner as a more global phenomenon, Abbas (2025) systematically conducted a research that concluded with a similar finding specifically in the field of management accounting. Similar non-verbatim

conclusions of research gap have also been identified in a literature study by Silaen (2024) and another by Purba (2023), although the two did not highlight the country-specific research gap. Thus, an imperative is derivable in justification of the alleviating nature of the inquiries presented in this research.

In departure from the more general discourse on the concept of digitization and automation, an underlying problem is identifiable in a more granular topic within the accounting discipline where the proposition for the use of artificial intelligence and/or machine learning is warranted for. Assessing the contemporary landscape of fraud detection, the current status quo on the use of Newcomb-Benford law of distribution has not presented itself to be entirely unchallenged. Benford (1938) independently reiterated a previously identified concept that mapped out a way for anomalous observations to be identified from the overall population of a dataset, generalizable over a diverse range of dataset, whereby conformity of the leading digits (and in its current form also extendable to the distribution of the rest of the digits) in a naturally generated dataset to a certain distribution pattern was observed. The concept has been extrapolated to be used within the context of accounting fraud detection, whereby methodologies are derived from under the assumption of conformity to Newcomb-Benford Law. A study by Druică, Oancea, and Vâlsan (2018) applied the method to a dataset of bank account balances and presented a finding that notwithstanding Newcomb-Benford's law (hereafter referred to as Benford's law) nonconformity supposedly suggesting possible existence of tampering, it was concluded instead that there were no strong discernible motives to indicate that the data would have

been tampered with as a result. In short, Druică, Oancea, and Vâlsan (2018) concluded that nonconformity to Benford's law can exist in naturally generated, non-tampered data and that the level of nonconformity is to be expected for the dataset-specific context. Expressing a similar albeit ultimately different vein of concern for Benford's law, Debreceny and Gray (2010) conducted a study by examining conformity to Benford's law in accounting journal entries. While the study does not refute but rather assumes Benford's law to be applicable for accounting datasets and that Benford's digit analysis can be used to granularly flag transactions, the study concluded that the use of Benford's law for granular identification in cases whereby the population-wise assessment using Benford's digit analysis lead to a wholistic conclusion of nonconformity, the granular implementation can not be justified. At the same time, given the nature of the more granular implementation of the method as introduced in Nigrini, M. (2022), to either flag the entire dataset, or even if narrowed down to subsets, the red flag system leads to impractical and expensive subsequent audit procedures on a very large number of flagged observations. In essence, the efficacy of Benford's law digit analysis alongside the new AI/ML paradigm presented an interesting point to be reexamined.

Concurrently, the use of artificial neural network models has presented itself to be a promising alternative to traditional statistical and parametric methods. Coakley and Brown (2000) identified non-verbatim the particular non-linear characteristics of artificial neural networks (ANN) to be of alleviating quality. The nature of neural network architecture has allowed itself to be

implemented in many dynamic procedures that previously require complex human cognition. Consequently, artificial neural network implementation in accounting has also seen extensive use for many different purposes including but not limited to within the accounting domain itself. Artene and Domil (2025) explored the use of neural network based financial forecasting and decision making using accounting data. Artificial neural network has also been implemented to digitize and automate the previously stated cognitively challenging aspect to automate that involves visual processing within the accounting domain itself. In particular, Liu (2024) investigated the use of recurrent convolutional neural network (R-CNN) architecture as an OCR methodology for bill and/or notes recognition within the context of automating financial transaction recording. In extrapolation on the identifiable research gap and the promising nature of neural network, this research proposes and explores the use of a particular neural network architecture, namely autoencoder as an alternative to Newcomb-Benford law in identifying anomalous observations within the context of financial fraud detection.

1.2. Formulation of Problems

The outline on which the research will be conducted, are indexed using a prescribed series of research questions provided below. In retrospect, the identified research problems are as follows:

1. As a consequence of the rise in digitization, automation, and affordable higher computing power, the immediate status quo of data complexity, volume, and potentials calls for the use of more robust data tools and

process automation (Computer Economics, 2020; GVR, 2025; IDC, 2023; Le Clair et al., 2017; Petroc, 2024; Villars et al., 2011)

2. The conceived (as identified in the cited literatures) compromising nature of current accounting anomaly detection vis-à-vis fraud and error detection system relying on Newcomb-Benford digit distribution law (Debreceny and Gray, 2010; Druică, et al., 2020);
3. The lack thereof applied research on the use of artificial intelligence and machine learning in accounting, notwithstanding the oversaturation of literature reviews on the topic. The research gap is even more pronounced with observable absence in the case of Indonesian accounting *zeitgeist*.

As such, the research design will be substantially defined with a particular emphasis in addressing the research questions:

1. How does a neural-network based autoencoder anomaly detection perform in conducting anomaly detection on accounting journal entry data?
2. Does a neural network-based autoencoder anomaly detection perform better in comparison to Newcomb-Benford's Law of Digit Distribution in identifying accounting anomalies?
3. Is it possible to integrate the concept to the new AI/ML paradigm?

The proposed research questions to be addressed, was appropriated given the several current status quo that entails the environment contemporary accounting and digitized economy exist in. As the intensification of the digitization trend in the economy in general and in accounting in particular has been, and will continue to be amplified, a state of urgency is presented as a current

matter at hand. In which factors of inefficiency and ineffectiveness in accounting processes need to be subsided to ensure a healthier informational representation.

1.3. Research Objectives

The objectives defined for the research, are defined in solvency to the inquiries proposed as the research problems formulated in the previous structure. As such, the objectives that can be reasonably defined are:

1. To design, train, deploy, test, and compare the neural network-based autoencoder accounting anomaly detection as well as to deliver a publicly accessible implementation along with relevant documentations for the purpose of future replications;
2. In delivering a comparative inquiry report on the effectiveness of neural network-based autoencoder accounting anomaly detection as opposed to the established Newcomb-Benford law best practices;
3. In investigating the possibility and feasibility of inventing a new approach that leverages both the autoencoder neural-network as well as Benford's law digit analysis.

1.4. Proposed Research Benefits

An occurring theme within the *zeitgeist* of Indonesian accounting research culture, including but not limited to the academically mandatory publications made by undergraduate students, is representable to be summarized as lagging behind their international counterparts. That is, as stated priorly, disparity exists

between conceptual/theoretical inquiries on the implementation of artificial intelligence/machine learning within the accounting domain and actual implementative and empirical research describing the technical procedure to do so, especially originating from Indonesia. The imperative derived from the research gap is also further exacerbated by the actual need to adopt a new paradigm in accommodating accounting processes under the rapid rate of digitization of economical transactions. Thus, the first research benefit is proposed as an attempt to alleviate the existing research gap

Concurrently, notwithstanding the general perception of inconspicuous buzzwords attributable to the term "Fourth Industrial Revolution", the current understanding on artificial intelligence and/or machine learning present within the Indonesian accounting *zeitgeist* largely falls short of a reasonable level in anticipation of the current need for further scalability. An *a priori* characterization can be reasonably postulated to be attributable to the 'black box nature' level of understanding present in the demographics. That is, rather than examining artificial intelligence and machine learning systems as a collective of granular procedures to automate different tasks and subtasks within the accounting domain, a 'black box' contradictory sentiment often arise to dismiss artificial intelligence/machine learning as a singular system that will not be able to automate accountancy. As such, the second research benefit is defined in relation to a corrective attempt to the current 'black box' level of understanding that can potentially contribute to sentiments of dismissal nature, by providing a tangible AI/ML solution to demystify the topic.

The third research benefit is proposed under a more pragmatic lens. That is, in interest to explore a potentially better way to conduct anomaly detection in the context of accounting, in providing a more dynamic, streamlined, and accurate alternative. The aforementioned problems along the line of addressing complexity and volume as well as the topic of failure found in several instances of non-conformity to Newcomb-Benford law that is widely used as the current mainstay method of fraud detection provide a much needed factor of importance on the inquiry.

1.5. Research Scope and Limitations

The research will be conducted under several scope and limitations. The nature and properties of which, are partially influenced by an a priori judgement in how the most realistic research design can be constructed given the implying circumstances. The very first of which is in how and where the observation on the characterisation of analytical processes to be conducted within the scope of the research. That is, what constitutes as the main analytical task, the goal of the task, and the definition of done.

At any rate, the first main analytical process goes along the way of testing the efficacy of how well a neural network-based anomaly detection procedure on accounting datasets perform in comparison to Newcomb-Benford's law digit analysis (the second analytical point of interest). In this case, model building will also involve an autoencoder based anomaly detection, with a particular emphasis on designing an implementable start-to-end procedure (a pipeline), applicable to a

broader range of accounting data types. As the architecture of the autoencoder model will follow the nature and shape of the dataset, it is not within the scope of the research to provide reusable saved models with weights and architecture, but rather a reusable start-to-end pipeline to produce a model that will be able to perform the anomaly detection procedure. The third analytical process is in relevance with regard to investigating the possibility of inventing a new approach that combines both the autoencoder neural-network alongside the Newcomb-Benford's law digit analysis and subsequently measuring the efficacy of said invention.

As is given with the confidential nature of accounting information, especially with journal entry datasets, the consideration of which promptly defines the limitations of the dataset to be analyzed. In this case, given the impracticality of obtaining an entire year worth datasets of accounting journal entry as well as other related accounting datasets that are to some degree confidential in nature, the datasets that will be analyzed consist of publicly available accounting dataset. Including but not limited to accounting journal entries. The availability of ground truth on the dataset (as in actual anomalies) will determine how the performance of the procedure will be evaluated (either supervised or unsupervised). Regardless of the availability of ground truth, model training will place an emphasis on unsupervised learning in consideration to real world scenarios where ground truth accuracy can not be feasibly obtained.

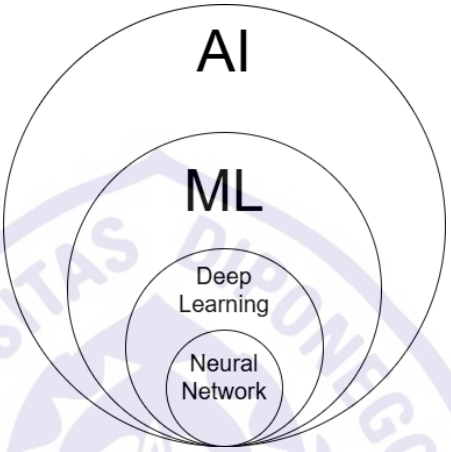
CHAPTER II
LITERATURE REVIEW

2.1. Theoretical Overview

2.1.1. Machine Learning and Artificial Intelligence

To define machine learning and artificial intelligence (hereafter will be referred conjointly as AI/ML) under a dichotomous juxtaposition is, both in general and within the specific context of this research, representing a potential factor of misalignment to one of the proposed research benefit namely in demystifying the topic for the broader accounting *zeitgeist* in Indonesia. It is within the interest of synchronization to the research benefit to conceive machine learning as a subcomponent that makes artificial intelligence possible (Figure 2.1). In a similar manner, the relationship between the recurring concepts of deep learning and neural networks reflects the same taxonomy in their nomenclature. Consequently, the choice of paradigm to define and to justify the nomenclature used in the context reflects this very nature.

Figure 2.1 Relationship between AI, ML, Deep Learning, and Neural Network



Source: Author

A more traditional paradigm that defines the nomenclature of the term artificial intelligence, in this case is a more semantic approach in defining the term. Chronologically speaking, the nomenclature was established within the context of the Dartmouth Conference in 1955 as is indicated in McCarthy, et al. (2006) by McCarthy himself. The workshop defined *non-verbatim,* a conjecture that all aspects of learning and/other components of intelligence, in theory, can be precisely described in a way that is configurable for a machine to simulate. Important scope was postulated that such a system (machine) will feature the ability of using human language, abstractions and concepts formulation, solve problems that at the time was within the sole domain of human cognition, and improve upon themselves. Thus, a caveat within the specific context of this research can consequently be postulated that pedantic compliance to this

particular paradigm is antithetical to the interest of justifying the lexical selection of the research title.

The nomenclature for the term machine learning (ML) is often attributable to the introduction of the word by Samuel, Arthur L. (1959). The generally accepted understanding of the term machine learning is defined as statistical algorithms that possess the ability to learn and subsequently perform tasks notwithstanding the absence of explicit instructions (programmed). However, the popular characterization of "without being explicitly programmed" was a *non-verbatim* paraphrasing of Arthur Samuel in Koza, et al. (1996). By following pedantic subscription to McCarthy's definition of AI, ML represents a subset of AI, a component that enables a system to possess the quality defined with the criteria but not necessarily representative of an AI system itself.

A single machine learning procedure can perform tasks that were previously described as to be within the exclusive domain of human cognition, for example in object classification, classifying whether an image contains depiction of a cat or a dog. In this case, several machine learning architectures can accomplish this task, namely convolutional neural network and its derivatives (CNN). However, whilst this example satisfy one of McCarthy's criterion for an AI system, and that machine learning in general and neural network in particular can be configured in many ways to satisfy the other criteria such as natural language processing oriented neural networks, full compliance to the criteria entails chaining together many different machine learning and non-machine

learning methods. Unless, a different and less pedantic approach is to be adopted, namely a taxonomy-based approach in harmonizing the two concepts.

The dichotomous division in statistical machine learning, and by extrapolating contextually, artificial intelligence (AI) models can be categorized as either generative models or discriminative models as is indicated in Jaakkola and Haussler (1998). Ran, et al. (2025) defined *non-verbatim* the dichotomy between discriminative and generative (AI) models whereby discriminative models focus on prediction and decision making. Concurrently, generative models synthetize new data based on the original reference.

Deriving from the dichotomous paradigm that divides and categorizes AI models, an argument can be made in justifying the word choice within the title of the research and the subsequent use of the term artificial intelligence. While machine learning represents a specific subset that makes AI possible within a broader, traditional, and more semantic definition paradigm of AI, machine learning procedures that follow the specific criteria that define the dichotomy can therefore, in and of itself be described as artificial intelligence thusly. That is, a neural network-based model following the autoencoder architecture for the purpose of anomaly detection can be described as an artificial intelligence model within this specific framework of definition.

Autoencoders by their very nature generate new sets of data as an attempt to reconstruct the original dataset, and therefore can be defined as a generative model. However, by subscribing to the school of thought that the use of autoencoder for the purpose of anomaly detection falls under the definition of

unsupervised learning (given the absence of ground truth) and therefore predictive in nature. Consequently, it can also be seen as discriminative. In another context, generative autoencoders that perform the goal of actually generating new sets of data can be seen such as in image generation with Variational Autoencoder (VAE).

Regardless, the fact that the model used within the research is attributable to the dichotomous categorization of the AI model, represents a reasonable justification for the use of the term AI to describe the model, as well as being an ML procedure in and of itself. That is, while the machine learning procedure itself satisfies the practical definition of machine learning and does not constitute as an AI implementation under traditional sense, an esoteric paradigm can be adopted to categorize it as such. It does not attempt to dilute the definition of artificial intelligence, but appropriates the term for the purpose of a more intuitive understanding for the target audience, the Indonesian accounting *zeitgeist*. Moreover, an observation can be made that many accounting literature reviews that aim to explore the use of AI in accounting often include scientific publications where some subcomponents of AI are described as an entire AI system. For example, by employing a heuristic conclusion on literatures that explore the use of neural networks in accounting as within the context of AI in accounting despite the partial nature of neural networks under McCarthy's criteria.

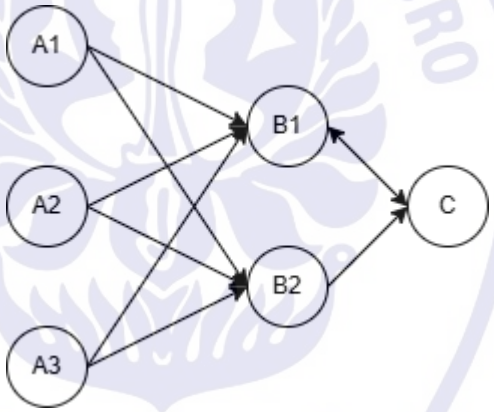
2.1.2. Artificial Neural Network

The cognitive process of the brain, more specifically the functionality in relation to the learning process is in a lot of ways describable as analogous, although through a degree of oversimplification that warrants a caveat, to how computers work. Observations concluding in linearity to the sentiment have been made predating the use of modern transistor-based computers as a testament to the conceptual similarity to the concept of computers rather than a specific architecture/paradigm of machines that apply the concept. Similar sentiments can be found on some of the earliest, accredited, works that introduce the concept of the perceptron used in artificial neural networks, such as in Rosenblatt (1958). The observation of similarity can be mainly explained in terms of the binary nature of the smallest unit of a computer that occupies the bits, either a state of on or off, the codification and interpretation of which determine the next defined steps to be performed. While the neurons of the brain do not occupy a binary state, rather continuous and analog within a range through graded potentials, they are usually described (albeit oversimplified) in terms of firing or not that usually do not occupy concurrently. The central point to be emphasized is that, in testament to the ability of neurons to learn (albeit not the only functionality), there is at least a theoretically analogous way to represent neural mechanisms that can potentially perform similar behaviors.

Artificial neural network is a computational model that is both analogous and historically has been inspired by the structure of the brain, more specifically in terms of the interactions of neurons, Puri, M., et al (2016). Likewise, the term

deep learning is usually used to describe the process of training artificial neural networks. In biological neurons, the activation, or rather the 'firing' of a neuron is typically influenced by the activation of other neurons. The architecture of artificial neural networks in this case reflects similar functionality to the layered nature of artificial neural networks, where the activation of a node (a neuron or a perceptron) in a layer influences the activation of neurons in the subsequent layer. This analogy is illustrated by the diagram below.

Figure 2.2 Simple Neural Network Architecture



Source: Author

Within the example of a simple artificial neural network architecture, the activation and contrastingly the state of which a neuron is not activated, is influenced by the neurons in the preceding layer behind it. In this case, the input layer (the first layer) which consists of the neuron a_1, a_2, and a_3, influences the activation of the hidden layer (can be several hidden layers chained together in a subsequent manner) subsequently after it. Likewise, the activation of the neurons

in the hidden layer, in this case consisting of the neuron b_1 and b_2, influence the activation of the neurons of the output layer, in this case the neuron c. The particular architecture elaborated within the literature is also known as the feedforward network, and forms the basis for the specific network architecture used in the paper later on.

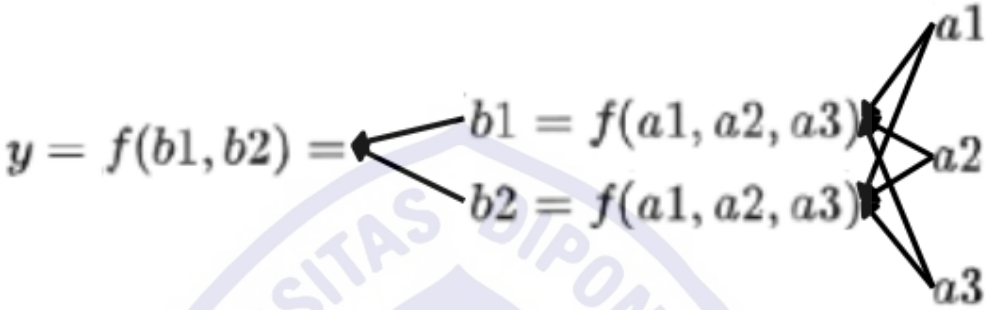
Discussed in Warner, B., and Misra, M. (1996), the connection between neurons can deviate from the unidirectional flow characteristic of the feedforward network. Various architectures exhibit distinct connection patterns. For instance, recurrent neural networks designed for processing sequential data, specifically mentioned first in Rumelhart, et al. (1986), incorporate feedback loops allowing information to persist over time. Rumelhart further elaborated on the ability of recurrent neural networks to perform computations that will take a feedforward network several times more computational power. Residual networks, first introduced in He, et al. (2016), employ skip connections, enabling the output of a layer to bypass intermediate layers and directly added to the output layer. Lastly, the main focus of the paper, autoencoders. While the architecture is typically characterized by being a subset of the feedforward network, the main idea of the network is a network that compresses and reconstructs the data through a bottleneck architecture. Notwithstanding the many special neural network architecture, the non-exhaustive nature of the compiled list is influenced by the constraint of time and scope.

In the interest of demystifying the black-box understanding of artificial intelligence in general and artificial neural network (henceforth will be referred to as ANN), it is within the bounds of reason to understand this mechanism in terms of its analogous simplification to a regression model. While the example provided below describes a linear regression, the analogy works the same with non-linear regression expressed in other notations of $y = f(x_1, x_2, \ldots, x_n)$ that map the relationship. Suppose a linear regression problem that models the relationship between the independent variables to the dependent variable, expressed with the linear regression notation of:

$$y = w_1.x_1 + w_2.x_2 + \ldots + w_n.x_n + \beta_0 + \varepsilon$$

The functionality of neurons/perceptrons in a multilayer perceptron ANN in Figure 2.2. above can be analogous to a regression problem that takes the input layer (the independent variables) of a_1, a_2, and a_3 and maps the aforementioned independent variables to the engineered variables in the hidden layer of b_1 and b_2, which are then mapped out again to the final output layer or the dependent variable of *c*. However, instead of the engineered variables being defined, the values that occupy these variables are learned through iterative trials similar to the mechanism of curve fitting in linear regression that tries to minimize the output of the loss function computed and are typically non-linear in nature. The quasi-engineered variables in a hidden layer can also be chained together to another layer of engineered variables for several layers preceding the output layer.

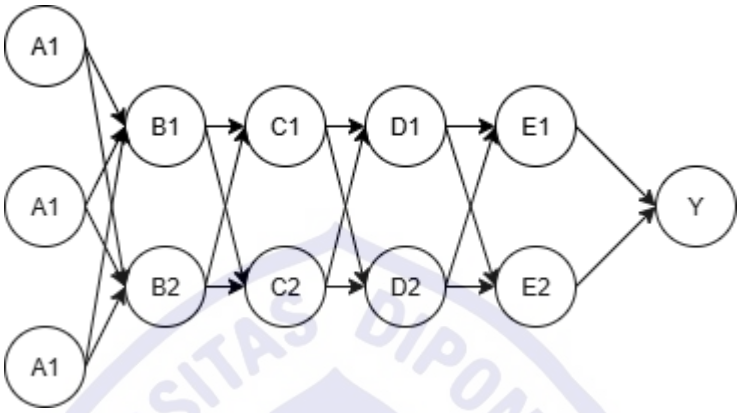
Figure 2.3. Simplified Mathematical Analogy of Artificial Neural Network to Regression Problem



Source: Author

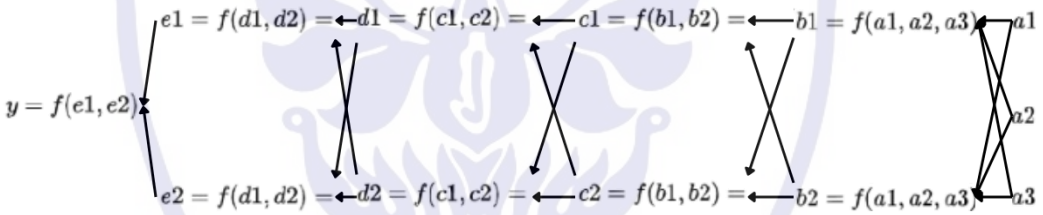
The number of perceptrons in a layer can be expanded *ad infinitum*. Similarly, the amount of hidden layers can also be further expanded with more hidden layers in between the input layer and the output layer up to infinity, such as illustrated below (Figure 2.4.) (although only up to 5 hidden layers with 2 perceptrons each, displayed). Generally speaking, the more complex the model the better they perform, albeit there exists a trade-off between complexity and the time it takes to compute and train the model. The output layer can also predict multiple values (represented with multiple perceptrons at the end) to predict several different values at once.

Figure 2.4. Artificial Neural Network with Multiple Hidden layers



Source: Author

Figure 2.5. Simplified Mathematical Analogy of Multilayer Artificial Neural Network to Regression Problems



Source: Author

In the paper that introduced the concept of perceptron (neuron) computation, Rosenblatt (1958), while it differs from the multilayer perceptron example shown above, Rosenblatt's perceptron still conceptually operates on a similar mechanism. That is, the operation in a perceptron is defined as the sum of the weights of the outputs of the perceptrons from the preceding layer it takes as inputs, and then transformed into binary representation of 0 or 1 through a Heaviside step function defining certain thresholds. In the example illustrated above (Figure 2.5.), suppose a neuron in the first hidden layer namely b_1 that

takes inputs from a_1, a_2, and a_3, the operation in the perceptron b_1 is in a theme of uniformity with a linear regression function, represented as $(\sum_{i=1}^{n}(w_i . a_i)) + \beta_0$ which is the same as $w_1 . a_1 + w_2 . a_2 + \ldots + w_n . a_n + \beta_0$. However, the product of the linear equation is then transformed using an activation function which can differ between layers. In the case of Rosenblatt's perceptron, the product of the linear operation typically represented as $z = (\sum_{i=1}^{n}(w_i . x_i)) + \beta_0$ is then transformed using a Heaviside step function of $\Phi(z)$ with a decision boundary predefined by a threshold number as a hyperparameter (Figure 2.6.). The use of activation functions transform the model to be able to capture nonlinear relationships between the inputs and the targeted variables.

Figure 2.6. Heaviside Step Function as the Activation Function of Rosenblatt's Perceptron

$$\phi(z) = \begin{cases} 0 \text{ if } z \leq \text{threshold} \\ 1 \text{ if } z > \text{threshold} \end{cases}$$

Source: Rosenblatt (1958)

The use and choice of activation function in a layer greatly influences the behavior of the neural network. In some implementations, it might be plausible to see a layer with linear activation function. In which case the product of the activation function will be equal to simply using the non-transformed sum of

weights of the inputs the perceptron takes in addition to a bias term, or $\Phi(z) = 1 . [\sum_{i=1}^{n} (w_i . a_i) + \beta_0]$, as discussed *non-verbatim* in Dubey, et al. (2021).

Reinterpreted from Ng, Andrew (2017), in circumstances whereby all hidden layers use a linear activation function, the behavior of the hidden layers will demonstrate an observable state of redundancy, as the behavior of the neural network will collapse into a direct linear mapping between the input layer (independent variables) to the output layer (dependent variable(s)). That is, it will behave the same way as a simple linear regression model and will accordingly be subjected to its inability to capture nonlinear complexity. Thus, there lies an imperative to at least include one transformative, nonlinear, activation function in at least one or more of the hidden layers. Consequently, the use of nonlinear activation functions influence the model's ability to capture nonlinear characteristics of the provided data.

To provide an exhaustive list of all activation functions that a neuron can have is beyond the scope of the research. However, a pattern of commonality is directly observable in considering the convergence on the choice of activation function(s) based on the problem that the model is trying to solve. The determinants of the selection of activation function in such a case are most apparent for the output layer, that is, the final value(s) that the inputs are trying to predict. For example, in problems related to predicting discrete, binary categorical problems that commonly employ logistic regression, the activation function for the neuron(s) of the output layer mainly reflects the need for a binary

classification. The sigmoid function is typically the most commonly used activation function that performs the binary transformation by providing an output in the form of probability between conformity and non-conformity to the positive state that is between 0 and 1. The sigmoid function is represented as $\sigma(z)$ elaborated below (Figure 2.7.). While by definition a sigmoid function maps out a binary category classification, in practice a network can be configured to perform multiclass classification where the output layer will have multiple sigmoid activated neurons, each corresponding to the probability of the observation belonging to each class. The final prediction decision is typically made with certain threshold numbers through a Heaviside step function in binary representation of positive or negative for each neuron.

Figure 2.7. Sigmoid Activation Function

$$\phi(z) = \sigma(z) = \frac{1}{1+e^{-z}} \{ z = (\sum_{i=1}^{n} w_i \cdot x_i) + \beta_0$$

or

$$\phi(z) = \sigma(z) = \frac{1}{1+e^{-z}} \{ z = W^T \cdot X + \beta_0$$

Where

$$W \in \mathbb{R}^{(n \cdot 1)} = \begin{bmatrix} w_1 \\ w_2 \\ \cdots \\ w_n \end{bmatrix}$$

$$X \in \mathbb{R}^{(n \cdot 1)} = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix}$$

Source: Ng, Andrew (2017)

Another common activation function is the hyperbolic tangent function, also known by its abbreviated name the tanh function. The tanh activation function maps out values to be between -1 and 1, and is represented in Figure 2.8. In prediction problems where the value of the dependent variable can't have negative values, such as for height, width, weight, and more, another prudent

configuration is to use the ReLU or rectified linear unit activation function. In essence, the ReLU formula simply returns the value if the value is more than zero, else it will return zero as defined with the formula $\phi(z) = ReLU(z) = max(0, z)$. The comparison between the initial linear prediction, and transformed value using the activation functions namely sigmoid, hyperbolic tangent (tanh), and rectified linear unit (ReLU) function is illustrated in Figure 2.9.

Figure 2.8. Tanh (Hyperbolic Tangent) Activation Function

$$\phi(z) = tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \{ z = (\sum_{i=1}^{n} w_i \cdot x_i) + \beta_0$$

$$or$$

$$\phi(z) = tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \{ z = W^T \cdot X + \beta_0$$

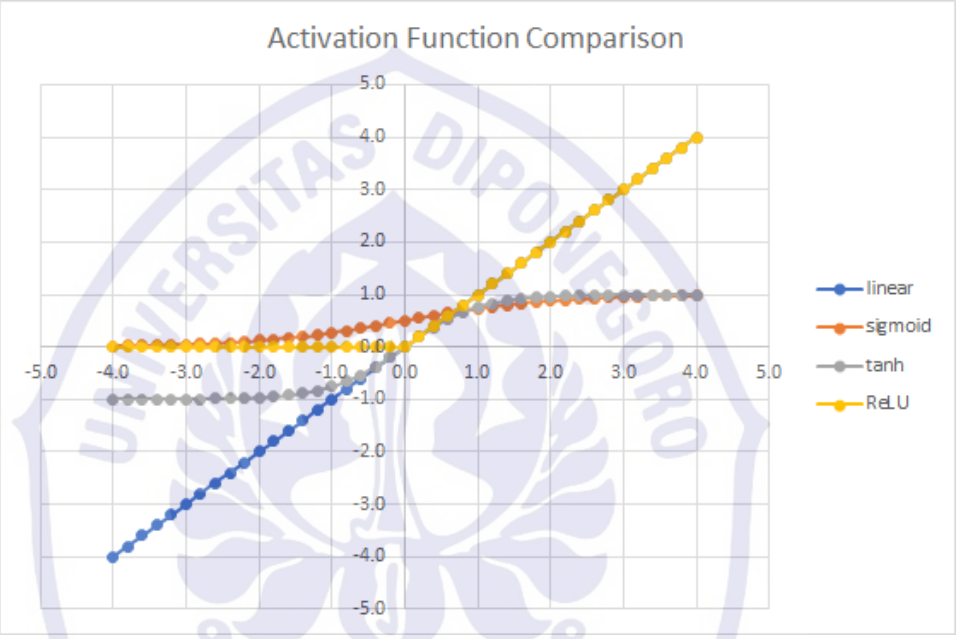
$$Where$$

$$W \in \mathbb{R}^{(n \cdot 1)} = \begin{bmatrix} w_1 \\ w_2 \\ \cdots \\ w_n \end{bmatrix}$$

$$X \in \mathbb{R}^{(n \cdot 1)} = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix}$$

Source: Ng, Andrew

Figure 2.9. Activation Function Comparison



Source: Author

Apart from the considerations of the problem that a model is trying to solve, the characteristics of activation functions also influence the choice of activation function, as indicated *non-verbatim* in Dubey, et al. (2021). Within the same literature, certain behaviors of an activation function and/or the lack thereof prompted the choice of activation functions. In this case, one of the highlighted consequences of said behavior was identified by Dubey to be the vanishing gradient problem. To understand the vanishing gradient problem, it is also important to establish an understanding of the subsequent processes in the iterations of neural network computation, namely the loss function and gradient

descent, whereby, within the interest of elaborating the two topics the use of the bifurcated framework of forward propagation and backward propagation (backpropagation) is a productive endeavour.

Forward propagation, though alluded to earlier, warrants a formal definition. Reinterpreted *non-verbatim* in Ng (2020), it is simply the entire process whereby the input variables move through the layers of a neural network sequentially within its architecture and produce a prediction after computing the output layer. The computation of which has been previously described; however, its formalization with regard to vectorized implementation, as summarized in Ng (2020) is expressed as:

For a neural network with L layers: the output for layer $Z^{[l]}$ given layer l:

$$Z^{[l]} = W^{[l]} \cdot A^{[l-1]} + b^{[l]}$$

where

$$A^{[l]} = g^{[l]}(Z^{[l]})$$

where $g^{[l]}$ is the activation function for the layer l.

After the forward propagation process predicts the value for $\hat{y}$ by completing the computation chain to the last (output) layer, a loss function is calculated by comparing the predicted value to the actual value. The concept of loss function in the context of deep learning and artificial neural network is largely describable by the definition of loss function used in mathematical optimizations, though the individual loss functions themselves might differ. As indicated by Hastie, et al. (2009), within statistical decision theory, for a function

$f(x)$ that accepts the input of x to predict y, the use of a loss function is necessary to penalize errors in prediction. In general, the aim of the model is to minimize the loss function. Similarly, as indicated by Warner, et al. (1996), this minimization is achieved by optimizing the weights within the computational chain of forward propagation. Warner, et al. (1996) also highlighted the sum of squared errors (SSE) as one of the most commonly used loss functions, defined in Figure 2.10.

Figure 2.10. Sum of Squared Errors (SSE) Loss Function

$$L(y, \hat{y}) = \frac{1}{2} \sum_{p=1}^{n} \sum_{k=1}^{O} (y_{pk} - \hat{y_{pk}})^2$$

Source: Warner, et al. (1996)

Here, p represents the subscript for the p-th observation (or row in data frame representation) up to n total observations(rows). Likewise, k represents the subscript for the k-th output of the model in cases where the model predicts more than one value up to O units. This particular loss function is mainly used in regression problems that typically do not aim to predict categorical variable(s).

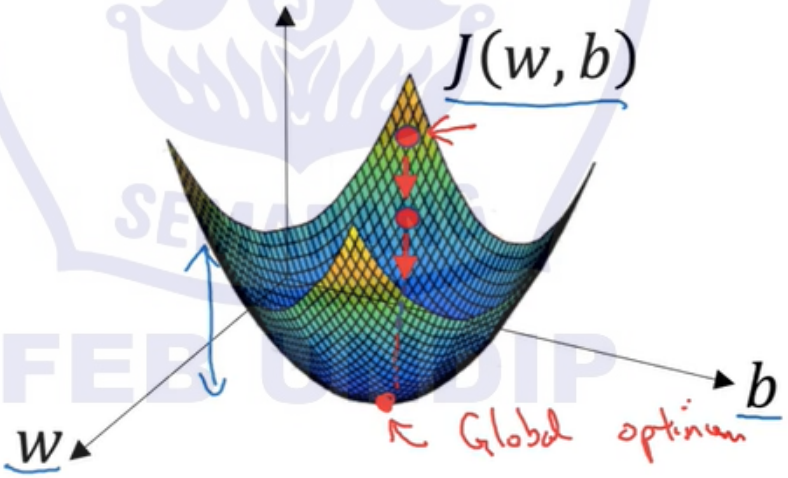
As previously stated, the aim of a model is to minimize the loss function and this is achieved by optimizing weights (and bias, though can be expressed in terms of weights as w_0) for the forward propagation computation. Consequently, a line of inquiry is in order in elaboration of what it means to optimize weights. In this case, for the weights of the features that each neuron

takes, iterative learning processes are done by trying to find the best possible weights that produce the most robust prediction, i.e. in minimizing the loss function. Adopting an analogous perspective from a linear regression system, this concept is similar to the optimization iterations in curve-fitting. In deep learning, however, the most commonly used algorithm to do this is known as the backpropagation algorithm. In Warner, et al. (1996), backpropagation is also specifically prescribed for feedforward neural networks, the concept of which forms a critical foundation for the model this research is trying to build.

Broadly speaking, the introduction of backpropagation as a concept for training neural networks is predominantly cited to be attributable to the work of Rumelhart, et al. (1986). To understand backpropagation, it is imperative to note its relation in regard to gradient descent. Gradient in the context of gradient descent refers to the application of the mathematical concept of gradient(slope) between the values of the weight to the values of the loss function. In Goodfellow, et al. (2016), the definition of backward propagation is narrowed to this process for computing the gradient. In Ng (2020), for the cost function (the sum of the outputs of the loss function) $J(w,\ b)\ =\ \frac{1}{n}\sum_{i=1}^{n} L(\widehat{y}\ ,\ y)$, gradient descent aims to find the value of w and b that minimize $J(w,\ b)$. Ng (2020) also indicated that this process utilize the chain rule in calculus to derive the gradients of the loss function with regard to the weights and biases (also describable as weight at w_0). To find the optimal value for w and b, gradient descent does not perform exhaustive grid search, instead the process guides the search by indicating the

direction w and b needs to go to find the next optimal value with regard to $J(w, b)$. Figure 2.11. illustrates this functionality (although dimensionally reduced for the sake of visualization), whereby rather than performing exhaustive grid search, gradient descent starts with random initialization, each iteration than indicate the direction the search needs to move until it reaches the global optimum (the minimum of the cost function $J(w, b)$), also known as convergence. The learning rate hyperparameter (symbolized as γ) governs how far away the 'step' each iteration takes, whereby trade-off exist in smaller α to be able to converge more optimally at the cost of higher computational time, and vice versa.

Figure 2.11. Gradient Descent Iteration with Dimensionally Reduced Visualization



Source: Ng, Andrew (2020).

The calculation of the gradient in backpropagation is closely linked to the cost function (the sum of the loss function). Within the implementation introduced by Rumelhart, et al. (1986) and expanded upon in Warner, et al. (1996), for the

general (not any specific) representation of the cost function *J(w, b)* or now simplified as E (sum of the loss function)

$$E = \frac{1}{n}\sum_{p=1}^{n} L(y_p, \widehat{y_p})$$

for the p-th observation or row up to n total rows. In Ng (2020), the gradient descent algorithm, starting with a random initialization, updates each feature as

$$w_1 := w_1 - \lambda \cdot \frac{\delta E}{\delta W_1}$$

$$w_2 := w_2 - \lambda \cdot \frac{\delta E}{\delta W_2}$$

$$\dots$$

$$w_n := w_n - \lambda \cdot \frac{\delta E}{\delta W_n}$$

$$b := b - \gamma . \frac{\delta E}{\delta b}$$

for γ learning rate, repeated until convergence (or practically speaking after a specified amount of iterations have been reached). For the vector Θ containing all the parameters of the model (w_1, w_2, ..., w_n, b), backpropagation handles the calculation of the gradient, which is a vector of the partial derivatives of the cost function with respect to each parameter, or in the vector $\nabla E(\Theta)$ represented in Figure 2.12.

Figure 2.12. The Gradient Vector

$$\nabla E(\theta) = \begin{pmatrix} \frac{\partial E}{\partial \theta_1} \\ \frac{\partial E}{\partial \theta_2} \\ \vdots \\ \frac{\partial E}{\partial \theta_n} \end{pmatrix}$$

Source: Ng, Andrew (2020)

The calculation of the partial derivatives follows the chain rule to derive the composite functions for each parameter, though defining the chain rule representation entails details that are stringent to the architecture, the loss function, and the activation function in a way that will be too extensive to cover. As the cost function and the activation functions of the preceding layers differ, the final expression in the chain differs, but the starting point of the chain rule derivation is the same. In retrospect, the problem of vanishing gradient is intricately linked to the gradient descent algorithm. In Dubey et al., (2023), vanishing gradient occurs where the gradient becomes very close to zero, especially with activation functions such as the sigmoid and the hyperbolic tangent, and the gradient descent algorithm does not learn much in such iterations and leads to poor convergence.

Given the complexity of deep learning and neural networks, this conceptual overview barely scratches the surface. The main goal of this preliminary introduction is to provide the necessary nomenclature understanding at an intuitive level on what it does and how it works. There has been an explosive rise in publications for neural network implementations, with differing architectures, activation functions, cost functions, training algorithms, and more. The specific and tailored nature of those publications are justifiable by their demonstrated performance in solving the specific problems the model is trying to solve.

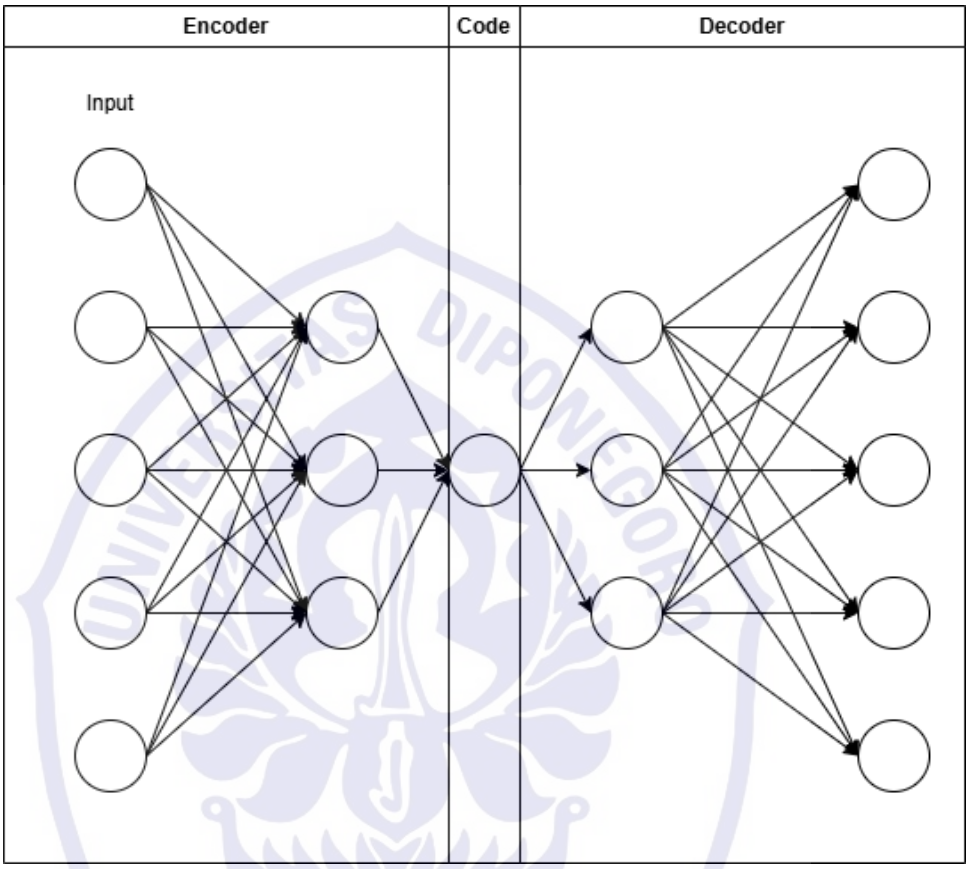
2.1.3. Autoencoders

Autoencoder is a very specific neural network architecture that has been demonstrated to be useful in several different applications, ranging from dimensionality reduction, anomaly detection, generative algorithms, and more. In Goodfellow, et al. (2016), an autoencoder is described as a specific neural network architecture that is trained with the purpose of copying its inputs to its outputs. That is, the output of the network is designed to be as close of a representation of the original input as possible. Goodfellow, et al. (2016) also indicated that the autoencoder architecture attempts to learn a dimensionally reduced representation of the input, and attempts to reconstruct the dimensionally reduced representation back to its original values (inputs) in a manner of preservative reconstruction.

The deficiency of formal institution of the nomenclature for autoencoder notwithstanding, Rumelhart, et al.(1986) introduced the concept described verbatim as the “encoding problem”. The conceptual inception of which calls for a

mapping of *N* input features to *N* output features, through a smaller set of hidden units. In departure from the topic of the chronological history of autoencoders, the main architecture that forms the backbone of autoencoders remains largely the same. Characterized *norverbatim* in Li, et al. (2023), the architecture of an autoencoder is characterized by its bottleneck configuration, that can mainly be bifurcated through the division of the encoder part and the decoder part, as seen in Figure 2.13. The same paper elaborated that the encoder part of the architecture encodes and in a way, compresses the data into smaller representations. This encoded part is also known as the code, the bottleneck layer, the latent space, and by other nomenclatures. Generally speaking, for a matrix of the input features of $X \in \mathbb{R}^{n \times m}$ for n rows and m columns, the bottleneck layer (code) is a smaller representation, suppose a matrix of $Z \in \mathbb{R}^{o \times p}$ for $o \leq n$ and $p \leq$ m. The same paper defines *nonverbatim* the decoder part of autoencoder as the part that tries to reconstruct the encoded representation back to its original shape as the input matrix, that is, back to a matrix with n rows and m columns.

Figure 2.13. Autoencoder Neural Network Architecture



Source: Author

The dimensionality of the bottleneck layer can be reasonably conjectured to have been established by a general consensus to be of smaller dimension in relation to the original representation. However, the degree of the dimensionality reduction in relation to the original dimension warrants discourses in and of itself. A paper published by Laakom, et al. (2024) highlighted the trade off between the dimensionality of the bottleneck layer to the distortion rate in decoding the representation given the complexity of the operations. On the other hand, higher dimensionality in the bottleneck layer warrants caveats on their own. The same paper highlights how despite the autoencoder architecture tries to approximate an

identity function of $f(x) \approx x$, the latent representation of the bottleneck layer is still a key task to perform especially in cases specifically requiring effective and reconstructible dimensionality reduction, rather than simply performing identity operations. In cases of explicit dimensionality reduction requirements, the line of inquiry is indeed in how autoencoders can learn the smallest possible bottleneck representation whereby the decoder can still reliably reconstruct the data. Nevertheless, even in different use cases such as in anomaly detection or in generative algorithms, a similar pattern of trade-off balancing remains true though to a lesser degree. Goodfellow, et al. (2016), implied *nonverbatim* how if the dimensionality of the bottleneck layer is too large, the behavior of the architecture will instead perform a task more similar to a copying function rather than an encoding-decoding chain.

In Li, et al. (2023), the loss function and consequently the cost function of conventional autoencoders is generally divided into two, the first being the Sum of Squared Errors (SSE) loss and the matrix implementation is

$$L(X, X^d) = \sum_{i=1}^{n} \sum_{j=1}^{m} (X_{ij}^d - X_{ij})^2 \; or \; ||X^d - X||^2$$

for $X \in \mathbb{R}^{n \times m}$ and $X^d \in \mathbb{R}^{n \times m}$, for a scalar loss, and subsequently the cost function is the average of the sum of the loss function

$$E = \frac{1}{T} \sum_{t=1}^{T} L(X, X_t^d)$$

$$E = \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{i=1}^{n} \sum_{j=1}^{m} (X_{ij}^d - X_{ij})^2 \right)$$

for T training iterations in an epoch (with batch implementation), whereby a training iteration completes one forward propagation and one backpropagation, in linear problems. Whereas in binary logistic problem, the binary cross-entropy loss function of $L(X,\ X^d)$ is defined as

$$L(X,\ X^d)\ =\ \sum_{i=1}^{n}\sum_{j=1}^{m} -\ (X_{ij}\ log\ X_{ij}^d\ -\ (1\ -\ X_{ij})\ log(1\ -\ X_{ij}^d))$$

and consequently, the cost function is the average of the loss function for up to T training iterations in a single epoch.

$$E\ =\ \frac{1}{T}\sum_{t=1}^{T} L(X,\ X_t^d)$$

$$E\ =\ \frac{1}{T}\sum_{t=1}^{T}[\ \sum_{i=1}^{n}\sum_{j=1}^{m} -\ (X_{ij}\ log\ X_{ij}^d\ -\ (1\ -\ X_{ij})\ log(1\ -\ X_{ij}^d))]$$

Nevertheless, different use cases especially in other complex implementations can employ other esoteric functions. The same thing can be said about the activation functions.

2.1.4. Anomaly Detection

Within the context of data analysis, the term anomaly in the concept of anomaly detection is often more commonly known by its statistical cognate as outlier. Reinterpreted nonverbatim from Lind, et al. (2019), the term outlier (hereby interchangeable and will be referred to as anomaly) refers to observations in a dataset that are atypically distant from other, otherwise normal observations. It is within the interest of this paper to warrant a caveat that such a definition is by its very nature characterized by a distance-based paradigm. Hence its subsequent practical description in the same literature to be accepted as values that are, in

relation to the first and third quartile of the data, to be observations that are 1.5 times the interquartile range less than the first quartile in the case of the prior, and greater than 1.5 times the interquartile range above the third quartile for the latter. This particular criteria in defining outliers holds true in data sets that subscribe to simpler assumptions, such as the case in univariate data. The main assumption of the method is that normal behavior in a data set is to follow a Gaussian distribution, as disclosed nonverbatim in Chandola, et al. (2009), characterizing it as a specific anomaly detection method rather than a broader, more generalized definition of anomaly. In providing a more generalized definition of anomaly, the aforementioned literature avoided this notion and instead defined anomalies as observations in a given dataset that are characterized by a pattern of nonconformity to the expected normal behavior of the dataset. Thus, in departure to the established definition mentioned by Lind, et al. (2019), the definition in Chandola, et al. (2009) provides a broader, more encompassing definition of anomalies.

Although omitted earlier, formal classification on the types of anomaly is warranted for including but not limited to the described distance based assumption in identifying anomalies. In Chandola, et al. (2009), the three main categories are point anomalies, contextual anomalies, and collective anomalies. The first and the simplest type of anomaly is the point anomaly. The concept refers to the previously disclosed distance-based assumption (although not limited to IQR based methodology). Contextual anomalies pertain to a type of anomaly that are characterized by their atypical behavior within the specific context they are

observed from. Provided in Chandola, et al. (2009) suppose an observation of daily temperatures in temperate region, a temperature of 15° celsius is considered normal if observed from an annual mean temperature as a benchmark, however, if the observation was found during the winter season the observation is characteristically anomalous with respect to the context, in this case by time and/or spatial context. In contrast, by omitting the time and/or spatial context of the data, a simple distance based assumption will fail to identify the anomaly. Lastly, the collective anomaly, refers to a type of anomaly that is individually not anomalous, however if observed over several instances the collective characterization of the group demonstrates anomalous characteristics.

The methodology of anomaly detection can be largely divided into supervised, semi-supervised, or unsupervised methods, as discussed nonverbatim in Goldstein and Uchida (2016). For supervised anomaly detection, in similar fashion to the concept of supervised learning in machine learning, the anomaly detection method utilizes data sets with labeled ground truths for anomalous observations, thus the algorithm mainly performs classification tasks. In a manner antithetical to the previous definition, the lack thereof ground truth labels consequently define unsupervised anomaly detection, as discussed in Farber and Al Rashdan (2025). Finally, as indicated in Goldstein and Uchida (2016), semi-supervised anomaly detection is similar to unsupervised anomaly detection in terms of the lack thereof ground truth labels especially for the training process, however, a degree of similarity to supervised anomaly detection is instead demonstrated in the bifurcation of the dataset into training and test sets. The main

implementation of autoencoder anomaly detection in this sense can mainly be classified as unsupervised, however, in a few datasets where ground truth labels are known, the training process does not utilize this ground truth and instead the ground truth is used to evaluate the performance with the test sets. Thus, in some perspectives, it can be characterized as semi-supervised.

Whilst distance based anomaly detection method that assumes some type of distribution as a benchmark of normality, such as by using the interquartile range or the Z-score, are typically categorized as point anomalies, another similar yet contrastingly occupying different category of anomaly detection is in terms of conformity to Benford's law/distribution. Central to the discourse of this research, Newcomb-Benford's law, or more commonly referred to as Benford's law, refers to an assumption on the distribution of digits although this particular research focuses more on the distribution of first digits. Benford (1938), reiterated this signal of normality whereby conformity of the leading digits of the data to a distribution of:

$$y = F_a = log(1 + \frac{1}{a})$$

for $y = F_a$ denoting the expected frequency of the leading digit a. Nonconformity of leading digits to this distribution can therefore categorize the collective instances as anomalous. By this definition, Newcomb-Benford's law of anomalous numbers can be characterized as a collective anomaly detection method.

An important caveat can be raised in regard to the multidimensionality of the data. Zimek, et al. (2012), established nonverbatim the challenges of anomaly

detection in multivariate data, the juxtaposition of which also implies that the simpler univariate based assumptions are easier to implement yet can be a subject of scrutiny. At the same time, multidimensionality can provide valuable contextualization to anomalies, such as by accounting for time, classes, possible unsupervised clusters, and more. Extensive plausibility on the discourse of multidimensionality in anomaly detection notwithstanding, the scope of the research delineated this excerpt exclusively to this extent.

The very premise of deviation from the expected normal behavior to be identified in the process of anomaly detection holds many potential utilities in a pervasive range of applications. Such as in medicine, computational network activity intrusion, and finance as indicated in Ahmed, et al. (2016) and in Chandola, et al. (2009), alongside industrial damages such as faults or defects, image processing, textual data, sensor network, as well as many more inexhaustible use cases. In relation to the topic of finance in general and accounting in particular, Schreyer, et al. (2017) explored the use of anomaly detection in accounting to identify abnormal accounting behaviors including but not limited to potential errors and/or fraudulent activities within the context of auditing.

2.2. Related Works

2.2.1. Machine Learning in Accounting

In the relevant context of discussion, publications on the integration between accounting and machine learning refers to both the cross-discipline integration between machine learning and by extension, artificial intelligence, in broader accounting usage. Within the Indonesian scientific publication scene, through a degree of heuristic evidentiality, a reasonable conclusion can be stated that there has been an inexhaustible growth in literature reviews compiling the topic, highlighting the contrast in comparison to actual technical papers on the topic. Furthermore, said publications on the topic originating from Indonesia are typically characterized by summarizing conceptual works rather than technical implementations, such as in Ghafar, I., et al. (2024); Purba, K.A., and Dewayanto, T. (2023); Silaen, R.P., and Dewayanto, T. (2024); Fadilla, A., et al. (2025); as well as Septiyanti, R.D., et al., (2025). The existence of compilations on technical implementation summary such as in Tarissa, B.V., and Dewayanto, T. (2024); as well as Hanin, G.F., and Dewayanto, T. (2024) notwithstanding, it is a reasonably justifiable argument to conclude that such works are atypical to the norm. Nevertheless, observations concluding the similarly growing number of publications disregarding this particular geospatial constraint is warranted for. It is however, important to elevate a higher degree of caution that extensive and deep exploration on the topic is not within the scope of this research.

The cross-disciplinary integration between accounting and machine learning practically spans across all accounting sub-disciplines, including but not

limited to accounting information systems. Although not typically associated with machine learning, transaction document processing has demonstrated a ubiquitous integration between machine learning and financial accounting, In particular but not limited to, the use of optical character recognition models (OCR) for such purpose. Arguably, the mundane nature of the usage has not warranted deeper accounting research, whereby publications on the topic are largely found on machine learning papers pertaining to the OCR model themselves as opposed to the accounting discipline itself. Such as in Auad, M., et al. (2024); Chen, Y., et al. (2024); and Liu, Z. (2024). In a similar vein, the results of the OCR procedures are often subjected to subsequent automation procedures, either rule-based data extraction or in relation to the dynamism present in real world datasets, machine learning and/or artificial intelligence-based solutions, particularly with natural language processing (NLP). For example, in extracting and contextualizing the dynamism present in financial statement datasets in XBRL extension, NLP based solutions have been implemented such as in Faccia, A., and Petratos, P., (2022); as well as Wang, R. Z. (2025). As demonstrated, the main advantage of machine learning and/or artificial intelligence based methodology as opposed to the established norm lies in its ability to process unstructured data, more importantly in image and textual data that often require a more tedious manual processing. In financial statement analysis, a study by Balata, P., and Breton, G. (2005) found a tendency (using textual analysis), that entities experiencing subpar financial performance demonstrated a propensity for significantly higher optimism as a disguising attempt in its narrative construction.

Within the tax management domain, machine learning and/or artificial intelligence have also made their way beyond data extraction, into critical assessments for both enterprises and governing entities alike. A study by Guo, S. (2022) elaborated upon an autoencoder-based neural network in combination to a classification procedure based on the output of said autoencoder in identifying enterprise tax risk classification. Another study by Lahann, J., et al. (2019) explored the use of several multi-class classification algorithms in identifying VAT non-compliance.

Building upon the ‘data is the new oil’ maxim, the lack thereof initiatives in embracing the methodology is not entirely warranted for. In general, concerns have been expressed particularly in relation to the topic of transparency, explainability, confidentiality, data security, alongside potential biases in the model. Such concerns can be found in Bhimani, A., and Willcocks, L., (2014); Munoko, I., et al. (2020); and Kokina, et al. (2025) . Especially given the ‘data hungry’ nature of machine learning and artificial intelligence to perform at their optimal utilization implies to some part, higher level of disclosures the level of which entities might be reluctant to disclose. At the same time, efforts have been made as an attempt to address these concerns. On the topic of data security and confidentiality, a federated-learning based model was introduced in Schreyer, M., Sattarov, T., and Borth, D. (2024). The existence of other ameliorating methodologies notwithstanding, the main narrative to be constructed given the constraint of the research scope is in commending the initiatives with regard to the

adoption of the technology in conjunction to addressing the valid concerns over the methodologies to be the more balanced position on the matter.

2.2.2. Anomaly Detection in Accounting

From a semantic point of view, the auditing process in accounting demonstrates a linearity in thematic goal to anomaly detection. That is, in identifying abnormal behaviours, in this case within the accounting context. In contrast to machine learning-based anomaly detection methodology, this excerpt focuses on non-ML implementations for identifying anomalies in accounting. In this case, defining anomalous behavior in accounting can extend anywhere on the broader extent of accounting data and processes. In terms of irregularities in data within the accounting process, is exemplified such as anomalous behaviours on journal entry data as defined in Schreyer, M., et al. (2017). Accounting anomaly can also refer to abnormality in accounting information as presented to external parties on the financial statement, as used in equity and financial research such as in Serrano-Cinca, C., et al. (2018); and Richardson, S.A., et al. (2010). Within the scope of this research however, the constraint adopted within the prior demonstrates higher relevance with regard to this paper.

Some of the works that are of great relevance to this discourse is the work by Debreceny, R.S., and Gray, G.L., (2010) that was built upon the foundational works in Nigrini, M. (1996), Nigrini, M. (2000), Nigrini, M. (2011), as well as Nigrini, M. (2022). These works delved on the use of conformity tests of the digits in accounting datasets to a distribution introduced almost a century earlier in Benford (1938) to indicate the existence of potential anomalies in accounting

data. In a similar vein, a prior work by Dlugosz, S., and Müller-Funk, U. (2009) elaborated upon the use of last digit analysis whereby a premise of normality can be identified in cases whereby the distribution of the last digits in a dataset is approximately uniform. In both cases, the use of digit analysis for the purpose of accounting anomaly detection presented a problem on the lack thereof granularity. That is, such methodologies identify whether the entire dataset, or subsets of the dataset, are contaminated with anomalies rather than granular identification of whether or not an observation is anomalous.

From a semantic point of view, the entire auditing process is by definition, an act of anomaly detection. Both statistical and nonstatistical audit sampling prior to the subsequent audit procedures presented a way to gauge the existence of anomalous behaviors in accounting data. However, such approaches can be reasonably characterized through an *a priori* reasoning to lead to higher expenditure on audit resources, as the subsequent audit procedures are not typically provided with a more directive search. This concern is also exacerbated with the higher degree of manual work associated with the process, as expressed in Adamov, A.Z., (2019). Moreover, the techniques employed on the so-called red flag tests in auditing are typically based on known anomalous (typically fraud) scenarios, and often does not generalize well to the more novel datasets, as expressed in Schreyer, M. (2017). On the other hand, digit analysis is typically associated with a lack of granularity in providing a singled-out identification of anomalous behaviors from a broader dataset.

2.2.3. Machine Learning Anomaly Detection

Anomaly (outlier) detection can be derived from the more traditional statistical and/or mathematical methodology, and increasingly as well, using machine learning procedures. In both cases, the anomaly detection procedure can be both bifurcated either as supervised (with existing ground truth) and unsupervised (given a lack thereof ground truth) method. Comparing statistical to machine learning anomaly detection methods, the difference mainly lies in the ability of the latter to better process unstructured data, such as textual, pictures, audio, and more. While an argument can be made that any unstructured data format can be preprocessed to be viably processed subsequently using statistical methods, the efficacy of which are significantly compromised in comparison to machine learning implementations.

2.2.4. Machine Learning Anomaly Detection in Accounting

Table 2.1. Related Works for Machine Learning Anomaly Detection Implementation in Accounting

No.	Author(s)	Title	Tools	Objective(s)	Result
1.	Schreyer, M. Sattarov, T., Borth, D., Dengerl, A., Reimer, B.	Detection of Anomalies in Large Scale Accounting Data using Deep Autoencoder Network (2017)	Autoencoder; Comparative testing to PCA, HDBScan, LOF, and OC-SVM	Accounting journal entry data anomalous transaction detection	0.56 precision score and 0.76 F1-score for autoencoder; Highest amongst all tested methods across all

					metrics
2.	Kurien, K.L., Chikkamannur, A.A.	Benford's Law and Deep Learning Autoencoder: An Approach for Credit Card Transactions in Social Media (2019)	Autoencoder and Benford Law (Separately, not combined); Comparative testing to Random Forest and Logistic Regression	Credit card transaction data anomaly (fraud) detection	0.92 precision score, 0.86 recall score, and 0.86 F1-Score; Highest amongst all tested methods except in precision (0.93 for random forest)
3.	Guo, S.	Intelligent Assessment Method of Enterprise Tax Risk Based on Deep Learning (2022)	Stacked Autoencoder with a classification layer; Compared to the results of two other papers	Classifying tax risk of entities	Lowest ROC deviation ratio and Root Mean Square Error between tested methods;
4.	Lahann, J., Scheid, M., Fettke, P.	Utilizing Machine Learning Techniques to Reveal VAT Compliance in Accounting Data (2019)	Testing several ML models, namely SVM, Random Forest, Naive Bayes, K-Nearest Neighbor, C3.5 Decision Tree, and Artificial	Classifying VAT Compliance with journal entries data from Germany, France, and Spain	C3.5 Decision Tree demonstrated the highest performance across all metrics (Accuracy, Precision, Recall, F1-Score, MCC, and AUC)

			Neural Network		
5.	Xie, Z., Huang, X.	A Credit Card Fraud Detection Method Based on Mahalanobis Distance Hybrid Sampling and Random Forest Algorithm (2024)	Mahanalobis Distance with SMOTE-ENN Hybrid Sampling and Random Forest; Comparative testing to other ML models and other sampling methodologies	Classifying credit card fraud on Taiwanese credit card default data	Highest performance across all metrics, achieved perfect score of 1 across recall, precision, F1-score, AUC, and MCC

In exploring the current status quo for publications on the topic of machine learning anomaly detection within accounting, the esoteric nature of the combination of the topics and sub-topics has lent itself to compromising nature to the number of publications on the matter. The discernable lack of relevant materials has been demonstrated in conducting the literature review notwithstanding, several works are at the very least, although stringent constraint to accounting needs to exercise higher level of flexibility, are foundational to this research in particular. In general, such studies follow the pattern of model building, configuration optimization, and comparison to baseline methods. Chief amongst them is the work by Schreyer, et al. (2017) in the use of an autoencoder neural network for anomaly detection with accounting journal entry data, using a

heaviside threshold function to define anomalies from the reconstructed error values. The work explored other baseline methods, such as reconstruction error based Principal Component Analysis (PCA), Density Based Clustering (DBSCAN), One-Class Support Vector Machine (OC-SVM), Local Outlier Factor (LOF), to the autoencoder model itself. Within the literature, the autoencoder model demonstrated the best performance out of all tested methods, with the highest precision, F1-Score, ROC-AUC, and recall.

The work by Schreyer, et al. (2017) is foundational to this research by virtue of research framework and paradigm, in building, configuring, and testing the performance of autoencoder-based accounting journal entry anomaly detection. However, we proposed a novel combined solution of autoencoder and Benford's law digit analysis to increase the performance in comparison to this baseline method. Although other publication has mentioned both autoencoder and Benford's digit analysis within the same work, such as in Kurien and Chikkamannur (2019), the use of Benford's digit analysis largely serves as preparatory exploratory data analysis in identifying whether or not the entire dataset is contaminated with anomalies, rather than a method that combines the two in singling out anomalous observations. Moreover, the work also demonstrated nonconformity to the topic of application in accounting, whereby the research was conducted on credit card transaction data. Another work by Xie, Z., and Huang, X. (2024) explored the use of machine learning anomaly detection within the similar context of credit card fraud detection, where a hybrid Mahalanobis distance based SMOTE-ENN sampling was used in handling class

imbalance for the main random forest classification model to predict credit card frauds, and achieved the highest performance in comparison to the other tested methods.

In the context of tax accounting, the use of anomaly detection, namely stacked autoencoders to classify entity tax risk classification was demonstrated in Guo, S. (2022). The method used a combination of the autoencoder encoding-decoding process to learn a representation that is then parsed to a classification layer to determine the tax risk category. The research compared to other baseline methods and found the autoencoder-based approach to be the best, by measuring ROC deviation ratio and Root Mean Square Error value. Within a similar context, Lahann, J., et al. (2019) tested several different machine learning algorithms to predict VAT compliance based on accounting journal entry data. The research found the C3.5 Decision Tree algorithm demonstrating the highest performance across all metrics, namely accuracy, precision, recall, F1-score, MCC, and AUC. In general, whilst the specificity with regard to the particular model used, the performance metric, the variables involved, and/or other signifying factors between these researches differ, the main silver lining for these collection of papers is the general research design and paradigm for applied machine learning research within the context of financial transactions in general, and accounting in particular, that serves as the main framework to be followed within this research.

2.3. Research Framework

This study aims to introduce a novel approach in leveraging AI/ML implementation within the context of accounting, in which case the specific implementation is in introducing a factor of novelty in conducting audit red flag tests. The thought process driving this synthesis is as illustrated in Figure 2.14 and Figure 2.15. below:

Figure 2.14. Research Framework Diagram

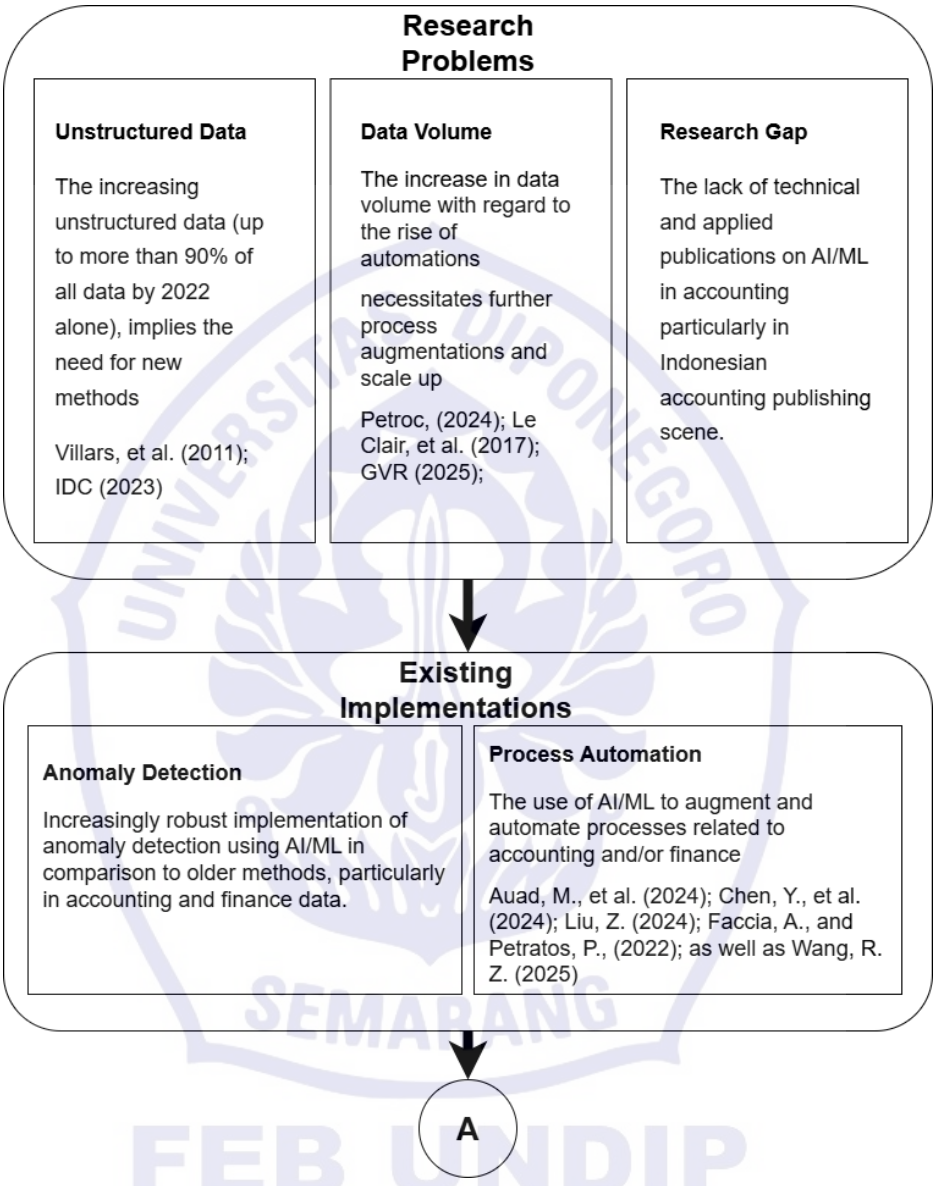
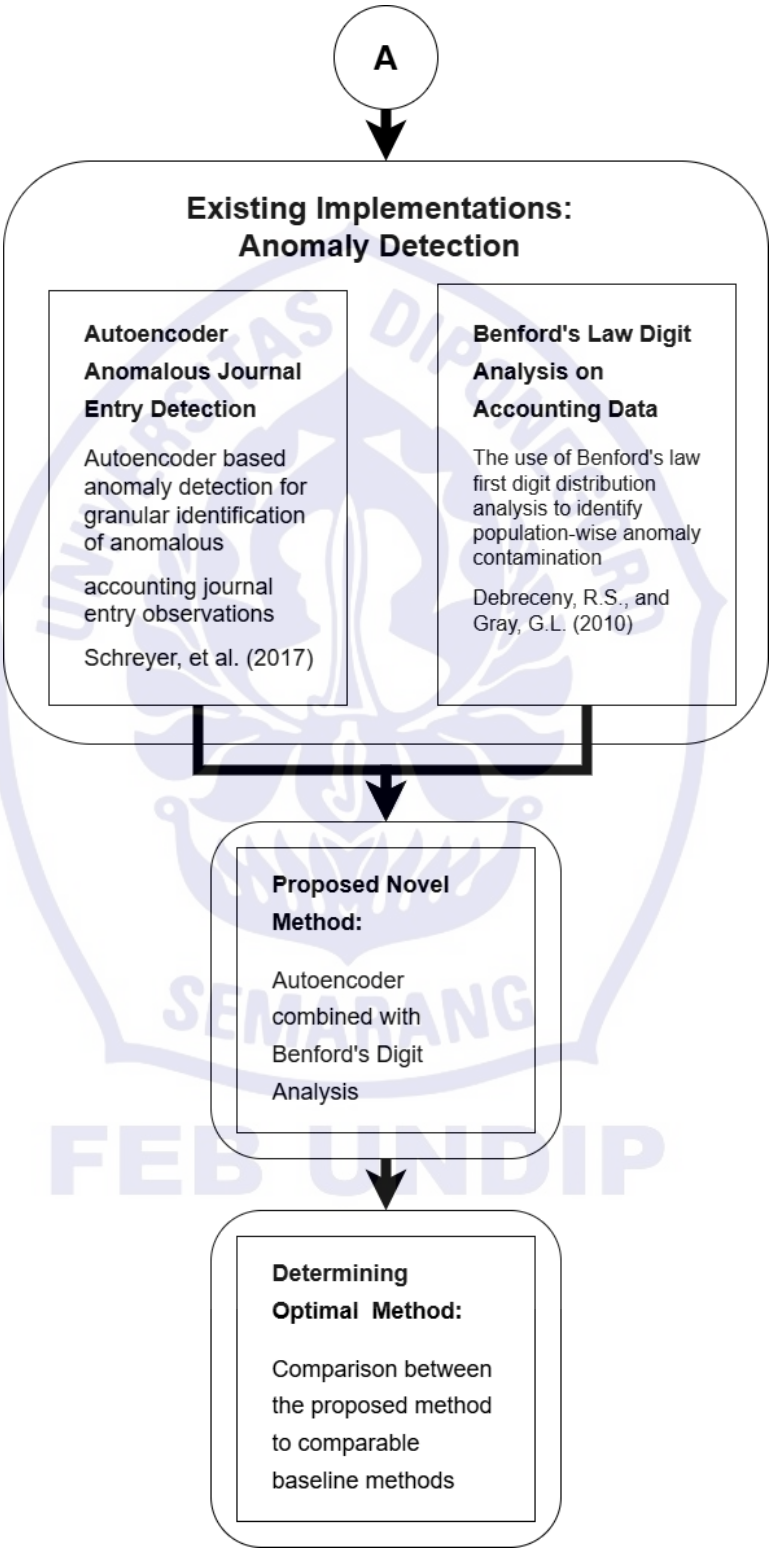


Figure 2.15. Research Framework Diagram Continuation



Source: Author

Adopting a deductive approach in synthesizing the research outline, deriving from a more general broad observation on the state of the datasphere, before narrowing down to current solutive implementations, and state-of-the-art publications on implementations within the context of accounting, the novel combined approach of autoencoder and Benford's law digit analysis is postulated to be a potentially better performing red flag test method in comparison to other baseline methods.

The current status quo of the datasphere, the global state of worldwide data, is characterized by increasing volume (Petroc (2024)), mostly consisting of complex unstructured data (Villars, et al. (2017); IDC (2023)), with regard to the rise of automation proxied by RPA statistics (Le Clair, et al. (2017); Computer Economics (2020); GVR (2025)). This characterization has prompted a growth in publications of ameliorative nature by leveraging artificial intelligence and/or machine learning, such as in Faccia, A., and Petratos, P., (2022); Auad, M., et al. (2024); Chen, Y., et al. (2024); Liu, Z. (2024); as well as Wang, R. Z. (2025). Further narrowing down to works implementing AI/ML solution in conducting anomaly detection procedures, are two foundational papers for this research. Namely, two papers introducing the use of artificial intelligence, machine learning, and data analytics for the purpose of anomaly detection within the realm of accounting (Schreyer, et al. (2017); Debreceny, R.S., and Gray, G.L., (2010)). Given the lack thereof satisfactory performance metrics for the autoencoder method introduced in Schreyer, et al. (2017) notwithstanding the method achieving the highest performance in comparison to other baseline methods, in conjunction with the lack of granularity for the purpose of red flag tests using

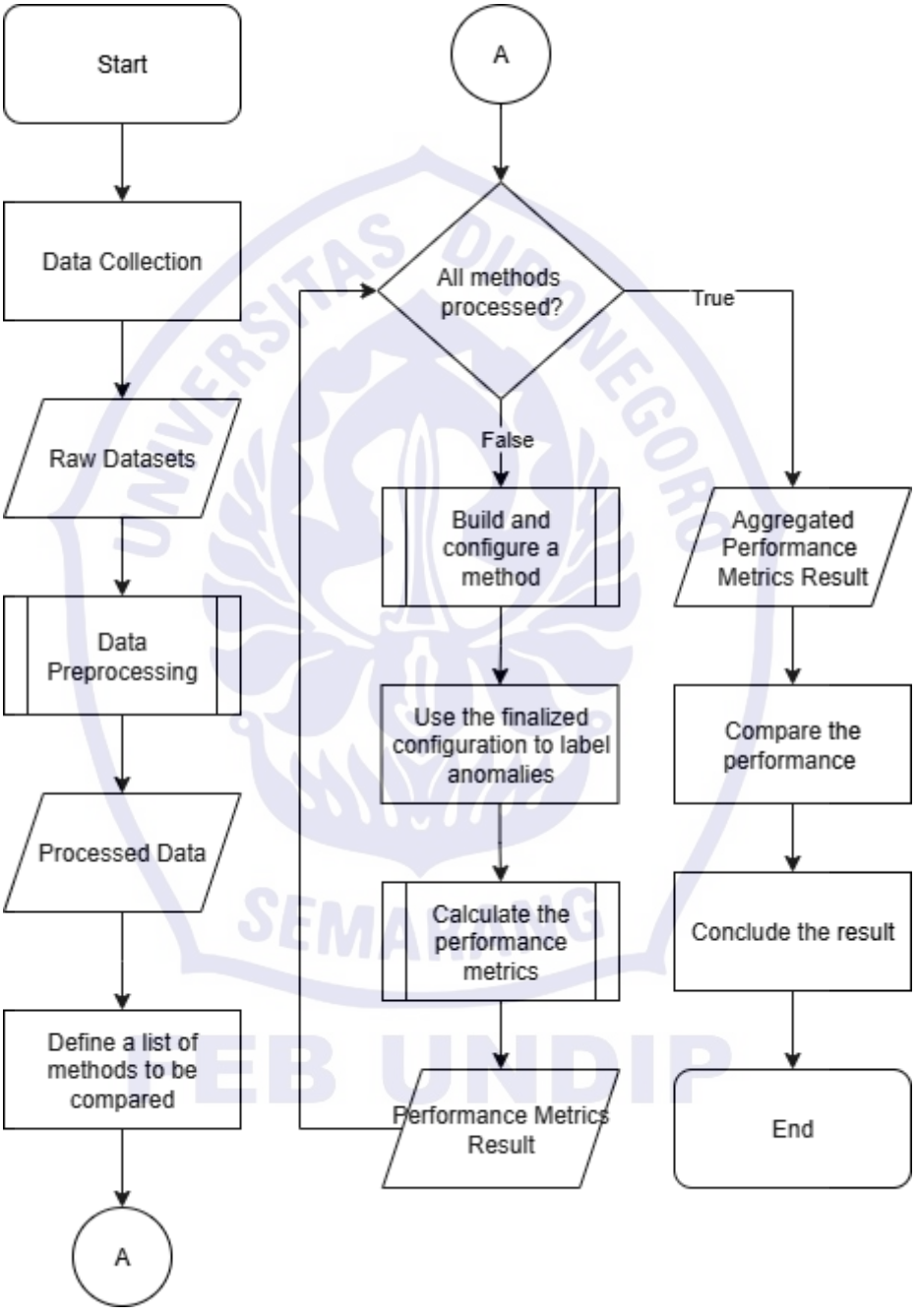
Benford's law digit analysis in Debreceny, R.S., and Gray, G.L., (2010), a new approach in combining the two is postulated to be of ameliorative nature to these previous works.

CHAPTER III
RESEARCH METHODOLOGY

3.1. Research Procedure Overview

This research mainly follows the framework of a case study research. In Sekaran, U., and Bougie, R. (2016), case studies refer to a research strategy with an emphasis in focusing on information collection for a specific object, activity, and/or event. Case studies can present both qualitative and quantitative data for both analyses and/or interpretation. Yin (2009) further elaborated the matter that case studies are typically characterized by the deployment of a diverse and inexhaustible range of data collection methods. In its entirety, the sequential procedures involved in conducting the search are as illustrated within Figure 3.1**.** below. In hindsight, the relevant procedures can be summarized as comparative testing between the proposed novel method of combined autoencoder neural network and Benford's first digit analysis in juxtaposition to a baseline autoencoder implementation, alongside a few additional baseline methods.

Figure 3.1. Broad Overview on Research Procedure



Source: Author

3.2. Population and Sample

To define population within the context of this research pertains to discourses on the nature of the target in generalizability upon which the population is subjected to by inferences made from the sample. As such, while the plausibility of processing the entire population to derive conclusions with regard to the population of interest is within a reasonable bound possibility for the method to achieve, the lack of necessary computing resources for this research functions as a factor of necessity for a sampling procedure to take place. This particular predicament demonstrates a theme of linearity to the purpose of sampling, as Lind, et al. (2019) defined a sample as a subset of the population of interest whereby inferences can be made from extrapolating the sample to the population. As such, the population for this research can be identified in relation to its aim to introduce a novel anomaly detection method that is theoretically applicable for any accounting journal entries data for a given fiscal year. In which case the sample by which a generalizable conclusion is derived from a dataset of 533,009 observations of accounting journal entry data for a single fiscal year from an anonymized business entity, obtained from a publicly available repository from the work of Schreyer, et al. (2017). The nature of the research necessitates for a realistic accounting journal entry data in large volume ideally even the entire journal entries for a given fiscal year. Due to the fact that the confidential nature of disclosures are really sensitive at that level, such data will be hard to come by especially for an undergraduate student. Schreyer's data provides data to that

extent of disclosure and volume, although the data has been anonymized using a one-way hash function.

This sample is then further divided into sub-samples, whereby the sub-samples follow a modified distribution preserving stratified random sampling method. In Lind, et al. (2019), stratified random sampling is described as a method in which the population is divided into subgroups referred to as strata, and each stratum is then randomly sampled from to ensure fair representation of each stratum. For this particular research, modifications are made in a way to ensure that an underlying distribution is preserved for the sub-sample. The sampling procedure for the sub-samples follow the sequences as detailed in Figure 3.2., Figure 3.3., and Figure 3.4.

Figure 3.2. Distribution Preserving Sampling Procedure

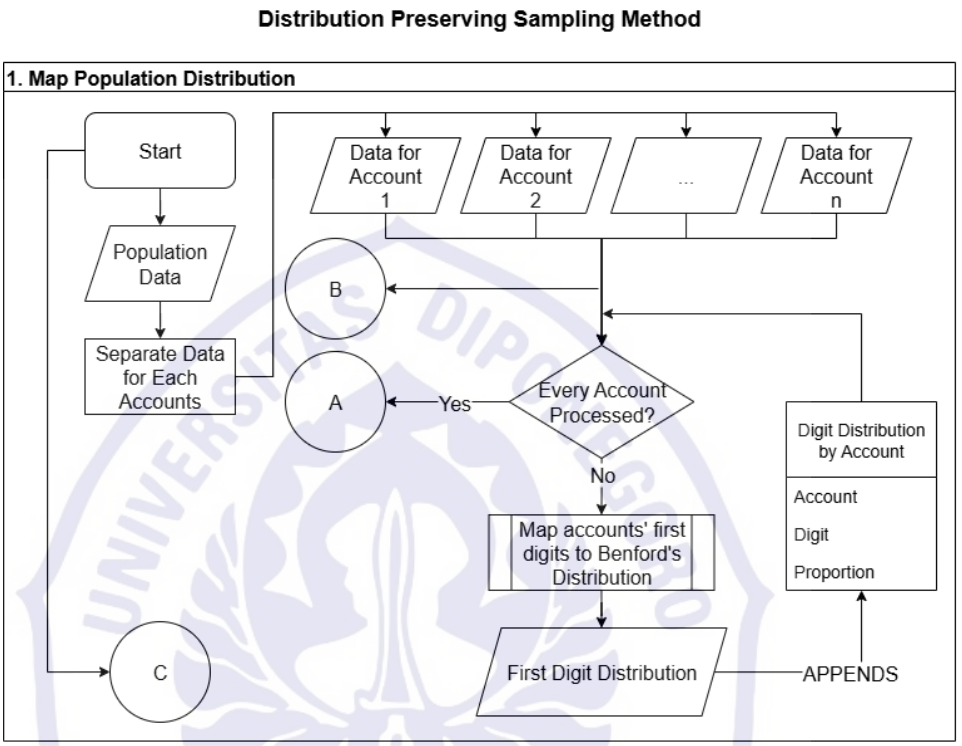


Figure 3.3. Distribution Preserving Sampling Procedure Continuation 1

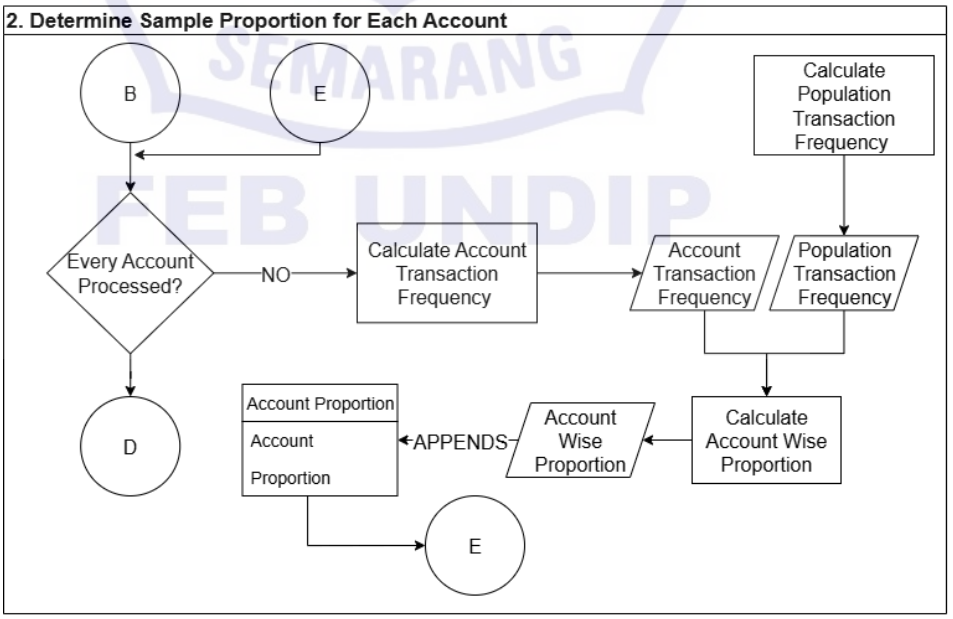
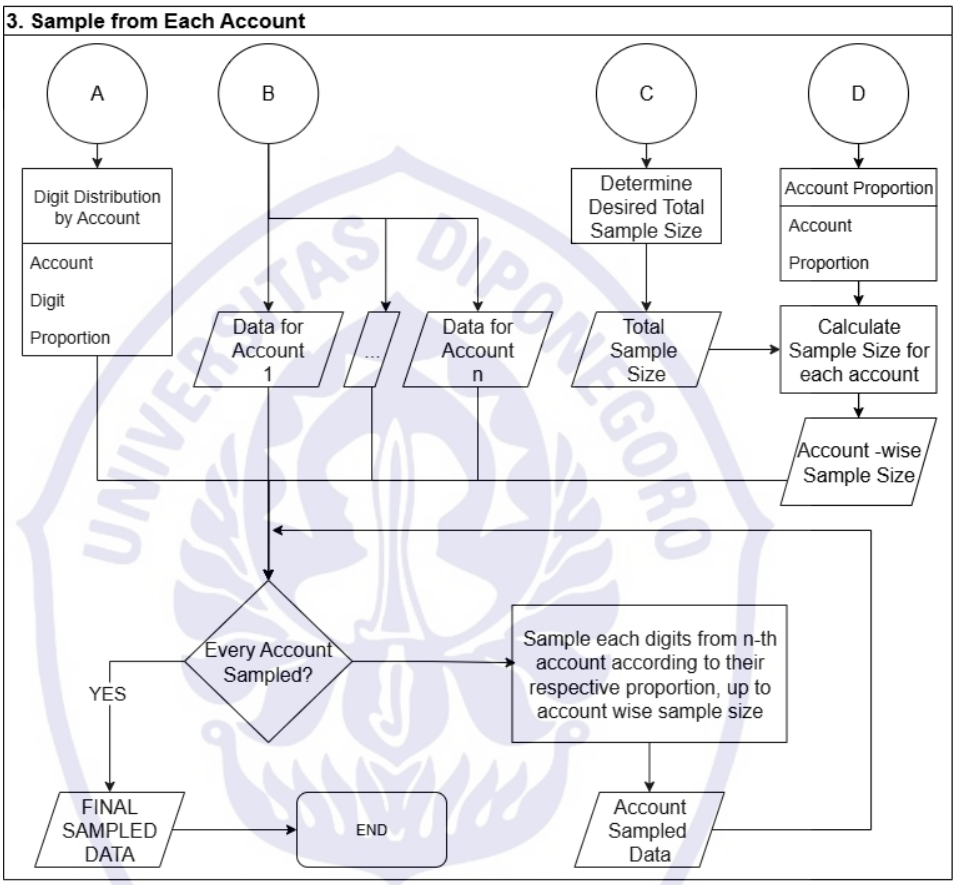


Figure 3.4. Distribution Preserving Sampling Procedure Continuation 2



Source for Figure 3.2., Figure 3.3., and Figure 3.4.: Author

3.3. Data Types and Sources

The types of data used in the research can be characteristically bifurcated dichotomously between qualitative and quantitative data. However, both of them are processed through the same input to output pipeline, thus they can be more accurately described as mixed data, containing both qualitative and quantitative data. The qualitative data refers to columns within the table identifiable as nominal categorical variables, such as:

1. Currency Key (WAERS);
2. Company Code (BUKRS);
3. Transaction Key (KTOSL);
4. Document Number (BELNR);
5. Posting Key (BSCHL);
6. General Ledger Account Number (HKONT);

While the quantitative, continuous numerical columns are:

7. Local Currency Amount (DMBTR);
8. Foreign Currency Amount (WRBTR);

The majority of the variables being of qualitative columns notwithstanding, the qualitative, nominal, categorical columns are processed quantitatively in conjunction to the quantitative variables by representing the categorical values by proxy of dummy variables.

With regard to the source of the data, the data was obtained from a publicly available repository from the work of Schreyer, et al. (2017). The data originated from a real world instance of accounting journal data for a single fiscal year from an anonymized entity, extracted from the entity's SAP ERP system. The data has been injected with synthetic anomalous journal entries.

The data is structured as tall data, whereby each column represents a single variable and each row represents an instance of observation. To understand how SAP represents the orthodox journal entry representation in normalized/tall data format, suppose the usual accounting representation of a normal compound journal entry below in Table 3.1.

Table 3.1. Normal Accounting Compound Journal Entry Representation

Date	Ref	Description	Account	Dr.	Cr.
2025/ 08/01	Inv-001	Purchase of Inventory	Inventory	1000	
			VAT-In	110	
			Act. Payable		1110
2025/ 08/02	Inv-002	Purchase of Inventory	Inventory	1000	
			VAT-In	110	
			Act. Payable		1110

Source: Author

Within a 'traditional' accounting format, a compound journal entry recording a single instance of a unique transaction is typically represented with multiple rows such as illustrated in Table 3.1. However, as denoted in Schreyer, et al. (2017), SAP ERP systems represent these attributes by utilizing two different tables namely the BKPF (*Buchungsbelegkopf*) table or the "Accounting Document Header" table, as well as the BSEG (*Buchungssegment*) or the "Accounting Document Segment" table.

Each row in the BKPF table represents a single transaction, however, a single transaction is further detailed under multiple rows within the BSEG table. Each unique instance of transaction is recorded as a single row within the BKPF table, using four unique primary keys namely the BELNR (Accounting Document Number), BUKRS (Company Code), GJAHR (Fiscal Year), and MANDT (Client

ID, not included), as presented in Table 3.2. While the relevant attributes to elaborate on each transaction are detailed using multiple rows in the BSEG table, as displayed in Table 3.3. The two tables can be joined using said primary keys of the BKPF table to the foreign keys of the BSEG table, such as by pivoting to the primary key columns to obtain a more traditional accounting representation.

Table 3.2. BKPF Table Representation

BELNR	BUKRS	GJAHR	BLDAT	BKTXT
Inv-001	A001	2025	2025/08/01	Purchase of Inventory
Inv-002	A001	2025	2025/08/02	Purchase of Inventory

Source: Contextualized by the Author

Table 3.3. BSEG Table Representation

BELNR	BUKRS	BLDAT	BUZEI	HKONT	DMBTR	SHKZG	BKTXT
Inv-001	A001	2025/08/01	1	Inventory	1000	S (Debit)	Purchase of Inventory
Inv-001	A001	2025/08/01	2	VAT-In	110	S (Debit)	Purchase of Inventory
Inv-001	A001	2025/08/01	3	Act. Payable	1110	H (Credit)	Purchase of Inventory
Inv-002	A001	2025/08/02	1	Inventory	1000	S (Debit)	Purchase of Inventory

Inv-002	A001	2025/08/02	2	VAT-In	110	S (Debit)	Purchase of Inventory
Inv-002	A001	2025/08/02	3	Act. Payable	1110	H (Credit)	Purchase of Inventory

Source: Contextualized by the Author

3.4. Data Collection Method

The data obtained and used for the research can be characterized as data collection through secondary sources. In Sekaran, U., and Bougie, R. (2016), secondary data are defined as data that have been collected by other parties different from the parties involved within the current research. The relevant data was made publicly available on a GitHub repository with regard to the work of Schreyer, et al. (2017). The anonymization of the data consequently implies the inability to identify the business entity of which the original accounting journal entries originate, as well as the actual fiscal year that the original data was in.

3.5. Data Analysis Method

The analytical procedures associated with the data analysis process of this research can be subsequently ramified into three consecutive steps, namely data preprocessing, model building and prediction (for both the baseline method and the proposed novel method), as well as model evaluation. Each step is associated with their respective input-output pairs, as summarized in Table 3.4.

Table 3.4. Data Analysis Steps Input-Output Pairings

Sequence Order	Process	Input	Output
1	Data Preprocessing	Raw Downsampled Data	Preprocessed Data (categorical columns are encoded with one-hot encoding; numerical columns are normalized using min-max scaling)
2	Model Building and Prediction	Preprocessed Data	Anomaly Binary Classification Labeling
3.	Model Evaluation	Anomaly Binary Classification Labeling, Ground Truth Anomaly Binary Classification Labels	Accuracy score, Precision score, Recall score, F1-score; for both baseline and proposed method

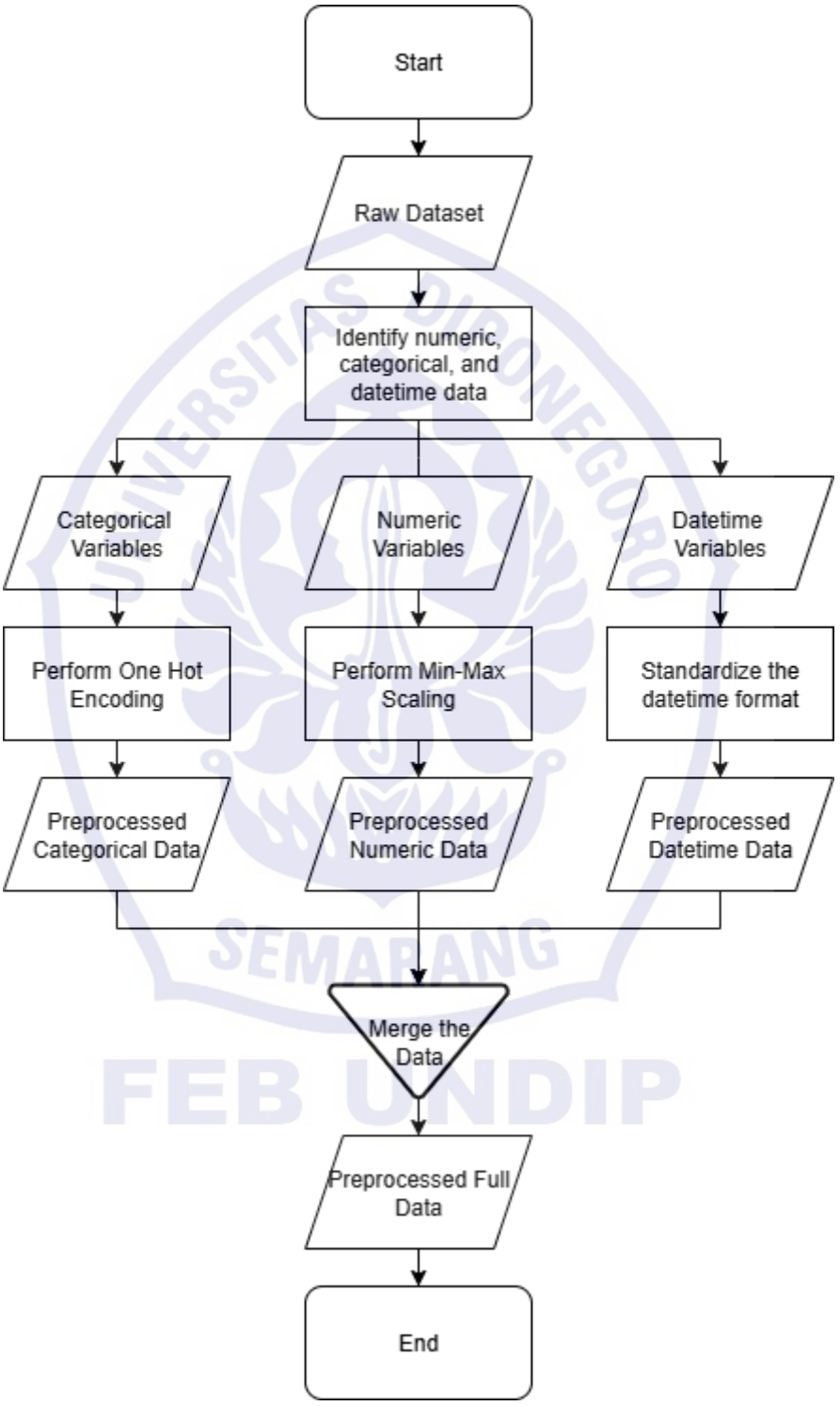
Source: Author

3.5.1. Data Preprocessing

The autoencoder model used in both the baseline implementation as well as the novel proposed method can technically work on qualitative and unstructured data. However, at its core, even within the context of unstructured implementations the model in its actuality performs mathematical matrix operation. Thus, implementations in processing unstructured data require numerical representation.

In general, the process of which data is transformed and/or prepared prior to analysis in order to ensure compliance to its processing constraints is associated with the data preprocessing phase, notwithstanding either the transformation is unstructured data to numerical data and/or otherwise (not limited to the description provided in this excerpt). Within the context of this particular framework of research, three distinct procedures are associated with the preprocessing phase, namely in preprocessing categorical, numerical, as well as timestamp data. The data preprocessing phase as illustrated below in Figure 3.5. provides a broad perspective on how the preprocessing phase is conducted.

Figure 3.5. Research Data Preprocessing Flowchart



Source: Author

In this context, categorical data (all categorical columns within the data) are processed using the one-hot encoding method. Within the context of machine learning, one-hot encoding refers to an encoding process for the data whereby a constraint is placed upon the data so that legal values are only represented in binary form, either 0 or 1, as defined in Harris, D., and Harris, S. (2012). Table 3.6. exemplifies the encoding process using one-hot encoding for the original representation in Table 3.5. One-hot encoding is in a lot of ways similar both in process and reasoning to dummy variables, with a key distinction in that one-hot encoding keeps all categories instead of dropping one category (as is done in dummy variable encoding to avoid multicolinearity).

Table 3.5. Before: Original Representation

Food Type	Quantity
Apple	54
Orange	15
Watermelon	25

Source: Contextualized by the Author.

Table 3.6. After: One-Hot Encoding Representation

Apple	Orange	Watermelon	Quantity
1	0	0	54
0	1	0	15
0	0	1	25

Source: Contextualized by the Author.

Numerical columns are processed using the min-max normalization function. Whereby for a matrix $A \in \mathbb{R}^{n \times m}$ of n rows and m columns, each cell is normalized as:

$$A_{(i,j)} = \frac{A_{(i,j)} - Min(A_j)}{Max(A_j) - Min(A_j)}$$

Timestamp data are preprocessed by dissecting the value into several variables of interest, typically still within the context of timestamp values. For example suppose the left-most column in Table 3.7. as the original representation while the subsequent columns after it are the extracted variables to represent the original timestamp format. Table 3.7. illustrates the preprocessing logic for timestamp data.

Table 3.7. Timestamp Data Preprocessing

Timestamp	Year	Month	Date	Day-of-the-week	Hour
Monday, 4 August 2025, 11.05	2025	8	4	1	11
Tuesday, 5 August 2025, 08.00	2025	8	5	2	8
Thursday, 25 December 2003, 09.00	2003	12	25	4	9

Source: Contextualized by the Author.

3.5.2. Model Building

3.5.2.1. Baseline Autoencoder

The first part of the predictive analysis pertains to the use of a baseline method to which the novel proposed approach will be benchmarked to. In this case, the first predictive tools to be used is based on the implementation of the autoencoder neural network as published in Schreyer, et al. (2017). With regard to the selection of the autoencoder as one of the main baseline predictive models to be analyzed, several reasons outlined this decision. Chief amongst them is based on an empirical reasoning that via a smaller pilot/replication study, the autoencoder neural network model has managed to identify synthetic anomalies within the accounting data with an accuracy score of about 99%.

To provide a general understanding on how the autoencoder neural network works in identifying accounting anomalies, a representation on how data is transformed throughout the process is used to visualize the process. The process starts with accounting general journal entries within the 'traditional format' as illustrated in Table 3.1. SAP representations on said data is then extracted, namely by means of the BKPF and BSEG table representation of the journal entries. The data is further processed either by the two tables, and subsequently performing feature selection on the joint data as in linearity to the implementation in Schreyer, et al. (2017). Ideally, the transformation is done in a way where each row represents a unique transaction. This data, henceforth will be referred to as the raw data, can be represented as the matrix $Raw\_Data \in \mathbb{R}^{n \times m}$, for rows n and m columns, where n also represents the number of transactions within the data.

This raw data contains a mixed set of columns, numeric or otherwise. The data is then subsequently preprocessed using the preprocessing method described within the previous part of the excerpt, to obtain a new numerical matrix as $Preprocessed\_Data \in \mathbb{R}^{n \times o}$, whereby the row dimension remains the same as n as well as to the amount of transactions, whilst the column dimension of o is changed into high-dimensional representation after the one-hot encoding procedures, and o will always be greater than m. In our pilot study the preprocessed, full or non-sampled, representation is a high-dimensional multivariate matrix, with n = 533,009 rows (representing unique transactions), and o = 618 columns. These preliminary processes are done to ensure compatibility for the mathematical operations associated with the autoencoder neural network.

After the data has been preprocessed, this data is then fed to the autoencoder neural network. The autoencoder model will learn to represent the $Preprocessed\_Data$ matrix into a smaller representation known as the code, the bottleneck layer, the latent space, or by other nomenclatures. This encoding process moulds the matrix into a smaller matrix of $Latent\_Space \in \mathbb{R}^{p \times q}$ for $p \leq n$ and $q \leq o$. The encoded representation is then reconstructed back into its original dimension of $Reconstructed\_Matrix \in \mathbb{R}^{n \times o}$. The degree of difference between each cell of the original matrix and the reconstructed matrix is then aggregated to a value known as the Reconstruction Error, where each row (each unique transaction) will be represented each with their own Reconstruction Error value.

The Autoencoder Reconstruction Error variable is the loss function variable used in training the autoencoder model which quantifies the extent to which the model fails to reconstruct the original representation as per the input layer. Suppose y matrix is used to represent the original *Preprocessed_data* matrix, and $\widehat{y}$ as the representation for the *Reconstructed_Matrix*, this loss function can be expressed in terms of the function $L(y, \widehat{y})$. The loss function $L(y, \widehat{y})$ used within this research is closely related to the Sum of Squared Errors (SSE) loss function in Warner, et al. (1996) as detailed within the preceding literature review section. However with regard to the aggregated value, the loss function used in each training iteration is the Mean Squared Error (MSE) variable, computed as:

$$L(y, \widehat{y}) = MSE = \frac{1}{n \times o} \sum_{i=1}^{n} \sum_{j=1}^{o} (y_{ij} - \widehat{y_{ij}})^2$$

for the entire data set representable as the matrix $A \in \mathbb{R}^{n \times o}$ with n rows and o columns (variable dimension after the preprocessing). Whereby y_{ij} stands for the original value as per the input data at row i and column j, and $\widehat{y_{ij}}$ stands for the reconstructed value as predicted by the autoencoder of the specified row and column index.

Consequently, the individual reconstruction error score per observation (row) is the uncapitalized *mse* computed across all columns (variables) for the i-th row as:

$$mse = \frac{1}{o}\sum_{j=1}^{o}(y_{ij} - \widehat{y_{ij}})^2.$$

Each value of the uncapitalized *mse* variable, i.e. the row-wise Reconstruction Error value, can then be stored into a column vector of $Reconstruction\_Error \in \mathbb{R}^{n \times 1}$ or simply the vector $Z \in \mathbb{R}^{n \times 1} = [z_1, z_2, ..., z_n]$. The uncapitalized *mse* variable will then be subsequently parsed to the Heaviside threshold function to determine the binary anomaly classification in the case of the baseline autoencoder implementation, or further operated upon with regard to the proposed method.

The final prediction i.e. the verdict on whether or not each transaction is predicted as anomalous is determined using a conditional heaviside step function, as illustrated in Figure 3.6.

Figure 3.6. Heaviside Step Function

$$\phi(z) = \begin{cases} 0 \text{ if } z \leq \text{threshold} \\ 1 \text{ if } z > \text{threshold} \end{cases}$$

Source: Rosenblatt (1958)

Whereby the value for threshold itself can be determined by auditors given their scope of sample size, through which the threshold value can be obtained. For example, if the auditors determined that the number of transactions they want to flag is 150 transactions, the threshold value is calculated as:

$$Target\ Percentile = \frac{A - B}{A} \times 100$$

Where A represents the total number of observations (transactions i.e. n), and B represents the desired number of observations to flag (in this case 150). The threshold value is simply the percentile value of the data at the target percentile. The final output of the model can then be evaluated using the prescribed performance metrics.

3.5.2.2. Benford's Digit Analysis

The second baseline method to be built upon and compared to the novel proposed method is an implementation of Benford's law on the distribution of digits introduced in Benford, F. (1938) particularly in application for anomaly detection in accounting based upon the foundational works of Nigrini, M. (1996), and Nigrini, M. (2011). In retrospect, Benford's law refers to an empirical law in natural data to have a tendency whereby the frequencies of digits tend to follow certain distributions. The most famous of all, is the Benford's distribution of first digits whereby the Benford's law distribution of first digits are as expressed with the function:

$$F(a) = log(1 + \frac{1}{a})$$

for $F(a)$ denoting the expected frequency of the leading digit a. The main reason as for why Benford's analysis is chosen for the second method of this research, particularly the implementation found in Nigrini, M., (2022), lies on its ability to granularly flag anomalous journal entries.

In Nigrini, M. (2022), an implementation was introduced for granular identification of problematic journal entries by means of Benford's digit analysis, as opposed to his initial works featuring population-wise contamination

indicator. Nigrini's implementation in granular identification is conducted by performing Benford's digit conformity for each sub-groups based on a category of the unique combinations of first-two digits. Observations belonging to problematic categories are therefore subsequently flagged as anomalous. In hindsight, Benford's digit analysis within this research follows a similar idea of sub-group evaluation. However, the main difference being that we examined the sub-groups by way of general ledger accounts. Within each sub-groups, conformity to Benford's law is analyzed in terms of the distribution of first digits in accordance with Benford's law. The degree of nonconformity is evaluated using Nigrini's recommended metrics of Mean Absolute Deviation (MAD).

The use of Mean Absolute Deviation as a way to measure the conformity of the distribution of positional digits (such as first digits, second digits, third digits, and so on) in accounting forensic analytics is mainly attributable to the author's exposure to the work in Nigrini, M.(2011) following the preceding work in Nigrini, M. (1996). Although Nigrini, M. (2017) has disclosed some caveats that MAD should still be taken with a grain of salt, the primary reason as to why Mean Absolute Deviation has demonstrated itself to still be a useful method in determining the conformity between distributions as opposed to other popular statistical tests such as the Z-Test, Chi-square Pearson's test, Kolmogorov-Smirnov test, Kuiper test, Chebyshev-distance test, Euclidian-distance test, etc., was touched upon in Druică, E., et al. (2018) and Nigrini, M.. (2011), primarily for issues related to the increase in sample size. The work by Nigrini, M. (2000) himself has also previously stated the misleading

nature of the Chi-square Pearson's test alongside the Z-Test in examining Benford's law conformity in higher sample size. Following this reason, the Mean Absolute Deviation test that can be computed as:

$$MAD = \frac{\sum_{i=1}^{n} \left| Obs_i - Exp_i \right|}{n}$$

for Obs_i representing the actual observed value to the expected value of Exp_i. Within the context of this research, the conformity test to Benford's law distribution refers to conformity between the observed and expected distribution of the first digits of the data, omitting subsequent digit analysis (second digits, third digits, and so on).

The work in Nigrini, M. (2011) provides a way to categorize conformity based on the value of MAD, as illustrated in Table 3.8. In our implementation, journal entries belonging to categories subscribing to MAD values higher than 0.0018 are therefore flagged as anomalous. For a real life deployment whereby this red flag anomaly categorization guides subsequent audit procedures, auditors can then sample again from this set of flagged accounts, providing capacity for integrating other techniques as well.

Table 3.8. Benford's Law Conformity Category based on MAD

MAD Score	Category
0.0000 to 0.0012	Close Conformity
0.0012 to 0.0018	Acceptable Conformity
0.0018 to 0.0022	Marginally Acceptable Nonconformity
Above 0.0022	Nonconformity

Source: Nigrini, M. (2011)

3.5.2.3. Proposed Combined Method

In retrospect, to develop a combined anomaly approach that leverages both the autoencoder accounting anomaly detection procedure as well as the Benford's law first digit analysis accounting anomaly detection pertains to experiments on different configurations derived from several heuristical designs. At the same time, the extent of which this research will explore different heuristically defined configurations is a delineating factor to the clarity of the research. As an alternative, notwithstanding the fact that the developmental process itself will not be described, the best performing heuristical approach, including the chosen heuristical mechanism, alongside the rationale for the mathematical and/or non-numerical operations associated with the proposed algorithm, will therefore be described within the fourth chapter of Results and Analysis.

3.5.3. Model Evaluation

For either the baseline autoencoder implementation, Benford's law digit analysis, as well as the proposed combined method, the final prediction output for

each row is a binary value denoting whether or not the observation is labeled as anomalous. The model evaluation phase of the data analysis pertains to discourses on the measurements used to evaluate how well this binary prediction reflects the ground truth. In which case, several metrics are used namely categorical accuracy, recall, precision, and F1-score.

3.5.3.1. Classification Accuracy

The accuracy score is a generalized score in measuring the performance of the categorical prediction with a lack thereof consideration to whether or not false positives or false negatives are more preferable. In continuation to the theme of linearity to the semi-supervised approach for autoencoder anomaly detection as demonstrated in Schreyer, M., et al. (2017), the model performs binary categorical prediction with regard to the status of anomalous behavior. However, the approach within the same literature trained the autoencoder model to the cross-entropy loss function as opposed to our MSE loss function. In this case, our approach differs, whereby binary classification is performed by parsing the reconstruction error to a deterministic Heaviside-step function with a grid search approach to find the best threshold, and only thusly can the categorical accuracy computation as can be calculated below can be performed:

$$\frac{(True\ Positives + True\ Negatives)}{(True\ Positives + True\ Negatives + False\ Positives + False\ Negatives)}$$

This difference in approach is related to the nature of the proposed novel method in combination between Autoencoder and Benford's digital analysis that will perform mathematical operations on the reconstruction error values of the autoencoder.

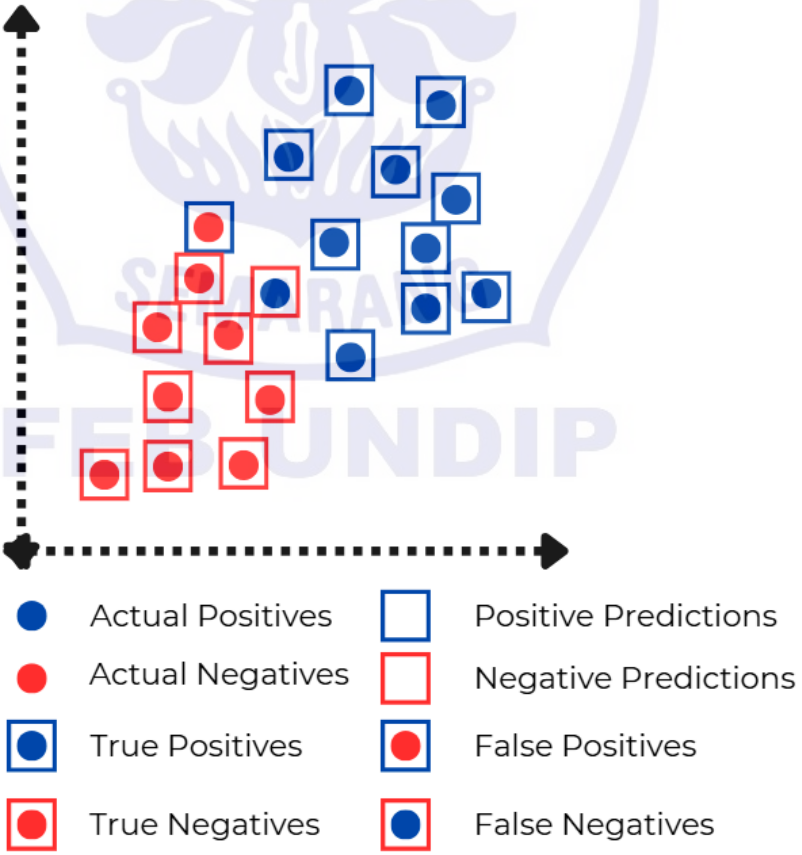
3.5.3.2. Classification Recall

Classification recall (or simply recall) is another metric used in gauging the performance of categorical prediction algorithms. As per Powers, D.M.W. (2007), classification recall is defined by its formulaic definition as the proportion between real positive cases that are correctly predicted by the model. Classification recall is also often referred to as the sensitivity value.

Referring to the diagram as illustrated in Figure 3.7., classification recall can be intuitively understood as:

$$\frac{True\ Positives}{Actual\ Positives} = \frac{True\ Positives}{(True\ Positives + False\ Negatives)}$$

Figure 3.7. Model Prediction Performance Variables



Source: Author.

3.5.3.3. Classification Precision

Classification precision refers to the ratio between positive predictions that are actual positives. It is also known by another name as the confidence value. Referring to Figure 3.7. above, it can be defined in two ways, as:

$$\frac{True\ Positives}{Positive\ Predictions} = \frac{True\ Positives}{(True\ Positives + False\ Positives)}$$

3.5.3.4. Classification F1-Score

The dynamic between precision and recall in binary categorical prediction can be largely understood as a trade-off relationship between either one of them. Whereby increasing the model's ability to improve on one metric tends to result in the decrease of the other metric. In general, classification recall is more preferable in cases where the costs of false positives are higher than the costs of false negatives, vice versa applies for classification precision, whereby the costs of false negatives are higher than the cost of false positives. To visualize this idea, it is beneficial to frame the two in an analogous context of accounting, whereby a red flag prediction is used to initialise subsequent audit procedures, positive predictions will prompt subsequent audit procedures on said observations. Higher precision is preferable in the context of public accounting audits, as false positives will lead to wasteful audit resources while public accounting auditors are not necessarily required to capture any and all errors, just an enough amount for materiality. Meanwhile, higher recall tends to be more preferable within the context of internal audit, as the cost of conducting subsequent audit procedures

are generally lower than the cost of misstatements due to the failure of internal validation.

However, there are instances whereby the need to balance between precision and recall are imperative. These instances necessitate the use of metrics such as the F1-score. The F1-score is simply the harmonic mean between precision and recall, that can be computed as:

$$F_1\ Score = 2 \times \frac{(Precision \times Recall)}{Precision + Recall}$$

CHAPTER IV
RESULTS AND ANALYSIS

4.1. Description of Research Object

This research is conducted to investigate the efficacy of two different anomaly detection methods namely using autoencoder neural networks based on the work of Schreyer, et al. (2017) and Benford's law granular first digit analysis based on the works of Nigrini, M. (1996), Nigrini, M. (2011), Nigrini, M., (2022), and Debreceny, R.S., and Gray, G.L., (2010), as well as developing a novel approach combining the two methods. The three methods, though different in procedure, aim to identify anomalous journal entries from a sampled dataset of about 533,009 observations of accounting journal entry data for a single fiscal year from an anonymized business entity, obtained from a publicly available repository from the work of Schreyer, et al. (2017). The data has been injected with synthetic anomalous journal entries, the main purpose of the models (methods) is in identifying whether or not a transaction is part of this anomalous group as a binary prediction of either 0 or 1 (1 denoting anomalous status). These binary predictions made by the different methods will then be compared to one another using several prescribed binary prediction performance metrics, namely classification accuracy, classification precision, classification recall, as well as the harmonizing score of F1-score to harmonize the different performance metrics.

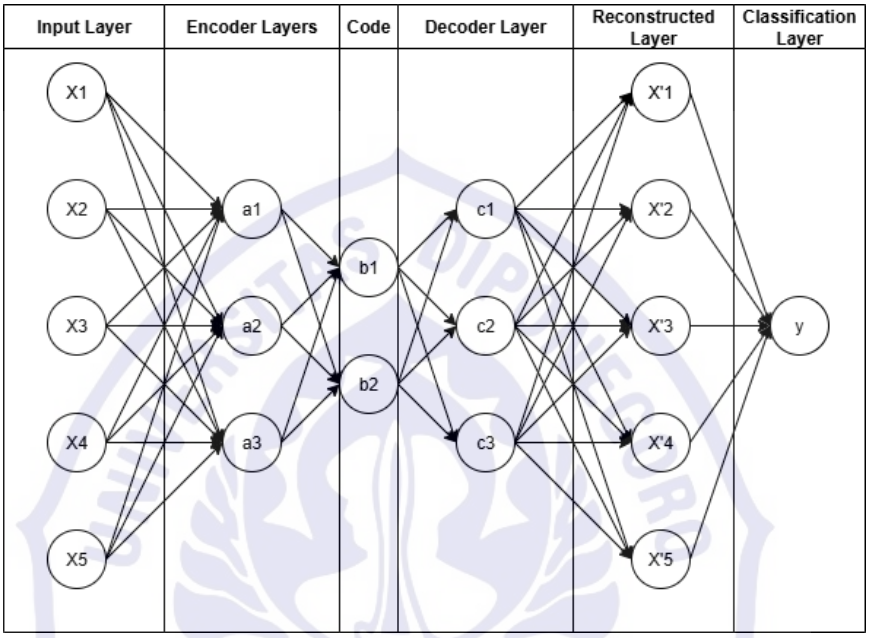
4.2. Data Analysis

4.2.1. Replication Studies

4.2.1.1. Autoencoder Neural Network Anomaly Detection

In retrospect the approach this research takes in setting up the prediction procedures for the autoencoder neural network is characterized by a theme of departure from the original implementation in Schreyer, et al. (2017). Schreyer's implementation (as illustrated in Figure 4.1.) appends an extra layer for binary prediction whereby the reconstructed outputs of the autoencoder are subsequently fed to, hence the model's inherent approach is strictly within the domain of supervised classification. In contrast, the approach adopted in this research trains the autoencoder model to predict by way of reconstructing the original representation of the data, such that a score of reconstruction error can then be attributed per row (i.e. per transaction) denoting the degree of difference to the original representation without training for a binary prediction. This score of reconstruction error is then fed to a heaviside step function whereby given a value of threshold set by the auditor (which can be configured to a percentile of the distribution of the reconstruction error), a decision is made to determine whether said observation is anomalous or not.

Figure 4.1. Simplified Overview on Schreyer's Autoencoder Architecture



Source: Author

The semi-supervised approach adopted in this research is the result of specific considerations on the nature of the intended deployment context for the autoencoder model in the real world. Where, in consideration to one of the research objectives is in developing and delivering a reusable anomaly detection procedure by way of a Python utility script, the context of which such deployment will be subjected to is a context characterized by an absence of ground truth error. In this context, the reusability of the autoencoder is not represented using saved models containing training weights since the model will have only learned the encoding-decoding procedure in relation to the norm of the particular data it is trained to, notwithstanding its inapplicability in different datasets. As such,

auditors will need to train the autoencoder specifically for their own dataset(s) whereby the lack of ground truth will be apparent. The reusability factor comes in the form of the interconnected start-to-end procedure itself, by way of a comprehensive start-to-end deliverable using reusable utility script(s).

Notwithstanding the departure from Schreyer's implementation by way of truncating the model up to the penultimate layer, the configuration of the model is still largely characterized by a theme of linearity to the implementation in Schreyer, et al. (2017). Schreyer's research investigated several shallow and deep autoencoder architecture configurations, in terms of how many layers they have as well as the number of neurons in each layer, as illustrated in Table 4.1. Schreyer's experiment found the deeper the layers and number of neurons, the better the model performance becomes. However, the configuration for the autoencoder implementation in this research is a deep autoencoder with 15 layers and an architecture configuration of [rows; columns]-256-128-64-32-16-8-4-*x*-8-16-32-64-128-256-[rows; columns] whereby auditors can set the size of the latent dimension of *x* (normally either 2 or 3). This choice of configuration is related to the constraints of the free Google Colab notebook environment in handling large computations that most auditors will have an immediate access to.

Table 4.1. Schreyer's Tested Autoencoder Configurations

E	Autoencoder Configuration
E1	[rows; columns]-3-[rows; columns]

E2	[rows; columns]-4-3-4-[rows; columns]
E3	[rows; columns]-8-4-3-4-8-[rows; columns]
E4	[rows; columns]-16-8-4-3-4-8-16-[rows; columns]
E5	[rows; columns]-32-16-8-4-3-4-8-16-32-[rows; columns]
E6	[rows; columns]-64-32-16-8-4-3-4-8-16-32-64-[rows; columns]
E7	[rows; columns]-128-64-32-16-8-4-3-4-8-16-32-64-128-[rows; columns]
E8	[rows; columns]-256-128-64-32-16-8-4-3-4-16-32-64-128-256-[rows; columns]
E9	[rows;columns]-512-256-128-64-32-16-8-4-3-4-8-16-32-64-128-256-512-[rows; columns]

Source: Schreyer, M., et al. (2017)

A hallmark of the truncated implementation of the autoencoder anomaly detection procedure adopted in this research is its compatibility with other audit sampling size determination procedures. Auditors will be able to set the threshold value to flag observations based on the value of reconstruction error in a way that only a determined number of observations will be flagged by the heaviside step function (as per the desired sample size) through a threshold calculation by means of a percentile value of the reconstruction error univariate distribution as defined in the research methods. In this research, a threshold value is calculated to flag 150 observations. For a sub-sampled dataset of 100,042 entries, the autoencoder neural network accounting anomaly detection achieved a performance as denoted in Table 4.2.

Table 4.2. Autoencoder Neural Network Performance Results

Performance Metric	Score
Classification Accuracy	99.96%
Classification Precision	93.33%
Classification Recall	80.46%
Classification F1-Score	86.42%

Source: Author

4.2.1.2. Benford's Law First Digit Analysis Anomaly Detection

The Benford's law first digit analysis implementation in this research, follows a specific approach to ensure granularity in classifying whether or not a transaction (in this case a row) is anomalous. In general, the foundational approach introduced in Nigrini, M. (1996) is a population-wise evaluation resulting in a somewhat unspecified existence of anomalous observations within the population. Notwithstanding the initially perceived lack of granularity, a new implementation in Nigrini, M. (2022) introduced an algorithm to leverage Benford's digit analysis to granularly identify where an observation is anomalous or not. This new implementation focuses heavily on grouped analysis, whereby Benford's analysis is conducted at different sub-population levels, i.e. whether observations belonging to a certain category of a categorical variable follow the expected distribution as detailed in Benford, F. (1938). This approach allows for identification of Benford's conformity (or in this case nonconformity) for observations belonging to that category.

In this research, the granular implementation for Benford's law digit analysis is conducted on the categorical variable of general ledger accounts. In simple terms, Benford's law conformity is tested for different accounts within the dataset. A result of MAD (Mean Absolute Deviation) can then be obtained for each category that scales up in representation of higher nonconformity. All observations belonging to that category (in this case account) will have a uniform value of MAD. Observations subscribing to an MAD value of more than 0.0018 will be flagged as anomalous. This threshold value is based on the determination of conformity category as introduced in Nigrini, M. (1996) as illustrated in Table 3.8., though simplified to a simple binary representation of either 0 and 1. Unlike the computational constraints of the autoencoder that necessitates the use of the prescribed downsampling, the Benford's law approach can accommodate the entire dataset of 533,009 unique transactions, of which the approach achieved performance metrics score as illustrated in Table 4.3.

Table 4.3. Benford's Law First Digit Analysis Anomaly Detection Performance Results

Performance Metric	Score
Classification Accuracy	99.99%
Classification Precision	100%
Classification Recall	70%
Classification F1-Score	82.35%

Source: Author

4.2.2. Developing and Testing the Novel Combined Approach

4.2.2.1. Developing the Novel Combined Approach

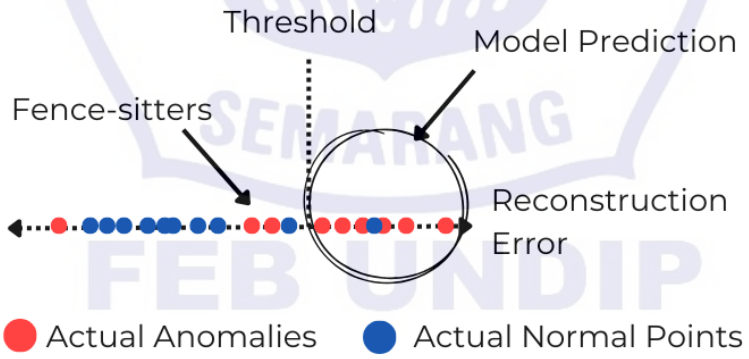
The relevant performance metrics used within the research including on but not limited to these baseline methods (autoencoder and Benford's digit analysis) discussed in the subsequent part of the Research Methodology chapter includes other measurements outside of accuracy, to which the business implications are applicable in lieu of and/or in addition to the scientific rigor of the research. In simple terms, we found out through the conducted replication studies that notwithstanding the fact that these baseline models throw out predictions on anomalies with great accuracy and precision, they missed out more than twenty percent of injected anomalies.

From the business perspective of public accounting firms' audit engagement, the lack thereof recall is not necessarily problematic, provided that precision is high and recall is simply 'good enough'. As they need not to identify any and all accounting anomalies. The audit needs of public accounting firms pertain to efficient audit expenses, in a way that flags raised by the predictive models in guiding subsequent audit procedures are not wasted on false flags of which the baseline methods have satisfied. However, the lack of recall is problematic in the case of internal auditing needs, whereby the cost of misstatements can be reasonably assumed to be higher than the cost of performing audit procedures.

The novel method that combines autoencoder neural network and Benford's law digit analysis proposed in this research aims to improve the recall

situation for better applicability in internal audit settings. This decision to combine the two methods stems from an *a priori* reasoning with regard to certain patterns of behavior particularly, in relation to the reconstruction error of the autoencoder. A postulate can be derived on the distribution of the reconstruction error along a univariate axis of itself, as illustrated in Figure 4.2., that observations demonstrating 'fence-sitting' behavior i.e. very close in value to the threshold but still less than or equal to the threshold should be reexamined through other perspective(s) for a reassessment of the flagging decision. In which case our preferred alternative perspective in conducting said reassessment is Benford's law digit analysis, given the performance as demonstrated within the pilot study and due to its very nature that it examines the data from a relatively distant perspective in relation to the autoencoder.

Figure 4.2. Reconstruction Error Flagging Decision



Source: Author

In retrospect, the autoencoder neural network maps out each transaction (represented as multivariate data) into a univariate vector matrix of $Reconstruction\_Error \in \mathbb{R}^{n \times 1}$ or simply referred to as the vector $Z \in \mathbb{R}^{n \times 1}$ = $[z_1, z_2, ..., z_n]$. The values in this vector represents the degree of deviation from the norm of said individual transaction, with regard to the failure for the model to reconstruct the data where given a low total loss across the entire data, the few observations that do not follow this trend typically exhibits anomalous behavior relative to the norm of the data. The heaviside step function then defines all observations above the threshold value to be anomalous. In this case, as illustrated in Figure 4.2., the predictions made by the autoencoder demonstrate a tendency for high precision in terms of low false flags though somewhat compromised in terms of recall that it missed some actual anomalies that are below the threshold criterion. The implementation within this research specifically aims to operate on observations below the threshold in an attempt to push more observations to be beyond the threshold by means of reassessment through Benford's law digit analysis.

Three criteria are used for the reassessment process, namely in terms of adjacency to threshold, Benford's law nonconformity, and data specific Benford's law nonconformity. While adjacency to threshold is to some degree self-explanatory as per why it is chosen as determining factor, the reasoning behind the two different Benford's law reassessment factors is justifiable by their difference in nature. Benford's law nonconformity is a binary score of 0 or 1, based on whether or not the Mean Absolute Deviation (MAD) value is greater

than or equal to 0.0018, stored as the vector matrix $benford\_label \in \mathbb{R}^{n \times 1}$. At the same time, the data specific Benford's law nonconformity factor is the normalized score of MAD specifically for that data. It is measured to ensure that even in Benford's law conforming data, some perspectives can still be valuable by examining the highest score of Mean Absolute Deviation values relative to that data. Suppose a vector matrix that stores the MAD values for individual observations as $MAD \in \mathbb{R}^{n \times 1} = [mad_1, mad_2, \ldots, mad_3]$, the normalized MAD values can be stored in a new vector matrix of $Norm\_MAD \in \mathbb{R}^{n \times 1} = [norm\_mad_1, norm\_mad_2, \ldots, norm\_mad_n]$ whereby the value at the *i*-th index can be measured with regard to the MAD matrix as:

$$norm\_mad_i = \frac{mad_i - Min(MAD)}{Max(MAD) - Min(MAD)}$$

The adjacency to threshold factor is a value that ranges from 0 to 1, with 0 denoting the furthest possible value from the threshold and 1 denoting the nearest possible value to the threshold. As with the two previous factors, this value is only calculated for observations less than or equal to the threshold. To compute the value, a distance metric to the threshold must first be calculated. Suppose a vector matrix that stores the adjacency score as $Adjacency\_Score \in \mathbb{R}^{n \times 1}$ and the distance vector matrix as $RE\_Distance \in \mathbb{R}^{n \times 1} = [re\_distance_1, re\_distance_2, \ldots, re\_distance_n]$ the distance value can individually be computed as:

$$re\_distance_i = Threshold - z_i$$

The adjacency to threshold factor is based on the normalized value of the distance calculation, that is then reversed given that the distance value ranges from 0 to 1 with 1 signalling the furthest value to the threshold. The vector matrix containing the individual values in $Adjacency\_Score \in \mathbb{R}^{n \times 1}$ can be computed as:

$$adjacency\_score_i = 1 - \left(\frac{re\_distance_i - Min(RE\_Distance)}{Max(RE\_Distance) - Min(RE\_Distance)}\right)$$

The three reassessment factors, each of them ranging from 0 to 1, can then be computed for the purpose of enabling a multiplier mechanism to the 'old' reconstruction error value. This multiplier mechanism is based on a heuristic approach developed through a trial-and-error development process. In simple terms, given a new matrix for the new transformed reconstruction error values as $new\_reconstruction\_error \in \mathbb{R}^{n \times 1}$, we need to first calculate the multiplier value to compute the new reconstruction error value stored as $Multiplier \in \mathbb{R}^{n \times 1}$, the multiplier value at the *i*-th index can be calculated as:

$$\frac{1}{3} + \frac{adjacency\_score_i + norm\_mad_i + benford\_label_i}{3}$$

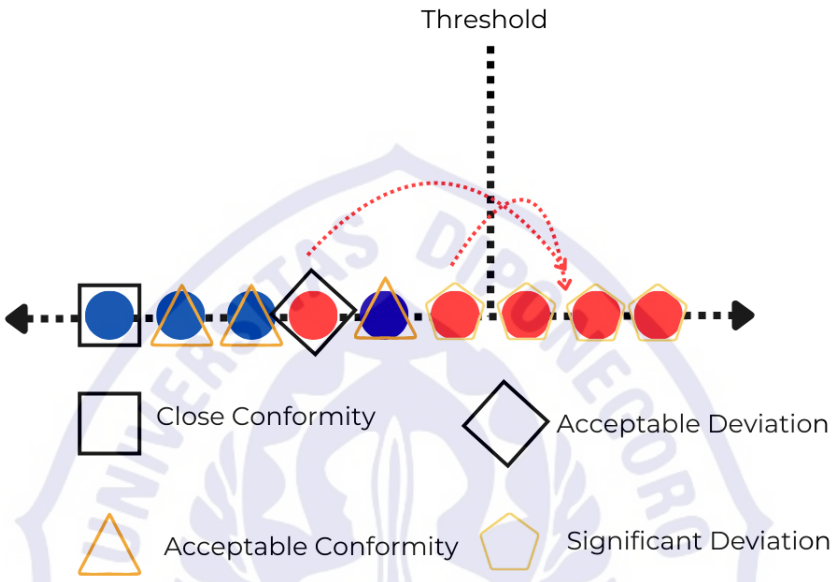
In retrospect, the multiplier value at its simplest terms is simply the average of the three reassessment factors. However, an additional operation is then conducted first by a constant value of 0.333. This addition is done so that the value of the multiplier will be close to or equal to 1 given that at least two criteria are satisfied. This operation is in relation to the calculation for the new reconstruction error, that is conducted by adding the old reconstruction error value to their respective distance to the threshold value after the multiplier is applied. If

the value of the multiplier is equal to 1, then this operation will shift the new reconstruction value to be equal to the threshold. A very small number of 1e-9 is then added to ensure that those values will be greater than the threshold value for anomaly flagging. If close to three criteria are met, then the new reconstruction value will be pushed even beyond the threshold (case in point for example with a 1.33 multiplier to the distance). In formal terms, the *i*-th value of the new reconstruction error can be calculated as:

$$z_i + (multiplier_i * re\_distance_i) + 1e - 9$$

The final output of the model will push more observations satisfying the criteria to some degrees that were previously below the threshold to be beyond the threshold, as illustrated in Figure 4.3. It is postulated that this operation will be able to improve the recall problem of both the autoencoder neural network as well as Benford's law digit analysis.

Figure 4.3. Novel Combined Approach Visualization



Source: Author

4.2.2.2. Testing the Novel Combined Approach

The combined approach ended up being more computationally demanding in comparison to either baseline methods, given its operations the computation for both the autoencoder neural network, the granular Benford's law first digit analysis, as well as the calculations associated with the heuristic multiplier mechanisms. As such, in a theme characterized by linearity to the autoencoder, a downsampling implementation was deemed necessary given the constraints of Google Colab's free environment, both in terms of reasoning as well as the downsampling scale downwards to 100,042 observations. The combined approach of both the autoencoder neural network that performs the encoding-decoding

operations to derive reconstruction error values that serve as a proxy that scales up for non-normativity alongside the two Benford's analysis reexamination that is then subsequently combined by means of the heuristic multiplier approach achieved a performance as detailed in Table 4.4.

Table 4.4. Novel Combined Approach Anomaly Detection Performance Results

Performance Metric	Score
Classification Accuracy	99.96%
Classification Precision	92.72%
Classification Recall	80.46%
Classification F1-Score	86.15%

Source: Author

4.3. Result Interpretation

In summary, the performance results for the three methods are as summarized in Table 4.5. In this case, the accuracy score is of no relevance given the highly imbalanced nature of the dataset. The other metrics are more important in evaluating the performance of the three tested approaches. First, with regard to precision, the precision value can be summarized as a measure of how pure the prediction of positives are. That is, given positive predictions, the percentage of true positives from this set of positive predictions is the value of precision. In terms of precision, the Benford's law digit analysis approach is the best performing model achieving a precision score of 100%, in comparison to the baseline autoencoder's 93.33% and the combined approach's 92.72% precision.

From this result, it can be inferred that all positive predictions flagged by the Benford's digit analysis are actual anomalies, that is, all observations subscribing to a Mean Absolute Deviation (MAD) value to the expected Benford's law distribution score higher than 0.0018 are actual anomalies. In comparison, the baseline autoencoder approach achieved 93.33% implying a false positive (false flag) rate of 6.67%, whilst the novel combined approach achieved a precision score of 92.72%, implying a false positive rate of approximately 7.28%.

Table 4.5. Performance Summary for the Three Methods

Anomaly Detection Method	Sub-Sample Method	Sub-Sample Size	Accuracy	Precision	Recall	F1
Autoencoder	Distribution Preserving Stratified Random Sampling	100,042 entries	99.96%	93.33%	80.46%	86.42%
Benford's Law Digit Analysis	None (Full Data)	533,009 entries	99.99%	100%	70%	82.35%
Proposed Novel Combined Approach	Distribution Preserving Stratified Random Sampling	100,042 entries	99.96%	92.72%	80.46%	86.15%

Source: Author

As discussed previously in justification of the chosen methodologies, higher precision scores are more preferable in a public accounting context rather than in an internal audit implementation. In both contexts, flags raised by the anomaly detection procedure(s) will result in subsequent audit investigations,

however public accounting auditing practices operate on a basis of sampling generalizability in any case such that precision is more valuable than recall. The business model of public accounting firms also highly disincentivizes exhaustive search into any and all anomalies given the time and resource constraints placed upon engagements. As such, the procedure most suitable for a public accounting context will be the one with the highest precision score (provided a reasonably 'good enough' level of recall) that minimizes false flag leads. In which case the Benford's law first digit analysis granular anomaly detection approach is highly suited for. The Benford's law digit analysis approach is also preferable as it is the implementation that requires the least amount of computational resources, can scale more efficiently to larger datasets, as well as providing the easiest implementation to intuit and understand (unlike the 'black box' nature of neural networks).

However, where Benford's first digit analysis approach shines with regard to precision, it is also the worst performing approach in terms of recall. The value of recall measures the percentage of true positives the model predicts with respect to the amount of actual positives. In which case the Benford's digit analysis approach only identifies 70% of the actual anomalies, missing out on 30% of actual anomalies. In comparison, both the autoencoder and the novel combined approach achieved a recall value of 80.46%. Higher recall is more preferable in situations whereby any and all anomalies need to be included within the mix of positive predictions. This statement holds true within the context of internal auditing, unlike public accounting, given the fact that the cost of subsequent audit

procedures are lower than the cost of a systematic inability to prevent misstatements given a failure of the anomaly detection method. In this case, both the baseline autoencoder as well as the novel combined approach are equally more preferable to the Benford's law first digit analysis. However, a caveat should subsequently be disclosed that the novel combined approach achieved the same level of performance at the cost of a higher computational resources for the same level of performance. Therefore, the baseline autoencoder method is the best-performing model in terms of recall.

Finally, with regard to the F1-score, the F1-score is the harmonic mean of the precision and recall values. The F1-score essentially measures how well the model performs in balancing the two performance metrics. That is, in an implementational context whereby both recall and precision are equally preferable, the F1-score should be the score used in determining the best-fit predictive model(s). In terms of F1-score, the baseline autoencoder approach, as introduced in and subsequently replicated from Schreyer, M., et al (2017), is the best performing approach. The baseline autoencoder approach achieved an F1-Score of 86.42%. The novel combined approach is the second best-performing model with an F1-Score of 86.15%, while at the same time the Benford's digit analysis approach is the worst-performing relatively speaking with an F1-Score of 82.35%.

A key consideration for both the baseline autoencoder as well as the novel combined approach is the desired number of observations to be flagged as determined by the auditor. Through empirical inquiries it is found that increasing

the desired number of observations to be flagged can potentially increase the performance in terms of recall. However, at the same time, it is imperative to note that the higher the desired number of observations to be flagged the lower the precision tends to be. As per the research conducted with the data made publicly available in Schreyer, M., et al (2017), this tradeoff mechanism between recall and precision is highly unfavorable with regard to the precision score. Whereby attempts to achieve miniscule gains in recall were met with detrimental compromises in precision. Outside of implementations whereby ground truths are available, configuring the threshold score by proxy of desired number of observations to be flagged to achieve the optimum performance of either precision or recall or even a balance between the two should not be of primary concern. It is advisable that auditors should still determine the amount of desired number of observations to be flagged through either conventional sampling size determination, in conjunction with considerations in terms of realistic subsequent procedures.

Pertaining to the novel combined approach, it is found that the combination between the baseline autoencoder (as replicated from Schreyer, et al. (2017)) and the Benford's law first digit analysis approach through the heuristic multiplier mechanism does not improve performance at all, at the very least as per the data used in this research. Although at the same time, it does not compromise the performance that much (0.61% decrease in precision and a 0.27% decrease in F1-Score in comparison to the baseline autoencoder). In retrospect, this novel combined approach was proposed on the basis of a postulate that observations that

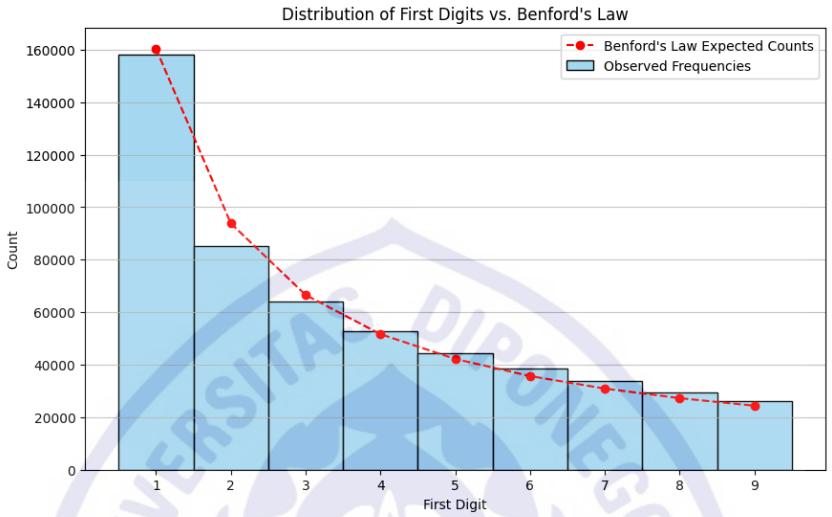
are very close in value to the threshold score (the so-called ‘fence sitters’) should be subjected to further scrutiny through other perspectives, of which the chosen tool for subsequent scrutinization was the Benford’s law first digit analysis implementation as introduced in Nigrini, M. (2022). This assumption is exemplified in Figure 4.3., whereby the values very close to the threshold though ultimately still below the threshold are actual anomalies that also belong to nonconformist groups with regard to Benford’s first digit distribution. The empirical conclusion to be made is that this assumption of supposed behavior was not found within the tested dataset. That data that has not been flagged by the autoencoder tends to be extreme outliers with relatively low reconstruction errors, suggesting an inherent limitation in scope of the autoencoder. Whereby the remaining unidentified anomalies of injected synthetic transactions were very well hidden and very close in nature to normal transactions to a point of being indistinguishable even after combining the analysis of the autoencoder neural network as well as the Benford’s law digit analysis. Although at the same time, the antithesis to the assumption has not been conclusive enough without further investigations in other data.

These findings, namely the performance of each distinct approach, are characterized by a theme of linearity to previous findings. The foundational work of Schreyer, M., et al (2017) investigated the effectiveness of autoencoder neural network in identifying anomalies on the same dataset of accounting journal entries and found that the autoencoder neural network is the best-performing model in comparison to the other benchmark methods. At the same time, although the

implementation in Kurien, K.L., and Chikkamannur, A.A. (2019) differ (implemented in detection of credit card frauds), autoencoder was similarly found to be the best-performing model. The research conducted by Guo, S. (2022) explored upon the use of stacked autoencoder neural networks with a classification layer in classifying VAT compliance and similarly concluded with a finding that testified upon the effectiveness of autoencoder neural networks relative to other tested methods.

On the other hand, notwithstanding the lack thereof literatures in congruence to the particular implementation of Benford's law first digit analysis for the purpose of granular identification of accounting journal entry anomaly detection as demonstrated in this research, this research reiterated the efficacy of the decades old concept introduced in Benford (1938) and Nigrini, M. (1996), alongside its derivative foundational works in Nigrini, M. (2022). With regard to the degree of population-wise conformity for the (mostly) naturally generated data made publicly available in relation to the work of Schreyer, et al. (2017) for the 533,009 unique transactions as illustrated in Figure 4.4. The proportion of transactions subscribing to each leading digit neatly conforms to its expected Benford's law distribution.

Figure 4.4. Population Wise Benford's Law First Digits Conformity



Source: Author's analysis based on data from Schreyer, M., et al (2017)

Concurrently, the concerns expressed priorly in Druică, Oancea, and Vâlsan (2018) with regard to the existence of Benford's law nonconformity in untampered, naturally generated data, that also consequently raised a point of skepticism with regard to the effectiveness of the method in identifying population-wise contaminations, can be antithetically addressed given the finding that this granular implementation whereby Benford's law conformity was tested at sub-group level presented a predictive performance of great precision in identifying anomalous observations. That is, Benford's digit analysis is still an effective method in performing anomaly detection in naturally generated data. Therefore, it can be concluded that the concerns expressed in Druică, Oancea, and Vâlsan (2018) are data specific characteristics that might not be as generalizable in comparison to Benford's law.

Meanwhile, another concern expressed in Debreceny and Gray (2010) that the use of Benford's law for granular identification in cases whereby the population-wise assessment using Benford's digit analysis lead to a wholistic conclusion of nonconformity will inevitably render granular implementations to be unjustifiable, is somewhat reiterated through a completely different rationale based on the findings that population-wise Benford's law conformity provides reasonable justification for the more granular analyses based on the success of granular identification of anomalous transactions in a population that conforms to Benford's first digit distribution.

CHAPTER 5
CONCLUSION

5.1. Conclusion

This research is conducted with the purpose of investigating and demonstrating the use of several analytical procedures in performing journal entry red flag tests within the context of accounting anomaly detection. Three analytical procedures were empirically explored upon encompassing well-established statistical tool within the realm of forensic investigation namely the Newcomb-Benford's law first digit analysis, alongside implementations that leveraged artificial intelligence and machine learning (as per the esoteric, specifically chosen, definition that is adopted to justify the use of the nomenclature) in replication of an established baseline approach using an autoencoder neural network, as well as an entirely novel approach in combination of the two by way of a heuristical multiplier mechanism. Inquiries were conducted in deployment of said approaches on a data of 533,009 journal entries obtained from an anonymized real world entity's journal entry data for a single fiscal period, extracted from its SAP BKPF and BSEG tables data that had been injected with synthetic data. The main goal of each analytical procedure is to correctly identify these synthetic entries, of which their respective performance can be summarized as:

1. The baseline autoencoder neural network approach was found to be the best performing model for being the best-performing model in terms of recall (in maximizing the identification of any and all anomalies) and F1-score (for a more balanced approach). A conclusion can be justifiably reached that the baseline autoencoder neural network approach can also be identified to be the approach that is most suited for in the context of internal auditing;
2. The Newcomb-Benford's law first digit analysis approach was found to be the best performing model in terms of precision (in maximizing the purity of positively flagged observations) while concurrently found to be the worst model in terms of recall. This approach is also the least computationally expensive approach as well as the best model by virtue of its relative quality for being the easiest to intuit and explain. Benford's first digit analysis anomaly detection was also identified to be of greater relevance within the deployment context of public auditing. These findings related to the efficacy of the granular implementation of the Benford's law of digit distribution highlight its conceptual truthfulness that datasets (particularly subsets of the data) demonstrating nonconformist behavior to the distribution can and should be justifiably viewed with a degree of skepticism;
3. By means of experimental studies, an optimal combinatory approach was finalized in the form of an initial autoencoder neural network implementation in reconstruction of the original representation that is then

subsequently reexamined through the lens of relative and absolute conformity to intra-group Benford's law first digit expected distributions using a heuristically defined multiplier mechanism. The novel approach did not improve upon the baseline methods though the decrease in performance was found to be miniscule. At the same time, it was found that the underlying postulate that established the rationale for the approach was not found to be true with regard to the tested data.

5.2. Limitation

In retrospect, several limitations characterize this particular research disregarding scope defined pre-conceived limitations, and highlighting findings on shortcomings unveiled upon throughout the various research processes, it is justifiable to synthesize that:

1. The proposed novel combined approach has not achieved its initial goal in introducing a new method that improves upon baseline methods;
2. The postulate of which the justification for the combined approach was established vis-à-vis the conjecture in terms of the so-called 'fence sitting' behavior was not found in the observed sample.

It can be inferred that these particular shortcomings can be directly traceable to:

a. The compromising nature of the lack of rigor with regard to the chosen reassessment method, namely that only one reassessment method was investigated (i.e. the Benford's law first digit analysis). At the reassessment process using Benford's law digit analysis was

also only conducted for the distribution of first digits, without further investigations on second digit, first-two digits, etc., the Benford's law flagging mechanism was also conducted by means of account-wide flagging, a more granular approach (which was not investigated) for example would be to flag per digit for each account;

b. The specific autoencoder architecture configuration was chosen not for the purpose of maximizing performance, but rather in consideration on the lack of computational resources;
c. Data specific behavior w.r.t. the absence of the previously postulated behavior of 'fence sitters' observations could not be conclusively ascertained as the specific findings.

5.3. Suggestion

In consideration of what have been empirically investigated throughout the research processes, several suggestions can be synthesized namely in:

A. Policy Implications

1. Implementations of artificial intelligence and machine learning in general, as well as neural networks in particular has been and will continue to be demonstrated as a promising paradigm and tools to augment the practice of accountancy especially within the domain of forensic analysis;
2. Audit red flag tests can integrate the use of the semi-supervised autoencoder neural network in amelioration of current limitations.

B. Suggestions for Future Researches

1. In approval and/or disapproval of the postulate that observations that are close in value to the threshold though ultimately still below the threshold can and should be subject to further scrutiny through different reassessment methods needs to be reexamined by way of investigating the use of more extensive as well as in investigation of the more rigorous, complementary analytical procedures. Such as including Benford's digit analyses for subsequent digits;
2. Future studies on the use of artificial intelligence and machine learning within the deployment context of this research can focus on on the existing research gap in terms of subsequent procedures, such as in performing vouching and/or tracing analysis on flagged observations;
3. The current research paradigm with regard to the integration of artificial intelligence and/or machine learning in accounting should focus more on empirically driven technically-implementative researches in lieu of the current trend of literature reviews or perceptual studies;
4. Future researches in exploration of the use of autoencoder neural networks within the context of accounting should strive to maximize performance if necessary computing resources are available.

BIBLIOGRAPHY

Abbas, K. 2025. "Management accounting and artificial intelligence: A comprehensive literature review and recommendations for future research.". Accessed at 18 March 2025, from Elsevier

Adamov, A.Z.,2019, "Machine Learning and Advanced Analytics in Tax Fraud Detection." 2019 IEEE 13th International Conference on Application of Information and Communication Technologies (AICT), IEEE, Baku, Azerbaijan

Ahmed, M., Mahmood, A.N., and Islam, M.R. 2016. "A survey of Anomaly Detection Techniques in Financial Domain." Future Generation Computer Systems Vol. 55, pp. 278-288. Accessed at 24 June 2025, from Elsevier

Aleccio, A., and Petratos, P.,2022, "NLP And IR Applications For Financial Reporting And Non-Financial Disclosure. Framework Implementation And Roadmap For Feasible Integration With The Accounting Process." 2022 6th International Conference on Natural Language Processing and Information Retrieval, pp. 117-124, ACM, New York, United States

Artene, E.A., and Domil, A.E., 2025. "Neural Networks in Accounting: Bridging Financial Forecasting and Decision Support Systems" Electronics Vol. 14, Issue 5, art. 993. Accessed at 18 March 2025, from MDPI

Auad, M., Alves, S., Kakizaki, G., Reis, J. C. S., and Silva, M. M.,2024, "A Filtering and Image Preparation Approach to Enhance OCR for Fiscal Receipts." 2024 37th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 1-6, IEEE, Manaus, Brazil

Autor, D., Levy, F., and Murnane, R. 2000. "Upstairs, Downstairs: Computers and Skills on Two Floors of a Large Bank." Industrial and Labor Review Vol. 55, No. 3, pp. 432-447. Accessed at 07 April 2025, from Cornell University

Balata, P., and Breton, G. 2000. "Narratives vs Numbers in the Annual Report: Are They Giving the Same Message to the Investors?" Review of Accounting and Finance, Vol. 4, Iss. 2, pp. 5-15. Accessed at 08 July 2025, from Emerald Insight

Benford, F.,1938, "The Law of Anomalous Numbers", Proceedings of the American Philosophical Society, Vol. 78, No. 4, pp. 551-572, University of Pennsylvania Press, Schenectady, New York

Bhimani, A., and Willcocks, L. 2014. "Digitization, Big Data and Transformation in Accounting Information." Accounting Business Research Vol. 44, Issue

4., pp. 469-490. Accessed at 09 April 2025, from London School of Economics

Bresnahan, T., Brynjolfsson, E., and Lorin, H. 1999. "Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence." The Quarterly Journal of Economics Vol. 117, No. 1, pp. 339-376. Accessed at 06 April 2025, from Oxford University Press

Chandola, V., Banerjee, A., and Kumar, V. 2009. "Anomaly Detection: A Survey." ACM Computing Survey, Vol. 41, No. 3, Article 15, pp. 1-58. Accessed at 23 June 2025, from ACM

Chen, Y., Zheng, F., Ouyang, J., Wang, Y., Xue, S., and Yu, H. 2024. "Intelligent Entry Method for Mobile Accounting Voucher Based on Image Recognition." Intelligent Computing Technology and Automation Vol. 47, pp. 454-461. Accessed at 07 July 2025, from IOS Press

Coakley, J.R., and Brown, C.E. 2000. "Artificial Neural Networks in Accounting and Finance: Modeling Issues." Intelligent Systems in Accounting, Finance and Management Vol. 9, Issue 2, pp. 119-144. Accessed at 18 March 2025, from John Wiley & Sons, Ltd.

Computer Economics. 2020. "Robotic Process Automation Adoption Trends and Customer Experience 2021", https://avasant.com/report/robotic-process-automation-adoption-trends-and-customer-experience-2021/, Accessed at 01 June 2025

Debreceny, R. S., and Gray, G. L. 2010. "Data mining journal entries for fraud detection: An exploratory study." International Journal of Accounting Information Systems, Vol. 11, Issue 3, pp. 157-181. Accessed at 18 March 2025, from Elsevier

Dlugosz, S., and Müller-Funk, U. 2009. "The Value of the Last Digit: Statistical Fraud Detection

with Digit Analysis." Advances in Data Analysis and Classification Vol. 3, pp. 281-290. Accessed at 08 July 2025, from Springer

Druică, E., Oancea, B., and Vâlsan, C. 2018. "Benford's Law and the Limits of Digit Analysis." International Journal of Accounting Information Systems, Vol. 31, December 2018, pp. 75-82. Accessed at 18 March 2025, from Elsevier

Dubey, S.R., Singh, S.K., Chauduri, B.B. 2021. "Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark." Neurocomputing Volume 503, pp. 92-108. Accessed at 12 May 2025, from Elsevier

Fadilla, A., Army, E., Rustam, Y.D.P., Indrijawati, A., and Pontoh, G.T. 2025.

"Peran Artificial Intelligence dalam Meningkatkan Kualitas Audit: Tinjauan Literatur Sistematis." Jurnal Akuntansi dan Governance Vol. 5, No. 2, 146-165. Accessed at 07 July 2025, from Universitas Muhammadiyah Jakarta

Farber, J.A., and Al Rashdan, A. Y. 2025. "Unsupervised Process Anomaly Detection and Identification Using the Leave-One-Variable-Out Approach †." Sensors, Vol. 25, Issue 7, art. 2098. Accessed at 24 June 2025, from MDPI

Ghafar, I., Perwitasari, W., and Kurnia, R. 2024. "The Role of Artificial Intelligence in Enhancing Global Internal Audit Efficiency: An Analysis." Asian Journal of Logistics Management Vol. 3, No. 2, pp. 64-89. Accessed at 18 March 2025, from Department of Business and Finance, Diponegoro University

Goldstein, M., and Uchida, S. 2016. "A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms." PLOS ONE 11(4), Article e0152173. Accessed at 24 June 2025, from Public Library of Science

Goodfellow, I., Bengio, Y., and Courville, A. 2016. "Deep Learning: An MIT Press Book." Chapter 14-Autoencoders, pp. 499. Accessed at 10 June 2025, from MIT Press

Goodfellow, I., Bengio, Y., and Courville, A. 2016. "Deep Learning: An MIT Press Book." Chapter 6-Deep Feedforward Network, pp. 200. Accessed at 03 June 2025, from MIT Press

Guo, S. 2022. "Intelligent Assessment Method of Enterprise Tax Risk Based on Deep Learning." Wireless Communications and Mobile Computing Vol. 2022, Article ID 5003935. Accessed at 08 July 2025, from Wiley Online Library

GVR. 2025. "GVR Report coverRobotic Process Automation Market Size, Share & Trends Report

Robotic Process Automation Market Size, Share & Trends Analysis Report By Type (Software, Services), By Operations, By End-use (BFSI, Pharma & Healthcare), By Deployment (Cloud, On-premise), By Enterprise Size, By Region, And Segment Forecasts, 2025 - 2030." Market Analysis Report. Accessed at 27 May 2025, from Grand View Research

Hanin, G.F., and Dewayanto, T. 2024. "Peran Machine Learning Dan Deep Learning Dalam Pendeteksian Pencucian Uang – A Systematic Literature Review." Diponegoro Journal of Accounting, vol. 13, no. 3, Jul. 2024, pp. 1-11. Accessed at 07 July 2025, from Universitas Diponegoro

Harris, D. and Harris, S. 2012. "Digital design and computer architecture (2nd ed.)" pp. 129. San Francisco, California: Morgan Kauffman.

Hastie, T., Tibshirani, R., and Friedman, J. 2009. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction." Springer Series in Statistics, 2nd ed. 2009, Corr., 3rd Printing 5th Printing. Accessed at 02 June 2025, from Springer

He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian,2016, "Deep Residual Learning for Image Recognition." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, IEEE, Las Vegas, NV, USA

IDC. 2023. "Untapped Values: What Every Executive Needs to Know About Unstructured Data." Whitepaper. Accessed at 27 May 2025, from International Data Corporation

Jaakkola, T., and Haussler, D. 1998. "Exploiting Generative Models in Discriminative Classifiers." Proceedings of the 12th International Conference on Neural Information Processing Systems, Cambridge, MA: MIT Press, 1998, 487–93. Accessed at 22 April 2025, from MIT Press

Kokina, J., Blanchette, S., Davenport, T., and Pachamanova, D. 2025. "Challenges and Opportunities for Artificial Intelligence in Auditing: Evidence from the Field." International Journal of Accounting Information Systems Vol. 56, art. 100734. Accessed at 09 April 2025, from Elsevier

Koza, J. R., Bennett, F. H., Andre, D., and Keane, M. A. 1996. "Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming." Artificial Intelligence in Design ’96, pp. 151–170. Accessed at 22 April 2025, from Kluwer Academic

Kurien, K.L., and Chikkamannur, A.A.,2019, "Benford’s Law and Deep Learning Autoencoders : An approach for Fraud Detection of Credit card Transactions in Social Media." 2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT-2019), May 17th & 18th 2019 , IEEE, Bangalore, India

L. Theodorakopoulos, A. Theodoropoulou, G. Kampiotis and I. Kalliampakou. 2025. "NeuralACT: Accounting Analytics Using Neural Network for Real-Time Decision Making From Big Data," IEEE Access Vol. 13, pp. 8621-8637. Accessed at 18 March 2025, from IEEE

Laakom, F., Raitoharju, J., Iosidifis, A., and Gabbouj, M. 2024. "Reducing Redundancy in the Bottleneck Representation of Autoencoders." Pattern Recognition Letters Vol. 178, pp. 202-208. Accessed at 10 June 2025, from Elsevier

Lahann, J., Scheid, M., and Fettke, P. 2019. "Utilizing Machine Learning Techniques to Reveal VAT Compliance Violations in Accounting Data.". Accessed at 18 March 2025, from IEEE

Le Clair, C., Cullen, A., and King, M. 2017. "The RPA Market Will Reach $2.9 Billion By 2021: While Large, It's Only A Subset Of The $48.5 Billion Broader AI Cubicle Spend." Trend Report. Accessed at 27 May 2025, from Forrester

Li, P., Pei, Y., and Li, J. 2023. "A Comprehensive Survey on Design and Application of Autoencoders in Deep Learning." Applied Soft Computing Vol. 138, 110716. Accessed at 10 June 2025, from Elsevier

Lind, D.A., Marchal, W.G., and Wathen, S.A. 2019. "Basic Statistics for Business and Economics, Ninth Edition". New York: McGraw-Hill Education.

Liu, Z. 2024. "Accounting-Oriented Research on Note Recognition Model based on Information Extraction Algorithm." WSEAS Transactions on Business and Economics Vol. 21, pp. 2640-2652. Accessed at 18 March 2025, from WSEAS

McCarthy, J., Minsky, M., Rochester, N., and Shannon, C.E. 2006. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955." AI Magazine, 27(4), 12. Accessed at 22 April 2025, from Association for the Advancement of Artificial Intelligence

Munoko, I., Brown-Liburd, H.L., and Vasarhelyi, M. 2020. "The Ethical Implications of Using Artificial Intelligence in Auditing." Journal of Business Ethics Vol. 167, pp. 209-234. Accessed at 08 July 2025, from Springer

Ng, Andrew. 2017. "Why Non-linear Activation Functions." Deeplearning Specialization, Course 1 Neural Networks and Deep Learning, Module 3 Shallow Neural Networks, Lecture 07, https://www.youtube.com/watch?v=NkOv_k7r6no&list=PLkDaE6sCZn6Ec-XTbcX1uRg2_u4xOEky0&index=31, Accessed at 22 May 2025

Ng, Andrew. 2020. "Forward and Backward Propagation." Deeplearning Specialization, Course 1 Neural Networks and Deep Learning, Module 4 Deep Neural Networks, Lecture 06, https://www.youtube.com/watch?v=-Lavz_I4l2U, Accessed at 02 June 2025

Ng, Andrew. 2020. "Gradient Descent." Deeplearning Specialization, Course 1 Neural Networks and Deep Learning, Module 2 Neural Network Basics, Lecture 04, https://www.youtube.com/watch?v=6dNWZZmA4fE, Accessed at 03 June 2025

Nigrini, M. 1996. "A Taxpayer Compliance Application of Benford's Law." The Journal of the American Taxation Association, 18, pp. 72-91. Accessed at 03 September 2025, from American Accounting Association

Nigrini, M. 2000. "Digital Analysis Using Benford's Law: Tests and Statistics for Auditors." EDPACS, Vol. 28 Issue 9, pp. 1–2. Accessed at 16 September 2025, from Taylor & Francis

Nigrini, M. 2011. "Forensic Analytics: Methods and Techniques for Forensic Accounting Investigations." . Hoboken, New Jersey: John Wiley & Sons, Inc..

Nigrini, M. 2017. "Audit Sampling Using Benford's Law: A Review of the Literature with Some New Perspectives." Journal of Emerging Technologies in Accounting, Vol. 14, Issue 2, pp. 29–46. Accessed at 16 September 2025, from American Accounting Association

Nigrini, M. 2022. "Using Benford's Law to Reveal Journal Entry Irregularities: Benford's Law can help uncover indicators of fraud - and anomalies that arise from legitimate business practices." Journal of Accountancy ,Vol. 234, Issue 3. Accessed at 03 September 2025, from American Institute of CPA's

Ning, X. 2022. "Neural Network Technology-Based Optimization Framework of Financial and Management Accounting Model." Computational Intelligence and Neuroscience Vol. 2022, Issue 1. Accessed at 18 March 2025, from John Wiley & Sons, Ltd.

Nwaimo, C.S., Adegbola, A.E., and Adegbola, M.D. 2024. "Predictive Analytics for Financial Inclusion: Using Machine Learning to Improve Credit Access for Underbanked Populations." Computer Science & IT Research Journal Vol. 5, Issue 6, pp. 1358-1373. Accessed at 18 March 2025, from Fair East Publishers

Petroc, Taylor. 2024. "Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2023, with forecasts from 2024 to 2028.", https://www.statista.com/statistics/871513/worldwide-data-created/, Accessed at 27 May 2025

Purba, K.A., Dewayanto, T. 2023. "Penerapan Artificial Intelligence, Machine Learning dan Deep Learning pada Kurikulum Akuntansi: A Systematic Literature Review." Diponegoro Journal of Accounting, Volume 12, No. 3, pp. 1-15. Accessed at 18 March 2025, from Universitas Diponegoro

Puri, M., Solanki, A., Padawer, T., Tipparaju, S.M., Moreno, W.A., and Pathak, Y. 2016. "Chapter 1 - Introduction to Artificial Neural Network (ANN) as a

Predictive Tool for Drug Design, Discovery, Delivery, and Disposition: Basic Concepts and Modeling." Artificial Neural Network for Drug Design, Delivery and Disposition 2016, pp. 3-13. Accessed at 10 May 2025, from Academic Press

Ran He, Jie Cao, and Tieniu Tan. 2025. "Generative Artificial Intelligence: a Historical Perspective." National Science Review, Volume 12, Issue 5, nwaf050 . Accessed at 22 April 2025, from Oxford University Press

Richardson, S.A., Wysocki, P.D., and Tuna, A.I. 2010. "Accounting Anomalies and Fundamental Analysis: A Review of Recent Research Advances." Journal of Accounting and Economics Volume 50, Issues 2–3, pp. 410-454. Accessed at 09 July 2025, from Elsevier

Rosenblatt, F. 1958. "The perceptron: A probabilistic model for information storage and organization in the brain." Psychological Review Vol. 62, No. 6, pp. 386–408. Accessed at 10 May 2025, from American Psychological Association

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. 1986. "Learning Internal Representations by Error Propagation." Tech. rep. ICS 8504. Accessed at 13 May 2025, from Institute for Cognitive Science, University of California

Samuel, A.L. 1959. "Some Studies in Machine Learning Using the Game of Checkers." IBM Journal of Research and Development, vol. 3, 1959, pp. 210-229. Accessed at 22 April 2025, from IBM Corporation

Schreyer, M., Sattarov, T., and Borth, D.,2024, "Federated and Privacy-Preserving Learning of Accounting Data in Financial Statement Audits," Proceedings of the 3rd ACM International Conference on AI in Finance, ICAIF 2022, pp. 105–113, Association for Computing Machinery, Inc.,

Schreyer, M., Sattarov, T., Borth, D., Dengel, A., and Reimer, B. 2017. "Detection of Anomalies in Large Scale Accounting Data using Deep Autoencoder Networks". Accessed at 17 March 2025, from ArXiv

Sekaran, U. and Bougie, R.. 2016. "Research Methods for Business: A Skill Building Approach Seventh Edition". West Sussex, United Kingdom: John Wiley & Sons, Ltd.

Septiyanti, R.D, Wany, E.W., and Prayitno, B. 2025. "Peran Akuntan, Integrasi Inovasi Artificial Intelligence (AI) Dan Internet of Things (IoT) Dalam Menghadapi Digitalisasi Ekonomi Menjelang Era Society 5.0." INCOME: Jurnal Akuntansi dan Keuangan Vol. 6, No. 1. Accessed at 07 July 2025, from Universitas Wijaya Putra

Serrano-Cinca, C., Gutiérrez-Nieto, B., and Bernate-Valbuena, M. 2018. "The Use of Accounting Anomalies Indicators to Predict Business Failure." European Management Journal Vol. 37, Issue 3, June 2019, pp. 353-375. Accessed at 09 July 2025, from Elsevier

Silaen, R.P., Dewayanto, T. 2024. "Penggunaan Berbagai Artificial Intelligence pada Proses Audit: A Systematic Literature Review." Diponegoro Journal of Accounting, Volume 13, No. 2, pp. 1-15. Accessed at 18 March 2025, from Universitas Diponegoro

Tarissa, B.V., and Dewayanto, T. 2024. "Penerapan Machine Learning dan Deep Learning Pada Peningkatan Deteksi Credit Card Fraud - A Systematic Literature Review." Diponegoro Journal of Accounting, vol. 13, No. 3, Jul. 2024, pp. 1-15. Accessed at 07 July 2025, from Universitas Diponegoro

Villars, F.L., Olofson, C.W., and Eastwood, M. 2011. "Big Data: What It Is and Why You Should Care." Whitepaper. Accessed at 27 May 2025, from IDC

Wang, R. Z. 2025. "Standardizing XBRL Tags with Natural Language Processing." Journal of Computer Information Systems, pp. 1–15. Accessed at 07 July 2025, from Taylor & Francis

Warner, B., and Misra, M. 1996. "Understanding Neural Networks as Statistical Tools." The American Statistician Vol. 50, No. 4, pp. 284-293. Accessed at 13 May 2025, from American Statistical Association

Xie, Z., Huang, X. 2024. "A Credit Card Fraud Detection Method Based on Mahalanobis Distance Hybrid Sampling and Random Forest Algorithm," in IEEE Access, vol. 12, pp. 162788-162798. Accessed at 18 March 2025, from IEEE

Yin, R. 2009. "Case Study Research: Design and Methods, 4th ed". Thousand Oaks, California, United States: Sage Publications.

Zimek, A., Schubert, E., and Kriegel, H.P. 2012. "A Survey on Unsupervised Outlier Detection in High-Dimensional Numerical Data." Statistical Analysis and Data Mining, 5(5), pp. 367-387. Accessed at 19 June 2025, from Wiley Online Library