Educational Administration: Theory and Practice
2024, 30(5), 5751-5770
ISSN: 2148-2403
https://kuey.net/

Research Article

Anomaly Detection In Telecommunication Networks: Leveraging Novel Big Data And Machine Learning Techniques For Proactive Fault Management

Dr. Shaker AbdulAziz Ali*

*Faculty of Computer Science

Citation: Dr. Shaker AbdulAziz Ali (2024), Anomaly Detection In Telecommunication Networks: Leveraging Novel Big Data And Machine Learning Techniques For Proactive Fault Management, *Educational Administration: Theory and Practice*, *30*(5), 5751-5770, Doi: 10.53555/kuey.v30i5.3849

1. Introduction

The telecommunication industry has experienced exponential growth, leading to complex and intricate networks designed to support the ever-increasing demands of a hyper-connected world. The emergence of technologies, such as the Internet of Things (IoT), 5G networks, and cloud computing, has made the performance and reliability of these networks more critical than ever. Li et al. (3351) posit that these cutting-edge innovations result in more complicated operations involving cellular networks, mobile nodes, and devices. Additionally, they generate a vast range of network measurements and parameters (volume) from numerous sources within and outside the network (variety), with rapid input and output (velocity) and potential data quality issues (veracity). These 4-V characteristics in big data cellular networks present both opportunities and challenges for network monitoring and management. As a result, anomaly detection in telecommunication networks has emerged as a crucial research area, as it enables early identification of potential faults and promotes optimal network performance. Karatepe and Zeydan (1) argue that a primary challenge for mobile service providers is identifying defects in user activity among millions of transactions from many users. Detecting a small percentage of anomalies in such vast datasets can prove difficult. A promising approach to addressing these network management challenges is the application of knowledge-based anomaly detection methods. This study aims to investigate the intersection of big data analytics and machine learning techniques to achieve proactive fault management in telecommunication networks.

Telecommunication networks are inherently susceptible to various faults, ranging from hardware failures to software glitches, configuration errors, and even malicious activities. Timely identification and remediation of such anomalies are essential in maintaining network performance, minimizing service disruptions, and reducing the negative impacts on user experience (Mdini 5). Li et al. (3352) note that traditional approaches to fault management rely on reactive methods, such as threshold-based alarms and manual inspection, which often fall short of addressing the challenges posed by modern, large-scale telecommunication networks. Consequently, there is a growing need for more sophisticated, proactive techniques to detect and mitigate network anomalies in real time. In recent years, big data and machine learning have emerged as promising avenues for addressing the complexities of anomaly detection in telecommunication networks. The deluge of data generated by these networks offers valuable insights into their operational dynamics, which can be harnessed through advanced analytical methods to facilitate proactive fault management (Nguyen et al. 19698). Machine learning provides a potent suite of algorithms for extracting meaningful patterns from large-scale, multi-dimensional data, enabling the identification of subtle anomalies that might evade conventional detection techniques.

The exigency of this research arises from the inexorable evolution of telecommunication networks, which are progressively becoming more convoluted and intertwined. The concomitant surge in network traffic and the escalating reliance on them for many critical applications engender a heightened susceptibility to faults and disruptions (Zeydan 1). Consequently, the ability to swiftly detect and mitigate such aberrations is integral for maintaining the integrity and robustness of the telecommunication infrastructure. Moreover, conventional fault management paradigms, which are predominantly reactive in nature, are inherently inadequate in addressing the multifarious challenges posed by contemporary telecommunication networks. The reactive approach to fault management is predicated on rectifying network anomalies only after they have manifested, often culminating in detrimental ramifications for service quality, network security, and customer satisfaction (Nguyen et al. 19697). The espousal of proactive fault management, in contradistinction, provides an avant-garde framework that empowers network operators to anticipate and avert potential faults through the early detection of glitches.

Copyright © 2024 by Author/s and Licensed by Kuey. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Furthermore, the promising confluence of big data and machine learning offers a compelling impetus to explore innovative techniques to address the necessities of proactive fault management. The voluminous and heterogeneous data generated by telecommunication networks is replete with latent information that telecommunication companies can harness to identify incipient fault signatures and discern aberrant patterns (Đorđević 73). With their inherent ability to adapt and learn from data, machine learning algorithms provide an effective means of deciphering the hidden and intricate relationships embedded within network data. Abbasi et al. (24) posit that amalgamating diverse machine learning approaches, encompassing supervised, unsupervised, and semi-supervised paradigms, presents a fecund ground for discovering novel and productive solutions. Besides, the choice of Python as the programming language for model development is contingent on its ubiquity, versatility, and the abundance of domain-specific libraries and frameworks catering to big data processing and machine learning. A Python-based model allows for simplistic experimentation with an array of algorithms and techniques while ensuring the reproducibility and scalability of the research outcomes. Overall, the rationale for this research lies on the pressing need for more effective and proactive approaches to fault management in telecommunication networks. It is also contingent on the transformative potential of big data and machine learning and the adoption of a versatile and widely supported programming language for model development.

1.1 The Research Methodology

This study adopts a rigorous and systematic research methodology to investigate the efficacy of big data and machine learning in anomaly detection and proactive fault management within telecommunication networks. The following sections outline the critical components of the methodological approach:

1.1.1 Data Collection and Preprocessing

The initial phase entails procuring and curating a copious and diverse dataset. The collected data will encompass a broad spectrum of telecommunication network data, such as call detail records, network logs, and performance metrics.

1.1.2 Feature Selection and Dimensionality Reduction

To optimize the efficiency and performance of the machine learning models, we will employ feature selection techniques. Furthermore, dimensionality reduction methods, including.

1.1.3 Machine Learning Model Development and Evaluation

A variety of machine learning algorithms, including supervised, unsupervised, and semi-supervised methods, as well as deep learning and reinforcement learning techniques, will be explored to identify the most suitable approach for anomaly detection in telecommunication networks.

1.1.4 Comparative Analysis and Model Selection

The performance of the various machine learning models will be compared and contrasted to determine the most effective approach for anomaly detection in telecommunication networks. This comparative analysis will consider not only the accuracy and reliability of the models but also their computational efficiency, scalability, and adaptability to evolving network dynamics.

This comparative analysis will serve to elucidate the relative strengths and weaknesses of the proposed model, thereby informing future research and development efforts in the realm of proactive fault management.

1.1.5 Interpretation and Implications

The findings and insights gleaned from the research will be meticulously analyzed and interpreted, drawing upon the relevant literature and contextualizing the results within the broader landscape of telecommunication network fault management. The implications of the study for network operators, service providers, and the academic community will be delineated, highlighting the potential contributions of the proposed model to enhancing network performance, security, and reliability, as well as fostering the ongoing evolution of proactive fault management strategies and techniques.

1.2 The Research Model

The ideal model is an ensemble model that combines the strengths of various algorithms and approaches, harnessing the power of automation for an efficient and accurate solution. This model will consist of multiple layers that address different aspects of the anomaly detection process, including data preprocessing, feature selection, model training, and evaluation.

2. Literature Review:

Telecommunication networks form the backbone of modern digital societies since they facilitate seamless global communication and data exchange. Ensuring the reliability and uninterrupted functioning of these networks is paramount, as network failures can have far-reaching consequences, including economic losses,

public dissatisfaction, and disruptions to critical services (Ahmed et al. 19). Consequently, proactive fault management and anomaly detection in telecommunication networks have garnered significant attention from both academia and industry. Himeur et al. (15) note that traditional fault management techniques, such as threshold-based, statistical, and rule-based methods, have been widely employed to detect and resolve network anomalies. However, the rapidly evolving nature of telecommunication networks, characterized by increased complexity, heterogeneity, and scale, has rendered these conventional techniques insufficient for detecting network anomalies effectively (Himeur et al. 14). Li et al. (3352) note that traditional approaches to fault management rely on reactive methods, such as threshold-based alarms and manual inspection, which often fall short of addressing the challenges posed by modern, large-scale telecommunication networks. Furthermore, the proliferation of big data generated by telecommunication networks presents both opportunities and challenges for anomaly detection. As a result, the implementation of novel, sophisticated approaches that leverage big data and advanced machine learning (ML) techniques have become indispensable in addressing the intricacies of contemporary telecommunication networks to ensure their resilience, security, and optimal performance in the face of unprecedented challenges and demands.

This literature review aims to provide a comprehensive overview of the evolving field of anomaly detection in telecommunication networks. The overriding focus is on the emergence of big data and ML techniques as potential solutions for tackling the challenges associated with traditional fault management methods. The review begins by examining the background and significance of anomaly detection in telecommunication networks, then by exploring these networks' key aspects, including their structure, the importance of reliability, and the challenges associated with conventional fault management techniques. Subsequently, it delves into the potential of big data and ML approaches in enhancing anomaly detection capabilities, discussing their role, benefits, and future prospects within the telecommunication domain. By synthesizing the current state of research and identifying knowledge gaps, the review contributes to a better understanding of the advances made in anomaly detection and guide future research efforts in developing more robust and reliable telecommunication networks.

2.1 Anomaly Detection in Telecommunication Networks

As the cornerstone of contemporary digital societies, telecommunication networks require reliable and robust anomaly detection mechanisms to manage their inherent susceptibilities and multifaceted challenges. Telecommunication networks underpin the seamless transmission of information and support the digital economy. Chaparro and Eberle (410) posit that with the exponential growth in network traffic and the increasing reliance on these networks for essential services, maintaining network reliability and minimizing service disruptions is critical. Besides being intrinsically vulnerable to various malfunctions, telecommunication networks are intrinsically disposed to issues spanning hardware failures, software aberrations, configuration discrepancies, and malevolent actions. The advent of advanced technologies, such as 5G and the Internet of Things (IoT), has also amplified the complexity and scale of telecommunication networks, which further underscores the need for sophisticated anomaly detection techniques. Prompt detection and rectification of these anomalies are indispensable in preserving network efficacy, mitigating service discontinuities, and attenuating the detrimental repercussions on user experience (Mdini 5). Consequently, Chaparro and Eberle (410) note that anomaly detection in telecommunication networks, which identifies deviations from normal behavior and potentially indicates the presence of faults, has gained substantial prominence in recent years. Effective anomaly detection plays a pivotal role in proactive fault management because it enables timely detection and resolution of network issues (Mdini 5). It prevents catastrophic failures and ensures high network performance and end-user satisfaction.

2.2 Overview of Telecommunication Networks

As vital conduits for information exchange, telecommunications networks rely on complex architectures and interconnected components to support diverse services. These networks can be conceptualized as intricate systems responsible for transmitting and exchanging information over considerable distances. They consist of numerous interconnected components, including switches, routers, base stations, and transmission media such as optical fibers, coaxial cables, and radio frequency links (Oest et al. 3). These interconnected components work in unison to ensure seamless data transmission, with switches and routers effectively directing traffic through the network, while base stations facilitate wireless connectivity. Transmission media provide the necessary physical pathways for data propagation, enhancing the efficiency and reach of telecommunication networks.

Telecommunication networks have undergone transformative advancements to meet ever-growing demands for enhanced capacity, coverage, and functionality. The shift from circuit-switched networks to packet-switched networks, epitomized by the widespread adoption of the Internet Protocol (IP), has notably augmented efficiency and versatility in network resource utilization (Van Heddeghem et al. 1). Moreover, Kim et al. (4) note that the emergence and implementation of cutting-edge wireless communication standards, such as 4G and 5G, have profoundly impacted mobile broadband services, facilitating rapid data transfer, reduced latency communication, and accommodating a diverse range of applications, including the IoT and autonomous systems. These technological advancements have expanded the potential use-cases for telecommunication networks and laid the groundwork for future innovations in the field. The relentless

progression of telecommunication networks will continue to propel the development of novel applications and services. It will bolster the capacity and versatility of communication infrastructure on a global scale.
Ensuring the reliability and performance of telecommunication networks is crucial, given their pervasive role in supporting modern societies' digital ecosystems. The increasing complexity and scale of these networks present network operators with mounting challenges, including capacity planning, traffic engineering, security, and fault management. In response to these evolving demands, researchers and industry practitioners have increasingly focused on devising innovative solutions that incorporate big data analytics and ML methodologies to tackle emerging obstacles and optimize telecommunication network operations.

2.3 Importance of Network Reliability and Fault Management

The significance of network reliability in telecommunication networks is undeniable. It profoundly influences users' quality of service and overall network performance. The meteoric rise of digital services, comprising e-commerce, e-health, and remote work, has rendered the continuous operation of telecommunication networks essential for individuals and organizations. In this context, Chaparro and Eberle (410) observe that dependable networks guarantee user satisfaction and bolster network operators' competitive advantage by curbing churn rates and cultivating customer loyalty. As a result, Baştuğ et al. (549) claim that network operators increasingly prioritize fault management – the detection, diagnosis, and resolution of network anomalies – as a critical component of network operations and maintenance. The authors also note that in the realm of wireless communication, big data brings numerous new information sets to network planning that can be interconnected to attain a better understanding of users and networks, such as location, user velocity, and social geodata. Efficacious fault management empowers network operators to minimize service disruptions, optimize resource utilization, and curtail operational costs (Keshavamurthy and Ashraf 1). Hence, prioritizing fault management in telecommunication networks is imperative for maintaining reliability, ensuring user satisfaction, and fostering a robust digital ecosystem for future growth.
The adoption of advanced techniques in fault management is crucial for managing the complex challenges in maintaining telecommunication networks' reliability and performance. Fault management, which encompasses anomaly detection, root cause analysis, and remediation, is pivotal in maintaining network performance and reliability (Yu et al. (349). Mdini (5) avers that anomaly detection, in particular, seeks to discern deviations from expected network behavior, which potentially signals the existence of faults or other issues impacting network performance. By expeditiously detecting network anomalies, fault management systems can initiate corrective measures and avert the exacerbation of issues that can culminate in network outages or service degradation. As telecommunication networks persistently grow in scale and complexity, Li et al. (3352) observe that harnessing sophisticated techniques like big data analytics and ML has become increasingly imperative to amplify the efficacy of fault management systems and safeguard the dependability of these essential infrastructures. Thus, integrating advanced techniques in fault management is crucial since it augments the resilience of telecommunication networks in the face of evolving challenges and maintains their indispensable role in supporting the digital ecosystem.

2.4 Challenges in Traditional Fault Management Techniques

Traditional fault management techniques, such as threshold-based, statistical, and rule-based methods, have been extensively employed to address network anomalies in telecommunication networks. However, Himeur et al. (15) observe that the ever-evolving nature of these networks, characterized by increased complexity, heterogeneity, and scale, has exposed several limitations, rendering them less effective for detecting network anomalies. For instance, threshold-based methods that rely on predetermined thresholds to identify anomalies suffer from low adaptability to dynamic network conditions and are prone to high false alarm rates, especially in the presence of non-stationary network behavior (Rastogi and Singh 952). They struggle to adapt to the ever-changing landscape of network conditions, leading to decreased effectiveness. Similarly, Ahmed et al. (24) postulate that statistical techniques, which employ various models to capture normal network behavior, struggle to accommodate the diverse and rapidly changing network patterns emerging in modern telecommunication networks. These approaches falter in efficiently adapting to and capturing the nuances of rapidly evolving network conditions, thereby limiting their effectiveness in anomaly detection. Considering these limitations, it is evident that traditional techniques are insufficient for modern telecommunication networks, which highlights the need for more advanced and adaptive anomaly detection methods.
As telecommunication networks evolve, traditional fault management methods face growing limitations. These constraints call for the exploration of advanced techniques for effective anomaly detection. Rule-based methods, which rely on expert-defined rules to detect network anomalies, are hindered by the need for extensive domain knowledge and the manual effort required for rule maintenance and update. These traditional techniques often fail to scale effectively with the increasing size and complexity of telecommunication networks, resulting in reduced accuracy and increased false alarms (Rastogi and Singh 952).

2.5 Emergence of Big Data and ML in Anomaly Detection

The emergence of big data and ML has engendered a paradigm shift in anomaly detection techniques. Accordingly, they provide more sophisticated and adaptive approaches to identifying and diagnosing network faults. The vast volume, velocity, and variety of data generated by modern telecommunication networks offer a wealth of insights that can be harnessed through big data analytics to facilitate improved decision-making in network management and fault detection (Abbasi, Amin Shahraki, and Taherkordi 20). Big data analytics can discern subtle correlations and hidden patterns within the massive datasets generated by telecommunication networks. It can enable more accurate and timely identification of network anomalies and foster more informed decision-making in network management and fault detection. ML techniques, such as supervised, unsupervised, and semi-supervised learning, have been leveraged to automatically learn and adapt to network behavior patterns, enhancing the ability to detect anomalies in a timely and accurate manner (Abbasi et al. 24). Incorporating these advanced techniques help network operators overcome the limitations of traditional fault management methods, such as reduced adaptability, scalability, and the reliance on expert-defined rules, ultimately leading to more robust and reliable telecommunication networks.

2.6 Role of Big Data in Telecommunication Networks

The role of big data in telecommunication networks has become increasingly critical as these networks generate and process massive amounts of data from diverse sources. Consequently, they necessitate the development of novel techniques for data management, analysis, and utilization. Thudumu et al. (1) posit that big data analytics, which encompasses a broad spectrum of advanced data processing techniques, enables network operators to extract valuable insights from vast data to facilitate informed decision-making in various aspects of network operations. The insights can help in performance optimization, resource allocation, customer experience management, and anomaly detection. By exploiting big data analytics, telecommunication providers can uncover patterns, trends, and correlations that may have been previously undetected (Elgendy and Elragal 214). This capability improves network performance, enhances user satisfaction, and increases competitiveness in the rapidly evolving telecommunication landscape.

2.7 Potential of ML for Improving Anomaly Detection

ML holds immense potential for improving anomaly detection in telecommunication networks. It offers powerful and adaptable techniques to model and analyze complex network behavior. ML algorithms, such as supervised, unsupervised, and semi-supervised learning methods, can be employed to automatically learn from vast amounts of network data, enabling the identification of intricate and evolving patterns that may signify network anomalies (Abbasi et al. 24). Notably, Dong and Wang (581) postulate that deep learning techniques, a subset of ML that leverages artificial neural networks, have demonstrated remarkable success in detecting previously unknown anomalies by capturing high-level abstractions in network data. Additionally, ML-based anomaly detection methods offer several advantages over traditional techniques, including enhanced adaptability to dynamic network conditions, scalability to handle big data, and reduced reliance on expert-defined rules and thresholds (Dong and Wang 581). Therefore, network operators can significantly improve the effectiveness and efficiency of fault management systems by capitalizing on the capabilities of ML. This premise implies that the technology can significantly contribute to more robust and resilient telecommunication networks. Incorporating ML in anomaly detection sets the stage for a new frontier in telecommunication fault management. It empowers network operators to effectively tackle emerging challenges and ensure their networks' sustained performance and reliability.

2.7 Anomaly Detection Techniques in Telecommunication Networks

Anomaly detection techniques in telecommunication networks encompass various methods designed to identify deviations from normal network behavior to facilitate timely and effective fault management. They broadly fall into three main groups: statistical methods, knowledge-based methods, and ML-based methods (Himeur et al. 15). Statistical methods, which include parametric and non-parametric approaches, rely on the analysis of historical network data to establish baseline behavior patterns and detect anomalies based on deviations from these patterns (Mason et al. 1).

2.8 Traditional Methods for Anomaly Detection

Threshold-based methods are among the earliest and most straightforward approaches to anomaly detection in telecommunication networks. They rely on establishing predetermined limits or thresholds for various network performance metrics. Wang (2) note that when monitored metrics exceed or fall below these thresholds, an anomaly is flagged, prompting further investigation or remediation efforts. The author also observes that these traditional methods are either complex in tuning or require prior knowledge or human interference. For example, neural network is one of the over-parameterized models; the scale of parameter is often much larger than that of training data. Therefore, validation is needed to determine whether the model learned from data, or simply remembered its training set. However, these methods are prone to high false alarm rates, particularly in non-stationary network behavior, which exacerbates the challenges associated with maintaining network reliability and performance (Wang 2).

Statistical methods, another category of traditional anomaly detection techniques, leverage historical network data to model normal network behavior and detect deviations from this baseline. Fan et al. (1124) note that these approaches make statistical assumptions on the underlying data distribution, such as Gaussian normal distribution, based on which scores are calculated for anomaly detection. A notable example is the generalized extreme studentized deviate (GESD)-based method. Previous studies have demonstrated the usefulness of GESD-based methods in identifying anomalies in building energy consumption profiles (Fan et al. 1124). Besides, as telecommunication networks evolve in complexity and scale, the efficacy of statistical techniques can be hindered by the emerging diverse and rapidly changing network patterns. This shortcoming necessitates the exploration of more advanced and robust anomaly detection methodologies.

Rule-based or knowledge-based methods involve using expert-defined rules, heuristics, or patterns to identify anomalies in network behavior. According to Asghar et al. (1682), these methods often necessitate the involvement of domain experts to develop and maintain the rule set, which can be time-consuming and labor-intensive. The authors also note that rule-based methods can struggle to adapt to evolving network conditions or identify previously unseen anomalies, as the scope of the predefined rules inherently limits them. Despite these limitations, these methods have been widely used in telecommunication networks due to their interpretability and ability to incorporate domain-specific knowledge (Asghar et al. 1689). However, the need for more adaptive and scalable anomaly detection techniques becomes apparent as telecommunication networks become more complex and dynamic. This evolution renders rule-based methods less effective in contemporary network environments. Therefore, while rule-based methods offer valuable insights and interpretability, their reliance on expert-defined rules and difficulty adapting to changing network conditions necessitate exploring alternative, more advanced techniques for anomaly detection.

2.9 Big Data Approaches for Telecommunication Networks

The advent of big data has revolutionized the telecommunication industry, presenting new opportunities for enhancing network performance, reliability, and anomaly detection. Big data approaches leverage large-scale, diverse, and high-velocity data generated by telecommunication networks, such as call detail records, network traffic data, and user behavior data, to derive insights and make data-driven decisions (Abbasi et al. 24). Thudumu et al. (1) claims that techniques such as data mining, distributed processing, and advanced analytics are employed to process and analyze this vast amount of data, which enables network operators to uncover hidden patterns, trends, and correlations that can be used for various applications, including capacity planning, demand forecasting, and intelligent fault management. As telecommunication networks continue to expand and evolve, big data approaches will continue to play an indispensable role in addressing the limitations of traditional anomaly detection methods (Habeed et al. 289). This application will pave way for more efficient, scalable, and adaptive fault management solutions. Simply put, integrating big data analytics into the telecommunication domain signifies a transformative shift in network management. It will foster more robust and resilient systems capable of meeting the challenges of an ever-changing digital landscape.

2.10 Characteristics of Big Data in Telecommunication Networks

The sheer volume of data in telecommunication networks presents formidable challenges and opportunities for anomaly detection and network management. Mardani (1) note that volume is a defining characteristic of big data in telecommunication networks, as these networks generate an enormous amount of data daily due to the exponential growth of connected devices and users and the increasing demand for high-bandwidth services. The author claims that this vast quantity of data, measured in petabytes or even exabytes, poses significant challenges in terms of storage, processing, and analysis. According to Lu et al. (9), this problem necessitates the adoption of distributed computing frameworks, such as Hadoop and Spark, to manage and process the data efficiently. Additionally, the proper handling of this data deluge requires the development of advanced analytics algorithms and data mining techniques that can effectively extract valuable insights from the massive datasets (Thudumu et al. 1). As telecommunication networks continue to expand, the volume of data they generate is expected to grow further, highlighting the importance of scalable big data solutions in addressing the ever-increasing data management demands (Mardani 1). Thus, the enormous volume of data in modern telecommunication networks obliges adopting robust and scalable big data technologies to ensure efficient, reliable, and adaptable network management and anomaly detection.

The multifaceted nature of data in telecommunication networks underscores the need for versatile analytics techniques and strategies. Variety is another key characteristic of big data in telecommunication networks, as the data generated encompasses a diverse range of formats, sources, and types, including structured, semi-structured, and unstructured data (Zaslavsky 2). Examples of data sources include call detail records, network traffic logs, user behavior data, and social media data, each presenting its own set of challenges in terms of data integration, preprocessing, and analysis (Zhao et al. 1). To effectively leverage this diverse data, Abbasi et al. (20) note that advanced analytics techniques, such as data mining and ML, are employed to uncover meaningful insights and correlations that can be used to inform network management decisions and improve anomaly detection. Furthermore, the ability to handle data variety necessitates the development of robust data integration methods, data quality assessment strategies, and domain-specific knowledge representation techniques (Zaslavsky 4). Addressing the variety aspect of big data in telecommunication

networks is crucial for capitalizing on the wealth of information available and fostering the development of more effective, adaptable, and comprehensive network management and anomaly detection solutions.

In an era of rapidly evolving telecommunication networks, managing the velocity aspect of big data is crucial for effective network management. Velocity refers to the high rate at which data is generated, transmitted, and processed, necessitating real-time or near-real-time analytics solutions to ensure timely and accurate decision-making (Zaslavsky 2). To address this challenge, big data solutions often incorporate stream processing technologies, such as Apache Kafka and Apache Flink, which enable the processing and analysis of high-velocity data streams in real-time or near-real-time (Salloum et al. 157). This premise implies that embracing the velocity aspect of big data allows network operators to be more responsive and agile in managing their networks. It enhances network resilience, optimized resource utilization, and improved customer satisfaction through proactive fault management and service assurance.

Addressing veracity is essential for maximizing the value of data-driven insights and enhancing the efficacy of anomaly detection methods. Veracity pertains to the quality, accuracy, and trustworthiness of the data, which can significantly impact the effectiveness of data-driven decision-making processes and anomaly detection techniques (Zaslavsky 2). Telecommunication networks often produce noisy, incomplete, or inconsistent data, which can undermine the reliability of the insights derived from the data (Yen et al. 199). Consequently, Gudivada et al. (3) aver that data preprocessing and cleansing techniques, such as data imputation, outlier detection, and data normalization, play a crucial role in ensuring the veracity of the data and enhancing the overall effectiveness of big data approaches in telecommunication networks. Network operators can foster more robust and reliable decision-making processes by addressing the veracity aspect of big data. This approach leads to improved network management, enhanced anomaly detection, and a better understanding of the underlying patterns and dynamics governing modern telecommunication networks.

2.11 Applications of Big Data in Telecommunication Networks

Network optimization is a critical application of big data in telecommunication networks. According to Zorzi et al. (1512), it capitalizes on the prodigious volumes of data generated by network devices and users to augment network performance, reliability, and resource utilization. Through thorough analysis of this data, Su et al. (172) observe that telecommunication operators can discern patterns, trends, and anomalies that inform decision-making processes related to network capacity planning, traffic management, and fault detection. Utilizing advanced analytics techniques, such as ML, in conjunction with optimization algorithms facilitates efficient network management and optimization (Mata et al. 43). The integration of disparate data sources, such as network traffic logs, call detail records, and user behavior data, also bolsters the comprehensiveness and accuracy of the insights gleaned (Zhao et al. 1). Thus, big data-driven network optimization contributes profoundly to the enhancement of telecommunication networks' overall quality of service and operational efficiency. It ensures the seamless delivery of critical communication services to a growing number of users and connected devices.

Customer experience management (CEM) constitutes an indispensable application of big data within telecommunication networks. Šipuš (513) postulates that it empowers operators to glean profound insights into their customers' needs, predilections, and behavioral patterns, thereby bolstering customer satisfaction and retention. The author also notes that big data analytics tools adeptly process various data sources, encompassing call detail records, social media data, and customer feedback, to engender actionable insights instrumental for personalizing services. Accordingly, their application helps orchestrate targeted marketing campaigns and anticipate customer attrition. Implementing ML techniques, including clustering, classification, and collaborative filtering, in conjunction with big data, helps comprehend customer behavior, customize their offerings, and proactively address customer concerns (Bogale 5). Furthermore, Bogale (6) postulates that real-time analytics enable operators to rapidly respond to emerging trends and shifting customer preferences. This outcome fortifies their competitive advantage within the rapidly evolving telecommunication landscape. In this regard, big data-driven CEM constitutes an invaluable asset for telecommunication operators since it fosters customer loyalty and differentiation in an increasingly saturated and competitive market.

Fraud detection is a critical application of big data in telecommunication networks. It addresses the mounting concern of fraudulent activities, such as subscription fraud, premium rate service fraud, and roaming fraud, which engender significant financial and security risks for both operators and customers (Kaur and Sharda 7453). Big data analytics harness the capacity to process immense volumes of network data in real-time, facilitating the identification of conspicuous patterns and anomalies that may signify fraudulent activities (Thudumu et al. 1). By incorporating advanced statistical techniques and ML algorithms, including supervised and unsupervised learning, Kibria et al. (32331) note that analysts are empowered to construct predictive models and detect fraud in telecommunication networks with greater accuracy, thereby mitigating both false positives and false negatives. Additionally, Zhao, Tong, et al. (2668) claim that integrating social network analysis and graph-based techniques further augments fraud detection capabilities by uncovering hidden relationships and collaborative fraud schemes. These studies demonstrate that leveraging big data-driven fraud detection methodologies can help telecommunication operators to effectively curtail revenue loss, safeguard their customers, and preserve the integrity and security of their networks.

2.12 Machine Learning Techniques for Anomaly Detection

ML techniques for anomaly detection have garnered significant attention in recent years, driven by the growing need for more accurate, efficient, and scalable solutions to identify and respond to irregularities. As opposed to traditional methods, Omar et al. (33) note that ML-based approaches enable the automated discovery of complex patterns and relationships in large-scale, high-dimensional data sets, which helps identify anomalies indicative of network faults, security breaches, or fraudulent activities. Abbasi et al. (24) posit that various ML algorithms, including supervised, unsupervised, and semi-supervised techniques, such as classification, clustering, and deep learning, have been employed to detect anomalies in telecommunication networks and demonstrated improved performance in terms of detection accuracy, false positive rates, and adaptability to dynamic network conditions. Integrating ML techniques with big data frameworks and tools, such as Hadoop and Spark, further enhances their applicability in telecommunication networks (Elshawi et al. 1). It enables real-time or near-real-time analysis of vast volumes of data for more effective and timely anomaly detection.

2.13 Supervised Learning Approaches

The advent of support vector machines (SVMs) has marked a significant advancement in supervised learning methodologies for anomaly detection. Amer et al. (2) claims that it has exhibited exceptional classification efficacy in high-dimensional spaces. According to Pestian et al. (943), SVMs have also demonstrated resiliency against overfitting, further enhancing their suitability for handling complex data structures prevalent in telecommunications. Using SMVs as a potent ML tool in telecommunication networks paves the way for more accurate and robust anomaly detection.

Decision trees and random forests, both supervised learning techniques, exhibit considerable potential in anomaly detection. Decision trees entail the formation of tree-like configurations, where internal nodes signify feature tests and leaf nodes delineate class labels, thereby presenting an intelligible and comprehensible model for network administrators (Mirzamomen and Kangavari 345). In contrast, Wu et al. (4) random forests, an ensemble learning approach, expand upon decision trees by constructing multiple trees and aggregating their outputs, effectively mitigating overfitting and enhancing generalization performance.

Neural networks, particularly deep learning architectures, have gained considerable attention for their applicability in telecommunication network anomaly detection, owing to their capacity for automatically learning intricate patterns and features from vast volumes of data. Chen et al. (485) observe that convolutional neural networks (CNNs) and recurrent neural networks (RNNs), including long short-term memory (LSTM) networks, have been employed to capture both spatial and temporal dependencies in network data, allowing for accurate identification of anomalies that may span multiple time frames or network layers. These deep learning models can hierarchically extract high-level features from raw data, eliminating manual feature engineering, a time-consuming and labor-intensive process (Chen et al. 485).

2.14 Unsupervised Learning Approaches

Clustering algorithms, such as K-means and Density-based spatial clustering of applications with noise (DBSCAN), are also employed as unsupervised learning approaches for anomaly detection in telecommunication networks. K-means iteratively assigns data points to their nearest cluster centroid , while DBSCAN identifies clusters based on high-density regions in the data space. Both techniques can effectively separate normal and abnormal network behaviors, with DBSCAN demonstrating robustness against noise and outliers, a crucial advantage in the context of telecommunication networks (Kremers et al. 2). Clustering algorithms offer valuable, unsupervised learning-based solutions for anomaly detection in telecommunication networks, enabling operators to identify and address potential issues without relying on predefined labels or training data.

These advancements have further enhanced autoencoders' applicability in telecommunication network anomaly detection, contributing to more accurate and efficient identification of abnormal behaviors.

2.15 Semi-supervised and Reinforcement Learning Approaches

Semi-supervised learning approaches are a promising alternative for anomaly detection in telecommunication networks. They offer a balance between the benefits of supervised and unsupervised learning methods. Reinforcement learning's ability to continuously learn and adapt to changing network conditions is a promising avenue for future research and development in telecommunication network anomaly detection.

2.16 Evaluation Metrics for Anomaly Detection Models

The evaluation of anomaly detection models in telecommunication networks necessitates the use of robust metrics, with precision, recall, and F1-score becoming prominent choices in this domain. Precision, also known as positive predictive value, measures the proportion of true positive predictions among all positive predictions made by the model, reflecting the model's accuracy in detecting actual anomalies (Saito and Rehmsmeier 2). Recall, or sensitivity, quantifies the proportion of true positive predictions among all actual positive instances, highlighting the model's ability to capture relevant anomalies without overlooking any

critical events. Lipton et al. (3) note that the F1-score represents the harmonic mean of precision and recall, which provides a single metric that balances the trade-off between these two measures. This equilibrium is crucial, as prioritizing one measure over the other may result in overlooking genuine anomalies or generating excessive false alarms. Consequently, it is critical to appraising the performance of anomaly detection models since it enables better decision-making regarding selecting and deploying suitable models in diverse network environments.

2.17 Leveraging Big Data and ML for Proactive Fault Management

The integration of big data and ML in telecommunication networks has created unprecedented opportunities for proactive fault management. It enables network operators to predict and prevent network anomalies before they escalate into service disruptions. Matsuno et al. (142) observe that ML techniques, including supervised, unsupervised, and semi-supervised learning, have been employed to process and analyze large volumes of heterogeneous network data, identifying patterns, trends, and correlations that would be impractical to discern using traditional methods. These approaches facilitate early detection of potential faults and prompt remediation by leveraging high-velocity streaming data, advanced analytics, and real-time visualization, thereby enhancing network reliability, performance, and customer satisfaction (Yan and Wang 1527).

2.18 Integration of Big Data and ML for Improved Anomaly Detection

The significance of rigorous data preprocessing and feature engineering in enhancing anomaly detection capabilities within telecommunication networks cannot be overstated.

2.19 Future Research Directions and Challenges

The potential advancements in big data and ML techniques for telecommunication networks hold significant promise for improving network performance, reliability, and security.

3. Methodology

The research focuses on the application of cutting-edge big data and machine learning (ML) techniques for anomaly detection in telecommunication networks. The central research questions probe the efficacy of these advanced computational paradigms in discerning anomalies that may signify potential faults or disruptions. The hypotheses posit that leveraging these techniques facilitates a more proactive approach to fault management, leading to improved network performance and customer satisfaction. Grounded in the pragmatist paradigm, the research design employs a mixed-methods approach, combining quantitative and qualitative analyses to ensure a comprehensive understanding of the problem space. It has been carefully chosen as it helps evaluate the performance of ML algorithms in a quantifiable manner while concurrently exploring and addressing the practical challenges inherent to deploying these algorithms within the dynamic context of telecommunication networks. The essence of this approach is to secure rigorous, actionable insights that can guide the development and implementation of next-generation network management strategies.

3.1 Exploration of Big Data Usage for Anomaly Detection

Examining the effective utilization of big data for anomaly detection in telecommunication networks necessitates a methodical, structured approach to the research. This objective is tackled by deploying robust big data analytics tools and techniques, which enable the extraction, processing, and analysis of vast volumes of telecom alarm indication signal data occurring at the E1/T1 port. In this vein, ML algorithms are integral to the methodology, as they are leveraged to identify patterns and anomalies within these big data sets. The successful execution of this objective further involves the rigorous preprocessing and cleaning of data and the application of feature selection and dimensionality reduction techniques to optimize the efficiency and performance of the ML models. This holistic methodology ensures that the research objective is addressed comprehensively, facilitating the detection of anomalies and the derivation of actionable insights that can inform proactive fault management strategies. Essentially, this objective charts a path for the transformative integration of big data and ML within the realm of telecommunication network management.

3.2 Evaluation of ML Algorithms for Anomaly Detection

The objective of assessing the performance of various ML algorithms to detect anomalies in telecommunication networks calls for a rigorous methodological approach. Accordingly, the research methodology applies various ML algorithms, including supervised, unsupervised, semi-supervised, and deep learning methods. These algorithms are trained and tested on a rich, curated dataset that captures a broad spectrum of network activities and potential anomalies. This approach gauges the predictive power of the algorithms and evaluates their computational efficiency, scalability, and adaptability to evolving network dynamics. Through this methodical examination, the research establishes an empirical basis for selecting optimal ML techniques for anomaly detection in telecommunication networks.

3.3 Optimization of ML
The research methodology incorporates a multifaceted approach to optimize ML models for anomaly detection through feature selection and dimensionality reduction techniques. The approach balances the need for model accuracy and predictive power with the practical considerations of computational efficiency and interpretability.

3.4 Research Methodology
The research methodology is a rigorous, six-step process. It commences with data collection and preprocessing, which involves collecting telecommunication network data and its subsequent cleaning and transformation, with feature engineering playing a vital role. The next stage focuses on feature selection and dimensionality reduction to identify and prioritize the most relevant features and reduce the computational complexity. This step is followed by developing and evaluating ML models, where suitable algorithms are identified, developed, and evaluated based on precision, recall, and F1-score. A comparative analysis of these models is conducted to identify the most effective one for anomaly detection. The model's performance is then evaluated and validated using rigorous metrics, and its performance compared with traditional methods. The research culminates in the interpretation of findings and discussion of implications.

3.5 Data Collection and Preprocessing
The initial phase delves into procuring and curating the dataset on the alarm indication signal data occurring at the E1/T1 port. This dataset encompasses alarm indication signal data occurring at the E1/T1 port. Such a broad spectrum of data sources aids in capturing the multifaceted nature of network activities and anomalies. Once collated, the raw alarm indication signal data undergoes rigorous preprocessing techniques to cleanse, normalize, and transform it, rendering the dataset suitable for subsequent analysis (Angehrn et al. 2). The derived features encapsulate the complex network details, providing a richer and more representative input to the ML models. Feature engineering substantively contributes to the efficacy and robustness of the anomaly detection process by using the most relevant and informative aspects of the data.

3.6 Feature Selection and Dimensionality Reduction
Feature selection techniques are vital in the research methodology, as they help identify the most relevant and informative subset of features from the original feature set. These techniques are critical in reducing the dimensionality of the alarm indication signal data while retaining the most discriminative features for anomaly detection. The methods optimize creating a compact, meaningful feature space to facilitate accurate and efficient anomaly detection in telecommunication networks.

3.7 Machine Learning Model Development and Evaluation
Random Forest is a formidable contender for anomaly detection within telecommunication networks. It amalgamates multiple decision trees for predictions, making it a versatile tool. Katuwal and Suganthan (2) note that random Forest constructs an ensemble of decision trees, each trained on a randomly sampled subset of the training data, a technique called bagging, which enhances the model's generalization capability and reduces overfitting. A decision tree represents a series of binary splits based on feature values to partition the data into different classes or categories. Random Forest's robustness against overfitting and noisy data lends itself to the task. It provides improved accuracy in anomaly detection by mitigating the impact of individual decision errors through the aggregation of multiple predictions.

3.8 Comparison of ML Models' Performance
The comparative analysis of ML models hinges on a rigorous assessment of their performance in detecting anomalies and their overall predictive capabilities. Accuracy gauges the overall correctness of the models' predictions, which epitomizes the proportion of correctly classified instances. Precision and recall, as explained previously, evaluate the models' competence in correctly identifying anomalies and capturing all actual anomalies, respectively. Computational efficiency, weighed regarding training time, memory requirements, and prediction speed, also constitutes a critical factor in model selection. A robust anomaly detection model should perform exceptionally on accuracy metrics and demonstrate high computational efficiency.

3.9 Criteria for Model Selection
Model selection for anomaly detection in telecommunication networks hinges on a balanced evaluation of several crucial factors. Firstly, performance serves as a paramount criterion, with the selected model expected to display high accuracy, precision, recall, and F1-score to reflect its efficacy in anomaly detection. By weighing these criteria against the strengths and limitations of each model, the most suitable ML model for anomaly detection in telecommunication networks is selected.

To conclude, the methodology delineated in this chapter provides a novel approach to the research objective of developing an effective ML-based anomaly detection system for telecommunication networks. The methodology's multifaceted nature, encapsulating data collection and preprocessing, feature engineering, model development, and evaluation, underscores its alignment with the research objectives and the

complexity of the study. Anticipated challenges include the data's high dimensionality, noise and redundancy, and the need for model interpretability and computational efficiency. These challenges are addressed through meticulous feature selection, dimensionality reduction, and the careful choice of interpretable ML models. This methodology will influence the overall research outcome by ensuring the development of an accurate, efficient, and interpretable anomaly detection system. It validates the efficacy of ML in telecommunications. Also, it explains its broader adoption in the telecommunication industry, thereby catalyzing a paradigm shift in how network anomalies are detected and managed.

4. Analysis & Discussion Conclusion

The research successfully elucidated the potential of harnessing big data for anomaly detection in telecommunication networks. It reveals the efficacy of specific preprocessing techniques to ensure data quality, relevance, and scalability. Regarding ML algorithm performance, the isolation forest model outperformed the One-Class SVM model across several metrics, demonstrating the diversity of model performance in tackling various network faults and their potential impacts. The study underscored the value of feature selection and dimensionality reduction techniques in enhancing the efficiency and accuracy of ML models within the context of anomaly detection. However, it also acknowledged the challenges of deploying big data and ML-based solutions in real-world telecommunication networks, pointing to a few mitigating strategies. Furthermore, the identified limitations, such as class imbalance and potential biases, signal a need for future research to further explore these aspects. By offering a granular understanding of anomaly detection in telecommunication networks, the study presents a strong foundation for future advancements in this critical field.

The findings bolster the theoretical underpinnings of the applicability of ML algorithms and big data in anomaly detection in telecommunication networks. They enhance the understanding of data preprocessing and feature selection techniques. The comparative analysis of the isolation forest and One-Class SVM models present a valuable benchmark for researchers and practitioners in selecting and optimizing anomaly detection models. It advances the understanding of the trade-off between different performance metrics, adding depth to the discussion on model selection and evaluation. From a policy perspective, the findings underline the importance of careful data handling practices, especially in handling class imbalances. The research's emphasis on potential limitations and biases affirms the necessity for robust and transparent methodologies, setting an example for future empirical studies. Recognizing the research constraints paves the way for further research endeavors, spotlighting the need for broader and more diverse datasets and the exploration of longitudinal study designs. Therefore, this study significantly contributes to academic discourse and industry practice in telecommunication network anomaly detection.

5.0 Research Contribution and Discussion

This chapter discusses the key research outcomes and elucidates their implications on academic theory and practical application in telecommunication networks. The primary objective is to establish the original contributions of this study, with a particular emphasis on the novelty of the developed framework, the empirical findings obtained, theoretical advancements made, and the practical implications that bear relevance for stakeholders. Further, the chapter interprets the results in light of existing literature, research objectives, and theoretical postulations. An essential part of this discourse will be acknowledging the study's limitations, highlighting the areas where caution must be exercised while interpreting the findings. Lastly, the chapter proposes future research directions and identifies limitations and unanswered questions arising from the research. This approach ensures the findings are fully contextualized within the broader scientific dialogue. It accentuates their significance and relevance in advancing knowledge on anomaly detection in telecommunication networks.

The research offers noteworthy contributions to anomaly detection within telecommunication networks. It contributes to the academic corpus with innovative framework development, empirical findings, theoretical advancements, and pragmatic implications. The developed framework integrates various theories, methodologies, and concepts and fills the research gap in the existing literature by providing an inclusive and novel approach to anomaly detection. A pivotal contribution arises from the empirical findings that shed light on the niche topic, which lends new perspectives and challenges prevailing paradigms. The theoretical advancements help enhance and broaden the understanding of anomaly detection in telecommunication networks through a critical synthesis of existing literature. Further, the study's outcomes have profound practical implications that may assist practitioners and policymakers in the telecommunication industry to make informed decisions. A detailed discussion segment will further delve into these contributions by critically analyzing the significance of the research findings, their accord with previous studies, and the potential limitations encountered during the study. Future research trajectories inspired by the unanswered questions and identified gaps will conclude this chapter, propelling further scholarly discourse.

5.2 Research Contribution

This research contributes significantly to anomaly detection in telecommunication networks, particularly in four domains. Firstly, the study innovates a unique framework that leverages big data and machine learning (ML) techniques for proactive fault management in telecommunication networks. This framework helps fill the existing literature gap, offering a comprehensive perspective on anomaly detection. Secondly, the empirical findings generated through rigorous data analysis and model evaluation using ML techniques are a notable contribution. These new insights foster a more profound understanding of anomaly detection, effectively enriching the empirical knowledge base. The third area of contribution lies in theoretical advancement. The study refines existing theories and models through its extensive literature review and analysis. This progression helps cultivate new conceptual frameworks and theoretical perspectives that augment comprehension of the research topic. Lastly, the practical implications derived from the findings have substantial relevance for industry professionals and policymakers in the telecommunication sector. These insights provide strategic guidance and actionable recommendations to effectively tackle real-world challenges.

5.3 Discussion

The significance of this research's findings lies in its multifaceted contribution to the broader ML field, particularly within anomaly detection. By rigorously evaluating different algorithms, such as the isolation forest and One-Class SVM models, the study delineates the optimal strategies and the underlying metrics that determine success in real-world applications. It transcends mere algorithmic comparison, extending into methodological advancements that enable scalable and accurate anomaly detection. Therefore, this research provides a vital steppingstone for future studies and offers practitioners a well-founded basis for implementing state-of-the-art techniques. Moreover, the clear articulation of metrics like precision, recall, and the F1 score creates a shared evaluation framework across various domains and applications. Considering the ever-expanding digital landscape, where anomalies and security breaches can have far-reaching consequences, this study is a beacon that guides both researchers and industry experts toward enhanced solutions. Its well-rounded exploration of theory and practice coalesces into a seminal piece that promises to resonate and influence the field of anomaly detection for years to come.

The alignment between the findings and the research objectives is evident in the systematic manner in which the study was conducted. A multi-layered methodology was harnessed to effectively explore the utilization of big data (Objective A) by deploying robust big data analytics tools and techniques. This application has culminated in transformative insights for telecommunication network management. The assessment of various ML algorithms (Objective B) was achieved by applying supervised and unsupervised learning methods, accompanied by comparative analysis, to select optimal techniques for anomaly detection. The optimization of ML models (Objective C), facilitated by techniques like LASSO regularization and PCA, demonstrated a balance between model accuracy and computational efficiency, which reflects an understanding of real-world applications. Lastly, addressing the practical challenges in deploying these advanced solutions within real-world networks (Objective D) was comprehensively explored, bridging theoretical advancements with industry applicability. The methodological synergy provides a robust foundation for the study's contributions to the field of telecommunication network management.

5.4 Alignment with Previous Studies

The findings accentuate the isolation forest model's prominence in anomaly detection within telecommunication networks. Accordingly, they corroborate with the assertions of Hariri et al. (1479) on the model's computational efficiency and effectiveness in anomaly identification. The precision metric of 0.863 illustrates an alignment with prior observations, which reveals a robust capacity to discern anomalies with significant accuracy. Comparatively, while the One-Class SVM model with a precision score of 0.832 is laudable, it is slightly overshadowed by the isolation forest's performance. Moreover, the examination of F1-scores provides a detailed analysis that emphasizes the importance of multifaceted evaluation – a point underlined by Tatbul et al. (7). The isolation forest model's F1-score of 0.901 signifies a balanced amalgamation of precision and recall, subtly contrasting the One-Class SVM model's F1-score of 0.908. The broader evaluation transcends mere reliance on singular metrics, furnishing insights into the subtleties of algorithm selection. By juxtaposing the empirical outcomes with existing literature, the study fosters a deeper comprehension of the relationship between big data and anomaly detection in telecommunication networks. It contributes to an enriched understanding of preprocessing, feature selection, and the intricate trade-offs inherent in algorithmic decisions. The alignment with previous studies illustrates continuity and innovation within the research landscape.

Integrating big data and ML for proactive fault management in telecommunication networks showcases a paradigmatic shift in enhancing anomaly detection capabilities. The current research findings significantly align with the works of Matsuno et al. (142) and Yan and Wang (1527), who have elucidated the symbiotic relationship between ML algorithms and large-scale heterogeneous data processing. The novelty lies in the stratified approach to data preprocessing and feature engineering, elements that Prakash et al. (335) and Khalid et al. (327) have deemed indispensable for the overall data quality. Meanwhile, the focus on model selection and optimization, resonating with Choi et al. (120060) and Greeshma and Sreekumar (3713),

illustrates the significance of customizing algorithms to suit network-specific requirements. The real-time monitoring and alerting mechanisms echo the perspective of Syafrudin et al. (2), emphasizing the necessity of agile network management. Moreover, the real-world implementations and case studies coalesce with the observations of Keshavarz et al. (2) and Yayah et al. (59), affirming the efficacy and applicability of big data and ML within diverse sectors, including telecommunications. Collectively, the convergence of these dimensions provides a robust analytical framework that is synergistic with previous studies and extends the discourse by highlighting interdisciplinary collaboration and strategic implementation, reflecting the evolving landscape of intelligent and resilient telecommunication networks.

5.5 Future Research Directions

The findings and limitations delineate several promising avenues for future research that merit exploration. First, a deeper examination of alternate resampling techniques to handle class imbalance may yield insights into enhancing predictive models' efficacy. Emphasizing the assimilation of larger, more diverse datasets could broaden analysis perspectives and reduce bias introduced by dataset specificity, thus enhancing the generalizability of the findings. Second, adopting longitudinal study designs could provide robust insights into the dynamism of anomaly detection in telecommunication networks, documenting subtle changes, trends, and variations over extended periods. An area demanding focus is an improved feature selection process that should be guided by domain expertise to ensure that all relevant variables are considered, potentially enhancing model accuracy. Finally, integrating interdisciplinary approaches encompassing statistical, ML, and domain-specific insights might lead to innovative methodologies that allow for a more comprehensive understanding of anomaly detection. These directions can lead to more adaptable and reliable anomaly detection models within telecommunication networks and extend the field's theoretical and practical frontiers.

Building upon the present study's findings requires a multifaceted approach that combines refinement, innovation, and collaboration. Future researchers should investigate the integration of interdisciplinary methods that fuses statistical analysis with ML and domain-specific knowledge. This approach could lead to more holistic models for anomaly detection within telecommunication networks. A commitment to rigorous feature selection, ensuring that all pertinent variables are thoughtfully included, could further fine-tune the accuracy of subsequent models. Collaborations with industry experts may offer an opportunity to validate models within real-world scenarios and enhance relevance and applicability. Moreover, future studies might explore the potential of employing diverse and extensive datasets that include various sources and time frames to mitigate biases and augment generalizability. Delving into longitudinal studies that monitor systems over protracted periods may unveil deeper insights into the subtleties and complexities of the subject matter. Together, these proposed directions provide a roadmap for advancing the field. They challenge future researchers to adopt innovative, reflective, and comprehensive strategies to yield a more profound understanding of telecommunication networks and the anomalies therein.

6. The Conclusion

Anomaly detection has increasingly become a key focus in telecommunication networks due to the ubiquity of big data and machine learning (ML) technologies. In other words, identifying and rectifying anomalies remains paramount, particularly in an era where network reliability can impact vast facets of daily operations and user experiences. This research sought to employ novel techniques to fortify proactive fault management within telecommunication networks. The crux of the thesis revolves around the central proposition that leveraging state-of-the-art big data analytics and ML algorithms can substantially enhance the detection of anomalies, thereby facilitating a more robust and resilient network infrastructure. To empirically substantiate this assertion, the study adopted a rigorous and systematic research methodology reinforced by data-driven analyses. The study provides actionable insights for network operators and policymakers through an exhaustive exploration that involves collecting and interrogating data sets and implementing novel ML models. This concluding chapter aims to sum up the research trajectory, crystallizing its key findings, implications, limitations, and potential avenues for future exploration.

The study reveals several pivotal findings in ML model development, evaluation, and comparative analysis. Initially, a robust preprocessing phase, which included transforming the 'Severity' class labels to binary and applying SMOTEENN, was crucial for mitigating class imbalance. The isolation forest algorithm was significant, especially when benchmarked against the One-Class Support Vector Machine (SVM). Its high precision, recall, and F1 scores illustrated its robust capability in anomaly detection within telecommunication networks. With an adroit precision metric of 0.863, the isolation forest model effectively corroborates Hariri et al.'s (1479) claims on its proficiency, outpacing the One-Class SVM despite its respectable precision score of 0.832. Concurrently, the dimensionality reduction and feature selection techniques heightened the ML model's efficacy, which encapsulated the essence of the third research objective. Further, Receiver Operating Characteristic (ROC) curve assessments and the subsequent AUC-ROC scores accentuated the disparity in performance between the models, with the isolation forest significantly outshining the One-Class SVM. Visual aids like confusion matrices provided detailed insights into the models' predictive performance and were invaluable tools for exhaustive comparison. Finally, the empirical analysis

showcased the isolation forest's computational efficiency and highlighted its prowess in discerning anomalies. These findings shed light on the varied dynamics of big data and ML models in anomaly detection.

Considering the detailed examination of multiple ML models for anomaly detection within telecommunication networks, this study illuminates the exceptional capability of the isolation forest model. Its inherent computational efficiency, as evidenced by the precision, recall, and F1 metrics, outstrips the performance of the One-Class SVM model. Moreover, the use of SMOTEENN for addressing class imbalance proves instrumental. It accentuates the importance of preprocessing steps in the predictive modeling workflow. AUC-ROC stands out as a salient performance indicator. The model's score of 0.861 emphasizes its superlative classification prowess. The juxtaposition of the ROC curve and confusion matrices reveals a multifaceted evaluation approach that is vital for a detailed understanding of model behaviors. Moving forward, future research endeavors can probe deeper into ensemble techniques and hybrid models. They can harness the strengths of multiple algorithms to further refine anomaly detection capabilities and ensure proactive fault management. Additionally, integrating domain-specific knowledge might amplify the efficacy of telecommunication network monitoring and optimize prediction accuracy and operational efficiency.

Although the current study has explored novel big data and ML techniques for proactive fault management in telecommunication networks, several potential research avenues remain unexplored. Notably, the reliance on SMOTEENN for class balancing presents a research opportunity to explore alternative data augmentation techniques. In other words, future research can examine their efficacy in enhancing model generalizability. Also, the study's cross-sectional design underscores the latent potential of employing longitudinal research designs. This direction could provide insights into the temporal dynamics and evolving nature of anomalies in telecommunication networks. As Chapter 4 highlighted the limited scope of the dataset used, future investigations can consider integrating multi-source, diversified datasets to enhance the robustness and comprehensiveness of the analysis. Moreover, given the rapid advancements in telecommunication technologies, examining the relationship between next-generation networks and anomaly detection could yield vital insights. These potential research directions could advance the understanding and efficacy of anomaly detection methods in the telecommunication industry.

Considering the insights derived from this research and the rapid advances in big data and ML, potential studies can build upon this work to advance anomaly detection in telecommunication networks. Future research could explore hybrid models incorporating traditional statistical techniques with advanced ML algorithms, potentially compensating for the limitations observed in the current study's methodologies. While the current model focused on specific features, future endeavors might utilize feature extraction and deep learning techniques to automatically identify and prioritize influential variables. With the proliferation of big data in telecommunications, there is an imperative to leverage these vast data streams more effectively. By harnessing sophisticated ML techniques like reinforcement learning or neural networks, future studies could better understand anomalous patterns, especially in real-time detection scenarios. Besides, incorporating federated learning approaches could allow for decentralized anomaly detection. They could address data privacy concerns and harness data directly from the source. Telecommunication infrastructures are ever-evolving and present increasing complexities. As such, integrating cutting-edge advancements in big data and ML will undoubtedly steer the direction of future research that promises more robust, adaptive, and real-time anomaly detection systems.

This thesis delineates a paradigmatic shift in anomaly detection methodologies for telecommunication networks. By synergizing novel big data and ML techniques, the study presents a sophisticated detection framework that is robust and adaptable. Notably, the innovative integration of the SMOTEENN resampling technique demonstrates an advanced approach to addressing the pervasive challenges of class imbalance and dimensionality. This approach subsequently paves the way for heightened model accuracy and predictive validity. The in-depth examination of the dataset's scope and specificity underscores the importance of comprehensive and diversified data sources for enhancing generalizability. The empirical findings augment the practical foundations of the field, which sets new benchmarks in model performance, especially the isolation forest algorithm. Moreover, the methodological rigor and holistic analysis espoused throughout the research are integral to the broader telecommunication sector. They signal a potential benchmark for future studies and practices. This thesis substantially contributes to the discourse on anomaly detection as it melds theory with practice. It cultivates a blueprint that promises enhanced network security and reliability and lays the groundwork for the next era of telecommunication network research. When juxtaposed against traditional methods, the findings offer concrete evidence of the thesis's advancements in practical applications.

The research anchors on ingenuity and comprehensive discussions. One monumental stride is introducing a contemporary framework that synergizes big data analytics and ML techniques. While traditional methodologies wrestle with the inherent complexities of high-dimensional datasets and class imbalances, this research presents a sophisticated solution. It starts with foundational descriptive analysis functions, such as `describe(),` `info(),` and `head(),` and employs a methodology that harnesses the power of visualization tools like box plots. It also implements such statistical methods as the interquartile range (IQR) to deal with outliers. Principal component analysis (PCA) facilitates adept dimensionality reduction that converts vast datasets into condensed yet information-packed structures. Further, the study addresses the ubiquitous challenge of class imbalance through the SMOTEENN to ensure data integrity for robust model training. The decision to use the random forest and One-Class SVM, benchmarked with precision, recall, and F1-score,

showcases the study's commitment to accuracy. Conclusively, by comparing the ROC curve and confusion matrices, the research augments the domain of anomaly detection and charts a visionary path for subsequent scholarly pursuits.

The study occupies a unique niche, not just because of its academic rigor but predominantly due to its interdisciplinary undertakings, ingenious collaborations, and groundbreaking techniques. This work has sculpted a narrative that, while rooted in anomaly detection in telecommunications, extends into realms as diverse as regulatory policy, software development, and academic discourse. The marriage of the isolation forest algorithm with the SMOTEENN technique highlights this synthesis. It shows the fusion of advanced mathematical tools and domain expertise. Besides, the delineation of metrics, namely precision, recall, and F1-score, does not merely serve as evaluative benchmarks but is posited as a gold standard. They beckon regulatory bodies to formulate cohesive assessment criteria across the sector. Similarly, the research is not just confined to theory; it invites software developers to integrate these empirical revelations into pragmatic solutions to fortify the telecom sector's defenses against vulnerabilities. Therefore, the thesis cohesively links academia, industry, and policy to augur a future of enhanced telecommunication infrastructures and the manifold socio-economic dividends they yield.

The pressing challenges of anomaly detection and real-time fault management in the telecommunication industry have consistently emerged as focal points that require innovative solutions. Through its rigorous exploration and empirical findings, this research illuminates the intertwined pathways of big data analytics and ML to proactively address anomaly detection concerns. The study's implication is even more pronounced as the telecommunication industry grapples with ever-increasing data volumes and multifaceted network dynamics. Proactive fault management through novel big data and ML methodologies is an indispensable strategy for preempting vulnerabilities and optimizing network performance. Furthermore, as the line between digital infrastructures and socio-economic progress becomes increasingly indistinct, ensuring the robustness and reliability of telecommunication networks is no longer just a technical imperative but a societal one. Thus, whereas this research is rooted in the specifics of telecommunications, it transcends its immediate domain. It is a clarion call to the broader industry to harness the might of big data and ML to develop resilient telecommunication networks.

ACKNOLOGEMENT

I am so grateful to Almighty Allah and my mother who passed away for her great love and prays. I am also so grateful to my supervisor Prof. Salvatore Faza for his continuous encouragement, support and his valuable time. I would like to extend my heartfelt gratitude to Bahrain Telecommunication Company for the incredible opportunity to serve as a professional engineer in the operation and maintenance sections for an extended period of time.

References

1. Abbasi, Mahmoud, Amin Shahraki, and Amir Taherkordi. "Deep Learning for Network Traffic Monitoring and Analysis (NTMA): A Survey." *Computer Communications,* vol. 170, 2021, pp. 19-41, doi:10.1016/j.comcom.2021.01.021.
2. Bergstra, James, et al. "Hyperopt: A Python Library for Model Selection and Hyperparameter Optimization." *Computational Science & Discovery,* vol. 8, no. 1, 2015, pp. 1-24, doi:10.1088/1749-4699/8/1/014008.
3. Đorđević, Valentina, Pavle Miloševic, and Ana Poledica. "Machine Learning Based Anomaly Detection as an Emerging Trend in Telecommunications." *Management: Journal of Sustainable Business and Management Solutions in Emerging Economies,* vol. 27, no. 2, 2022, pp. 71-82, doi:10.7595/management.fon.2020.0002.
4. Kabir, Md Faisal, Tianjie Chen, and Simone A. Ludwig. "A Performance Analysis of Dimensionality Reduction Algorithms in Machine Learning Models for Cancer Prediction." *Healthcare Analytics,* vol. 3, no. 100125, 2023, pp. 1-9, doi:10.1016/j.health.2022.100125.
5. Morang'a, Collins M., et al. "Machine Learning Approaches Classify Clinical Malaria Outcomes Based on Haematological Parameters." *BMC Medicine,* vol. 18, no. 1, 2020, pp. 1-16, doi:10.1186/s12916-020-01823-3.
6. Nguyen, Giang, et al. "Deep Learning for Proactive Network Monitoring and Security Protection." *IEEE Access,* vol. 8, 2020, pp. 19696-716, www.digital.csic.es/bitstream/10261/221908/1/deeprotec.pdf.
7. Karatepe, Ilyas Alper, and Engin Zeydan. "Anomaly Detection in Cellular Network Data Using Big Data Analytics." *European Wireless 2014; 20th European Wireless Conference,* VDE, 2014, pp. 1-5, www.researchgate.net/publication/286519918_Anomaly_detection_in_cellular_network_data_using_big_data_analytics.
8. Ahmed, Mohiuddin, Abdun Naser Mahmood, and Jiankun Hu. "A Survey of Network Anomaly Detection Techniques." *Journal of Network and Computer Applications,* vol. 60, 2016, pp. 19-23, doi:10.1016/j.jnca.2015.11.016.

9. Aladaileh, Mohammad Adnan, et al. “Renyi Joint Entropy-Based Dynamic Threshold Approach to Detect DDoS Attacks against SDN Controller with Various Traffic Rates.” *Applied Sciences,* vol. 12, no. 12, 2022, pp. 1-17, doi:10.3390/app12126127.
10. Al-Musawi, Bahaa, et al. “Identifying OSPF LSA Falsification Attacks Through Non-linear Analysis.” *Computer Networks,* vol. 167, 2020, pp. 107031, doi:10.1016/j.comnet.2019.107031.
11. Amer, Mennatallah, Markus Goldstein, and Slim Abdennadher. “Enhancing One-class Support Vector Machines for Unsupervised Anomaly Detection.” Master Thesis*,* German University in Cairo, 2013, www.madm.dfki.de/_media/theses/ma_thesis_amer.pdf.
12. Asghar, Ahmad, Hasan Farooq, and Ali Imran. “Self-healing in Emerging Cellular Networks: Review, Challenges, and Research Directions.” *IEEE Communications Surveys & Tutorials,* vol. 20, no. 3, 2018, pp. 1682-709, doi:10.1109/COMST.2018.2825786.
13. Baker, Thar, et al. “A Secure Fog-based Platform for SCADA-based IoT Critical Infrastructure.” *Software: Practice and Experience,* vol. 50, no. 5, 2020, pp. 503-18, doi:10.1002/spe.2688.
14. Bansal, Aasthaa, and Patrick J. Heagerty. “A Comparison of Landmark Methods and Time-dependent ROC Methods to Evaluate the Time-varying Performance of Prognostic Markers for Survival Outcomes.” *Diagnostic and Prognostic Research,* vol. 3, 2019, pp. 1-13, doi:10.1186/s41512-019-0057-6.
15. Baştuğ, Ejder, et al. “Big Data Meets Telcos: A Proactive Caching Perspective.” *Journal of Communications and Networks,* vol. 17, no. 6, 2015, pp. 549-57, doi:10.1109/JCN.2015.000102.
16. Berrar, Daniel. “Cross-Validation.” *Encyclopedia of Bioinformatics and Computational Biology*, vol. 1, 2019, pp. 1-8, doi:10.1016/B978-0-12-809633-8.20349-X.
17. Bhattacharyya, Dhruba Kumar, and Jugal Kumar Kalita. *Network Anomaly Detection: A Machine Learning Perspective*. CRC Press, 2013.
18. Bogale, Tadilo Endeshaw, Xianbin Wang, and Long Bao Le. “Machine Intelligence Techniques for Next-generation Context-aware Wireless Networks.” 2018, pp. 1-10, doi:10.48550/arXiv.1801.04223.
19. Campbell, Aaron, Kyle Caudle, and Randy C. Hoover. “Examining Intermediate Data Reduction Algorithms for Use with t-SNE.” *Proceedings of the 2019 3rd International Conference on Compute and Data Analysis,* 2019, p. 36-42, doi:10.1145/3314545.3314549.
20. Carter, Jane V., et al. “ROC-ing Along: Evaluation and Interpretation of Receiver Operating Characteristic Curves.” *Surgery,* vol. 159, no. 6, 2016, pp. 1638-45, doi:10.1016/j.surg.2015.12.029.
21. Castro, Diogo, et al. “Apache Spark Usage and Deployment Models for Scientific Computing.” *EPJ Web of Conferences*, vol. 214, EDP Sciences, 2019, doi:10.1051/epjconf/201921407020.
22. Chaparro, Cameron, and William Eberle. “Detecting Anomalies in Mobile Telecommunication Networks Using a Graph-based Approach.” *The Twenty-Eighth International Flairs Conference,* 2015, pp. 410-15, www.ailab.wsu.edu/adgs/pdfs/ChaparroFLAIRS2015.pdf.
23. Chen, Cen, et al. “Gated Residual Recurrent Graph Neural Networks for Traffic Prediction.” *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 33. no. 01, 2019, pp. 485-92, doi:10.1609/AAAI.V33I01.3301485.
24. Cheng, Lefeng, and Tao Yu. “A New Generation of AI: A Review and Perspective on Machine Learning Technologies Applied to Smart Energy and Electric Power Systems.” *International Journal of Energy Research,* vol. 43, no. 6, 2019, pp. 1928-73, doi:10.1002/er.4333.
25. Chicco, Davide, and Giuseppe Jurman. “The Advantages of the Matthews Correlation Coefficient (MCC) Over F1 Score and Accuracy in Binary Classification Evaluation.” *BMC Genomics,* vol. 21, 2020, pp. 1-13, doi:10.1186/s12864-019-6413-7.
26. Chicco, Davide, Niklas Tötsch, and Giuseppe Jurman. "The Matthews Correlation Coefficient (MCC) is More Reliable Than Balanced Accuracy, Bookmaker Informedness, and Markedness in Two-class Confusion Matrix Evaluation." *BioData Mining,* vol. 14, no. 1, 2021, pp. 1-22, doi:10.1186/s13040-021-00244-z.
27. Choi, Kukjin, et al. “Deep Learning for Anomaly Detection in Time-series Data: Review, Analysis, and Guidelines.” *IEEE Access,* vol. 9, 2021, pp. 120043-65, doi:10.1109/ACCESS.2021.3107975.
28. Das, Arun, and Paul Rad. “Opportunities and Challenges in Explainable Artificial Intelligence (xai): A Survey,” 2020, pp. 1-24, doi:10.48550/arXiv.2006.11371.
29. Dierckens, Karl E., et al. “A Data Science and Engineering Solution for Fast K-means Clustering of Big Data.” *2017 IEEE*, 2017, doi:10.1109/Trustcom/BigDataSE/ICESS.2017.332.
30. Dong, Bo, and Xue Wang. “Comparison Deep Learning Method to Traditional Methods Using for Network Intrusion Detection.” *2016 8th IEEE International Conference on Communication Software and Networks (ICCSN)*. IEEE, 2016, pp. 581-85, doi:10.1109/ICCSN.2016.7586590.
31. Elgendy, Nada, and Ahmed Elragal. “Big Data Analytics: A Literature Review Paper.” *Advances in Data Mining. Applications and Theoretical Aspects: 14th Industrial Conference, ICDM 2014, St. Petersburg, Russia, July 16-20, 2014, Proceedings 14*. Springer International Publishing, 2014, www.researchgate.net/profile/Ahmed-Elragal/publication/264555968_Big_Data_Analytics_A_Literature_Review_Paper/links/541e9b9a0cf203f155c0655a/Big-Data-Analytics-A-Literature-Review-Paper.pdf.

32. Elshawi, Radwa, et al. “Big Data Systems Meet Machine Learning Challenges: Towards Big Data Science As a Service.” *Big Data Research,* vol. 14, 2018, pp. 1-11, doi:10.1016/j.bdr.2018.04.004.
33. Engle, Ryan D. *A Methodology for Evaluating Relational and NoSQL Databases for Small-Scale Storage and Retrieval*. Air Force Institute of Technology Wright-Patterson AFB OH Wright-Patterson AFB United States, 2018, www.apps.dtic.mil/sti/pdfs/AD1063484.pdf.
34. Fan, Cheng, et al. “Analytical Investigation of Autoencoder-based Methods for Unsupervised Anomaly Detection in Building Energy Data.” *Applied Energy,* vol. 211, 2018, pp. 1123-35, doi:10.1016/j.apenergy.2017.12.005.
35. Ghazi, Mohd Rehan, and Durgaprasad Gangodkar. "Hadoop, MapReduce and HDFS: A Developers Perspective." *Procedia Computer Science,* vol. 48, 2015, pp. 45-50, doi:10.1016/j.procs.2015.04.108.
36. Greeshma, K. V., and K. Sreekumar. “Hyperparameter Optimization and Regularization on Fashion-MNIST Classification.” *International Journal of Recent Technology and Engineering (IJRTE),* vol. 8, no. 2, 2019, pp. 3713-19, doi:10.35940/ijrte.B3092.078219.
37. Gudivada, Venkat, Amy Apon, and Junhua Ding. “Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations.” *International Journal on Advances in Software,* vol. 10, no. 1, 2017, pp. 1-20, www.personales.upv.es/thinkmind/dl/journals/soft/soft_v10_n12_2017/soft_v10_n12_2017_1.pdf.
38. Habeeb, Riyaz Ahamed Ariyaluran, et al. “Real-time Big Data Processing for Anomaly Detection: A Survey.” *International Journal of Information Management,* vol. 45, 2019, pp. 289-307, doi:10.1016/j.ijinfomgt.2018.08.006.
39. Himeur, Yassine, et al. “Artificial Intelligence-based Anomaly Detection of Energy Consumption in Buildings: A Review, Current Trends, and New Perspectives.” *Applied Energy,* vol. 287, 2021, pp. 1-26, doi:10.1016/j.apenergy.2021.116601.
40. Hira, Zena M., and Duncan F. Gillies. “A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data.” *Advances in Bioinformatics,* no. 2015, 2015, pp. 1-13, doi:10.1155/2015/198363.
41. Huang, Chang-Jiang, et al. “Realization of A Quantum Autoencoder for Lossless Compression of Quantum Data.” *Physical Review,* vol. 102, no. 3, 2020, pp. 1-15, doi:10.1103/PhysRevA.102.032412.
42. Islam, S. N. “ShellAg: Expert System Shell for Agricultural Crops.” *2013 International Conference on Cloud & Ubiquitous Computing & Emerging Technologies*. (pp. 83-6) IEEE, 2013, doi:10.1109/CUBE.2013.24.
43. Kanagavelu, Renuga, et al. “Two-phase Multi-party Computation Enabled Privacy-preserving Federated Learning.” *2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID),* Melbourne, VIC, Australia, 2020, pp. 410-19, doi:10.1109/CCGrid49817.2020.00-52.
44. Kaur, Ramandeep, and Neeru Sharda. “Utilizing ICT to Fight Against Crime: Emerging ICT Tools, Forms of Crime and Its Solutions.” *International Journal of Computer Science and Information Technologies,* vol. 5, no. 6, 2014, pp. 7452-7, www.academia.edu/9328911/Utilizing_ICT_to_Fight_against_Crime_Emerging_ICT_Tools_Forms_of_Crime_and_Its_Solutions.
45. Keshavamurthy, Bharath, and Mohammad Ashraf. “Conceptual Design of Proactive SONs Based on the Big Data Framework for 5G Cellular Networks: A Novel Machine Learning Perspective Facilitating a Shift in the Son Paradigm.” *2016 International Conference System Modeling & Advancement in Research Trends (SMART)*. IEEE, 2016, doi:10.1109/SYSMART.2016.7894539.
46. Keshavarz, Hassan, et al. “The Value of Big Data Analytics Pillars in Telecommunication Industry.” *Sustainability,* vol. 13, no. 13, 2021, pp. 1-38, doi:10.3390/su13137160.
47. Khalid, Samina, Tehmina Khalil, and Shamila Nasreen. “A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning.” *2014 Science and Information Conference,* London, UK, 2014, pp. 372-378, doi:10.1109/SAI.2014.6918213.
48. Kibria, Mirza Golam, et al. “Big Data Analytics, Machine Learning, and Artificial Intelligence in Next-generation Wireless Networks.” *IEEE Access,* vol. 6, 2018pp. 32328-38, doi:10.1109/ACCESS.2018.2837692.
49. Kim, Junhyeong, et al. “5G-ALLSTAR: An Integrated Satellite-cellular System for 5G and Beyond.” *2020 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*. IEEE, 2020, www.ieeexplore.ieee.org/abstract/document/9124751.
50. Kotenko, Igor, et al. “Detection of Stego-insiders in Corporate Networks Based on a Hybrid NoSQL Database Model.” *The 4th International Conference on Future Networks and Distributed Systems (ICFNDS),* 2020, pp. 1-8, doi:10.1145/3440749.3442612.
51. Kremers, Bart JJ, et al. “Two Step Clustering for Data Reduction Combining DBSCAN and K-means Clustering.” *arXiv preprint arXiv:2111.12559* (2021), pp. 1-20, doi:10.48550/arXiv.2111.12559
52. Li, Bing, et al. “Anomaly Detection for Cellular Networks Using Big Data Analytics.” *IET Communications,* vol. 13, no. 20, 2019, pp. 3351-59, doi:10.1049/iet-com.2019.0765.
53. Lipton, Zachary Chase, Charles Elkan, and Balakrishnan Narayanaswamy. “Thresholding Classifiers to Maximize F1 Score.” *arXiv preprint arXiv:1402.1892,* 2014, pp. 1-16, doi:10.48550/arXiv.1402.1892.

54. Liu, Xiufeng, Nadeem Iftikhar, and Xike Xie. “Survey of Real-time Processing Systems for Big Data.” *Proceedings of the 18th International Database Engineering & Applications Symposium*, 2014, pp. 356-61, doi:10.1145/2628194.2628251.
55. López, Victoria, et al. “Cost-sensitive Linguistic Fuzzy Rule Based Classification Systems Under the MapReduce Framework for Imbalanced Big Data.” *Fuzzy Sets and Systems,* vol. 258, 2015, pp. 5-38, doi:10.1016/j.fss.2014.01.015.
56. Lourenço, João Ricardo, et al. "Choosing the Right NoSQL Database for the Job: A Quality Attribute Evaluation." *Journal of Big Data,* vol. 2, no. 1, 2015, pp. 1-26, doi:10.1186/s40537-015-0025-0.
57. Lu, Xiaoyi, et al. “Accelerating Spark with RDMA for Big Data Processing: Early Experiences.” *2014 IEEE 22nd Annual Symposium on High-Performance Interconnects*. *2014 IEEE 22nd Annual Symposium on High-Performance Interconnects*, Mountain View, CA, USA, 2014, pp. 9-16, doi:10.1109/HOTI.2014.15.
58. Maimó, Lorenzo Fernández, et al. “A Self-adaptive Deep Learning-based System for Anomaly Detection in 5G Networks.” *Ieee Access,* vol. 6, 2018, pp. 7700-12, doi:10.1109/ACCESS.2018.2803446.
59. Mardani, Morteza. *Leveraging Sparsity and Low Rank for Large-scale Networks and Data Science*. Dissertation, University of Minnesota, 2015, www.proquest.com/openview/246e157003963021e1c23f6549ac848a/1?pq-origsite=gscholar&cbl=18750.
60. Mason, Andrew, et al. “Online Anomaly Detection of Time Series at Scale.” *2019 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (Cyber SA)*. (pp. 1-8) IEEE, 2019, doi:10.1109/CyberSA.2019.8899398.
61. Mata, Javier, et al. "Artificial Intelligence (AI) Methods in Optical Networks: A Comprehensive Survey." *Optical Switching and Networking,* vol. 28, 2018, pp. 43-57, doi:10.1016/j.osn.2017.12.006.
62. Matsuno, Ivone P., et al. “Aspect-based Sentiment Analysis Using Semi-supervised Learning in Bipartite Heterogeneous Networks.” *Journal of Information and Data Management,* vol. 7, no. 2, 2016, pp. 141-41, www.periodicos.ufmg.br/index.php/jidm/article/view/342.
63. Mdini, Maha. *Anomaly Detection and Root Cause Diagnosis in Cellular Networks*. Diss. Ecole Nationale Supérieure Mines-Télécom Atlantique Bretagne Pays de la Loire, 2019, www.theses.hal.science/tel-02304602/file/2019IMTA0144_Mdini-Maha.pdf.
64. Min, Erxue, et al. “Su-ids: A Semi-supervised and Unsupervised Framework for Network Intrusion Detection.” *Cloud Computing and Security: 4th International Conference, ICCCS 2018, Haikou, China, June 8–10, 2018, Revised Selected Papers, Part III 4*. Springer International Publishing, 2018, doi:10.1007/978-3-030-00012-7_30.
65. Mirzamomen, Zahra, and Mohammad Reza Kangavari. “Evolving Fuzzy Min–max Neural Network Based Decision Trees for Data Stream Classification.” *Neural Processing Letters,* vol. 45, 2017, pp. 341-63, doi:10.1007/s11063-016-9528-8.
66. Nguyen, Thien Duc, et al. “DÏoT: A Federated Self-learning Anomaly Detection System for IoT.” *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019, pp. 1-12, doi:10.1109/ICDCS.2019.00080.
67. Oak, Rajvardhan, et al. “Malware Detection on Highly Imbalanced Data Through Sequence Modeling.” *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security,* 2019, pp. 37-48, doi:10.1145/3338501.3357374.
68. Oest, Frauke, et al. “Evaluation of Communication Infrastructures for Distributed Optimization of Virtual Power Plant Schedules.” *Energies,* vol. 14, no. 5, 2021, pp. 1-20, doi:10.3390/en14051226.
69. Omar, Salima, Asri Ngadi, and Hamid H. Jebur. “Machine Learning Techniques for Anomaly Detection: An Overview.” *International Journal of Computer Applications,* vol. 79, no. 2, 2013, pp. 33-41, www.citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=0278bbaf1db5df036f02393679d485260b1daeb7.
70. Pestian, John, et al. “A Machine Learning Approach to Identifying Changes in Suicidal Language.” *Suicide and Life-Threatening Behavior,* vol. 50, no. 5, 2020, pp. 939-47, doi:10.1111/sltb.12642.
71. Prakash, Andrea, Narem Navya, and Jayapandian Natarajan. “Big Data Preprocessing for Modern World: Opportunities and Challenges.” *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*. Springer International Publishing, 2019, pp. 335-43, doi.org/10.1007/978-3-030-03146-6_37.
72. Purwanto, Yudha et al. “Traffic Anomaly Detection in DDos Flooding Attack.” *2014 8th International Conference on Telecommunication Systems Services and Applications (TSSA)*, Kuta, Bali, Indonesia, 2014, pp. 1-6, doi:10.1109/TSSA.2014.7065953.
73. Rastogi, Shikha, and Jaspreet Singh. “A Systematic Review on Machine Learning for Fall Detection System.” *Computational Intelligence,* vol. 37, no. 2, 2021, pp. 951-74, doi:10.1111/coin.12441.
74. Saito, Takaya, and Marc Rehmsmeier. “The Precision-recall Plot is More Informative Than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.” *PloS One,* vol. 10, no. 3, 2015, pp. 1-21, doi:10.1371/journal.pone.0118432.
75. Salloum, Salman, et al. “Big Data Analytics on Apache Spark.” *International Journal of Data Science and Analytics,* vol. 1, 2016, pp. 145-64, doi:10.1007/s41060-016-0027-9.

76. Shenbakapriya, R., M. Kalimuthu, and P. Sengottuvelan. "Improving Clustering Performance on High Dimensional Data using Kernel Hubness." *International Journal of Computer Applications (IJCA),* 2014, pp. 27-30, www.citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=24b6baee7012126918581503348ed529a84f4c54.
77. Singh, Akhilendra Pratap, et al. "Evolution of Wireless Sensor Network Design from Technology Centric to User Centric: An Architectural Perspective." *International Journal of Distributed Sensor Networks,* vol. 16, no. 8, 2020, pp. 1-24, doi:10.1177/1550147720949138.
78. Šipuš, Danijel. "Big Data Analytics for Communication Service Providers." *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2016, pp. 513-17, doi:10.1109/MIPRO.2016.7522198.
79. Su, Ruoyu, et al. "Resource Allocation for Network Slicing in 5G Telecommunication Networks: A Survey of Principles and Models." *IEEE Network,* vol. 33, no. 6, 2019, pp. 172-9, doi:10.1109/MNET.2019.1900024.
80. Sun, Yuanyuan, et al. "Semi-supervised Deep Learning for Network Anomaly Detection." *Algorithms and Architectures for Parallel Processing: 19th International Conference, ICA3PP 2019, Melbourne, VIC, Australia, December 9–11, 2019, Proceedings, Part II 19*. Springer International Publishing, 2020, 383-90, doi:10.1007/978-3-030-38961-1_33.
81. Syafrudin, Muhammad, et al. "Performance Analysis of IoT-based Sensor, Big Data Processing, and Machine Learning Model for Real-time Monitoring System in Automotive Manufacturing." *Sensors,* vol. 18, no. 9, 2018, pp. 1-24, www.mdpi.com/1424-8220/18/9/2946#.
82. Taherizadeh, Salman, and Marko Grobelnik. "Key Influencing Factors of the Kubernetes Auto-scaler for Computing-intensive Microservice-native Cloud-based Applications." *Advances in Engineering Software,* vol. 140, 2020, pp. 1-11, doi:10.1016/j.advengsoft.2019.102734.
83. Thudumu, Srikanth, et al. "A Comprehensive Survey of Anomaly Detection Techniques for High Dimensional Big Data." *Journal of Big Data,* vol. 7, 2020, pp. 1-30, doi:10.1186/s40537-020-00320-x.
84. Van Heddeghem, Ward, et al. "Power Consumption Evaluation of Circuit-switched Versus Packet-switched Optical Backbone Networks." *2013 IEEE Online Conference on Green Communications (OnlineGreenComm)*. IEEE, 2013, www.ieeexplore.ieee.org/abstract/document/6731029.
85. Vardakas, John S., et al. "Towards Machine-learning-based 5G and Beyond Intelligent Networks: The Marsal Project Vision." *2021 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)*. Athens, Greece, 2021, pp. 488-93, doi:10.1109/MeditCom49071.2021.9647671.
86. Wang, Shi-Yang. "Ensemble2: Anomaly Detection via EVT-Ensemble Framework for Seasonal KPIs in Communication Network." 2022, pp. 1-17, doi.org/10.48550/arXiv.2205.14305.
87. Wu, Dazhong, et al. "A Comparative Study on Machine Learning Algorithms for Smart Manufacturing: Tool Wear Prediction Using Random Forests." *Journal of Manufacturing Science and Engineering,* vol. 139, no. 7, 2017, 1-9, doi:10.1115/1.4036350.
88. Yan, Jun, and Xiangfeng Wang. "Unsupervised and Semi-supervised Learning: The Next Frontier in Machine Learning for Plant Systems Biology." *The Plant Journal,* vol. 111, no. 6, 2022, pp. 1527-38, doi:10.1111/tpj.15905.
89. Yayah, Fauzy Che, Khairil Imran Ghauth, and C. Ting. "Adopting Big Data Analytics Strategy in Telecommunication Industry." *Journal of Computer Science & Computational Mathematics,* vol. 7, no. 3, 2017, pp. 57-67, doi:10.20967/jcscm.2017.03.002.
90. Yen, Ting-Fang, et al. "Beehive: Large-scale Log Analysis for Detecting Suspicious Activity in Enterprise Networks." *Proceedings of the 29th Annual Computer Security Applications Conference*, 2013, pp. 199-208, doi:10.1145/2523649.2523670.
91. Yu, Xianwen, et al. "VAEGAN: A Collaborative Filtering Framework based on Adversarial Variational Autoencoders." In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*. 2019, pp. 4206-12, www.ijcai.org/Proceedings/2019/0584.pdf.
92. Yu, Yinbo, et al. "Fault Management in Software-defined Networking: A Survey." *IEEE Communications Surveys & Tutorials,* vol. 21, no. 1, 2018, pp. 349-92, doi:10.1109/COMST.2018.2868922.
93. Zaslavsky, Arkady, Charith Perera, and Dimitrios Georgakopoulos. "Sensing As a Service and Big Data." 2013, pp. 1-8, doi:10.48550/arXiv.1301.0159.
94. Zhao, Tong, et al. "Action Sequence Augmentation for Early Graph-based Anomaly Detection." *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2021, pp. 2668-78, doi:10.1145/3459637.3482313.
95. Zhao, Ziliang, et al. "Understanding the Bias of Call Detail Records in Human Mobility Research." *International Journal of Geographical Information Science,* vol. 30, no. 9, 2016, pp. 1-26, doi:10.1080/13658816.2015.1137298.
96. Zhou, Yuxun, et al. "Non-parametric Outliers Detection in Multiple Time Series a Case Study: Power Grid Data Analysis." *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 32. no. 1, 2018, pp. 4605-12, doi:10.1609/aaai.v32i1.11632.

97. Zoidi, Olga, et al. "Graph-based Label Propagation in Digital Media: A Review." *ACM Computing Surveys (CSUR,)* vol. 47, no. 3, 2015, pp. 1-35, doi:10.1145/2700381.
98. Zorzi, Michele, et al. "Cognition-based Networks: A New Perspective on Network Optimization Using Learning and Distributed Intelligence." *IEEE Access,* vol. 3, 2015, pp. 1512-30, doi:10.1109/ACCESS.2015.2471178.
99. Al-Madani, Basem, Ahmad Shawahna, and Mohammad Qureshi. "Anomaly Detection for industrial Control Networks Using Machine Learning with the Help from the Inter-arrival Curves." 2019, pp. 1-14, doi:10.48550/arXiv.1911.05692.
100. Angehrn, Zuzanna, et al. "Artificial Intelligence a nd Machine Learning Applied at the Point of Care." *Frontiers in Pharmacology,* vol. 11, no 759, 2020, pp. 1-12, doi:10.3389/fphar.2020.00759.
101. Ayesha, Shaeela, Muhammad Kashif Hanif, and Ramzan Talib. "Overview and Comparative Study of Dimensionality Reduction Techniques for High Dimensional Data." *Information Fusion,* vol. 59, 2020, pp. 44-58, doi:10.1016/j.inffus.2020.01.005.
102. Budholiya, Kartik, Shailendra Kumar Shrivastava, and Vivek Sharma. "An Optimized XGBoost Based Diagnostic System for Effective Prediction of Heart Disease." *Journal of King Saud University-Computer and Information Sciences,* vol. 34, no. 7, 2022, pp. 4514-23, doi:10.1016/j.jksuci.2020.10.013.
103. Cai, Jie, et al. "Feature Selection in Machine Learning: A New Perspective." *Neurocomputing,* vol. 300, 2018, pp. 70-79, doi:10.1016/j.neucom.2017.11.077.
104. Fernando, K. Ruwani M., and Chris P. Tsokos. "Dynamically Weighted Balanced Loss: Class Imbalanced Learning and Confidence Calibration of Deep Neural Networks." *IEEE Transactions on Neural Networks and Learning Systems,* vol. 33, no. 7, 2021, pp. 2940-51, doi:10.1109/TNNLS.2020.3047335.
105. Folch-Fortuny, Abel, et al. "Enabling Network Inference Methods to Handle Missing Data and Outliers." *BMC Bioinformatics,* vol. 16, no. 1, 2015, pp. 1-12, doi:10.1186/s12859-015-0717-7.
106. Ghafoori, Zahra, et al. "Efficient Unsupervised Parameter Estimation for One-class Support Vector Machines." *IEEE Transactions on Neural Networks and Learning System,* vol. 29, no. 10, 2018, pp. 5057-70, doi:10.1109/TNNLS.2017.2785792.
107. Görtler, Jochen, et al. "Uncertainty-aware Principal Component Analysis." *IEEE Transactions on Visualization and Computer Graphics,* vol. 26, no. 1, 2019, pp. 822-31, doi:10.1109/TVCG.2019.2934812.
108. Katuwal, Rakesh, and Ponnuthurai N. Suganthan. "Enhancing Multi-class Classification of Random Forest Using Random Vector Functional Neural Network and Other Oblique Decision Surfaces." *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1-8, doi:10.1109/IJCNN.2018.8489738.
109. Lin, Yi Zhou, Zhen Hua Nie, and Hong-Wei Ma. "Structural Damage Detection with Automatic Feature Extraction Through Deep Learning." *Computer Aided Civil and Infrastructure Engineering,* vol. 32, no. 12, 2017, pp. 1025-46, doi:10.1111/mice.12313.
110. Osman, Musa, et al. "Artificial Neural Network Model for Decreased Rank Attack Detection in RPL Based on IoT Network." *International Journal of Network Security,* vol. 23, no. 3, 2021, pp. 496-503, doi:10.6633/IJNS.202105 23(3).15.
111. Ren, Hansheng, et al. "Time-series Anomaly Detection Service at Microsoft." *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,* 2019, pp. 3009-17, doi:10.1145/3292500.3330680.
112. Wassermann, Sarah, et al. "Vicrypt to the Rescue: Real-time, Machine-learning-driven Video-qoe Monitoring for Encrypted Streaming Traffic." *IEEE Transactions on Network and Service Management,* vol. 17, no. 4, 2020, pp. 2007-23, doi:10.1109/TNSM.2020.3036497.
113. Hariri, Sahand, Matias Carrasco Kind, and Robert J. Brunner. "Extended Isolation Forest." IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 4, 2019, 1479-89, doi: 10.1109/TKDE.2019.2947676.
114. Iftikhar Hussain, A. D. I. L., and Asad Zaman. "Outliers Detetcion in Skewed Distributions: Split Sample Skewness Based Boxplot." Economic Computation and Economic Cybernetics Studies and Research, no. 3, 2020, 279-96, doi:10.24818/18423264/54.3.20.17.
115. Tatbul, Nesime, et al. "Precision and Recall for Time Series." Advances in Neural Information Processing Systems, vol. 31, 2018, 1-11, www.proceedings.neurips.cc/paper/2018/file/8f468c873a32bb0619eaeb2050ba45d1-Paper.pdf.
116. Yang, Fangyuan, et al. "A Hybrid Sampling Algorithm Combining Synthetic Minority Over-sampling Technique and Edited Nearest Neighbor for Missed Abortion Diagnosis." BMC Medical Informatics and Decision Making, vol. 22, no. 1, 2022, 1-14, doi:10.1186/s12911-022-02075-2.
117. Zhou, Lina, et al. "Machine Learning on Big Data: Opportunities and Challenges." Neurocomputing, vol. 237, 2017, 350-361, doi:10.1016/j.neucom.2017.01.026.