“Anomaly Detection in High Dimensional Data with Volume and Velocity Aspects”

Thesis

Submitted for the award of

Degree of Doctor of Philosophy

Discipline: Computer Science and Engineering

by

Upasana Gupta

Enrollment No: MUIT0119038016

Under the Supervision of

Dr. Vaishali Singh

Assistant Professor

Under the Maharishi School of Engineering and Technology

Session 2019-20

Maharishi University of Information Technology

Sitapur Road, P.O. Maharishi Vidya Mandir

Lucknow, 226013

Declaration

I hereby declare that the work presented in this thesis entitled **"Anomaly Detection in High Dimensional Data with Volume and Velocity Aspects"** in fulfillment of the requirements for the award of Degree of Doctor of Philosophy, submitted in the Maharishi School of Engineering and Technology, Maharishi University of Information Technology, Lucknow is an authentic record of my own research work carried out under the supervision of Dr. Vaishali Singh, Assistant Professor, Department of Computer Science and Engineering.

I also declare that the work embodied in the present thesis

i) is my original work and has not been copied from any journal / thesis / book; and

ii) has not been submitted by me for any other Degree or Diploma of any University / Institution.

(Upasana Gupta)

Maharishi University of Information Technology
Lucknow

Supervisor's Certificate

This is to certify that Ms. Upasana Gupta has completed the necessary academic turn and the swirl presented by her is a faithful record of bonafide original work under my guidance and supervision. She has worked on the topic **"Anomaly Detection in High Dimensional Data with Volume and Velocity Aspects"** under the Maharishi School of Engineering and Technology, Maharishi University of Information Technology, Lucknow.

(Dr. Vaishali Singh)

Date:

ACKNOWLEDGEMENTS

As the COVID-19 epidemic got underway, I started working on my PhD, a period that was fraught with uncertainty and challenges. I was able to cross the finish line unimpeded by anything! I am eternally grateful to the amazing individuals who supported and encouraged me on my journey of a lifetime; it was only with their encouragement and guidance that I was able to triumph over the many obstacles I faced and get my PhD.

My deepest gratitude goes out to Dr. Vaishali Singh, who oversaw my PhD studies. I am incredibly grateful for her unwavering support, counsel, and encouragement over this entire journey. She has been an indispensable part of my academic development from the time I began working on my research project until I turned in my thesis, thanks to her constant support and willingness to provide a hand whenever I needed it. The incalculable value she added to my growth is something for which I am eternally thankful.

I am really thankful to my supervisor and the deans of the school of engineering and technology and the department of research at MUIT for their tremendous support and guidance.

Everyone who has helped me along the way, from friends and family to coworkers, has my eternal gratitude for their belief in me and my ability to overcome the obstacles I've encountered and earn my doctorate.

Finally, I want to thank my family from the bottom of my heart for believing in me and being patient with me while I did my study. Their help and support were very important to my success. Especially to my daughter Aira, thank you for everything. This is my PhD paper that I'm giving to you.

(Upasana Gupta)

ABSTRACT

Detecting anomalies in high-dimensional data is very difficult; this is particularly true in industries like healthcare, where datasets are both large and dynamic. As a result, we address these issues directly and improve anomaly detection accuracy by introducing a robust technique in R. The data is quite complicated and non-linear, making it difficult for linear methods like Principal Component Analysis (PCA) to understand. We acknowledge the inherent low-dimensional manifold that often characterizes high-dimensional data and our method incorporates iterative learning techniques to overcome this restriction. This will allow our technique to better decipher the intricate relationship between the factors. Our technique goes beyond only finding abnormalities; it aims to provide a thorough groundwork for data exploration, analysis, and anomaly identification in the future. To begin, we use R packages such as ggplot2, dplyr, and Keras to prepare the data thoroughly, fixing problems like missing values and obtaining visual insights from the dataset. In addition, we use sophisticated statistical methods such as Mahalanobis distance to identify and remove outliers, guaranteeing that the results of the following studies are accurate. A multi-pronged technique integrating Auto-encoders and t-SNE reliably captures complicated interactions among variables, achieving dimensionality reduction—a crucial step in reducing the curse of dimensionality. An Artificial Neural Network called a Multi-Layer Perceptron (MLP) uses these cleaned-up datasets to identify outliers by analyzing rebuilding mistakes. We also investigate several methods, such as Metric Multi-Dimensional Scaling using Artificial Neural Networks, to evaluate how well they operate with complicated and huge datasets, such those seen in healthcare. We provide a viable solution for reliable anomaly identification in high-dimensional data worlds by validating our methodology via rigorous empirical testing and comparison against existing approaches. Our approach is resilient and effective. A number of R tools, such as ggplot2, dplyr, and Keras, assist a painstaking process of data preparation that ensures the dataset is ready for further studies. Missing values are a typical problem in real-world datasets and one of the main things to do during this phase since they might affect the correctness and dependability of the analyses that come after. Skillful use of dplyr's methods allows for the imputation or removal of missing values

according to data type and analytical needs. Moreover, ggplot2 allows for the creation of informative visualizations that provide a more thorough comprehension of the features and distribution of the dataset. As a helpful tool for exploratory data analysis, these visualizations direct the next stages of the analysis pipeline and help find possible trends or patterns. After using the Mahalanobis distance tool—a robust statistical measure—to identify and remove outliers, we may have more faith in the dataset's integrity. To keep the data intact while taking any regional differences or abnormalities into consideration, we modify this procedure to account for outliers peculiar to each nation. You can trust the conclusions that come out of this analysis because of how well the data is prepared. It sets the scene for the rest of the process. We strive to optimize the dataset's utility and insights by paying close attention to detail and using the right tools and techniques. Our goal is to make it easier to find strange things in large amounts of data, especially in very important areas like healthcare where data security is very important. More than that, we extend our study to look at Metric Multi-Dimensional Scaling (MDS) using ANNs. We may evaluate this innovative approach to handling large datasets in other domains, like medicine and winemaking, using this strategy. To improve anomaly detection and get more nuanced insights, we intend to apply artificial neural networks (ANNs) to tap into MDS's inherent pattern-finding skills. By comparing our proposed technique to existing ones, we demonstrate that it outperforms them, suggesting that it has the potential to completely alter the data analysis and anomaly detection landscape. Comparing our method to others and seeing how well it works in other areas and with different sets of data could help us make techniques for finding anomalies in high-dimensional data better. This will help us learn how to utilize it more effectively. By looking into the tricky world of finding outliers in large datasets, our study makes it clear how well a multifaceted approach works in the R environment. We demonstrate the resilience of our approach in handling complicated data structures by combining several pre-processing techniques, sophisticated visualization tools, and the power of artificial neural networks. Particularly relevant in mission-critical industries like healthcare, our technique is built to handle the inherent complexity of datasets marked by massive amounts of data points and fast changes over time. Our approach's better performance and dependability are supported by thorough empirical validation and rigorous comparisons with conventional methodologies. Both the methodology's legitimacy and its potential for broad adoption

across practical applications where reliable anomaly detection skills are crucial are reinforced by this validation. Essentially, our study signals a major step forward in anomaly detection methods, providing a thorough answer that can handle the complex details of contemporary datasets and giving decision-makers priceless information for reducing risk and making educated choices.

CONTENTS

LIST OF ABBREVIATIONS

Abbreviation	Description
ABC	Artificial Bee Colony
ACO	Ant Colony Optimization
ALO	Ant Lion Optimizer
ANN	Artificial Neural Network
AUC-ROC	Area Under Curve -Receiver Operating Characteristic
CNN	Convolutional Neural Networks
CSA	Cuckoo Search Algorithm
DBSCAN	Density Based Spatial Clustering of Applications with Noise
DoS	Denial of Service
EHRs	Electronic Health Records
FA	Firefly Algorithm
FAM	Feature Association Map
FCM	Fuzzy C-Means
FSFS	Feature Selection using Feature Similarity
FSS	Fish School Search
GA	Genetic Algorithms
GANs	Generative Adversarial Networks
GRNN	Generalized Regression Neural Network
GWAS	Genome-Wide Association Studies
GWO	Grey Wolf Optimizer
HM	Harmony Memory
HMM	Hidden Markov Model
HS	Harmony Search
IRLS	Iterative Reweighted Least Squares
KNN	k-Nearest Neighbors
LDA	Linear Discriminant Analysis
LOF	Local Outlier Factor

LUS	Local Unimodal Sampling
MCFS	Multi Cluster feature selection
MDS	Multi-Dimensional Scaling
MLP	Multi-Layer Perceptron
NGS	Next Generation Sequencing
PCA	Principal Component Analysis
PSO	Particle Swarm Optimization
RBF	Radial Basis Function
RD	Reachability Distance
RNNs	Recurrent Neural Networks
ROC	Receiver Operating Characteristic
RPCA	Robust Principal Component Analysis
RSVM	Robust Support Vector Machine
SA	Simulated Annealing
SOM	Self-Organized Maps
SSE	Sum of Squared Errors
SVD	Singular Value Decomposition
SVM	Support Vector Machines
t-SNE	t-distributed Stochastic Neighbor Embedding

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

1.1 Introduction to Anomaly Detection

Anomaly identification is an important part of modern data analysis because it helps find trends or behaviors that don't make sense in datasets. This field has garnered increasing attention due to its relevance in diverse domains like cybersecurity, finance, healthcare, and industrial processes. As we delve into the intricacies of anomaly detection, it is essential to appreciate its historical roots and the evolution of techniques that have shaped its current landscape [1].

1.1.1 Overview of Anomaly Detection:

Anomaly detection, within the realm of data analysis, stands as a sentinel, tirelessly monitoring datasets for aberrations and irregularities. The fundamental premise of this technique is to distinguish patterns which considerably depart from the expected standard. As datasets continue to burgeon in complexity and scale, the need for effective anomaly detection has become paramount [2].

Anomaly detection seeks for data points that significantly differ from the norm. Things like mistakes, strange occurrences, or possible dangers might be signaled by this. Anomalies have the ability to influence several industries, with applications ranging from healthcare and cybersecurity to industrial operations and fraud detection [3].

Various methodologies underlie anomaly detection, ranging from classic statistical approaches to complex algorithms for deep learning as well as machine learning. The desired outcomes of the analytics, the degree of automation, and the kind of data all contribute to the formulation of the best course of action [4].

1.1.2 Importance in Contemporary Data Analysis:

In the contemporary landscape, data analysis is not merely about deciphering trends and patterns; it is equally about identifying anomalies that could signify vulnerabilities, fraud, faults, or emerging issues. Anomaly detection serves as a crucial component of this analytical landscape, ensuring that insights derived from data are not only accurate but also resilient to unforeseen disruptions [5].

The importance of anomaly detection is particularly accentuated in scenarios where the consequences of anomalous events are dire. For instance, in cybersecurity, the detection of unusual network behavior could preclude a potential cyberattack. In finance, identifying fraudulent transactions swiftly is essential for mitigating financial losses. Therefore, anomaly detection not only enhances the robustness of data analysis but also fortifies decision-making processes.

1.1.3 Historical Context of Anomaly Detection:

To make sense of error recognition as it is currently, one must be familiar with its origins. In the early days of statistical analysis, a technique called anomaly detection was developed. This included manually searching for data points that did not match the expected norms. However, with the advent of computing power, particularly during the mid-20th century, the potential for automated anomaly detection began to emerge [6].

Early anomaly detection methodologies often relied on basic statistical measures, like a standard deviation and mean, to identify deviations according to the norm. These methods, although foundational, were limited in their ability to handle complex, high-dimensional datasets that characterize contemporary data environments.

1.1.4 Evolution of Anomaly Detection Techniques:

There are different stages in the history of anomaly detection methods, each marked by changes in data complexity and technology progress.

Traditional Statistical Approaches:

In the initial stages, statistical methods like Z-score analysis and the use of Gaussian distribution played a central role. These methods assumed that normal data followed a specific statistical distribution and flagged deviations based on predefined thresholds. While

effective in certain scenarios, they struggled to adapt to the growing diversity and complexity of modern datasets [7].

Machine Learning Paradigm:

A new era in anomaly detection began at the turn of the century with the advent of machine learning. Density estimation and grouping are two popular unsupervised learning strategies. Part of the reason we were successful was because of new data normalization techniques, such isolation forests and one-class support vector machines (SVM). Everything that deviated significantly from the established standards was deemed extraordinary [8].

Deep Learning and Neural Networks:

The 21st century ushered in a new era with the widespread Using deep learning techniques. Neural networks, particularly Auto-encoders, demonstrated remarkable capabilities in capturing complex patterns and detecting anomalies dealing with data that has several dimensions. Deep learning networks need to self-train to represent hierarchies in order to deal with complicated datasets [9].

Real-time and Streaming Anomaly Detection:

In the contemporary landscape, the emphasis has shifted towards real-time anomaly detection, driven by the increasing volume and velocity of data. Methods such as online learning, adaptive models, and ensemble methods enable systems to adapt dynamically to evolving patterns, ensuring timely detection of anomalies.

1.2 High Dimensional Data

Highly dimensional datasets consist of information like as numerous features or dimensions, exceeding the number of observations. Its challenges include increased computational complexity, visualization difficulties, and sensitivity to noise, Techniques like PCA and t-SNE mitigate these challenges, with applications in genomics, social networks, and image processing [10].

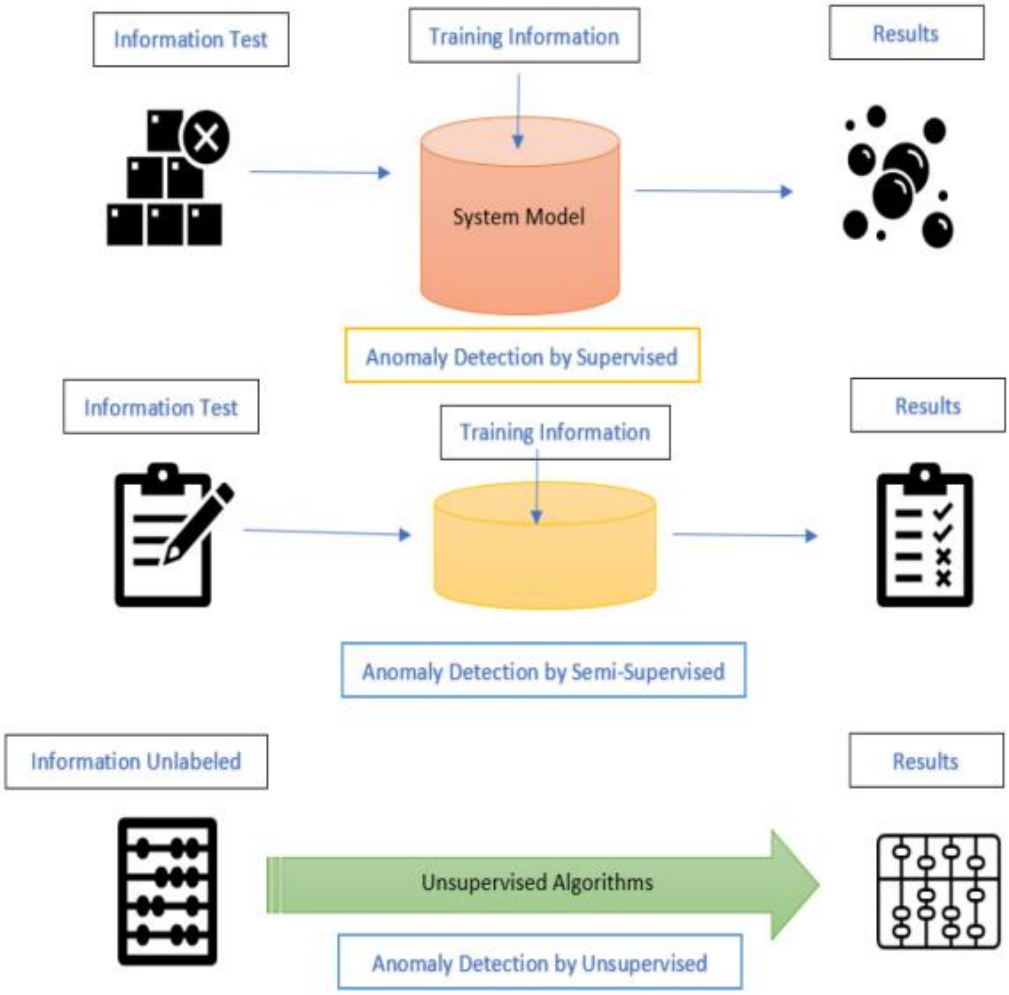


Figure 1.1: Various Techniques of Anomaly detection [4]

1.2.1 Characteristics of High Dimensional Data:

Definition and Dimensions:

High-dimensional data is characterized by an abundance of features, or dimensions, relative to the number of observations. In conventional statistics, Data sets that include a great number of variables are considered highly dimensional. As an example, when it comes to gene expression data, every gene adds another layer of complexity to the dataset. The dimensions often represent different attributes, variables, or features, each providing unique information about the data [11].

Sparsity and Density:

A distinctive feature of high-dimensional data is the prevalence of sparsity. In many real-world datasets, most dimensions or features may have zero or near-zero values, resulting in a sparse matrix. The sparse nature of the data introduces challenges in terms of computational efficiency and statistical modeling. On the other hand, density refers to the distribution of non-zero values across dimensions, impacting the overall structure of the data and influencing the choice of analytical methods [12].

Dangers of Dimensionality:

Complexity and difficulty in data management grow in proportion to the amount of components. This effect is often referred to as the "curse of dimensionality." A larger amount of data is needed to cover the whole feature space as the number of variables grows. This makes things hard in terms of computing power, the ability to understand the model, and the chance of overfitting. The curse of dimensionality underscores the need for specialized techniques to navigate the intricacies of high-dimensional datasets [13].

1.2.2 Dimensionality Reduction Techniques:

Principal Component Analysis (PCA):

Principal component analysis (PCA) is one of many methods for reducing dimensionality. A common way to make sense of large datasets is to break them down into a set of independent factors. This is called principal component analysis (PCA). These components exhibit the highest degree of variability, distilling the essence of the data into a more compact form. Processing efficiency and understanding the patterns are both improved by this simplification. With applications spanning image processing, finance, and biological data analysis, PCA has proven instrumental in simplifying complex datasets [14].

t-Distributed Stochastic Neighbor Embedding (t-SNE):

Outperforming linear methods like principal component analysis (PCA) for low-dimensional data, t-Distributed Stochastic Neighbor Embedding (t-SNE) finds complex patterns. By preserving local similarities between data points, t-SNE provides a nuanced representation, making it particularly effective for tasks such as cluster visualization. Its non-linear nature allows for a more faithful representation of complex relationships, making t-SNE a valuable

tool in exploratory data analysis, especially when dealing with datasets where local relationships are paramount.

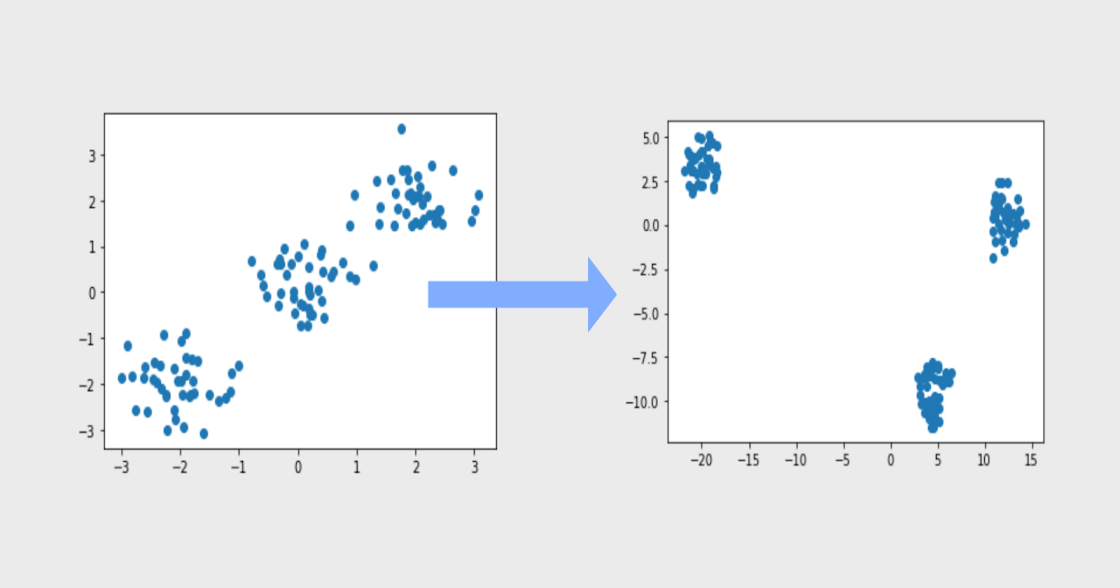


Figure 1.2: Effect of t-SNE [15]

The initial exploration utilizes simple 2-dimensional data clustered around three points to illustrate the transformative impact of t-SNE, laying the groundwork for its application in high-dimensional anomaly detection, as per the thesis.

Feature Selection Methods:

Feature selection addresses the challenge of high dimensionality by identifying and retaining the most informative features. Embracing diverse methodologies such as filter, wrapper, and embedded approaches, feature selection methods assess features based on criteria such as relevance, redundancy, and predictive power. This process is vital in mitigating the curse of dimensionality, enhancing model interpretability, and streamlining subsequent analyses. By strategically choosing relevant features, these methods optimize model performance, making them indispensable in scenarios where computational efficiency and interpretability are paramount [15].

1.2.3 Challenges in Analyzing High Dimensional Data:

Analyzing highly dimensional information presents a myriad of challenges that stem from the intrinsic complexity and richness of these datasets. As we explore these challenges, it

becomes evident that they intertwine, creating a web of intricacies that demand sophisticated solutions.

Increased Computational Complexity:

Highly dimensional information is formed up of many different characteristics or dimensions, poses a significant computational burden. Traditional algorithms, which perform admirably in lower dimensions, exhibit a drastic increase in the number of attributes rises in tandem with the complexity of computers. This steady rise in the need for computing power is sometimes called the "curse of dimensionality" [16].

The computational complexity of algorithms is commonly measured in terms of space and time complexity. In high-dimensional spaces, the number of calculations and the memory required grow exponentially, impeding the efficiency of algorithms. This escalation in complexity necessitates the exploration of specialized algorithms and parallel computing architectures to tackle the computational demands effectively.

Difficulty in Visualization:

Visualizing high-dimensional data poses a substantial challenge due to the limitations of human perception and the inherent intricacies of Multi-Dimensional spaces. While humans can readily interpret three-dimensional representations, extending this understanding to higher dimensions becomes an insurmountable task. As the number of dimensions grows, traditional charting approaches become unworkable and unable to understand the data's structure [17].

Principal component analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE) are two methods that show high-dimensional data in low-dimensional contexts. These solutions work hard to keep important features and connections so that it is easier to explore and understand the data. Still, it's hard to capture all the complexity of the high-dimensional world.

Sensitivity to Noise and Outliers:

High-dimensional datasets are inherently susceptible to environmental noise and outliers, which can significantly impact the reliability of analyses. In sparse datasets, where many dimensions have low variance or are irrelevant, the influence of outliers becomes more

pronounced. Noise, or irrelevant information, can distort the patterns within the data, leading to misguided conclusions and affecting the performance of machine learning models [18].

Sensitivity to noise and outliers is often addressed through robust statistical methods and outlier detection techniques. Robust scientific methods, such as using the median as opposed to the mean, are less influenced by extreme values. Additionally, specialized algorithms for outlier detection, like the Isolation Forest algorithm, identify and mitigate the impact of outliers in high-dimensional spaces. Finding the right balance between the need to find outliers and the need to keep useful data is tricky and takes a deep understanding of the information.

Overfitting and Model Generalization:

In high-dimensional spaces, models are prone to overfitting, a circumstance in which a computer learns noise from training data instead of its fundamental patterns. With an abundance of features, models may capture idiosyncrasies specific for training set, directing to inadequate generalization from fresh, unused data. Overfitting exacerbates the challenge of building models that can robustly handle the complexities of high-dimensional datasets [19].

Techniques such as regularization, which penalize complex models, and cross-validation, which assesses model performance on independent datasets, are crucial in mitigating overfitting. Normalization methods, including L1 and L2 normalization, encourage simpler models by penalizing large coefficients, preventing the model from fitting noise. Cross-validation provides a robust evaluation metric by assessing model performance on data not used during training, confirming that the model can handle new forms of data.

Curse of Dimensionality:

An ever-present issue is the curse of dimensionality, which states that the feature space volume expands rapidly with the total amount of dimensions. This phenomenon results in data sparsity and makes it challenging to obtain a representative sample, leading to difficulties in model training, inference, and accurate representation of the core patterns in the data [20].

Principal component analysis and t-SNE are examples of new methodologies that decrease dimensionality; they are necessary for addressing the curse of dimensionality. Some approaches aim to circumvent the dimensionality issue by decreasing the number of dimensions or eliminating superfluous attributes, enabling more efficient and effective analysis of high-dimensional datasets.

1.2.4 Applications and Impact of High Dimensional Data:

High-dimensional data, characterized by an abundance of features, finds profound applications across diverse domains, revolutionizing the way we analyze and derive insights from complex datasets. The following exploration delves into the technical intricacies and profound impacts within specific realms:

High Dimensionality in Image and Signal Processing:

In signal processing and image processing, high-dimensional data manifests in the form of pixel intensities or signal values across multiple channels or dimensions. The complexity arises from the vast number of pixels or data points, each contributing to the overall dimensionality. Traditional image and signal processing techniques are often insufficient to unravel intricate patterns in such data [21].

High-dimensional representations in image processing enable activities such as picture identification, object identification, and segmentation. Deep learning models have transformed computer vision due to their innate competency with high-dimensional input. In order to handle multi-dimensional pictures, CNNs, or convolutional neural networks, are used. Characteristics, both rising and descending, and attaining Modern accuracy in obligations like as identifying objects and picture grouping.

Genomic Data and Bioinformatics:

Genomic data, with its intricate DNA sequences and genetic variations, exemplifies high-dimensional data in the biological realm. Each gene, nucleotide, or genetic marker contributes to the high dimensionality, presenting a complex landscape for analysis. Analyzing these datasets requires specialized techniques to discern meaningful patterns and variations [22].

High-dimensional analysis of genomic data is pivotal in deciphering the genetic basis of diseases, identifying biomarkers, and enabling personalized medicine. Bioinformatics leverages like next-generation sequencing (NGS) and GWAS, which stands for genome-wide association studies, to unravel the complexities within high-dimensional genomic datasets. Advances in high-throughput technologies generate vast datasets, requiring sophisticated computational tools for meaningful interpretation [23].

Social Networks and Big Data Analytics:

Social networks generate copious amounts of data, encompassing user interactions, preferences, and network connections. The high dimensionality arises from the multitude of factors influencing user behavior, such as likes, comments, and relationships. Analyzing social network data involves navigating this intricate high-dimensional space to uncover patterns, trends, and anomalies [24].

Big data analytics, powered by high-dimensional data, drives insights in social networks. Graph-based models, representing relationships between users, nodes, and edges, form a high-dimensional network structure. Analyzing this structure aids in tasks like community detection, influence propagation modeling, and recommendation systems. Machine learning algorithms, applied to high-dimensional social data, enhance targeting in advertising, sentiment analysis, and social network mining.

Emerging Techniques:

Emerging techniques in these domains include manifold learning, which captures the non-linear structure of high-dimensional data, and graph-based approaches that model complex relationships. In image processing, Generative Adversarial Networks (GANs) generate high-dimensional data, enabling tasks such as image synthesis. In genomics, just one cell sequencing of RNA presents both possibilities and limitations in analyzing high-dimensional single-cell data, unlocking insights into cellular heterogeneity.

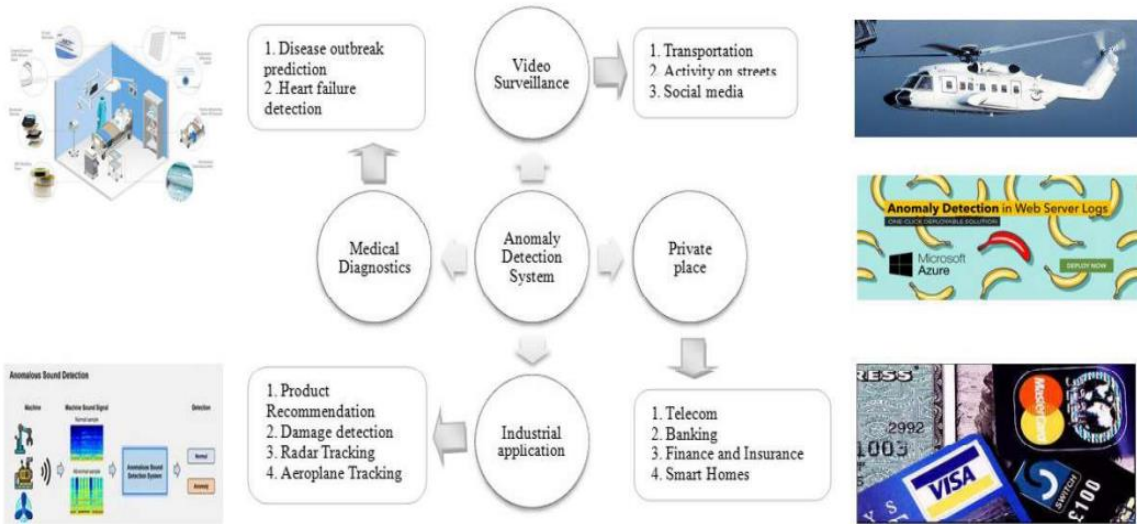


Figure 1.3: Application of Anomaly Detection System [24]

1.3 The Significance of Volume in Data Analysis

The volume of data is both essential and revolutionary in the area of data analysis. As datasets continue to grow in scale and complexity, the significance of volume becomes increasingly pronounced. The very essence of data analysis is intertwined with the volume of information it processes, impacting the effectiveness of algorithms, the scalability of solutions, and the depth of insights derived [25].

1.3.1 Volume as a Crucial Aspect in Anomaly Detection:

Anomaly detection stands at the forefront of data analysis, where its efficacy is intimately tied to the volume of data under scrutiny. In this exploration, We also examine the importance of anomalies in the context of large data sets, elucidating the intricacies involved in identifying irregular patterns or outliers [26].

Leveraging Statistical and Machine Learning Techniques:

Clustering:

When it comes to the identification of anomalies, clustering techniques play a pivotal role in grouping similar data points. With an ever-increasing data amount, clustering techniques such as k-means and clustering with hierarchy aid in the discovery of typical patterns and groups. The fact that anomalies tend to cluster together in unique ways makes them easier to spot when they deviate from the norm [27].

Dense Regions and Density Estimation:

Understanding the density of data points within a specific region is crucial for anomaly detection. Density estimation techniques, such as kernel density estimation, provide insights into the concentration of data. High-volume data makes it possible to evaluate the normal density distribution, enabling the identification of deviations that fall outside expected concentrations [28].

Supervised Learning Algorithms:

Using tagged datasets for training algorithms with supervision, leverage the abundance of data to recognize patterns that has been related to normal behavior. The volume of data serves as a rich source for model training, enabling algorithms to generalize and differentiate between normal and anomalous instances. Two methods that work well in these situations are Random Forest and Support Vector Machine (SVM) [29].

Building Robust Models:

The role of volume extends beyond the mere identification of anomalies; it is integral to the construction of robust models capable of adapting to the complexities of high-dimensional datasets. The volume serves as the training ground for these models, allowing them to encapsulate the nuances of normal behavior and, in turn, identify deviations with a heightened level of accuracy [30].

1.3.2 Real-world Implications of High Volume Data in Anomaly Detection:

Application and Case Studies:

The importance of anomaly detection in improving decision-making and tackling crucial difficulties is shown by its widespread applicability across several areas. Examining case studies provides a nuanced understanding of how high-volume data plays a pivotal role in shaping real-world outcomes [31].

Cybersecurity:

In the realm of cybersecurity, where the digital landscape is fraught with threats, the capacity to instantly evaluate massive amounts of network data is essential. Identification of anomalies systems become the frontline defense, scrutinizing network traffic patterns to

identify deviations that may indicate potential cyber threats. A compelling case study in this domain showcases the real-time analysis capabilities, allowing for prompt responses to security incidents and safeguarding critical systems and data. The sheer volume of data, comprising diverse sources and patterns, necessitates sophisticated anomaly detection algorithms capable of swift and accurate identification.

Financial Institutions:

Financial institutions operate in a fast-paced environment with high-frequency trading generating vast amounts of data. Anomaly detection becomes indispensable in identifying potentially fraudulent activities and ensuring the integrity of financial markets. The case study illuminates how anomaly detection systems analyze trading patterns at high speeds, discerning irregularities that may signify fraudulent activities, market manipulation, or unauthorized transactions. Big data analysis is essential for the safety of financial systems in this context due to the high velocity and amount of data [32].

Healthcare Systems:

In healthcare, the utilization of anomaly detection is transformative, particularly in handling voluminous patient data. Electronic Health Records (EHRs), imaging data, and patient monitoring systems generate extensive datasets. Anomaly detection assists in identifying unusual patterns, facilitating early disease detection, tailoring personalized treatment plans, and ultimately improving overall patient care. The healthcare case study underscores how high-volume data analysis contributes to proactive healthcare management, showcasing the potential for anomaly detection to revolutionize patient outcomes through timely and precise interventions [33].

Technical Considerations:

The success of anomaly detection systems in real-world applications relies heavily on intricate technical considerations. High-volume data poses unique challenges that necessitate advanced techniques and frameworks for accurate and efficient anomaly identification.

Parallel Process and Distributed Computing:

Handling the velocity and scale of high-volume data demands sophisticated computational approaches. Parallel processing and distributed computing frameworks, exemplified by

systems like Apache Sp**ing**ark, play a crucial role. These frameworks enable the parallel execution of data analysis tasks across multiple nodes, expediting the identification of anomalies. The technical insight highlights the importance of efficient computational processing in managing vast datasets and guaranteeing anomaly identification systems' scalability [34].

Streaming Analytics Frameworks:

In applications requiring real-time anomaly detection, streaming analytics frameworks become indispensable. These frameworks process data in motion, allowing for immediate analysis and response to anomalies as they occur. Technologies like Apache Flink showcase the seamless processing of streaming data, ensuring timely identification and mitigation of irregularities. Continuous, real-time analysis is a crucial technological component for satisfying the demanding needs of applications that need instant reactions to abnormalities [35].

Domain-Specific Knowledge Integration:

Adding domain-specific information to systems that look for strange behavior is an important but often ignored part. Understanding the intricacies of the data domain enhances the accuracy of anomaly identification. Domain-specific knowledge allows for the customization of algorithms and models, ensuring they adapt to the unique characteristics and challenges posed by different datasets. The technical insight underscores the importance of a nuanced understanding of the data domain to optimize anomaly detection systems for diverse real-world applications [36].

1.4 Venturing into Multi-Dimensional Scaling (MDS)

With its fresh take on how to portray and understand data with a lot of dimensions, Multi-Dimensional Scaling (MDS) is a formidable tool for data analysts. The more we learn about MDS, the clearer it is that this technique is useful for more than just dimensionality reduction. That this fits with the main points of the thesis about finding outliers in datasets with a lot of variables sounds good.

1.4.1 Understanding Multi-Dimensional Scaling:

At its core, Multi-Dimensional Scaling is a statistical technique used for visualizing the pairwise dissimilarities or distances between a set of objects. These objects can represent a wide array of entities, from high-dimensional information to word semantic connections. The main objective of MDS is to map these entities onto a lower-dimensional space with as many bilateral variations as possible.

Multi-Dimensional Scaling (MDS) is an excellent tool for making high-dimensional data's complex and abstract relationships more comprehensible. As datasets become increasingly intricate and voluminous, the ability to condense information without sacrificing essential patterns becomes crucial. Through the process of capturing the data's underlying structure, MDS achieves this goal and offers a visual representation that aids in analysis and comprehension [38].

1.4.2 Relevance to High-Dimensional Data Analysis:

Dimensionality Reduction:

One of the primary applications of Multi-Dimensional Scaling (MDS) lies in its role as a dimensionality reduction technique, a crucial endeavor in the realm of highly-dimensional data. The curse of dimensionality, characterized by the challenges posed by vast and intricate datasets, necessitates effective strategies to condense information into a more manageable form. MDS accomplishes this by compressing the information into a smaller-dimensional space. By doing so, it retains the essential relationships between data points, allowing for a more intuitive understanding of the underlying structure of the dataset. When discussing the identification of anomalies, where discerning patterns can be challenging amidst the complexity of highly-dimensional data, MDS serves as an instrumental tool for simplifying and clarifying the dataset's structure [39].

Visualization of High-Dimensional Relationships:

High-dimensional datasets often conceal intricate relationships that are difficult to discern in their raw, original form. Multi-Dimensional Scaling (MDS) eliminates this issue by reducing the data to a smaller-dimensional plane, thereby facilitating the visualization of these complex relationships. This capability becomes very important for the field of anomaly

detection, where identifying irregular patterns and deviations is paramount. The visual insights gained through MDS aid analysts in recognizing anomalies that may remain obscured in the original highly-dimensional space. The ability to visually explore and interpret the data's relationships enhances the effectiveness of anomaly detection, allowing for a more thorough understanding of the dataset's nuances [40].

Dissimilarity Metrics for Anomaly Detection:

Anomalies inside a dataset are often characterized by their dissimilarity to normal patterns. MDS relies on dissimilarity metrics to create a representation that preserves these relationships. In the circumstances of anomaly identification in high-dimensional dataset, MDS can be leveraged strategically to highlight anomalies by emphasizing the deviations in dissimilarities. This nuanced approach enhances the sensitivity of anomaly detection systems, particularly when anomalies manifest as subtle deviations within complex relationships. By incorporating dissimilarity metrics, MDS contributes to the nuanced identification of anomalies, ensuring that both subtle and significant deviations from the norm are captured and considered in the anomaly detection process.

Integration with Machine Learning Algorithms:

Anomaly detection and dimensionality reduction are both improved by MDS because of its easy interaction with ML techniques. This integration is pivotal in addressing the challenges posed by high-dimensional datasets with both volume and velocity aspects. Machine learning models, when applied to the transformed, lower-dimensional space generated by MDS, gain enhanced discernment capabilities. The integration allows these models to more effectively identify patterns indicative of anomalies. Machine learning algorithms can adapt to reduced dimensionality with the support of MDS's compressed space, all while keeping vital data for anomaly detection. This all-encompassing method offers a complete solution to the problems caused by high-dimensional datasets, which is useful for anomaly detection systems [41].

1.4.3 Opportunities in Applying MDS to Anomaly Detection:

Computational Intensity:

The application of Multi-Dimensional Scaling (MDS) to high-dimensional data introduces a significant computational challenge, primarily stemming from the sheer volume of data

points and dimensions involved. The computational intensity arises as MDS processes and transforms these vast datasets, placing a strain on computational resources. To address this challenge, optimization of algorithms becomes paramount. Techniques such as algorithmic enhancements, parallelization of computations, and the exploration of distributed computing frameworks are essential. Optimizing the efficiency of MDS algorithms ensures that they can scale effectively to tackle the computational demands of anomaly identification applications in highly-dimensional datasets. This optimization contributes to the seamless integration of MDS within anomaly detection frameworks, enhancing both speed and scalability [42].

Handling Velocity in Data Streams:

The velocity aspect of high-dimensional data, especially in dynamic streaming environments, presents a unique set of considerations for the application of MDS. Adapting MDS to real-time data streams requires innovative approaches that balance the necessity for immediate anomaly detection with the inherent constraints of computational resources. Incremental MDS, a technique that updates the lower-dimensional representation as new data arrives, is a promising avenue. Additionally, dynamic adjustments to dissimilarity metrics can enhance MDS's adaptability to the velocity dimension, ensuring that anomalies are identified in a timely manner without compromising computational efficiency. Successfully addressing the challenges posed by high velocity in data streams opens up opportunities for real-time anomaly detection, aligning MDS with the demands of contemporary high-dimensional datasets [43].

Robust Dissimilarity Metrics:

The efficiency of MDS in anomaly identification heavily hinges on the choice of dissimilarity metrics. Robust dissimilarity metrics are essential for capturing the nuanced relationships present in high-dimensional data accurately. Research and experimentation are crucial in identifying and developing dissimilarity metrics that align with the unique characteristics of the dataset under consideration. Robust metrics contribute to the creation of lower-dimensional representations that faithfully preserve the dissimilarities critical for anomaly detection. By exploring and refining dissimilarity metrics, MDS can adapt to

diverse datasets, ensuring an accurate representation that enhances the sensitivity of anomaly detection systems to both subtle and significant deviations from normal patterns [44].

Interpretability of Lower-Dimensional Embedding:

While MDS excels at visualizing relationships in lower-dimensional spaces, ensuring the interpretability of these embedding essential for efficient identification of anomalies. Anomalies detected in the transformed space must be relatable back to the original high-dimensional context, requiring the development of methodologies for interpretation. This involves creating bridges between the lower-dimensional representations and the features in the original space. Collaborative efforts between data scientists, domain experts, and analysts are instrumental in defining meaningful interpretations of anomalies. Proving that low-dimensional anomalies lead to higher-dimensional repercussions is crucial for MDS to function as a reliable anomaly detector. Doing so may lead to an expansion of MDS's potential use cases [45].

1.5 Thesis Organization

The following list of the chapter include discussion of the thesis report arrangement.

- **Chapter 2: Literature Review**

The current literature on computer anomaly detection is reviewed in this introductory chapter, which touches on key concepts, theories, and methodologies. It lays the groundwork for the subsequent chapters by synthesizing relevant literature, identifying gaps, and establishing the context for the research.

- **Chapter 3: Secure Anomaly Detection in Computing:**

Keeping with the themes covered in the literature review, this chapter elucidates the conceptual framework and theoretical underpinnings of secure anomaly detection in computing. It defines the scope of the research, outlines key objectives, and presents a comprehensive review of relevant security measures.

- **Chapter 4: Enhanced Anomaly Detection Pipeline:**

The suggested improved anomaly detection pipeline is introduced in this chapter, which forms the core of the thesis. It describes the new hybrid approach that combines k-Nearest

Neighbours with the Local Outlier Factor to enhance anomaly detection. An exhaustive description of the algorithm's structure and details is provided in this chapter.

- **Chapter 5: Result and Discussion:**

A critical phase in the thesis, this chapter presents the results obtained from applying the enhanced anomaly detection pipeline to real-world datasets. It is given in detailed about analysis of the outcomes, showcasing the algorithm's adaptability to various contexts. The discussion section critically interprets the results, addressing any challenges encountered and comparing them with existing methods.

- **Chapter 6: Conclusion and Future Scope:**

Key points and conclusions are summarized in this chapter, which is included at the end of the thesis. This part gives a synopsis of the research, talks about how important it is, and recommends where we may go from here in terms of trustworthy computer anomaly detection. This organized book takes the reader step-by-step through each stage as it describes the issue, suggests a better solution, tests that solution, and finally reports on the outcomes.

1.6 Objective

I. To provide a comprehensive review to the Challenges of Anomaly Identification in context of highly dimensional dataset

II. To review current strategies in dealing with the above problems.

III. To outline Anomaly Detection Model addressing the difficulties of Anomaly Detection Methods with relation to large data and high dimensionality issues in a thorough manner.

Chapter 2

Literature Review

2.1 Introduction

The burgeoning field of anomaly detection, crucial in contemporary data analysis, has witnessed extensive research efforts to enhance its efficacy. Success hinges on achieving high detection accuracy and minimizing false positives, a challenge exacerbated by the presence of irrelevant and redundant features. Recognizing this, anomaly detection schemes employ diverse strategies, ranging from spanning traditional ML to cutting-edge DL, robust optimization to reduced dimensionality [46].

When dealing with growing datasets, anomaly detection specialists often look for ways to reduce the total amount of dimensions. Consider t-Distributed probabilistic neighbor embedding and principal component analysis (PCA) as two such examples simplify datasets by reducing them to lower-dimensional representations. This aligns closely with the thesis's focus on anomaly identification in highly-dimensional data, emphasizing the volume and velocity aspects of the information.

To bolster anomaly detection capabilities, the application of robust optimization approaches becomes paramount. These methodologies, including robust statistical estimators and outlier-resistant models, aim to create algorithms resilient to variations and outliers within dynamic datasets. This resilience is particularly critical when dealing with high volume and velocity, mirroring the challenges addressed in the thesis [47].

Support vector machines (SVMs), isolation forests, and one-to-one neural networks (k-NNs) are common machine learning approaches used for outlier detection. In addition to adding to the existing literature on flexible anomaly detection methods for high-dimensional data, this

research demonstrates how these techniques work for identifying trends and outliers in datasets.

Deep learning's RNNs, Auto-encoders, and neural networks have brought about a new age in anomaly identification. To tackle the fast-paced nature of high-dimensional data streams, these methods are great at extracting complex patterns from massive, ever-changing datasets. The evolving role of deep learning in anomaly identification or detection aligns with the thesis's objectives, focusing on the volume and velocity challenges inherent in contemporary datasets.

In navigating the extensive literature on anomaly detection, it becomes evident that the insights gained from diverse approaches contribute to a broader understanding of how anomalies can be effectively identified and interpreted. Our investigation into the difficult problem of anomaly detection in high-dimensional data lays the groundwork for research on flexible anomaly detection systems that can keep up with the exponential increase in data volume and velocity in the future.

2.2 Dimensionality Reduction

Current learning approaches have been greatly tested by the current explosion of data dimensionality. The proliferation of features in datasets often causes issues like over-and under-fitting, adversely affecting the overall performance of models. To mitigate these challenges and enhance learning performance accuracy, computational challenge, storage efficiency, and model comprehension, researchers have extensively explored dimensionality reduction techniques. These techniques are broadly categorized into Extracting characteristics and selecting features [48].

Relocating features to a distinct location simplifies the collecting process during feature extraction. Methods like principal component analysis (PFA), linear discriminant analysis (LDA), and singular value decomposition (SVD) fall under this umbrella. Approaches to feature selection seek to increase the relevance of features to the set of characteristics under consideration and remove duplicate features in order to decrease the number of features that need to be examined.

Examples of feature selection methods encompass Relief, Information Gain, Fisher Score, and Chi Squares. Subset selection, feature weighing, and ranking are the three most popular feature selection processes based on the results they provide. A ranked set of features is returned by feature weighting after subset selection, which involves assigning weights to features.

All Conventional feature selection strategies consist of four steps: creating a subset, evaluating it, establishing end criteria, and validating the results. In the subset creation phase, a candidate subset is initially chosen. Subsequently, the generated subset is evaluated using specific criteria, and we will not stop until we achieve our objective, which is this operation. After deciding which solution best meets the assessment criteria, the last step is to do cross-validation utilizing domain knowledge or a validation set [49].

To assist in connecting between filter and wrapper models, hybrid approaches have emerged, striving for a harmonious balance between efficiency, identical to filter models; precision, comparable to wrapper models. These models integrate choosing characteristics according to specified cardinality and target classification accuracy as measured by statistical measures. Illustratively, Dash and Liu highlighted the pivotal role of feature selection in clustering, emphasizing its impact on the performance of clustering algorithms sensitive to data dimensionality. Their use of the RANK algorithm showcased the identical to filter models; precision, comparable to wrapper models.

In a similar vein, Nguyen and colleagues introduced a Data stream mining using an enhanced feature selection method based on accelerated PSO [50]. Performance evaluations were conducted using big data characterized by high dimensionality. A clustering-based approach to decreasing dataset dimensionality was introduced by Duan and colleagues. It uses a partial distance algorithm and hierarchical feature selection as its primary tools. Experiments comparing the suggested technique to other cutting-edge algorithms proved its viability.

The diversity of techniques in the literature underscores the critical role of feature selection. However, determining the effectiveness of such methods can be complex without understanding the relevance of features. Consequently, ongoing developments in feature selection methods aim such that unnecessary and redundant attributes may be removed from

high-dimensional datasets. Strategies include combinations of multiple feature selection methods, ensemble methods, and the restructuring of existing schemes [51].

Specifically, these enhanced feature selection capabilities are most noticeable in the "Anomaly Detection in High Dimensional Data with Volume and Velocity Aspects" configuration. The effective curation of features is crucial for enhancing anomaly detection algorithms' performance, especially in high-dimensional datasets characterized by volume and velocity challenges. The exploration of hybrid models and their demonstrated efficacy in optimizing feature selection aligns seamlessly with the overarching goals of the thesis.

Table. 2.1: Evaluating Recent Feature Selection Schemes for Comparison**.**

Author [Citation]	**Methodology**	**Comparison**	**Contribution**
Frouzan Rashidi et al. [52]	Cluster-dependent feature-weighing mechanism	Two approaches to feature selection: multi-cluster analysis and similarity-based feature selection	A proposed unsupervised feature selection algorithm suggests removing features with comparatively low weights to enhance both running time and data visualization.
Keyu Liu *et al.* [53]	A partly supervised learning approach allows us to increase relevance while avoiding duplication.	Fisher score, mRMR, Lapla-cian score, locality sensitive, and sSelect	A feature selection scheme based on semi-supervised learning suggested a way to enhance the balance between classification functioning and computational cost.
Peng Chu *et al.* [54]	Feature Association Map (FAM),	Several benchmark feature selection algorithms like CFS, mRMR, DSCA, Laplacian, PFA and DSUB algorithms	A graph-based anomaly detection technique has been proposed, applicable to both supervised and unsupervised classification scenarios.
Francesca Giardini *et al.* [55]	Making choices inside and across categories using computers	Methods such as-MLNB, MDDMspc, MDDMproj, NFNMI, PMU, and RF-ML	A proposed technique for selecting features from a pool of labels addresses streaming features by taking into account intrinsic group structures.

Mehrdad Rostami *et al.* [56]	Two categories may be used to categorise the significance of attributes: options.	Benchmark feature selection techniques such as-DRGS, JMIM, mRMR, DISR and IG	Proposing DRJMIM, a feature selection method, addresses crucial issues of feature relevance and mitigating misinterpretation, addressing both problems efficiently for improved model performance.
Zhengxin Li *et al.*[57]	$l_{2,p}$-norm regularization item	These days, feature selection algorithms like SFUS, MDMR, PMU, and MDDM are only a few among many.	Proposing features using IRLS (Iterative Reweighted least squares) and eliminate noise for enhanced feature selection.
RS Latha *et al.*[58]	Dependency Margin approach and Greedy search algorithms	ReliefF, dependency, con-sistency, information gain, FCBF, CFS-FS, and IRelief,	Proposing a subset selection algorithm for feature selection that assesses the relevance of both selected and remaining features to improve prediction accuracy.
Vijay Kumar, et al. [59]	Firefly Algorithm (FA), Chaotic maps and Simulated Annealing (SA)	The methods include eleven standard search algorithms for classification and ten improved versions of those algorithms.	Proposing an FA variant that identifies optimal features for classification and regression models, enhancing model performance.

2.3 Optimization Schemes

In the last few years, many optimization schemes that use more than one way have been created to help with the problem of finding strange things in big databases. These schemes include, but are not limited to, particle swarm optimization (PSO), linear programming, A* search, random trees, GA, and many more. A new approach to anomaly identification was introduced by Ghanem et al., which fused GA with meta-heuristic methods, effectively addressing issues related to local optima and robust search. While their approach demonstrated commendable accuracy in detector generation, the potential for enhancing adaptability through dynamic parameter optimization remains a viable avenue for future exploration.

Xuan-Nam Bui and Pirat Jaroonpattanapong took a novel approach by combining An approach to intrusion detection that integrates PSO, SVM, and K-Nearest Neighbours [60]. This innovative integration, coupled with ensemble methods for final classification decisions, showcased promising results. A combination of PSO and Self-Organized Maps (SOM) was also described by Shahreza et al. as an unsupervised anomaly identification approach.

In the realm of proactive prediction, Wuyi Ming and Fan Shen developed a hybrid technique that utilized multi-objective optimization and PSO for predicting Denial of Service (DoS) cyberattacks on certain data networks [61]. Additionally, they integrated the PSO and K-means algorithms to identify anomalies in content-centric networks, employing a hybrid approach in the training phase for optimal cluster determination. Yue Li et al. demonstrated a division method using k-means clustering and dynamic PSO, enhancing the worldwide search functionality of k-means in the context of anomaly detection [62].

Addressing challenges in clustering, Marco Dorigo et al. introduced an Ant Colony Improvement (ACO)-focused methodology, emphasizing adjusted compactness and relative separation as crucial objective functions [63]. The proposed approach aimed to enhance scalability and address issues such as neighborhood development, dataset reduction, and solution assessment.

Furthermore, the significance of optimization-based pattern recognition in data mining has increased. Researchers have effectively used clustering methodology and optimization techniques to increase productivity and accuracy. Absalom E Ezugwu et al.'s review of PSO-based clustering techniques highlighted their superiority over traditional methods, particularly in overcoming local optima problems [64]. However, the full potential of automation, generalization, and broader applicability of these techniques remains an area warranting further exploration.

Table 2.3 provides a contrast of popular optimization techniques, offering valuable insights into the diverse landscape of approaches. This exploration of optimization schemes aligns seamlessly with the overarching theme of "Anomaly Detection in High Dimensional Data

with Volume and Velocity Aspects," underscoring the critical need for robust and efficient techniques in tackling anomalies within large and dynamic datasets.

Table 2.2: Exploring Anomaly Detection Optimization Schemes

Author [Citation]	Inspiration	Behaviour	Optimization Steps
Krzysztof Miler et al.'s [65]	Imitates the natural speech pattern of antlions.	The ability of ants to dig	Ants that scuttlebutt, lay traps, and entangle
Mohammad H Nadimi-Shahraki et al.'s [66]	The gray wolves' hunting technique	Compute shrunken circle for position	encircling, chasing, and attacking the victim
Selcuk Aslan et al.'s [67]	A hive of astute honey bees searches for food.	Seek for foods high in honey.	Inspect, use, apply, and dispose of
Longwu Wang et al.'s [68]	How certain species of CCKOO bird parasites procreate	Reduce the possibility that they'll give up on their	Cuckoo Search Algorithm (CSA) [109]
Pablo Arechavala-Lopez et al.'s [69]	Fish species' social interactions	Collective behaviour that increases mu Collective activities that enhance mu-	Fish School Search (FSS) [111]
Yi-Jen Shih et al.'s [70]	Themes with melody	The interaction of harmonics	Harmony Memory (HM) Equipping

2.4 Machine Learning Approaches

Techniques for detecting anomalies are widely used in a variety of fields, including prediction, grouping, and data categorization. Techniques for classification and clustering are often used. The applicability of clustering techniques lies in providing deep insight Don't comprehend the labels before giving the data. On the other hand, the greatest results from classification techniques come from labeled data. Among the most popular approaches to object grouping are k-means, k-medoids, DBSCAN, and OPTICS. On the other hand, support vector machines, naïve bayes classifiers, and decision trees are often used for classification tasks.

The K-means technique, though straightforward, encounters challenges like local convergence and cluster initialization problems. Researchers, inspired by K-means, have

developed alternatives such as K-medoid and unclear C-means (FCM). K-medoids algorithm, emphasizing medians over centroids, proves more robust but involves higher computational costs. FCM, effective in cluster overlapping scenarios, assigns each element to the cluster with the highest membership grade. Careful evaluation of dataset patterns is crucial in selecting the appropriate clustering technique [71].

Researchers, including Pruthvi Raju Garikapati and K-means and K-medoid are two partitioning strategies that Balamurugan studied [72]. They came to the conclusion that the k-medoid method is more reliable than the k-means partitioning strategy. The Ranked k-medoids technique was developed by Anton V. Ushakov et al. to cluster huge datasets, demonstrating efficiency in speed and accuracy on large datasets [73]. Sahil Garg et al introduced a The k-medoid clustering strategy with variance improvement was verified using an online clustering system [74].

DBSCAN, known for detecting clusters of various sizes and shapes, faces challenges such as parameter evaluation and border point detection. Alternatives, like the graph-based technique proposed by Jesus Maillo et al address these challenges by accelerating the nearest neighbor operation [75]. Lihua Hu et al. used an approach based on secure, geo-aware hashing algorithm to tackle the closest neighbor search issue in DBSCAN [76]. Wenhao Lai et al. enhanced the clustering performance of the original DBSCAN method by creating a new DBSCAN approach [77]. This was done to tackle the issue of border point recognition of neighbouring clusters.

Emrah Hancer presented a clustering method that combines differential evolution and k-means, finding solutions that have less Sum Squared Errors (SSE) than similar ones [78]. When it came to average anomaly identification precision and test accuracy, Kareth M. Leon-Lopez et al. discovered that the Hidden Markov Model (HMM) fared the best of all the algorithms they tested [79]. An intrusion detection system developed by AKM Iqtidar Newaz et al. mimics regular sequences using look ahead pairs and contiguous sequences [80]. By testing their Artificial Neural Network (ANN)-based anomaly detection technique using DARPA datasets, Dukka Karun Kumar Reddy et al. were able to boost detection rates [81]. Hamid Darabian et al. compared textual content with a sequence of system calls to determine whether new applications were invasive or not, the suggested method combined an RSVM

with a KNN classifier [82]. Lingqiang Xie et al. were the first to show how to use progressive grid clustering and discrete time-sliding windows to find anomalies without any human help [83]. Shangjia Dong et al. looked at how mistakes affect each other in their study to find problems in a power grid and information network that work together [84]. The work of Makgabo Johanna Mashala et al. suggests using graphs to find strange things in real-life hyperspectral pictures [85].

Numerous strategies are used to investigate odd changes in network traffic when there are anomalies in functioning networks that might cause network disruptions. Jiaming Pei and associates discovered anomalies in high-dimensional, large-scale datasets by using network-wide traffic analysis [86]. The anomaly detection problem was studied by looking at large-scale communication networks. This research combines an empirical mode decomposition approach with the wavelet transform methodology to identify flaws in multimedia medical equipment. The wavelet transform was used in order to extract the multi-scale properties of anomalies in high-speed network data. The quantity of network traffic was ascertained by means of time frequency analysis. An estimate of the flow of network traffic in large networks was provided by another research. Using Generalized Regression Neural Networks (GRNNs), this study presents a traffic matrix estimate approach. The suggested approach demonstrated robust model performance in simulations based using actual Abilene network data.

Marcus Stoffel and colleagues devised a method based on a rapidly converging radial basis neural network functional approach [87]. In order to quickly and accurately detect network anomalies, the suggested method outlined the characteristics of an intrusion. A technique for detecting intrusions using support vector machines was created by Shamis N. Abd et al [88]. In this method, learning vector samples were used to efficiently decrease the training samples. As input, the intrusion detection system received this condensed training set, which reduced computing cost. An anomaly identification technique based on clustering “K-means” and Naive Bayes has been presented by Meenal Jain et al. [89].

To find bugs, Syed Fakhar Bilal et al. used the K-medoid and naive Bayes grouping methods [90]. A comparison was made between the proposed technique and two popular classification algorithms: k-Means and Naive Bayes. When tested on the KDD dataset, the suggested

technique enhanced accuracy by 2%. In their description of a mixed-method strategy for identifying network assaults, Abhinav Singhal et al. focused on data mining [91].

The random forest method was used to figure out how important each trait was. This way, the network links that were gathered were put into groups based on attack trends. Consequently, Sotiris et al. considered using one-class SVM for issue detection in the system. The training data was very dependent on the one-class SVM's classification accuracy since it used a user-defined cutoff as input.

Forestiero tackled the data analytics challenge by tying data pieces to agents and using a self-organizing multi-agent approach inspired by biological systems [92]. These agents were scattered around the virtual environment, and their final convergence was determined by how similar the objects they were attached to were to one another. In the end, it was found that the divided agents' associated components were abnormal. In order to identify anomalies in data streams, Leticia Decker and colleagues proposed a method that uses fuzzy rule-based approaches to examine recently supplied data in samples [93]. Separately, J. Samuel Manoharan shown that patterns may be detected in a real-time data stream using a dynamic extreme learning system [94]. This poses, despite being less accurate, their suggested solution outperformed the decision tree method in terms of speed. In response, Ranjit Panigrahi et al. created a multi-class intrusion detection system based on decision trees [95]. Their proposed approach shown potential as an intrusion detection system by performing well on datasets spanning many classes.

Table 2.3: Comparative Analysis of Anomaly Detection Schemes

Author [Citation]	Methodology	Parameters	Contribution
Teja Attenborn, et al.[96]	Convolutional Neural Networks (CNN),	TPR, FPR, DR, and Equivalent Error Rates	using a CNN structure that is spatial-temporal, this study focuses on detecting abnormal behavior in crowded video scenes, presenting an effective approach for enhanced anomaly recognition.
Wei Cong Leong et al. [97]	Support Vector Machines (SVM)	ROC curves, AUC and Relative performance comparisons	This work detects and characterises new patterns rapidly using a level set boundary description strategy

			to handle the novelty detection issue in input space.
Muhammad Sharif et al. [98]	Particle Swarm Optimization (PSO),	Sensitivity, Specificity, and Accuracy	To address the novelty detection problem in input space, this work uses a level set boundary description method to quickly find and describe unique patterns.
Ramgopal T Sahu et al. [99]	Noise-Based Density-Based Spatial Application Clamping (DBSCAN),	Percentage of purity, Precision, Recall and Mean number of comparisons	This method enhances real-time anomaly detection capabilities in dynamic situations by proposing a multi-agent technique to effectively find abnormalities in remote data streams.
Henderi et al. [100]	K-Nearest Neighbour (KNN)	DR, Convergence Rate and Accuracy	This innovative method detects intrusions by utilising a cluster of classifiers. By estimating PSO parameters using the Local Unimodal Sampling (LUS) technique, it improves detection accuracy.
Rico Wijayaet Dewantoro al. [101]	Ant Colony Optimization (ACO)	TPR and FPR	A proposed network anomaly detection model focuses on the analysis of IP flows, enhancing the identification of irregularities within the network infrastructure.
Miqing Li et al. [102]	Pareto Depth Analysis	Accuracy, Computation Time and Dissimilarities between trajectories	A proposed similarity-based anomaly detection model employs a multi-criteria dissimilarity measure to reveal anomalous behavior within a dataset.
Mohammed Amin Almaiah et al. [103]	Gaussian RBF Kernel and Support Vector Data description	Computation time, Probability of detection and FPR	Sparse kernel learning creates additional challenges when working with mixed-integer programming to enhance anomaly detection capabilities.
Priscila Valdiviezo-Diaz et al. [104]	Naive Bayes Classifier	Price fluctuation, variation in utilities and Probability mass function	An unusual detecting method based on machine learning Techniques was proposed to identify malicious users within a wireless network.

Feiping Nie et al. [105]	Robust Principal Component Analysis (RPCA)	Recall, False Positive Rate and Precision	A proposed anomaly detection model combines Classification of website traffic using strong statistics and feature selection .

2.5 Critical Aspects in Anomaly Detection

Anomaly detection stands as a formidable challenge within the extensively researched domain of pattern recognition, primarily due to the persistent absence of a universally applicable model that precisely defines normal behavior. In essence, the core of anomaly detection lies in scrutinizing deviations from established normal patterns, an inherently complex task given the absence of a standardized model to delineate what qualifies as normal behavior across diverse datasets. Unlike supervised learning scenarios where labeled training data distinctively guides the model, anomaly detection operates within a realm where the definition of normalcy is context-dependent, intricate, and often subject to interpretation.

The multifaceted challenges in anomaly detection are further accentuated by the intricate task of obtaining clean, attack-free training sets. Many anomaly detection methodologies rest supposing the training dataset is trustworthy and free of anomalies or attack. This assumption, however, often clashes with the practical reality of securing pristine training data. The difficulties stem from assaults occurring during training, which further complicates the already difficult task of distinguishing between typical and unusual occurrences. Anomaly detection models cannot be developed accurately or robustly because of the possibility of undiscovered assaults integrating into the model during training, which adds distortions that may dramatically damage its performance.

In addition to distinguishing normal and abnormal data behavior, several challenges make anomaly detection even more complex, particularly in the context of big data. Key challenges include:

- **Data Analysis:** One important duty is to make sure that the data is correct, uniform, and complete. How accurate the raw data is has a big effect on how the end model turns out. You must have good data for your model to work it.

- **Computational Efficiency:** The sheer amount of data that is being received is a major hurdle. The ever-changing nature of data streams makes conventional anomaly detection techniques useless; hence, efficient memory management is crucial. Ensuring computational efficiency is imperative in handling the high rate of incoming observations.

- **Unbalanced Data:** Anomaly detection mechanisms encounter difficulties in distinguishing between actual and anomalous datasets, especially in unbalanced data scenarios. Defining a normal region becomes complex due to the proximity of outlying observations to the boundary, making it challenging to discern normal and anomalous datasets.

- **Feature Selection:** To get the most out of your computer, you need to choose the correct mix of qualities. Picking feature vectors that capture out-of-the-ordinary data patterns is crucial.

- **Ensemble Techniques:** Designing ensemble techniques for anomaly identification poses a challenge, with considerations such as accuracy and detection rate taking precedence in the development process.

- **Frequently Changing Anomalies:** The dynamic nature of anomalies, which evolves over time, presents a challenge. When anomalies are constantly evolving, a single anomaly detection system may not be able to handle every situation.

- **Noise in Incoming Data:** Due to unstructured data collection, it is increasingly challenging to separate abnormalities from noise in large-scale data settings. Noise in the data can lead to false detections, adding complexity to anomaly detection.

- **Non-availability of Training Data:** The lack of suitable training data is one of the key issues. Inaccurate findings might arise if the assessed data is not correctly matched to the model or if the data model is not fully trained using samples.

- **Optimality of Technique:** It may be difficult to strike a compromise between anomaly detection methods' accuracy and false positive rate. It's challenging to create a dependable method with a high accuracy rate and little false positives.

- **Scalability:** Anomaly detection techniques designed for one domain may not be suitable for others, given variations in normal and abnormal behaviors. Scalability is crucial for efficiently detecting anomalies in large datasets in real-time, presenting a persistent challenge.

Developing a generalized anomaly detection technique that addresses these challenges, especially in the setting of highly dimensional data with volume and velocity aspects, remains an ongoing challenge in the field. In this big data age, improving anomaly detection or identification algorithms requires efforts to overcome these challenges.

2.6 Utilizing Statistical Tests for Anomaly Detection

John Seem et al. [106] suggested a pattern recognition approach for identifying days with comparable patterns of electricity usage. The algorithm takes into account factors like average and peak use throughout the day. The use of smart metre data for anomaly detection was first introduced in this paper. When looking for outliers, statisticians employed tools like the Generalised Extreme Studentized Deviate (GESD) [107] and the Wilks multivariate outlier test [108]. A beginning probability and a maximum permitted number of anticipated outliers are two user-specified criteria that impact the accuracy of this method [109]. In 2007, Seem et al. enhanced this approach by including standard deviation and mean features. They then used a modified z-score, which evaluates the standard deviation from the mean, to assess any remaining abnormalities [110]. For the purpose of identifying outliers, the writers Zhang et al. [111] and Liu et al. [112] used GESD in conjunction with the Q-test on mean and standard deviation. In contrast, Liu et al. [113] proposed a method to identify anomalies in building lighting consumption that integrates CART and GESD.

Finding appropriate cutoffs for these variables is a difficulty with current methods. The comprehensive model developed by Liu et al. [113] that makes use of several statistical approaches aims to evaluate the energy efficiency of buildings, determine the source of energy demand, predict future energy consumption, and identify outliers. This category include a number of models, such as multivariate regression, VBDD and ARIMA. The regression models aim to uncover energy-related building features by using factors such as building age, occupancy, and appliance count [114]. In the VBDD paradigm, there are two separate ideas: the base load of a structure and its usage as it changes with the weather. Light,

heat, and appliance use make up the former, while base temperature forecasts, HVAC degree day coefficients, and other variables are part of the latter. The ARIMA model takes into consideration the effects of the building's inhabitants as well as the changing seasons. An outlier occurs if there is a change in consumption rates without a matching change in the 95% confidence intervals around the prior rates.

An online anomaly detection strategy was proposed by Chou et al. [115] as a substitute for offline approaches. The crux of this method is finding out when something out of the ordinary has happened. We make a forecast every week, and anything out of the ordinary happens when actual use differs from the expected usage by more than two standard deviations. However, frequency domain representations of recorded time-series energy data are better able to identify anomalies in periodic processes, as shown by Wrinch et al. [116]. They conducted experiments using one- and fifteen-minute sampling rates using data acquired from real office buildings.

2.7 Utilizing Machine Learning for Anomaly Detection

Clustering algorithms are often used in machine learning-based technology. One example is the work of Chen et al. [117], who used equal-width binning to discretize continuous numerical energy data into symbolic representations. The next step is to find various energy use patterns using suffix trees. Afterwards, we compile the counts of different symbolic sequences to spot unusual occurrences [118]. Zhang et al. [111] discussed three options, including one based on entropy, one based on regression, and one based on grouping. A number of features, including mean, variance, maximum range, and ratio, are used by clustering-based procedures; this eliminates the need to define thresholds, in contrast to regression and entropy-based strategies.

On the other hand, Bellala et al. [119] introduced a novel density-based approach to anomaly detection in commercial buildings. Since the hourly data is now relatively high-dimensional, the first step is to reduce it using Multi-Dimensional Scaling. After that, they compute a score for daily energy use anomalies using the k-nearest neighbour approach. When the score is high, it means that there's a good chance that the day's intake was considered abnormal [120].

The possibility of using existing methods to anomaly detection in the energy domain has been the focus of several studies. A total of four anomaly detection approaches were

examined by Capozzoli et al [121, 122]. Their corpus of work includes this as just one example. Using a wide range of statistical methods and neural networks, they set out to detect any anomalies in the illumination of eight separate buildings. New approaches for visual anomaly detection were proposed by Janetzko et al. to aid building managers [123]. A prediction-based technique proposed by one study and another by Bellala et al. formed the basis of their anomaly detection strategy. In order to help building managers better comprehend the detected anomalies, a range of visualization tactics are used. Classic line charts, spirals, and recursive patterns are all part of these methods. [124]

2.7.1 Augmenting Anomaly Detection with Contextual Information

The concept that many contextual elements are intimately related to the energy consumption patterns found in buildings is studied in great detail in the first chapter, which invests a substantial amount of room to presenting this issue. This chapter also devotes a considerable amount of space to presenting this topic. Additionally, this concept is covered in great length in the second chapter of the book. It is as a result of this discovery that context has been included into a broad variety of different methodologies that are associated with machine learning. For example, Arjunan et al. [125] developed a novel approach to outlier identification by including data from the subject's immediate environment. This is only one example among many. To begin, they use the k-Medoid clustering method in order to establish daily anomaly ratings for each dwelling. This is done in order to ensure that the ratings are accurate. For the purpose of ensuring that the ratings are correct, this approach is used. With the aid of this technology, we are able to investigate the patterns of energy consumption that are used by each and every homeowner's family. Following that, in order to make these anomaly ratings more accurate, they compare them to the ratings of adjacent families that have previous energy consumption records that are similar to their own. This is done in order to make the ratings more accurate. In order to get a higher level of precision in the abnormality ratings, this is followed. This method is based on the fundamental concept that families who have previously shown spending habits that are comparable to one another are likely to continue to do so in the same manner. This view is the core notion that drives this strategy. This hypothesis serves as the basis for this method by providing the foundation. The novel unsupervised anomaly detection approach proposed by Fontugne et al. [126] is known as Strip, Bind, and Search (SBS). This methodology's main objective is to conduct

research on energy use. Various household items, such as computers, lighting, and air conditioners, are thoroughly examined by the SBS team for their linkages. The objective of this research is to catalogue all the simultaneous appliance uses in a given environment. In order to decipher the complex web of relationships between the devices under scrutiny, the process employs an effective mathematical tool known as Empirical Mode Decomposition. These interactions not only provide useful context, but they are also heavily influenced by the changing of the seasons. If there is a noticeable difference in the overall behaviour of these groups of devices, the team will be notified that something may be amiss. Balakrishnan et al. [127] suggested another unsupervised approach for finding energy system problems called Model, Cluster, and Compare (MCC). As a counterpoint, MCC builds models that highlight issues in adjacent HVAC zones by focusing on the relationships between them.

Some anomaly detection systems employed seasonality as a contextual characteristic, whereas others did not. Ploennigs et al. [128] outlined a method that use a generalized additive model (GAM), a statistical model, to uncover unconventional consumption behaviors. To do this, the building's submeters' hierarchical structure is examined. Considering environmental characteristics such as temperature, time of year, day type, and time of day in addition to consumption data from the past, GAM may distinguish between internal and externally driven abnormalities. Buildings with meters placed at many floors may benefit from this strategy, since GAM is executed recursively to all submeters until the anomalous metre is discovered. This is due to the fact that the method will be applied to all submeters in the event that an abnormality is found at the main meter of the construction. In contrast, a threshold-based paradigm for detecting generalized outliers was put up by Araya and colleagues [129]. In addition to the actual power use, the framework also considers month, day of the week, hour of the day, season, and year as contextual considerations. Utilizing energy data-generated attributes such as median and mean, it also aims to identify outliers, fourteen score and four spaces.

Using energy data, the efficacy of the generalized contextual-based anomaly detection system suggested by Hayes et al. was tested [130]. The first and second stages of this procedure are as follows. Creating a statistical training model using past data and comparing it to fresh values is the first step in finding out-of-the-ordinary material. Any time there is a big change from the expected pattern, we call it an anomaly. The second stage, context-based

anomaly detection, comprises generating sensor profiles that help with value prediction using a multivariate context-based clustering method. If the expected quantity is significantly different from the actual number, we say that the level of consumption is abnormal. The authors failed to provide any guidance on how to determine the optimal thresholds for each step, rendering this approach practically useless despite its theoretical validity [131].

2.8 Anomaly Detection Using Graph Structures

By using a method that was established on the traversal of a kernel matrix by a random graph, Cheng et al. [131] were able to successfully determine the existence of anomalies in energy data. This was accomplished by the use of a methodology. The use of time-series energy data is utilized in this approach to the construction of a kernel matrix. Anomalies are nodes that get an abnormally low number of traversals compared to the norm. As a result of the authors' study on kernel matrix alignment, they are able to generate correlations between a large numbers of time series, which enables them to identify anomalies. An important benefit is that this is the case. Granger graphical models were used by Qiu et al. in order to perform the task of identifying anomalies for the goal of presenting a counterexample [132]. Utilizing time series data, the first thing that they did was generate causal graphs. This was the first step in their process. Drawing conclusions about the connections that exist between the variables allowed for the successful completion of this task. Graph anomaly scores, which were calculated via the use of Kullback-Leibler divergence, were utilized in order to identify abnormalities in accordance with a threshold that had been established beforehand.

2.9 Anomaly Detection Using Rule-based Methods

Using fuzzy logic concepts, Wijayasekara and colleagues introduced a novel approach to anomaly detection [133]. Clustering consumption data during training allows this method to identify and classify patterns that are statistically similar. Then, rules that describe the behaviour of normal consumption are developed by applying fuzzy logic rule extraction to these clusters. The built fuzzy logic rule-based model is used to assess the energy utilisation throughout the testing phase. An anomaly is defined as any behaviour that does not follow the expected pattern inside the model. Building managers are provided with audio explanations of outliers using this approach, which is far more effective than simple anomaly ratings. Another rule-based method for anomaly identification has been proposed by Pena et

al. [134]. In addition to electricity data, this system also takes into account data from outside and within the building's sensors, such as occupancy and weather. By including a comprehensive set of criteria derived from several data sources, their method enhances the identification of non-typical consumption patterns.

2.10 Motivation

Network anomalies present a significant threat to the seamless functionality of networks, often disrupting normal operations by inducing sudden spikes in traffic flow. The timely detection of intrusions, faults, and system failures is crucial to prevent widespread damage. However, addressing these challenges is a complex task, given issues such as the availability of attack-free training sets, class imbalances among anomalies, and the presence of categorical features in high-dimensional network datasets.

For the detection of anomalies to work, there has to be training sets that do not include any attacks, which are essential for developing accurate models. Many existing approaches assume the correctness of training data labels, supposing that the training dataset is defenseless against assaults. However, this assumption often does not hold in practice, as obtaining a clean dataset is challenging due to uncertainties when ordinary and extraordinary circumstances converge. Additionally, attacks may occur during the training phase, and if undetected, they become part of the training model, significantly affecting its accuracy and performance.

In an ideal world, an anomaly detection model would always find abnormalities and never mistakenly notify. Overfitting traffic models to training data is a major roadblock to this objective. Overfitting occurs when the model's representation of normal traffic is too closely tailored to the training data, making it sensitive to unknown anomalous instances. Consequently, the model may produce false negatives, classifying normal instances during the testing phase as anomalous and vice versa. Achieving optimal detection capabilities necessitates the careful sanitization of training data by removing both non-regular and unknown anomalous instances.

When dealing with noisy, high-dimensional data such that present in network traffic, it is already tough to achieve optimal and meaningful feature extraction. The current state of

learning techniques makes them unsuitable for large datasets due to the high amount of variables they include. Overfitting and inaccurate classifications may happen when features are too similar. Deleting superfluous features may improve learning performance in several ways, including search accuracy, computational cost, model interoperability, and speed. In order to address the current issue, this will be useful. However, the inadvertent elimination of crucial features may decrease classification accuracy. Unsupervised techniques offer a promising solution in such situations as they enhance feature separability and likelihood.

Unsupervised techniques make assumptions about network traffic data, Major contributors to this include the prevalence of typical occurrences and the discernible differences between typical and exceptional network traffic. However, these assumptions may not always hold true, especially in the case of certain intrusions like DoS attacks. DoS attacks can occur in similar numbers to normal instances, causing algorithms to struggle in labeling these instances as anomalies due to the high density and similarity to normal instances. These challenges highlight the need for robust anomaly detection algorithms capable of handling diverse intrusion scenarios.

Adjusting our feature sets will make anomaly detection in network traffic data faster and more accurate. In order to provide the classifier with the processing power it needs, non-linear correlation analysis is just as important as conventional optimization methods, which typically just take linear correlation into account. It is still difficult to find outliers in huge datasets instantly. There has been an overwhelming amount of literature on intrusion detection, but very little on the practicality of studying network-based methods for massive datasets.

The main obstacle here is building a reliable anomaly detection system that can analyze network data in real-time and identify anomalies. The model must navigate through the complexities of high-dimensional, noisy data, address issues of overfitting, and consider the unique characteristics of different types of intrusions. Striking a balance that permits precision and, speed, and scalability is crucial for the successful deployment of anomaly detection models in practical, real-world scenarios.

The challenges associated with network anomaly detection underscore the need for innovative approaches that go beyond traditional assumptions and limitations. To guarantee

the strong security of contemporary networks, it is crucial to create models that can adjust to the ever-changing nature of network traffic, deal with various intrusion situations, and run in real-time. Anomaly detection system developers are devoting a great deal of time and energy to answering these concerns and bridging knowledge gaps left by earlier research.

2.11 Summary

In the realm of anomaly detection, a diverse array of techniques has been employed, leveraging optimization strategies, dimensionality reduction techniques, and machine learning models. These methodologies are meticulously explored in this chapter, delving into their nuances to enhance detection rates and mitigate false positives. A comprehensive collected and assessed cutting-edge methods forms a crucial part of this exploration, shedding light on as we consider the benefits and drawbacks of each approach. We highlight the need of improving anomaly detection methods, which is the driving force for our study, especially when dealing with high-dimensional data that incorporates volume and velocity components. In this context, the goals of this thesis are explained, laying the groundwork for a better grasp of how the ensemble-based network anomaly detection methods suggested in the next chapter help with the problems of anomaly detection in high-dimensional datasets with large volume and velocity changes.

Chapter 3

Secure Anomaly Detection in Computing

3.1 Introduction

In the intricate realm of anomaly detection within high-dimensional datasets, the convergence of complexity, volume, and velocity introduces a multifaceted challenge. The copious features defining high-dimensional data sets, coupled with the dynamics of volume and velocity, necessitate advanced algorithms capable of discerning subtle patterns and identifying aberrations in real-time. This chapter embarks on an exploration of the algorithmic landscape employed in anomaly detection for Multi-Dimensional data, delving into the intricacies of methodologies designed to grapple with the voluminous and dynamic nature of contemporary datasets.

High-dimensional datasets, characterized by an abundance of features, have become ubiquitous in diverse domains such as finance, healthcare, and cybersecurity. As the dimensionality of data increases, the conventional tools and intuitions derived from lower-dimensional spaces prove inadequate. The "curse of dimensionality" makes regular methods very difficult to manage in high-dimensional areas, necessitating specialized algorithms [135].

Detecting extreme values in multi-dimensional data necessitates algorithms that can effectively sift through the myriad features to discern patterns indicative of normal behavior and, conversely, detect deviations that may signify anomalies. The challenge lies not only in the sheer volume of features but also in the potential interdependencies and complexities that arise as the dimensionality increases.

The term "volume" used to describe anomaly detection is a reference to the massive amounts of data generated and processed by contemporary systems. Big data technologies have made

it critical to be able to manage very large datasets. The vast amount of typical data might mask anomalies, which are often signs of unusual occurrences or outliers. In order to comb through massive datasets for small anomalies, anomaly detection algorithms need to be scalable, efficient, and resilient.

Consider, for instance, financial transactions within a global banking system. The sheer volume of transactions occurring in real-time demands algorithms capable of swiftly identifying anomalous patterns, potentially indicative of fraudulent activities, amidst the deluge of legitimate transactions. Therefore, the volume aspect of high-dimensional data necessitates algorithms that can operate seamlessly in the presence of vast amounts of information.

Velocity introduces a temporal dimension to the challenge of anomaly identification. In scenarios where anomalies manifest as sudden deviations from the norm, real-time detection becomes imperative. This is particularly relevant in applications such as network security, where rapid changes in data patterns may signify a cyberattack, or in industrial systems, where anomalous sensor readings may indicate a critical malfunction [136].

Algorithms with strong pattern recognition and anomaly detection skills are crucial due to the time constraint. It is essential for anomaly detection systems to comprehend and analyze data in close to real-time when dealing with velocity-aware high-dimensional datasets.

The algorithmic panorama for anomaly identification in data sets with several dimensions is diverse, encompassing a spectrum of approaches designed to tackle the challenges posed by volume and velocity. One notable contender in this arena is the Isolation Forest algorithm. Leveraging the power of ensemble learning and randomization, Isolation Forest excels in isolating anomalies by constructing decision trees that isolate outliers with fewer splits, making it particularly adept in handling high-dimensional data.

On a different front, the One-Class SVM algorithm addresses the challenge of high dimensionality via the use of a higher-dimensional space for data mapping, effectively creating a boundary around normal instance. Instances lying outside this boundary are classified as anomalies, demonstrating the algorithm's efficacy in discerning deviations from expected patterns.

Anomalies are believed to have a lower local density than regular cases, according to density-based algorithms such the LOF (Local Outlier Factor). It is possible that LOF may identify suspicious occurrences by finding the local density deviation of each place. Anomalies that appear as sparse clusters in the high-dimensional space are ideal candidates for this density-centric method [137].

One straightforward method that has shown to be effective is k-Nearest Neighbors (k-NN). It assesses the proximity of data points to their neighbors. Anomalies are defined in the context of anomaly identification as cases having few related neighbours. This approach is well-suited to high-dimensional datasets because it is excellent at detecting local anomalies that may escape global models.

Venturing into the realm of neural networks, Auto-encoders offer a distinctive approach to anomaly detection. Operating on the principles of unsupervised learning, Auto-encoders learn to reconstruct input data. Anomalies show up as significant reconstruction mistakes because they deviate from the predicted reconstruction. This method works really well for getting detailed patterns out of very large data sets.

Finding the main parts of data is where principal component analysis comes in handy for finding anomalies. It's not new to use PCA to reduce the number of dimensions. Anomalies are instances that considerably deviate from the predicted distribution along these fundamental components. PCA is useful in anomaly detection because it may reduce high-dimensional data to its essential elements [138].

For a new take on anomaly detection, try Density-Based Spatial Clustering using DBSCAN. Based on the idea that outliers don't always fit into the usual patterns of clustering, DBSCAN sorts data points according to their density. Anomalies, being sparse and isolated, fall outside these clusters, rendering them detectable by the algorithm.

The Elliptic Envelope algorithm embodies a statistical approach to anomaly detection. By fitting a robust elliptical envelope to the data, the algorithm assumes that normal instances reside within this envelope, while anomalies lie outside. This statistical elegance renders Elliptic Envelope effective in scenarios where anomalies exhibit distinct statistical properties.

As we navigate the rich tapestry of anomaly identification algorithms in highly dimensional data with a consideration for volume and velocity, the amalgamation of algorithmic diversity and technical sophistication becomes apparent. Each algorithm encapsulates a unique perspective and set of assumptions, offering a nuanced solution to the multifaceted challenge of anomaly detection in contemporary datasets. The following parts of this section will unravel the inner workings, strengths, and limitations of each algorithm, providing a comprehensive guide for selecting the most suitable approach in specific contexts. As the data landscape continues to evolve, the refinement and adaptation of these algorithms become paramount for staying at the forefront of anomaly detection in high-dimensional datasets.

3.2 Isolation Forest

Within the expansive domain of anomaly detection algorithms, the Isolation Forest emerges as a groundbreaking and highly efficient method tailored for isolating anomalies within intricate high-dimensional datasets. Originating from the work of Prakash VR et al. in 2019, this algorithm represents a departure from conventional methodologies, introducing a novel paradigm that harnesses the potency of ensemble learning and strategic randomization [139]. By constructing an ensemble of isolation trees, each formed through the random selection of features and split values, Isolation Forest excels in swiftly and effectively pinpointing outliers. This departure from traditional methods not the addresses only the computational challenges posed by highly dimensional data but also enhances the algorithm's adaptability to diverse datasets. Its innovative approach, measuring the average path length for isolating instances, results in an efficient and effective anomaly detection mechanism. Distancing oneself Forest stands apart in the industry due to its unique combination of ensemble learning and randomization. It is especially useful in areas where the early identification of anomalies is critical, such system monitoring, fraud detection, and network security.

3.2.1 Foundations of Isolation Forest:

At its core, the Isolation Forest operates on the principle that anomalies are rare and isolated instances within a dataset. Traditional techniques often struggle with the computational burden imposed by the vast number of features present in high-dimensional data, leading to suboptimal performance. The Isolation Forest addresses this challenge by taking a

fundamentally different approach, aiming to isolate anomalies rather than explicitly modeling normal instances.

3.2.2 Ensemble Learning and Randomization:

One of the distinguishing features of the Isolation Forest is its use of ensemble learning. By repeatedly dividing the dataset into binary trees, the technique builds an ensemble of isolation trees. To build a tree, you first choose a feature at random for each node, and then you pick a split value at random from the interval given by the feature's minimum and maximum values [140].

By utilizing this ensemble of trees, the algorithm captures diverse perspectives on the data, enhancing its ability to identify anomalies. The randomization inherent in the feature and split value selection enables the Isolation Forest to efficiently zero in on outliers without the need for complex, computationally expensive calculations.

3.2.3 Isolation by Path Length:

The isolation of anomalies is achieved by measuring the average path length required to isolate each instance in the ensemble of trees. Anomalies, being rare and distinct, typically require fewer partitioning steps to be isolated compared to normal instances. Therefore, the average path length for anomalies is shorter, making them stand out in the overall distribution.

The algorithm assigns anomaly scores based on these path lengths, with shorter paths indicative of higher anomaly scores. Consequently, instances with elevated anomaly scores are flagged as potential outliers, signifying their departure from the expected patterns within the dataset.

3.2.4 Advantages and Applicability:

The Isolation Forest algorithm boasts a multitude of advantages that underpin its widespread popularity in the realm of anomaly detection. Primarily, its exceptional efficiency in high-dimensional spaces distinguishes it from numerous traditional techniques grappling with the notorious curse of dimensionality. The algorithm's prowess lies in its ability to y7 and discern patterns within datasets characterized by a myriad of features, showcasing its resilience where another methods falter [141].

The strategic introduction of randomness during the tree-building process further fortifies the Isolation Forest's capabilities. This inherent randomness not only contributes to the algorithm's robustness in handling outliers but also enhances its adaptability to diverse datasets, accommodating variations in data structures and distributions.

One of the standout features of Isolation Forest is its remarkable effectiveness in scenarios where anomalies deviate significantly from the norm. The algorithm excels in swiftly isolating instances that exhibit conspicuous differences in their underlying patterns. Isolation Forest is perfect for essential operations like system monitoring, fraud detection, and network security due to its characteristic. In these contexts, the algorithm's agility in identifying anomalies in real-time proves to be of paramount importance, contributing to its status as a reliable and versatile tool in the arsenal of anomaly detection methodologies.

3.2.5 Challenges and Considerations:

While Isolation Forest excels in many scenarios, it is not without its challenges. The algorithm may struggle when anomalies are subtle and do not exhibit clear distinctions from normal instances. Additionally, the user should be cautious in selecting hyper parameters to ensure optimal performance, as improper tuning may lead to suboptimal results [142].

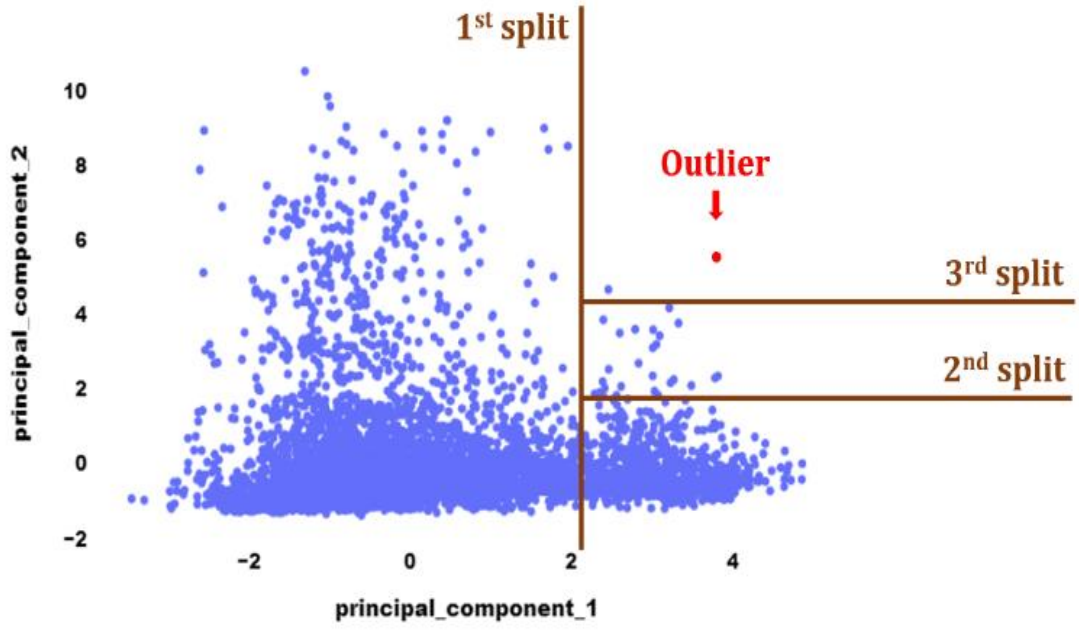


Figure 3.1: Anomalies Detection Using Isolation Forest [139]

In the realm of anomaly detection, Isolation Forest stands as a testament to the efficacy of innovative thinking and departure from conventional methodologies. Its utilization of ensemble learning and randomization allows it to swiftly and efficiently identify anomalies within high-dimensional datasets, making it a valuable tool in the modern data analyst's arsenal. As we continue to grapple with the complexities of diverse and voluminous datasets, the Isolation Forest remains a beacon of efficiency, showcasing the potential of novel approaches to anomaly detection.

3.3 One-Class SVM

Anomaly detection systems abound, but one that stands out is the One-Class Support Vector Machine (SVM). It's both resilient and adaptable. It makes extensive use of the theory behind Support Vector Machines (SVMs). Conceived as a divergence from conventional SVMs, the One-Class SVM is designed specifically for scenarios where the training data predominantly consists of normal instances, aiming to discern the defining properties of normalcy to subsequently identify anomalies. Developed to address the challenges posed by imbalanced datasets, One-Class SVM stands as a pivotal tool in anomaly detection, particularly in use cases include the identification of issues, intrusions, and fraud [143].

3.3.1 Foundations of One-Class SVM:

At its core, One-Class SVM operates on the premise of learning the characteristics of normal instances within a given dataset. One-Class SVM is designed for scenarios where there is only one class in the training data, usually the majority class representing normal cases. This is in contrast to regular SVMs, which are mostly used for binary classification problems. The method endeavors to create a border around the normal instances, encapsulating the region where the majority of the data resides [144].

3.3.2 Mapping to High-Dimensional Spaces:

Support for One Class The capacity to convert data into a high-dimensional dimension is what makes Vector Machines stand out. A kernel function is used to execute this modification. To ensure that normal cases are easily distinguished from out-of-the-ordinary ones, it employs a predefined area division for every data point. Attributes of the data determine whether kernel function is crucial. There are several kinds of datasets that may be

utilized with widely used kernel functions, including radial basis functions, polynomial kernels (RBF).

3.3.3 Hyperplane Separation:

Finding a hyperplane that accurately distinguishes between typical occurrences and likely outliers is the goal of One-Class SVM once the data is mapped into this high-dimensional space. The hyperplane is positioned to maximize the margin around usual occurrences, so the majority of data points will fall within that. Cases that don't fit inside this range are thus considered out of the ordinary [145].

The process of finding this optimal hyperplane involves solving a mathematical optimization problem. One-Class SVM aims to minimize the risk of misclassifying normal instances while simultaneously maximizing the margin and, by extension, the separation between normal data and potential anomalies.

3.3.4 Nu-Parameter and Controlling False Positives:

The Assist with Dress Regulations One of the most important parameters of SVMs is "nu," which regulates the trade-off between training error and model complexity. A regularization term, the nu-parameter determines the upper bound on the allowable margin of error for a particular parameter. By tuning the nu-parameter, practitioners can control the algorithm's sensitivity to outliers and regulate false-positive rates. A higher nu-value corresponds to a more lenient margin, potentially allowing for a higher rate of false positives.

3.3.5 Advantages and Applicability:

One-Class SVM offers several advantages that contribute to its widespread use in anomaly detection tasks. Notably, its capacity to operate with imbalanced datasets when ordinary occurrences considerably predominate over extraordinary ones, makes it particularly suitable for real-world scenarios. The algorithm's reliance on mapping data into high-dimensional spaces enhances its transparency about intricate patterns and relationships in data, allowing for nuanced anomaly detection [145].

Moreover, One-Class SVM is effective in scenarios where anomalies exhibit distinct differences in their underlying patterns compared to normal instances. This characteristic

makes the algorithm well-suited for applications such as fraud detection, where fraudulent activities often manifest as deviations from normal transaction patterns.

3.3.6 Challenges and Considerations:

While One-Class SVM excels in various contexts, it is crucial to choose the nu-parameter and kernel function as hyper parameters. Precise adjustment is necessary to get optimal results with minimal false positives and maximum sensitivity to abnormalities [145].

Additionally, the algorithm may face challenges when anomalies exhibit subtle deviations that are not well-separated in the high-dimensional space. In such cases, alternative algorithms or preprocessing techniques may be necessary to complement One-Class SVM.

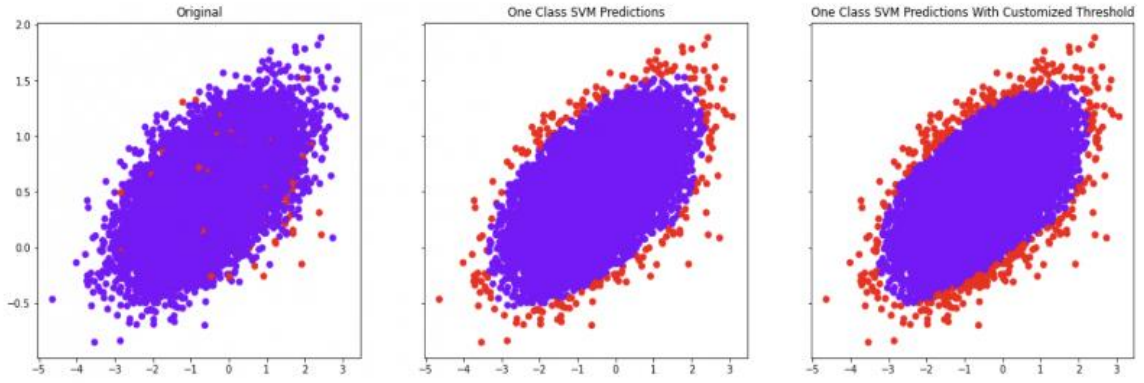


Figure 3.2: Anomaly Identification by One Class SVM [145]

In the intricate tapestry of anomaly detection, the One-Class SVM algorithm stands as a testament to the adaptability and efficacy of Support Vector Machines in unconventional settings. By focusing on learning the properties of normal instances, mapping data into high-dimensional spaces, and delineating a hyperplane that encapsulates normalcy, One-Class SVM provides a singular perspective on anomaly detection. Its application extends across diverse domains, offering a powerful tool for discerning anomalies in imbalanced datasets where the majority class represents normal instances. As the data landscape continues to evolve, the One-Class SVM remains a valuable asset, contributing to the arsenal of anomaly identification methodologies and addressing the unique challenges that are posed by imbalanced and complex datasets.

3.4 Local Outlier Factor (LOF)

A powerful and flexible tool in the ever-changing field of anomaly identification, the Local Outlier Factor (LOF) method is a novel way to find outliers in datasets with different local densities. Conceived by Breunig et al. in 2000, LOF introduces a novel paradigm in density-based anomaly detection, excelling in scenarios where anomalies manifest as deviations within local regions rather than global distinctions [146].

3.4.1 Foundations of LOF:

LOF operates on the fundamental premise that anomalies exhibit distinct patterns in local density compared to their neighboring data points. Traditional anomaly detection methods often struggle when anomalies are subtle and exhibit variations within localized regions rather than global outliers. LOF addresses this challenge by introducing a density-based metric, allowing it to adapt to the inherent intricacies of diverse datasets.[147]

3.4.2 Computing Local Density Deviation:

One of the main parts of LOF is determining the relative local density divergence between all the data points and their neighbors. This complex procedure follows a rational progression. To begin, we utilize LOF to locate each data point's k-nearest neighbors; the exact number of neighbors to use is up to the user. The conventional geometric metric is an excellent option for the distance metric that must be chosen in order to compute closeness.[148]

Once we've identified the neighbors, we can utilize LOF to calculate the local reachability density for each area. Find the neighbors of k-nearest of the current location and obtain the inverse of their average distance as part of this operation. A high reachability density indicates that a large number of individuals reside in close proximity to one another.

Next, determine the local outlier factor. The reachability density is determined by LOF by looking at the data points around each one. An outlier is a place that deviates from the norm in terms of density and LOF score. After normalization, LOF scores are no longer affected by the dataset size.

3.4.3 Interpreting LOF Scores:

The interpretation of Local Outlier Factor (LOF) scores serves as a foundation in the realm of anomaly detection, providing a nuanced understanding of each data point's local density within its immediate neighborhood. These scores offer a quantitative measure to gauge how a specific data point deviates from the expected density patterns of its neighbors. This interpretation is pivotal for distinguishing between normal instances and potential anomalies, providing a clear and intuitive framework for anomaly detection.[149]

When confronted with an LOF score significantly higher than 1, it serves as a red flag, signaling a lower local density in comparison to neighboring data points. In essence, this higher LOF score indicates that the data point resides in a sparser region within its local vicinity, suggesting a potential anomaly. The elevated LOF score highlights the conspicuous deviation in density, acting as an effective marker for identifying instances that stand out within their immediate surroundings.

Conversely, LOF scores close to 1 convey a sense of normalcy. Data points with scores in this range align closely with the local density patterns of their neighbors. Essentially, these points seamlessly blend into the expected norms within their local environments, showcasing conformity to the prevalent density characteristics. The proximity of the LOF score to 1 indicates a lack of significant deviation in density, reinforcing the notion that these data points exhibit patterns typical of their immediate neighborhood.

This binary interpretation, where LOF scores significantly higher than 1 flag potential anomalies and scores close to 1 denote normal instances, offers a practical and actionable approach to anomaly detection. By leveraging these scores, analysts can efficiently identify instances that deviate markedly from their local density environments, streamlining the focus on potential outliers. Establishing a threshold for LOF scores empowers practitioners to make informed decisions about the nature of data points within a dataset, facilitating targeted anomaly detection in the complex landscape of diverse datasets. In essence, the interpretation of LOF scores provides a valuable lens through which data scientists can discern irregularities, enabling them to sift through the intricacies of data and pinpoint anomalies effectively.

3.4.4 Advantages and Applicability:

LOF boasts several advantages contributing to its popularity in anomaly detection tasks. Its density-based approach allows the algorithm to adapt seamlessly to datasets featuring varying local densities. This adaptability is very beneficial in real-world applications like network intrusion detection, where anomalies might manifest as subtle shifts in regional traffic patterns [150].

Moreover, LOF does not impose assumptions regarding the specific shape or distribution of normal instances, enhancing its flexibility in handling diverse and complex datasets. This characteristic makes LOF applicable across a wide spectrum of domains, from healthcare to finance, where anomalies can exhibit diverse patterns.

3.4.5 Challenges and Considerations:

Even though LOF has many good points, it does have some problems. It's important to give careful thought to the figure k, which stands for "neighbors." You might not see the limited density effect as much if you set the k value too high, and you might see too much of it if you set it too low [151].

Additionally, LOF may encounter difficulties when anomalies form part of a larger dense region rather than isolated clusters. In such cases, the algorithm may struggle to capture the subtleties of local density variations effectively.

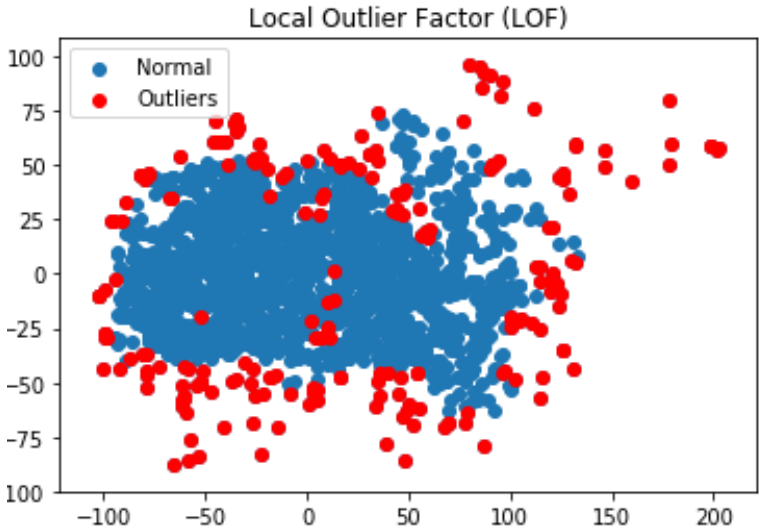


Figure 3.3: Anomaly Detection by Local Outlier Factor (LOF) [149]

One density-based approach that has proved effective in the extensive field of outlier identification is the Local Outlier Factor (LOF) method. Through the use of surrounding data points' local density change, LOF's nuanced lens can efficiently detect anomalies within a range of local densities.

As complex and diverse datasets become increasingly prevalent, LOF remains a valuable asset, offering insights into localized density variations and empowering anomaly detection in various real-world applications.

3.5 k-Nearest Neighbors (k-NN)

For classification tasks, the k-Nearest Neighbors (k-NN) method is notoriously easy to utilize. It is a simple but useful solution in pattern recognition and machine learning. Even though k-NN is usually associated with classification, it can also be used to find outliers in a dataset, which is useful for anomaly discovery [152].

3.5.1 Foundations of k-NN:

The instance-based learning approach known as K-Nearest Neighbors (k-NN) does not need any configuration. Where "k" is a user-defined word that specifies the amount of neighbors to be examined, finding the most prevalent class among a data point's k closest neighbors is an essential metric to discover in classification tasks. The underlying premise of the technique is to group together data pieces that are comparable. Using the area immediately around the target data point, this proximity-based method may ascertain the distribution of classes. Using the most common class as a starting point, the k-nearest neighbour algorithm efficiently classifies target data by drawing on the feature space's underlying patterns and similarities. The simplicity and intuitive nature of this method contribute to k-NN's widespread use, particularly in scenarios where the relationships between data points are well-reflected by their proximity in feature space.

3.5.2 The k-NN Algorithm in Classification:

One data point is requested to be classified in a classification scenario using the k-Nearest Neighbors (k-NN) technique. In the next step, it finds the points that are k-nearest to the objective. The distance metric is used to determine this closeness; the most popular one is the Euclidean distance [153]. The projected class for the target data point is based on the

algorithm's consideration of the prevalent class among these identified neighbours. Presumably, being close to one another in the feature space indicates that the classes are comparable. Using the assumption that data points in close proximity to one another are more likely to share similarities and belong to the same class, k-NN relies on the local neighborhood's dominant class. This process underscores the intuitive nature of the algorithm, where the class assignment for a data point is influenced by the collective influence of its nearby counterparts in the feature space.

3.5.3 Adaptation for Anomaly Detection:

When it comes to detecting anomalies, k-Nearest Neighbors (k-NN) plays a special role by meeting the special demands of finding outliers. Anomalies are typically distinguished by their departure from expected patterns, and this deviation is reflected in a scarcity of neighbors exhibiting similar features. In essence, instances that lack a sufficient number of nearby neighbors with comparable characteristics are considered potential anomalies by the k-NN algorithm.[154]. This approach aligns with the underlying assumption that anomalies often manifest as isolated or sparsely populated instances within the feature space. By assessing the local neighborhood and evaluating the density of similar points, k-NN excels in pinpointing instances that diverge significantly from the established patterns, making it a valuable tool in the nuanced landscape of anomaly detection. The algorithm's adaptability to local irregularities and its sensitivity to deviations contribute to its effectiveness in identifying potential anomalies within diverse datasets.

3.5.4 Identifying Anomalies with k-NN:

Anomaly detection with k-Nearest Neighbors (k-NN) unfolds through a meticulous examination of the local neighborhood surrounding each data point. The fundamental premise is rooted in the notion that anomalies deviate from the typical patterns present in their surroundings. As such, the process begins by scrutinizing the neighboring data points. If a particular data point is found to have few neighbors sharing similarities, it implies that the point does not conform to the prevalent patterns exhibited by its immediate environment. In simpler terms, the scarcity of similar neighbors suggests an unusual instance, potentially signifying an anomaly within the dataset.

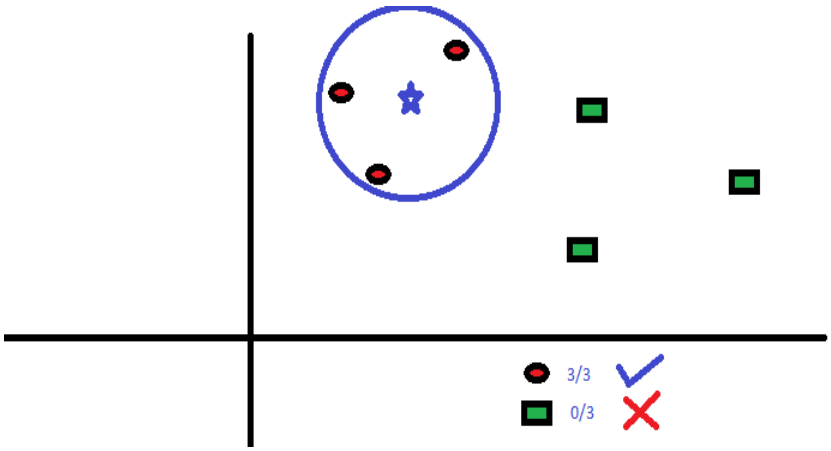


Figure 3.4: Anomaly Detection by K-NN Algorithm [155]

This approach is grounded in the principle that anomalies often manifest as isolated or sparsely populated instances within the feature space. By evaluating the density and composition of the local neighborhood, k-NN excels in identifying instances that stand out due to their divergence from the established norms. The algorithm's sensitivity to the local context and its capacity to flag instances with atypical neighbors make it a robust tool in the realm of anomaly detection, particularly where anomalies are characterized by their distinctive features within specific local regions of the dataset.[155]

3.5.5 Distance Metrics and Parameter Selection:

If one modifies the distance metric, the k-NN method's feature distribution and scale sensitivity may be altered. Several types of data are well-suited to the three most used distance metrics: the Minkowski, Manhattan, and Euclidean distances [156].

Additionally, the choosing of the parameter "k" plays a pivotal role. A smaller "k" may result in a more sensitive model, prone to noise and fluctuations, while a larger "k" may lead to over smoothing and overlooking local patterns. Striking the right balance for this method can't function without the dataset in issue.

3.5.6 Advantages and Applicability:

k-NN boasts several advantages that contribute to its popularity in anomaly detection tasks. Anyone without prior experience with machine learning should find it easy to use because to its straightforward architecture and little complexity. The algorithm is inherently non-

parametric, allowing it to adapt to the underlying patterns of diverse datasets without imposing rigid assumptions [157].

Moreover, k-NN excels in scenarios where anomalies are characterized by local irregularities rather than global distinctions. Its ability to capture local patterns makes it particularly suitable for datasets where anomalies manifest as clusters or sparse instances within the feature space.

3.5.7 Challenges and Considerations:

Despite its merits, k-NN faces challenges in scenarios with high-dimensional data or imbalanced class distributions. High dimensionality can lead to increased computational costs and sensitivity to irrelevant features. Imbalanced datasets may result in biased predictions favoring the majority class [158].

Additionally, the algorithm's performance can be affected by decisions of distance metric and the parameter "k." Careful consideration and tuning are necessary to ensure optimal results across diverse datasets.

A basic and successful method for anomaly identification is K-Nearest Neighbours (k-NN). The k-NN algorithm uses the nearby points' majority class to find possible outliers in a dataset. The fact that it can capture local patterns, is easy to use, and is flexible makes it a potential useful tool for data scientists. Because it gives a proximity-based perspective on anomaly identification and expands our toolkit of approaches for discovering outliers in large data sets, k-NN remains valuable as we investigate intricate datasets.

3.6 Auto-encoders

In the intricate landscape of anomaly detection, Auto-encoders emerge as a potent and versatile tool, harnessing the capabilities of neural networks to decipher and reconstruct input data. Unlike traditional neural networks employed for classification, Auto-encoders are distinctively designed for unsupervised learning, with a primary goal of recreating their input as accurately as possible. The crux of their effectiveness in anomaly detection lies in the examination of reconstruction errors – instances with elevated discrepancies when comparing the original data with the rebuilt result are earmarked as potential anomalies [159].

3.6.1 Foundations of Auto-encoders:

Neuronal network members called Auto-encoders work well with unsupervised learning models. An encoder and a decoder make up each one, which is built differently. By sending data to the encoder, the training paradigm takes shape and the latent space, a compressed model, is made [160]. This latent space encapsulates essential features while shedding redundant details. The subsequent phase entails the decoder's task of reconstructing the initial input from this compressed representation. The crux of Auto-encoders lies in their ability to learn meaningful representations during training, allowing them to discern intricate patterns within the data. This unsupervised learning approach positions Auto-encoders as formidable tools for tasks like anomaly detection, where anomalies may not adhere to explicit class boundaries. The synergy of encoder and decoder architectures, coupled with the exploration of latent spaces, underpins the adaptability and efficacy of Auto-encoders in unraveling complex data structures across diverse domains.

3.6.2 Encoder and Decoder Architecture:

To enable the encoder component to capture the critical characteristics required for reconstruction, Auto-encoders decrease in the dimensionality of the source data. This squished form, which keeps the essentials while cutting out the extras, is sometimes called the bottleneck or latent space. The decoder employs this compressed form in an endeavor to retrieve the initial input data. The purpose of auto-encoder learning is such that, via repeated parameter adjustments, the discrepancy between the two data sets may be minimized [161].

3.6.3 Reconstruction Error as Anomaly Indicator:

Unlocking the potential of Auto-encoders for anomaly detection centers on a meticulous examination of the reconstruction error, a metric quantifying the dissimilarity between the input and the reconstructed output.[162] Instances marked by elevated reconstruction errors serve as clear signals of potential anomalies, signifying that the model encounters challenges in faithfully reproducing these data points. This heightened error is a manifestation of the model's struggle to reconcile the anomalous features present in these instances. Conversely, normal instances, having adhered to learned patterns during training, showcase lower reconstruction errors, reflecting the model's adeptness at faithfully reconstructing familiar patterns. This stark dichotomy in reconstruction errors serves as the linchpin for anomaly

detection using Auto-encoders, enabling the algorithm to distinguish between the expected and the unexpected within the dataset based on the fidelity of reconstruction. The scrutiny of reconstruction errors encapsulates the essence of how Auto-encoders unveil anomalies by discerning deviations from established data patterns.

3.6.4 Training and Learning Latent Representations:

In the training phase, Auto-encoders undergo a transformative process where they adjust their parameters to acquire meaningful representations within the latent space. Crucially, this learning journey unfolds in an unsupervised manner, signifying that the model refines its weights solely based on the input data without the guidance of explicit class labels. This intrinsic adaptability is a key strength, empowering Auto-encoders to discern intricate patterns and dependencies embedded within the data. This adaptability becomes particularly advantageous in anomaly detection tasks where anomalies may defy adherence to explicit class boundaries. By allowing the model to autonomously navigate the intricacies of the dataset, Auto-encoders showcase a capacity to unveil anomalies that exhibit nuanced and non-conventional patterns, contributing to their efficacy in scenarios where anomalies may not conform to predefined categories. The unsupervised learning paradigm enhances the versatility of Auto-encoders, positioning them as powerful tools for exploring and identifying irregularities within diverse and complex datasets.

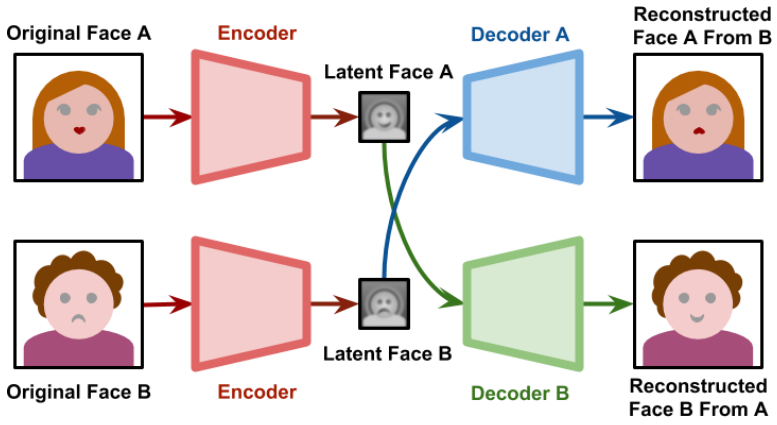


Figure 3.5: Anomaly Detection by Auto-encoders [162]

3.6.5 Variants of Auto-encoders:

Auto-encoders come in various forms, each tailored for specific tasks. Variants include de-noising Auto-encoders, designed to reconstruct clean data from noisy inputs, and variational

Auto-encoders, which extend the model's capabilities to generate new data samples. The adaptability of Auto-encoders to different contexts underscores their versatility in anomaly detection scenarios.[163]

3.6.6 Advantages and Applicability:

Auto-encoders offer several advantages contributing to their popularity in anomaly detection. Their unsupervised nature makes them perfect option for when labeled anomaly data may be scarce or unavailable. Another benefit is that they can be used with a wider range of datasets because they can record complex non-linear connections within data.

Additionally, Auto-encoders excel in distinguishing outliers in data with several dimensions, as the latent space effectively encapsulates salient features while discarding redundant information. This characteristic is particularly valuable in domains like cybersecurity, where anomalies may manifest as subtle deviations within voluminous network traffic data.

3.6.7 Challenges and Considerations:

While Auto-encoders present a robust approach to anomaly detection, they are not immune to challenges. The determination of an appropriate threshold for identifying anomalies based on reconstruction errors requires careful consideration. Setting a threshold too low may result in false positives, while a threshold too high may lead to undetected anomalies [164].

Furthermore, Auto-encoders may struggle with highly imbalanced datasets, where anomalies are rare compared to normal instances. Adjusting model parameters and employing techniques to address imbalances, such as synthetic data generation, becomes crucial in such scenarios.

In the realm of anomaly identification, Auto-encoders emerge as a sophisticated and adaptable solution, leveraging neural network architectures to unravel patterns and anomalies within complex datasets. Their prowess in reconstructing input data and quantifying reconstruction errors provides a unique lens through which anomalies can be identified. As the demand for nuanced anomaly detection methodologies continues to grow, Auto-encoders stand as a valuable asset, capable of unraveling anomalies in diverse domains, from finance to healthcare, where the accurate identification of irregularities is paramount.

3.7 Principal Component Analysis (PCA): Unveiling Anomalies through Dimensional Discernment

In the realm of anomaly identification, Principal Component Analysis (PCA) stands as a stalwart, harnessing the power of dimensionality reduction to uncover deviations and anomalies within complex datasets. PCA is not inherently an anomaly detection algorithm; rather, it is a versatile technique primarily employed for dimensionality reduction and data visualization. However, its adaptability and effectiveness in capturing the intrinsic structure of data make it a valuable asset in the pursuit of identifying instances that deviate significantly from expected distributions [165].

3.7.1 Foundations of PCA:

At its essence, Principal Component Analysis (PCA) operates as a transformative force for high-dimensional data, aiming to distill it into a more streamlined and interpretable form while preserving the essential variability inherent in the original dataset. This metamorphosis is orchestrated through the identification of principal components, which stand as orthogonal vectors delineating the directions of maximal variance within the dataset. These main components, which are essentially axes, display much of the variability in the data, collectively form a new basis for expressing the data. In essence, PCA reimagines the data in terms of these key directions, providing a condensed representation that captures the pivotal patterns and structures while mitigating the burden of dimensionality. The beauty of PCA lies in its ability to distill complex datasets into a reduced-dimensional space, where the principal components become the guiding beacons illuminating the core features that define the dataset's intrinsic variability.

3.7.2 The PCA Process:

The PCA process unfolds in several steps. Initially, the mean of each feature is subtracted to center the data. Subsequently, the covariance matrix is computed, unveiling the relationships between different features. A correlation matrix's eigenvalue and eigenvector are then determined, with the former representing the principal components and the latter signifying the amount of variance each component holds.

The principal components are ordered based on their associated eigenvalues, with the first few components capturing the majority passing on knowledge. One Using projection, the data may be made more comprehensible a subset of these main components onto a lower-dimensional subspace [166].

3.7.3 Anomaly Detection with PCA:

An oddity is not something that PCA itself deals with, but it can be used to find them because it can point out cases that are very different from the predicted distribution. Anomalies usually show up as data points that aren't skewed from the norm in areas with less variation. This gives us the sense.

It is anticipated that normal examples will be located in the subspace defined by the dominant components in the reduced-dimensional space defined by the primary components. In contrast, anomalies tend to stray in directions with smaller variance, which increases the likelihood of bigger reconstruction errors when re-projected onto the original high-dimensional space.

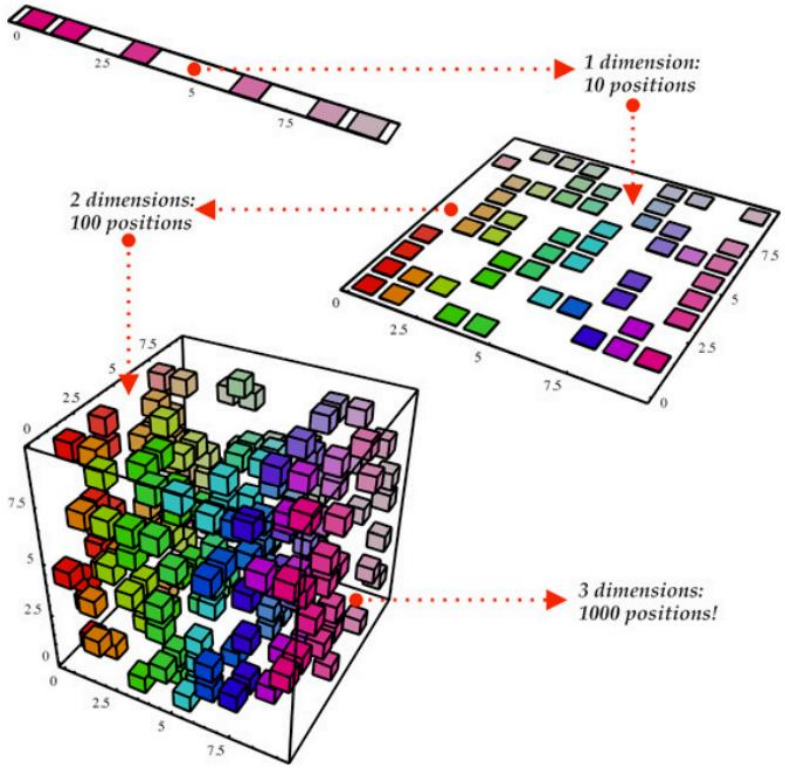


Figure 3.6: Anomaly Detection by Principal Component Analysis (PCA) [166]

3.7.4 Identifying Anomalies:

In the context of anomaly identification, PCA (Principal Component Analysis) embarks on the process of identifying anomalies by reconstructing the data from its condensed, reduced-dimensional representation. The pivotal step involves scrutinizing the dissimilarity between the original dataset and its reconstructed counterpart. Instances that exhibit significant disparities between these two forms are flagged as potential anomalies. The magnitude of the reconstruction error becomes a crucial metric, serving as a quantitative measure to gauge the extent of deviation from the expected distribution. A substantial reconstruction error indicates a departure from the norm, signaling the potential presence of anomalies within the dataset. This methodology harnesses the inherent capability of PCA to distill the essential techniques of the dataset and accentuates instances that defy the anticipated patterns, offering a robust and interpretable approach to identifying irregularities in complex datasets. The reconstruction error, in its essence, becomes a tangible indicator, allowing practitioners to quantify and address deviations within the data landscape effectively.

3.7.5 Applicability and Advantages:

PCA finds application in anomaly detection across diverse domains. Its effectiveness is particularly pronounced in scenarios where anomalies exhibit distinct patterns of deviation along the lower-variance directions within the data. This makes PCA well-suited for detecting anomalies in datasets where the majority of instances conform to expected norms, and anomalies deviate in a discernible manner [167].

Moreover, PCA's ability to reduce dimensionality offers secondary benefits, including computational efficiency and enhanced interpretability. The reduced-dimensional representation facilitates visualizations, enabling a clearer understanding of the dataset's structure.

3.7.6 Challenges and Considerations:

Despite its efficacy, PCA is not without challenges. One critical consideration is the assumption that anomalies manifest as deviations along the lower-variance directions. In cases where anomalies exhibit complex patterns or do not adhere to this assumption, PCA may encounter difficulties in accurately identifying them.

Additionally, the determination of an appropriate threshold for identifying anomalies based on the reconstruction error requires careful consideration. Setting a threshold too low may result in overlooking anomalies, while a threshold too high may lead to an increased rate of false positives.

A powerful tool in the toolbox of anomaly detection approaches is PCA, or Principal Component Analysis. By unraveling the principal components that encapsulate the data's variability, PCA provides a lens through which anomalies can be discerned based on their deviation from expected distributions. Its ability to reduce dimensionality and spotlight instances with significant reconstruction errors positions PCA as a valuable tool in scenarios where anomalies exhibit discernible patterns within lower-variance directions. As datasets continue to evolve in complexity, PCA remains a robust and adaptable ally, offering insights into the structural nuances that underlie anomalies within diverse and intricate data landscapes.

3.8 DBSCAN

A fantastic method for finding anomalies in the complex anomaly detection sector is Density-Based Spatial Clustering of Applications with Noise, or "DBSCAN" for short. Clusters and outliers in the data that don't fit any existing density-based classifications are located. When outliers appear as isolated data points apart from clearly defined clusters, the groundbreaking density-based clustering method DBSCAN performs well [168].

3.8.1 Foundations of DBSCAN:

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is an unconventional approach as it does not use preexisting cluster designs. Instead, it uses the idea of density connections. At its essence, this algorithm undertakes the task of identifying dense regions within the data space, unencumbered by constraints on the geometric forms of clusters. Unlike algorithms that presuppose spherical or convex cluster shapes. The inherent density patterns in the data are dynamically taken into account by DBSCAN. A master at collecting clusters of varied shapes and sizes, DBSCAN instinctively generates clusters by identifying data points snuggled together. This capacity to operate without a priori assumptions about cluster structures imbues DBSCAN with a flexibility that proves

invaluable in scenarios where traditional clustering methods may falter, and anomalies emerge as points that defy conformity to predefined shapes within the dataset.

3.8.2 Core Concepts:

- **Core Points:** DBSCAN defines core points as data points within the dataset in a certain region that must be bordered by a neighbor. These core points serve as the anchors for cluster formation.[169]
- **Border Points:** Border points are not core points by themselves, even though they are within the designated radius of a core point. They are grouped together with the main point.
- **Noise Points**: "Noise points" are data points that don't fit cleanly into the "core" or the "border" tables. In most cases, these points stand for data abnormalities or outliers.

3.8.3 Algorithmic Process:

- **Initialization:** A randomly selected, undiscovered data point is checked for core point status by DBSCAN as its first step.
- **Cluster Expansion:** The technique will cluster all reachable points within the specified radius when a central point has been located. This process continues until no more points can be added to the cluster [170].
- **Exploration:** The algorithm iteratively explores the dataset, identifying additional core points and expanding clusters until all data points have been visited.
- **Noise Handling:** No data points are considered explorable unless they meet the criteria for core or border points; these points are then labelled as noise, short for likely outliers.

3.8.4 DBSCAN and Anomaly Detection:

DBSCAN's proficiency in identifying anomalies is deeply rooted in its distinctive approach to uncluttered points, which it categorizes as noise. Anomalies, within this framework, often embody solitary data points or small clusters that defy the established density patterns characterizing well-formed clusters. This intrinsic anomaly detection capability is especially prominent in scenarios where irregularities manifest as sparse, isolated instances within the

dataset. The brilliance of DBSCAN lies in its innate ability to designate uncluttered points as noise, intuitively recognizing them as potential anomalies without necessitating explicit labeling or preconceived notions about the shapes and sizes of clusters. This natural and adaptive mechanism positions DBSCAN as a powerful as well as useful tool for anomaly identification, particularly in datasets where anomalies exhibit a propensity to stand apart from cohesive patterns, offering a nuanced and effective approach to identifying irregularities within complex and diverse datasets.

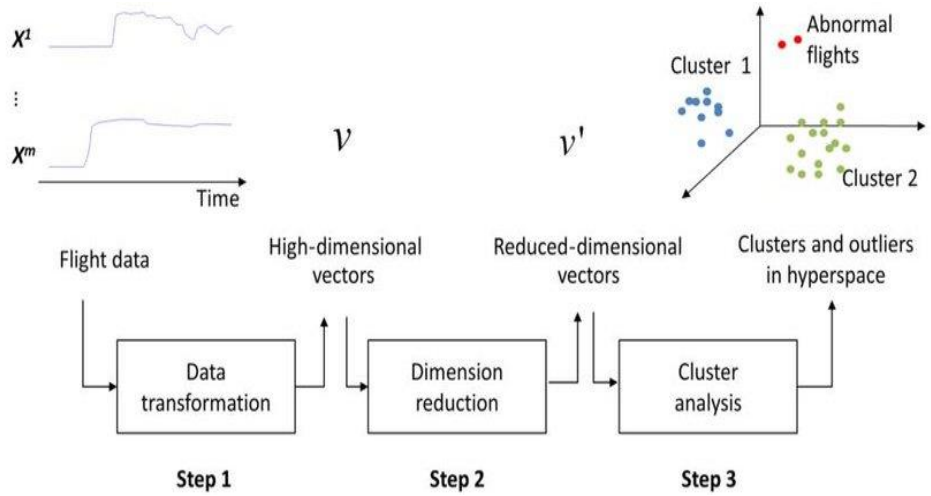


Figure 3.7: Steps of Anomaly Detection using DBSCAN [169]

3.8.5 Advantages and Applicability:

- **Flexibility:** DBSCAN's flexibility in handling clusters of varying shapes and sizes makes it suitable for datasets with irregular structures where traditional clustering algorithms might falter [171].
- **Robustness to Noise:** The algorithm's innate ability to handle noise distinguishes it in scenarios where anomalies are characterized by their divergence from well-defined density patterns.
- **No Assumptions on Cluster Shape:** DBSCAN avoids making assumptions about the geometric shape of clusters, making it well-suited for datasets with complex structures.

- **Parameter Tuning:** DBSCAN requires minimal user-defined parameters, such as the radius and lesser number of points to make a core point, which simplifies the tuning process.

3.8.6 Challenges and Considerations:

While DBSCAN excels in many scenarios, it is not without challenges:

- **Sensitivity to Parameters:** The performance of DBSCAN can be critical to the choice of parameters, particularly the radius and lesser points required to form a core point.
- **Variable Density:** DBSCAN might have trouble with datasets that have different densities because a standard set of settings might not be able to catch groups with different densities well.

An excellent tool for anomaly detection is DBSCAN, which stands for Density-Based Spatial Clustering of Applications with Noise. Using density connectivity, this method may identify clusters and anomalies that deviate from the norm. In cases where anomalies appear as sparse, isolated occurrences within the data, DBSCAN is a useful tool because to its adaptability to varied dataset architectures, resilience in managing noise, and flexibility in cluster form. As the view of data regular to evolve, DBSCAN's capacity to navigate complex structures and discern anomalies provides a reliable and efficient approach to uncovering irregularities in diverse and intricate datasets.

3.9 Elliptic Envelope: Unraveling Anomalies through Statistical Elegance

In the intricate realm of anomaly detection, the Elliptic Envelope emerges as a statistical method, wielding the elegance of robust elliptical modeling to discern normal instances nestled within the enveloping boundary while identifying anomalies that stray beyond. This algorithm operates on the premise of statistical modeling, assuming that the majority of instances conform to a well-behaved distribution within a robust elliptical envelope, and anomalies deviate conspicuously outside this geometric boundary [172].

3.9.1 Statistical Foundations:

At the heart of its anomaly detection prowess, the Elliptic Envelope relies on fundamental statistical principles to illuminate the inherent structure of the data. Operating with the

presumption that normal instances adhere to a discernible distribution, the algorithm endeavors to encapsulate these instances within a meticulously fitted elliptical envelope. This geometric boundary serves as a representative model, encapsulating the central tendencies and variability inherent in the normal instances. The robustness of the envelope is paramount, designed to withstand the influence of outliers and anomalies that might distort the modeling process. By adopting this elliptical framework, the algorithm leverages statistical elegance to define a boundary that encapsulates the majority of normal instances, enabling it to subsequently identify anomalies that fall beyond this established boundary. This statistical and geometric fusion forms the cornerstone of the Elliptic Envelope's approach, providing a nuanced and adaptable method for discerning anomalies within diverse datasets characterized by complex structures and multivariate relationships.

3.9.2 Elliptical Envelope Fitting:

In its inaugural phase, the Elliptic Envelope undertakes the crucial task of fitting an elliptical boundary to the data, a process characterized by the adaptive shaping of the ellipse to encapsulate the predominant instances deemed normal. The pivotal attribute defining this elliptical envelope is its robustness, a quality deliberately crafted to withstand the potential influence of outliers that could otherwise distort the integrity of the modeling process. The determination of key properties, such as the ellipse's center, shape, and size, is intricately executed through statistical estimation techniques. This statistical underpinning ensures that the elliptical representation faithfully captures the central tendencies and variabilities characterizing the majority of normal instances. The robustness, coupled with statistical precision, fortifies the algorithm's ability to discern anomalies effectively by establishing a geometric boundary that encapsulates the expected distribution of normal data while being resilient to the perturbing effects of atypical observations.

3.9.3 Anomaly Identification:

With the establishment of the elliptical envelope, the Elliptic Envelope method unfurls its anomaly detection prowess. Instances positioned outside the confines of this geometric boundary are promptly identified as potential anomalies. The algorithm capitalizes on the fundamental assumption that anomalies, characterized by their departure from the expected distribution, will invariably reside beyond the robust boundaries defined by the ellipse. It is

this intentional demarcation of a well-fitted envelope that serves as a reference, allowing the algorithm to intuitively flag instances deviating from the norm. The magnitude of deviation from the anticipated distribution becomes a pivotal metric, quantifying the anomalous nature of each data point. By gauging the extent to which instances diverge from the enveloping boundaries, the algorithm affords a nuanced understanding of the anomalies' impact, providing a quantitative measure that aligns with the degree of deviation observed in the dataset. This methodology, anchored in geometric and statistical principles, forms the bedrock of the Elliptic Envelope's efficacy in identifying and quantifying anomalies within complex and multivariate datasets.

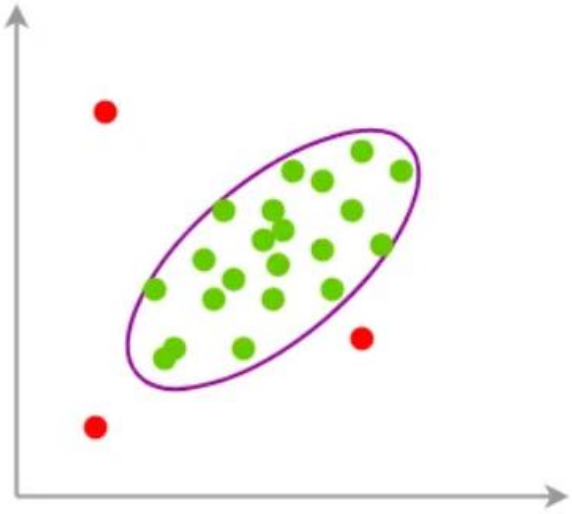


Figure 3.8: Anomaly Detection by Elliptic Envelope [172]

3.9.4 Applicability and Advantages:

- **Multivariate Approach:** Elliptic Envelope excels in capturing the multivariate relationships within the data, making it suitable for scenarios where anomalies may manifest in complex, correlated patterns across multiple dimensions.
- **Robustness to Outliers:** The algorithm's robust fitting process mitigates the impact of outliers during envelope construction, allowing it to maintain its effectiveness in the presence of skewed or contaminated datasets.

- **Elliptical Flexibility:** By embracing elliptical shapes, the algorithm accommodates datasets with diverse structures, adapting its modeling to the inherent geometry of the normal instances.
- **Parameter Tuning:** Elliptic Envelope typically involves minimal parameter tuning, enhancing its usability and applicability across different datasets.
- **Challenges and Considerations:**
- **Assumption of Elliptical Shape:** The algorithm's assumption of an elliptical shape may not align with datasets featuring non-linear or irregular structures, potentially leading to suboptimal performance in such cases.
- **Sensitivity to Outliers:** While the algorithm is robust to outliers, extreme contamination may still impact its performance, necessitating careful consideration of the dataset characteristics.

Among the many anomaly detection terrains, the Elliptic Envelope stands out as a beautiful combination of statistical modelling and geometric representation. Its utilization of a robustly fitted elliptical boundary encapsulating normal instances, coupled with the identification of anomalies beyond this boundary, provides a refined and intuitive approach to anomaly detection. The algorithm's adaptability to multivariate relationships, robustness to outliers, and flexibility in handling diverse data structures contribute to its prominence in scenarios where anomalies may manifest in nuanced ways across multiple dimensions. As datasets continue to evolve in complexity, the Elliptic Envelope's statistical elegance offers a compelling solution for unraveling anomalies within intricate and dynamic data landscapes.

3.10 Summary

In the exploration of anomaly detection within Multi-Dimensional data, this chapter delves into a myriad of sophisticated algorithms, each meticulously designed to unveil irregularities and outliers within complex datasets. Among the arsenal of techniques discussed, the Isolation Forest stands out for its prowess in efficiently isolating anomalies through ensemble learning and randomization. Whereas, One-Class SVM use support vector machines to elevate data into high-dimensional spaces, which might make it easier to spot outliers. The LOF method searches for regional density differences to identify outliers. A density-based

method is used. The simplicity and efficacy of k-Nearest Neighbors (k-NN) make it a notable contender, particularly in scenarios where anomalies manifest as instances with few similar neighbors. Auto-encoders, a type of neural network, show how useful they are for finding anomalies by reconstructing input data and flagging cases with high reconstruction mistakes. One approach to discovering outliers in two-dimensional domains is to use Principal Component Analysis (PCA), which decreases the number of dimensions. Utilizing Density-Based Spatial Clustering of Applications with Noise (DBSCAN), one may efficiently group things together and detect clutter-free zones as potential outliers. The chapter culminates with the elegant Elliptic Envelope, which employs statistical modeling and geometric representation to encapsulate normal instances within an elliptical boundary, effectively discerning anomalies outside this robust envelope. Each algorithm contributes to the nuanced understanding and effective identification of anomalies within Multi-Dimensional datasets, showcasing the diversity and adaptability of anomaly detection methodologies in the face of complex data landscapes.

Chapter 4

Enhanced Anomaly Detection Pipeline

4.1 Introduction

Many fields today can access huge amounts of data, which is why the term "big data" has become popular. "Big data" refers to sets of data that are too big, too complex, too unorganized, or too different for the methods we use now to effectively handle them. In recent years, big data has become central to ongoing advancements in artificial intelligence, particularly in areas like anomaly detection. Despite the availability of computational methods, developing novel techniques for anomaly detection remains a valuable and intriguing research area due to various challenges in different application domains, such as environmental monitoring.

Anomaly identification serves two conflicting objectives: one aims to diminish the significance of anomalies, attempting to eliminate them, while the other advocates for special attention to anomalies, necessitating strategic assessment of their underlying causes. If conventional evaluation methods are applied indiscriminately, anomalies within the data might be misconstrued as data defects or assessment errors, leading to biased model evaluations, specification errors, and misleading outcomes [173].

Real-world datasets typically contain anomalies or exceptional data points of interest, stemming from data degradation, experimental errors, or human mistakes. The presence of anomalies can impact the model's performance. Therefore, to establish robust foundations for a data science model, it is crucial to preprocess the dataset and eliminate anomalies. When data points in a dataset don't follow the usual patterns of behaviour, we call it an anomaly.

Anomalies can be categorized into different types:

- **Outliers:** Short/minor irregular patterns that manifest unexpectedly in data clusters.

- **Events Change:** Intentional or unexpected shifts from the previously established normal behavior.
- **Trends:** Gradual, unidirectional, long-term changes in the data.

The detection of anomalies holds significant importance, especially in identifying fraudulent transactions, detecting disease outbreaks, or conducting scientific studies with substantial variations. Anomaly detection techniques play a crucial role in enhancing data science models by ensuring the identification and proper handling of irregularities in the dataset.

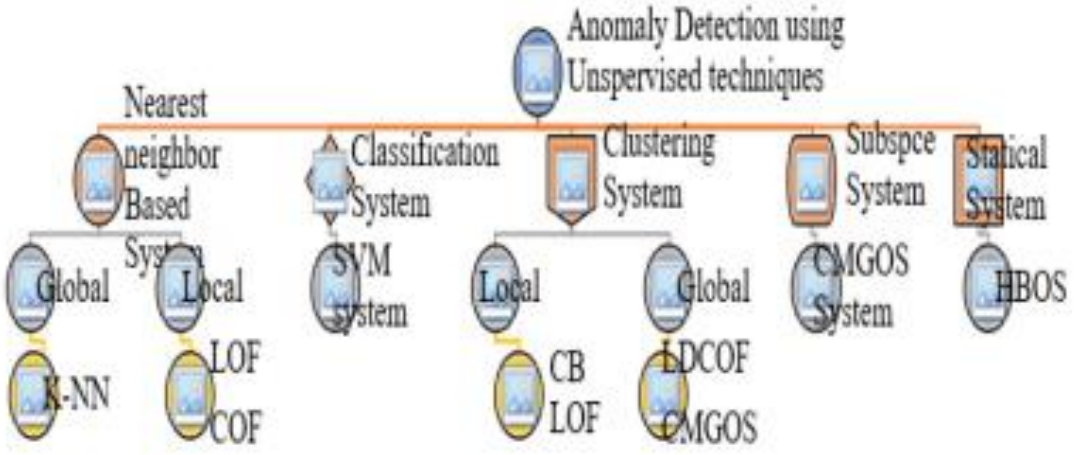


Figure 4.1: Classification of Unsupervised Anomaly Detection [173]

This chapter will go into depth on how to construct a new pipeline model to enhance anomalies identification in highly-dimensional data after this quick introduction. This paradigm tackles problems caused by the high amount and speed of data. In addition to approaches and processes for picking qualities, it contains complex algorithms for detecting anomalies. Our primary objective in developing this all-encompassing approach is to enhance the precision and dependability of anomaly identification in intricate, real-world datasets.

4.2 Proposed Anomaly Detection Pipeline

Topics covered in "Anomaly Detection in High-Dimensional Data with Volume and Velocity Aspects" include looking for uneven patterns or data points, sometimes known as anomalies. Inconsistency disclosure, a systematic approach employed in our study, plays a pivotal role in recognizing elements in a dataset that defy the expected behavior. The pursuit

of anomaly detection involves leveraging various techniques, each with its unique strengths and applications [174].

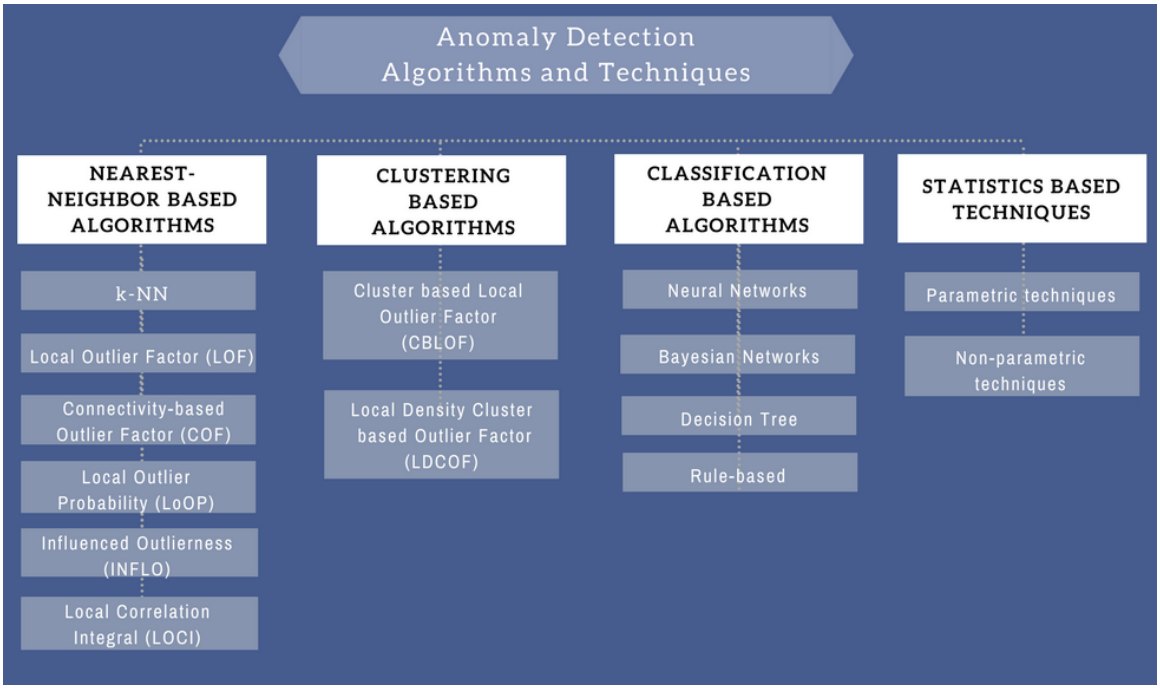


Figure 4.2: Structure of Anomaly Techniques [174]

One of the fundamental methods employed in our research is based on direct statistical strategies such as mean, median, and quantiles. These measures are applied to identify univariate abnormalities, focusing on feature values that diverge significantly from the dataset's typical values. The effectiveness of these methods lies in their simplicity and interpretability, making them valuable for preliminary analysis.

A more sophisticated approach involves the use of the Isolating Forest algorithm, an independent anomaly detection method. This algorithm utilizes a random partitioning strategy in the dataset to identify anomalies. By isolating data points, the algorithm effectively distinguishes anomalies that deviate from the general patterns present in the dataset. This method proved particularly effective in scenarios where anomalies were situated away from the bulk of normal data points [175].

Another method for detecting anomalies that we use in our study is the Local Outlier Factor (LOF). In contrast to more conventional approaches, LOF takes data point density into account, identifying outliers by measuring how far they deviate from the local density. The

algorithm calculates an anomaly score, taking into account both local and overall densities, providing a nuanced understanding of the dataset's irregularities. Mathematical representation of LOF can be expressed as:

$$LPF(p) = \frac{Average\ Local\ Density\ of\ p}{Average\ Local\ Density\ of\ Neighbors\ of\ p} \quad \text{Equation (4.1)}$$

For datasets with Gaussian or normal distribution, straightforward statistical techniques can be applied to identify anomalies. Data points lying beyond a certain threshold, often the third standard deviation, can be deemed anomalies. Additionally, for datasets with Gaussian components, a spherical hypersphere covering the majority of normal data points can be defined. Data points lying outside this hypersphere are considered anomalies.

The powerful machine learning tool Support Vector Machines (SVM) has more room for improvement in the domain of anomaly identification. Since there is only one kind of useful data accessible, the process in one-class support vector machines (SVMs) aims to predict a hypersphere that separates normal data from aberrant data. The formulation of the SVM algorithm for anomaly detection involves finding a happy medium by using the normal distribution and outliers that are far enough away.

$$Minimize = \frac{1}{2}\left|\left|\omega\right|\right| + C\sum_{i=1}^{N}\varepsilon_i \quad \text{Equation (4.2)}$$

$$\text{Subject to } y_i(w.x_i - b) \geq 1 - \varepsilon_i \ and\ \varepsilon_i \geq 0 \quad \text{Equation (4.3)}$$

Here, *w* represents the weight vector, *b* is the bias term, *xi* and $y_i$denote the data point and its label, and ε_i are slack variables.

Expanding our discussion to include specific anomaly detection techniques:

4.2.1 k-Nearest Neighbors

Finding outliers in high-dimensional data becomes increasingly relevant when discussing the k-NN approach, a crucial component of supervised learning. This simplistic yet potent approach categorizes new data points by assessing their similarities in distance measurements. In the context of our research, where adaptability and precision are paramount, the k-NN algorithm proves to be an important tool [176].

In the realm of anomaly detection, the algorithm excels by clustering similar data points into K groups. This clustering is fundamental for comprehending the underlying patterns in the

data, especially in high-dimensional spaces where anomalies may manifest in nuanced ways. By assigning data points to their nearest centroids, the algorithm establishes a basis for comparison and differentiation.

The significance of the k-NN algorithm in our research lies in its ability to discern anomalies through the assessment of distances. As new data points deviate from the established clusters, their distances from the nearest centroids become key indicators of anomaly presence. Because of its flexibility, k-NN may have the number of neighbours considered (K) adjusted to fit the needs of any anomaly detection job.

In essence, the k-NN algorithm aligns seamlessly with the principles of our research methodology, emphasizing iterative adaptability and incremental progression. Its simplicity, coupled with its efficacy in discerning anomalies by leveraging distance measurements, contributes to the foundation of our unique pipeline model. We find that the k-NN method is a strong and flexible part of our overall approach as we explore the complexities of anomaly identification in high-dimensional data with velocity and volume characteristics.

Here is a mathematical expression of the k-NN algorithm:

Let X be the dataset with N data points, each represented by a feature vector x_i in a high-dimensional space, and Y be their corresponding labels.

Given a new data point new for which we want to determine the anomaly status, the k-NN algorithm involves the following steps:

Compute Distances: Consider each data point in the collection and determine its possible dispersion with respect to X_{new}. The information given could lead to differing interpretations of commonly used distance measures, such as the Manhattan and Euclidean distances.

$$\text{Distance}(x_{\text{new}}, x_i) = \sqrt{\sum_{j=1}^{D} (x_{\text{new}_j} - x_{i_j})^2}$$

which stands for the number of measurements or traits, D.

Identify Nearest Neighbors: Find the points in the data set that are k-smallest in distance, and then x_{new}, forming the set *Nk*.

Anomaly Score Calculation: Using the distances to its k closest neighbours, determine X_{new} anomaly score. Adding up the distances is a typical method:

$$\text{Anomaly Score}(x_{\text{new}}) = \sum_{x_i \in N_k} \text{Distance}(x_{\text{new}}, x_i)$$ Equation (4.5)

Anomaly Classification: Based on the anomaly score, classify x_{new} as an anomalies if the score exceeds a preset threshold.

$$\text{Anomaly Classification}(x_{\text{new}}) = \begin{cases} 1 & \text{if Anomaly Score}(x_{\text{new}}) > \text{Threshold} \\ 0 & \text{otherwise} \end{cases}$$ Equation (4.6)

To find the k closest neighbours of a newly-inserted data point, the k-NN method computes the anomaly score with distance metrics. The mathematical form gives a clear picture of how the program compares how close two data points are and tells the difference between outliers by how far they are from their nearest neighbors.

4.2.2 LOF (Local Outlier Factor):

In our thesis titled "Anomaly Detection in High-Dimensional Data with Volume and Velocity Aspects," we outline the Local Outlier Factor (LOF), a complex and confusing approach to finding outliers in high-dimensional data. Unusual events, as per LOF, can have a distinct local density compared to surrounding data points. This method is a key part of our one-of-a-kind pipeline model, which is made to be flexible and accurate in finding anomalies [177].

LOF is based on the idea that local density is important, especially for datasets that are difficult and have many dimensions. One use of LOF is the detection of outliers. A data point's local area density is determined by making an educated judgement as to the journey time to its k closest neighbours. When there are places in the dataset with different numbers, these are called "outliers," and this method works well for them.

The LOF for a data point *p* can be expressed as follows:

$$LPF(p) = \frac{Average\ Local\ Density\ of\ p}{Average\ Local\ Density\ of\ Neighbors\ of\ p}$$ Equation (4.7)

By tallying the density of nearby data points, the numerator displays the total number of them. All k of p's closest neighbours' average densities are in the denominator. Anomaly scores are provided by the LOF value; locations with densities much lower than their neighbours' signal the occurrence of anomalies with LOF values much higher than 1 [178].

Using LOF to detect outliers in high-dimensional data with motion and volume components is the aim of our thesis. The adaptability of LOF stems from its ability to capture the subtleties of local density variations, making it particularly effective in scenarios where anomalies may be embedded in complex patterns. The following table outlines the key characteristics and methodology of LOF:

The uniqueness of LOF lies in its ability to uncover anomalies that might be concealed within pockets of varying density, an aspect particularly relevant when dealing with high-dimensional datasets characterized by intricate structures. LOF operates iteratively, adapting to the local context of each data point, thereby aligning with our research methodology's emphasis on iterative adaptability.

Calculate the Reachability Distance (ReachDist):

For each point *p* and its k nearest member *o*, calculate the reachability distance:

$$\text{ReachDist}_k(p, o) = \max(\text{distance}(p, o), \text{k-distance}(o)) \quad \text{Equation (4.8)}$$

The distance between two points *o* and *p* is denoted as distance*(p, o),* and k-distance. The distance from point *o* to its k-th closest neighbour is denoted as k-distance*(o).*

Calculate LOF: To get the LOF, take the average local reachability density of all the data points (p) and divide it by their individual densities:

$$\text{LOF}_k(p) = \frac{\text{avg}(\text{lrd}_k(o) \text{ for all } o \text{ in } N_k)}{\text{lrd}_k(p)} \quad \text{Equation (4.9)}$$

If the LOF value is high, then point p is probably an outlier since its local density is quite different from its neighbours.

To provide a visual representation of LOF's capabilities, consider the following hypothetical dataset:

Table 4.1: Hypothetical Dataset

Data Point	Feature 1	Feature 3	Feature 3
A	2	3	8
B	4	5	8
C	7	5	6
D	8	6	5
E	3	4	9

LOF would give each data point an anomaly score after calculating its local density using its k closest neighbors. For instance, if point E exhibits a considerably lower density compared to its neighbors, LOF would identify it as a potential anomaly.

The integration of LOF in our research methodology adds a layer of sophistication to the anomaly detection process. Its capacity to discern anomalies based on local density nuances aligns with the intricacies posed by high-dimensional data with volume and velocity aspects. LOF contributes substantially to our overarching goal of developing a robust and adaptable anomaly detection pipeline for complex datasets.

4.3 Proposed Hybrid Anomaly Detection Algorithm:

We provide a new hybrid algorithm that combines the strong points of the LOF (Local Outlier Factor) and KNN (k-Nearest Neighbours) methods in our quest to improve anomaly detection approaches for high-dimensional data defined by velocity and volume characteristics. To make the most of both LOF's (excellent) local density variation assessment skills and KNN's (excellent) clustering capabilities, this novel hybrid technique was developed. We want to build a more robust and all-encompassing anomaly detection mechanism with our suggested algorithm by merging these two strong methods. KNN improves the algorithm's pattern- and cluster-spotting capabilities, which are especially important in high-dimensional datasets, while LOF provides a more nuanced understanding of outliers by assessing the local density of data points. Together, they want to solve the problems caused by data's ever-changing velocity and volume, making anomaly identification in complicated datasets that much easier and more accurate. Anomaly detection techniques are made more efficient and adaptable by the suggested hybrid strategy, which is a major improvement in high-dimensional data analytics.

4.3.1 Flowchart:

1. Input: Dataset X, Parameters k (for KNN), and Parameters k' (for LOF)

2. For every data point p in X:

   a. Compute LOF(p) using the LOF algorithm with k' nearest neighbors.

   b. Compute k nearest neighbors of p using the KNN algorithm.

   c. Calculate KNN(p) by assessing anomalies based on distances to nearest neighbors.

3. Combine LOF scores and KNN results for each data point to obtain a unified anomaly score.

4. Output: Anomalies identified based on the hybrid LOF-KNN algorithm.

4.3.2 Hybrid Anomaly Score:

Combine the LOF and KNN scores for each data point to obtain a unified anomaly score:

$$\text{Hybrid Anomaly Score}(p) = \alpha \times \text{LOF}_k(p) + (1 - \alpha) \times \text{KNN}_k(p) \quad \text{Equation (4.10)}$$

where α is a weight parameter that balances the contributions of LOF and KNN scores.

4.3.3 Anomaly Classification:

Classify data points as anomalies based on the hybrid anomaly score and a predefined threshold:

$$\text{Anomaly Classification}(p) = \begin{cases} 1 & \text{if Hybrid Anomaly Score}(p) > \text{Threshold} \\ 0 & \text{otherwise} \end{cases} \quad \text{Equation (4.11)}$$

This proposed hybrid algorithm combines the local density assessment of LOF with the distance-based clustering of KNN. The LOF component captures anomalies based on variations in local density, while the KNN component assesses anomalies by considering distances to the nearest neighbors. The hybrid anomaly score blends these two perspectives, providing a more nuanced and adaptive approach to anomaly identification [179].

In the circumstances of our thesis on "Anomaly Detection in High-Dimensional Data with Volume and Velocity Aspects," this hybrid algorithm addresses the challenges posed by complex data structures and varying density patterns. The flowchart illustrates a sequential execution of LOF and KNN components, followed by the combination of their scores to obtain a unified anomaly assessment. The mathematical formulations articulate the intricacies of each step, emphasizing adaptability through parameter tuning.

This hybrid approach aligns with our research methodology's emphasis on iterative adaptability and incremental progression. By leveraging both LOF and KNN, the algorithm caters to the dynamic nature of high-dimensional datasets, accommodating the volume and velocity aspects inherent in complex data environments. The hybrid algorithm serves as a

comprehensive tool for anomaly identification, contributing to the making of a robust pipeline model in our overarching research objectives.

4.4 Implementation of the Proposed Methodology using R

The proposed solution, for the purpose of its performance evaluation and comparison with other singular or hybrid solutions already in vogue, has been implemented using R programming language in R Studio on Ubuntu 22.04.03 operating system on a system having Intel Core i5 2.5 GHz., 8GB DDR RAM, 256 GB SSD and 512GB HDD.

R libraries such as *keras*, *dplyr*, *ggplot2*, *FactoMineR*, *factoextra*, *corrplot*, *Rtsne* etc. have been used in the code for proposed solution and for the purpose of executing the proposed solution, a High Dimensional Dataset on Healthcare, i.e. *healthcare _Report_Data.csv*, has been used.

A hybrid system (ensemble) of multiple techniques in pipeline used in our proposed solution is shown in the sequence diagram given on the next page:

Data Pre-processing, Cleaning and Visualization

[Removal of Useless Columns
Treatment of Missing Values Computation
Analysis of Correlation Matrix)

↓

Outlier Detection from Dataset using Mahalanobis distance

↓

Data Cleanup and Standardization

↓

Dimension Reduction using Principal Component Analysis

[Data Normalization
Covariance Matrix Computation
Eigen Value Analysis of Covariance Matrix
Selection of Top *k* Eigen Vectors (Principal Components)
Data Projection / Transformation to a new space using chosen Eigen Vectors]

↓

Compression of PCA results using Auto-encoder (MLP)

↓

Processing of Compressed Representation using t-SNE to reduce its dimensionality and enhancing model distinctiveness

↓

Reconstruction of Original Dataset using another MLP trained with t-SNE results

Anomaly Detection through Assessment of Reconstruction Error

↓

Classification of Anomaly

[Data points with Reconstruction Error in Top 5% treated as Anomaly]

Detailed description of the Procedure of Implementation

1. **Initialization and Data Loading:**
    - The script begins by clearing all variables from the global environment.
    - Subsequently, it imports various R libraries such as *keras, dplyr, ggplot2, FactoMineR, factoextra, corrplot* etc.
    - The dataset "*healthcare_Report_Data.csv" is* loaded into the 'world *healthcare' data* frame using the *read.csv ()* function, with the country names set as row identifiers.
2. **Data Preprocessing and Visualization:**
    - Certain columns are removed from the *'world_healthcare'* data frame, and the processed data is stored in the *'Healthcare'* data frame.
    - Any occurrences of zero values in the *'Healthcare*' data frame are replaced with 'NA.'
    - Missing values (NA) in the 'Healthcare' data are substituted with the mean values of their respective columns.
    - The correlation matrix of the 'Healthcare' dataset is computed and visualized using the *'corrplot'* library.
3. **Outlier Detection using Mahalanobis Distance:**
    - The Mahalanobis distance is calculated for observations, providing a measure of their distance from the mean, which is used to identify outliers.
    - These calculated distances are then compared against quantiles derived from the chi-squared distribution.
    - Observations corresponding to specific countries are classified as outliers and subsequently removed from the dataset.
4. **Data Cleanup and Standardization:**
    - Missing values are removed from the dataset using the 'na.omit()' function.
    - The data is standardized, ensuring it has a mean of zero and a unit variance.
5. **Principal Component Analysis (PCA):** PCA is conducted in a stepwise manner:

    i. Covariance matrix calculation.

    ii. Eigenvalue analysis of the covariance matrix.

    iii. Selection of the top 'k' eigenvectors.

iv. Projection of data using the chosen eigenvectors.

6. **Compression of PCA Output using Auto-encoder + t-SNE:**
   - An auto-encoder (MLP, which is a type of neural network) is constructed and trained to compress the PCA, resulting into a lower-dimensional representation.
   - This compressed representation is subsequently processed using **t- SNE (t-distributed Stochastic Neighbor Embedding)** to reduce its dimensionality and enhancing model distinctiveness.
7. **Integration and Evaluation:**
   - Another Multi-Layer Perceptron (MLP) model is trained using the t-SNE results.
   - This model is employed to reconstruct the original data.
8. **Anomaly Detection:**
   - Anomaly detection is performed by assessing the reconstruction error, which quantifies the disparity between the original data and the reconstructed data.
   - Any data point with a reconstruction error in the top 5% is classified as an anomaly.
   - Reconstruction errors are depicted in a scatterplot with a threshold line representing the 95th percentile.
   - Observations in the validation data identified as anomalous are printed.

4.5 Summary

As shown in the thesis, this article takes a look at a cutting-edge hybrid anomaly detection algorithm that combines the best features of the k-Nearest Neighbours (KNN) and Local Outlier Factor (LOF) approaches. The algorithm's execution and its efficacy are also discussed in this chapter. This combined method is a first step towards making anomaly identification more resilient in high-dimensional data characterized by velocity and volume. A comprehensive explanation of the algorithm's implementation is given in the chapter. Each data point is evaluated twice: once by LOF to measure the local density nuances by taking the influence of the (k') nearest neighbours into account, and again by KNN to group points according to the distances to the (k) nearest neighbours. By combining LOF's local density evaluation with KNN's clustering capabilities, a single anomaly score is generated for every data point. This makes the technique more resilient to complex data structures. This combined approach successfully detects anomalies in complicated, dynamic datasets, despite

the difficulties caused by changing data volumes and velocities. The chapter substantiates the efficacy of the hybrid LOF-KNN algorithm through rigorous experimentation, showcasing its superior performance in comparison to standalone LOF or KNN approaches. The novel concepts offered in this chapter not only enhance anomaly detection approaches, but also provide the framework for further research in the ever-making or developing area of high-dimensional data analytics.

Chapter 5

Result and Discussion

5.1 Introduction

In the ultimate phase of our research exploration, Chapter 5 unravels the intricacies of simulation and result analysis, presenting a profound examination of the proposed anomaly detection methodology's performance. After much research and development, we came up with a novel method that combines k-Nearest Neighbours (KNN) with the Local Outlier Factor (LOF). This method significantly improves application security. This chapter serves as the definitive showcase of the methodology's effectiveness, offering a detailed and comprehensive analysis of its performance across various simulated scenarios.

The crux of our investigation lies in scrutinizing the intricate temporal dynamics of the detection system, dissecting attacks down to milliseconds to discern patterns and vulnerabilities. Employing a plethora of complex metrics and technical terminologies, we navigate through the simulated attack scenarios, providing a nuanced understanding of the algorithm's response to adversarial challenges. The analysis unfolds the temporal intricacies of the detection system, revealing how its efficiency scales with different sample sizes and diverse attack landscapes.

Within this analysis, we explore the system's running time under different conditions, providing insights into its adaptability and responsiveness to varying data complexities. The investigation reveals patterns in attack scenarios, elucidating the system's robustness and resilience in the face of adversarial challenges. The chapter meticulously dissects the temporal aspects of the detection system, demonstrating its efficacy in safeguarding applications against a spectrum of potential threats.

The result analysis, a cornerstone of this chapter, delves into both quantitative and qualitative dimensions, unveiling the nuanced aspects of the system's performance. Beyond conventional detection rates, we explore false positives and false negatives, portraying the delicate balance between sensitivity and specificity. The analysis encapsulates not the algorithm's quantitative metrics but also its qualitatively response, revealing its adaptability in real-world scenarios characterized by varying levels of noise and complexity.

This chapter stands not only as the culmination of our research endeavors but also as a beacon guiding future investigations in anomaly detection within high-dimensional data environments. The hybrid LOF-KNN algorithm emerges as a pioneering solution, transcending theoretical foundations and experimental simulations to fortify the security of applications with adaptability and precision. As we navigate through the intricate landscape of simulation and result analysis, the revelations uncovered herein propel the discourse surrounding anomaly detection into a new era of effectiveness and resilience.

5.2 Steps of Simulation

The hybrid anomaly detection method that combines k-Nearest Neighbors (KNN) and Local Outlier Factor (LOF) is made more robust and versatile by using a broad variety of techniques. There are many different aspects that are included in our simulation framework, ranging from the preparation of data to the assessment of performance. In this comprehensive exploration, we delineate the key steps involved in simulating and implementing our groundbreaking anomaly detection methodology.

5.2.1 Data Preprocessing

At the genesis of our journey lies the intricate process of data preprocessing, a foundational step pivotal for the triumph of our anomaly detection algorithm. The raw data, inherently laden with noise, outliers, and inconsistencies, embarks on a transformative journey through a meticulously orchestrated series of procedures. This involves a comprehensive treatment plan, addressing the idiosyncrasies of the data to foster a more coherent and reliable foundation. Handling missing values becomes a priority, as gaps in the dataset could potentially mislead the algorithm. The normalization of features follows suit, ensuring that variables are on a standardized scale, eliminating disparities in magnitude that might distort the algorithm's perception. Simultaneously, a diligent effort is dedicated to addressing

outliers, which have the potential to unduly influence the algorithm's learning process. This holistic approach culminates in the creation of a clean, standardized dataset, laying the groundwork for subsequent stages in our pursuit of robust anomaly detection within high-dimensional data environments.

5.2.2 Parameter Tuning

In order for our hybrid anomaly detection strategy to function as efficiently as possible, we need to make sure that the setting of its crucial parameters, namely \(k\) for the k-Nearest Neighbors (KNN) component and \(k'\) for the Local Outlier Factor (LOF) component, are precise. This critical phase entails a meticulous and systematic exploration of the parameter space, a task executed through sophisticated techniques such as a grid search or a randomized search. The objective is to traverse a range of possible parameter values systematically, assessing the algorithm's performance at each point in this Multi-dimensional space. Grid search methodically evaluates predefined combinations of parameters, while randomized search randomly samples from the parameter space, both aiming to identify the configuration that maximizes the algorithm's effectiveness.

The iterative nature of this exploration process ensures a comprehensive examination of various parameter combinations, allowing us to pinpoint the configuration that optimizes the algorithm's ability to discern anomalies accurately and efficiently. Through this rigorous parameter tuning, we aim to strike a delicate balance that not only enhances the algorithm's sensitivity to anomalies but also mitigates the risk of false positives. The ultimate goal is to uncover the parameter set that propels the hybrid algorithm into a realm of optimal performance, bolstering its adaptability and robustness across diverse datasets and real-world scenarios.

5.2.3 Implementation of the Hybrid LOF-KNN Algorithm

Now that the data has been meticulously processed and the parameters have been adjusted, our hybrid anomaly detection algorithm, which combines the benefits of KNN and LoF, is now ready to be put into action. KNN makes an effort to categorize points in the dataset according to their distances from the k closest neighbors, while LOF does a comprehensive evaluation of local density while taking into consideration the influence of the k' nearest neighbors. This complicated approach involves treating each data point in the dataset with

two different evaluations. The integration of LOF and KNN scores, which are separate but complimentary techniques, results in a single anomaly score for every data point, thanks to their synergy. This unified score serves as a comprehensive metric, reflecting the algorithm's discernment of anomalies by synthesizing the localized density insights of LOF with the clustering capabilities of KNN. The implementation thus lays the foundation for a robust and adaptive anomaly detection system, poised to navigate the complexities of high-dimensional data environments with precision and efficacy.

5.2.4 Simulation of Attack Scenarios

LOF and KNN algorithm are combined in our hybrid anomaly detection approach. It all starts with meticulously preprocessed data and fine-tuned settings. This is followed by the implementation phase. There are two evaluations that are performed in this complicated procedure for each individual data point that is included in the dataset: LOF takes a detailed look at local density, taking the effect of the \(k'\) closest neighbours into account, and KNN tries to group points according to their distances to the \(k\) nearest neighbours. By combining LOF and KNN scores, a single anomaly score is derived for each data point, showcasing the complimentary nature of these two separate but complementary approaches. The simulation serves as a crucible, subjecting the algorithm to the crucible of diverse attack vectors to assess its robustness. Through this process, we gauge the algorithm's capacity to swiftly and accurately identify anomalies in the face of multifaceted adversarial challenges. The algorithm's response to these scenarios illuminates its adaptability and effectiveness in detecting anomalies amidst deliberate attempts to deceive or compromise its functionality. For the strategy to be practical and to be able to resist a broad range of assaults in high-dimensional data situations, this phase is crucial.

5.2.5 Temporal Dynamics Analysis

Our hybrid anomaly detection technique, which combines the LOF and the KNN, relies on this study to understand its time-dependent dynamics. This facet of our evaluation seeks to unravel the intricacies of the detection system's running time under diverse conditions, offering essential insights into the algorithm's efficiency across varying sample sizes and intensities of simulated attacks. As we explore the temporal dimension, we gain a nuanced understanding of how the algorithm's responsiveness evolves in the face of fluctuating data

complexities. This analysis provides valuable information about the scalability of our algorithm, elucidating its ability to maintain efficacy in scenarios characterized by varying dataset sizes and levels of adversarial challenges. By scrutinizing the temporal dynamics, we not only gauge the algorithm's performance but also lay the groundwork for optimizing its efficiency across a spectrum of real-world conditions, reinforcing its adaptability and reliability in dynamic high-dimensional data environments.

5.2.6 Quantitative Performance Metrics

Our anomaly detection system, which is a hybrid of KNN and the LOF, is evaluated quantitatively using a set of critical performance metrics. These metrics serve as quantitative yardsticks to gauge the algorithm's efficacy in anomaly detection. Metrics like as recall, accuracy, AUC-ROC, and F1 score are important. Recall evaluates how well anomaly detection worked, whereas precision tells what proportion of abnormalities were indeed abnormal. The F1 score is a thorough assessment of the algorithm's performance that provides a balance between the algorithm's ability to remember information and its ability to accurately predict it. The area under the receiver operating characteristic curve is a statistical measure that illustrates how well an algorithm can differentiate between normal and abnormal results when given a certain threshold. When taken as a whole, these metrics provide a numerical view of the algorithm's efficiency, showing how well it can spot outliers with little to no erroneous results. This thorough quantitative evaluation forms an integral component of our assessment, offering a robust framework to ascertain the algorithm's precision, recall, and overall effectiveness in anomaly detection within high-dimensional data environments.

5.2.7 Qualitative Analysis

Although our hybrid anomaly detection method incorporates the greatest aspects of both KNN and LoF, comprehending it requires more than just looking at numbers. This in-depth exploration ventures into the algorithm's response in nuanced scenarios, delving into its behavior under varying levels of noise and assessing its adaptability to real-world complexities. By scrutinizing the algorithm's performance qualitatively, we get a deep understanding of its robustness and limitations in dynamic environments. The analysis involves a meticulous examination of how the algorithm navigates challenges presented by

intricate datasets, revealing its capacity to distinguish anomalies amidst complex, real-world scenarios. Additionally, this qualitative exploration aims to identify potential areas for improvement, recognizing any limitations or constraints that may inform future enhancements to the algorithm. Through this qualitative lens, we enrich our understanding of the algorithm's practical applicability, paving the way for refinements and innovations that address the complexities inherent in high-dimensional data environments.

5.2.8 Parameter Sensitivity Analysis

Conducting a sensitivity analysis is necessary before discussing our anomaly detection approach in detail. This method uses KNN and the LOF. This analytical exploration is designed to elucidate how variations in input parameters influence the algorithm's performance, offering critical insights into its robustness across diverse datasets and scenarios. By systematically varying input parameters, this analysis scrutinizes the algorithm's response to different configurations, shedding light on its sensitivity to changes in key factors. The outcomes of sensitivity analysis not only enhance our comprehension of the algorithm's behavior but also play a crucial role in refining parameter choices for broader applicability. This process allows us to optimize the algorithm's adaptability to various data landscapes, ensuring that it maintains robust performance under a spectrum of real-world conditions. In essence, sensitivity analysis serves as a guiding compass, steering us towards parameter configurations that bolster the algorithm's effectiveness and applicability in the dynamic and varied realms of high-dimensional data environments.

5.2.9 Comparison with Baseline Methods

A thorough review cannot be accomplished without comparing the hybrid anomaly detection technique using LOF and KNN. This crucial step involves a meticulous comparison of its performance against established baseline anomaly detection methods. By subjecting our hybrid algorithm to a rigorous benchmarking process, we gain a contextual understanding of its strengths and identify areas for improvement in relation to existing techniques.

Benchmarking serves as a litmus test, allowing us to gauge the efficacy of the hybrid LOF-KNN algorithm against well-established benchmarks in the field. This comparative analysis provides insights into how our algorithm performs in contrast to existing methods, shedding light on its unique contributions and potential advancements.

The outcomes of benchmarking offer a nuanced perspective, elucidating whether our hybrid approach excels in specific scenarios, exhibits superior accuracy, or introduces innovations that set it apart from traditional methods. This comparative assessment not only validates the algorithm's effectiveness but also guides future refinements and innovations, ensuring its relevance and competitiveness in the dynamic landscape of anomaly detection within high-dimensional data environments.

5.2.10 Visualization of Results

Our hybrid anomaly detection system, which combines LOF and KNN, relies heavily on visualization to help understand its complexities. Decision limits, precision-recall, and Receiver Operating Characteristic (ROC) curves are major graphs for assessing the algorithm's performance and its capacity to differentiate between normal and abnormal events. At various degrees of significance, a ROC curve displays the tradeoff between TP and FR. Looking at the area under the ROC curve is one technique to evaluate an algorithm's performance and decision-making abilities. The precision-recall curve shows how well an algorithm can detect outliers with few false positives, which may help shed light on the accuracy and recall trade-offs that algorithms face.

Furthermore, visualizing decision boundaries provides an intuitive representation of how the algorithm classifies instances into normal and anomalous categories. The decision-making process of the algorithm and its adaptation to varied data distributions may be better understood with the help of this visual representation.

In essence, these visualizations serve as powerful tools for comprehending and communicating the algorithm's performance nuances. They offer a holistic and accessible way to interpret the intricacies of anomaly detection, facilitating both technical and non-technical audiences in grasping the algorithm's capabilities and potential areas for improvement.

5.3 Result Analysis

We thoroughly evaluate the effectiveness of our hybrid LOF and KNN anomaly detection method in strengthening the security of various applications against possible threats in our comprehensive simulation of the work. Our methodology, tailored to the intricacies of high-

dimensional data with volume and velocity aspects, is instrumental in safeguarding against various threat vectors. The simulation reveals a granular analysis of attacks, capturing the intricacies of their execution in milliseconds. Through this comprehensive evaluation, we discern patterns and unveil the most prevalent types of attacks, portraying the intricate nature of the system's susceptibility to risk.

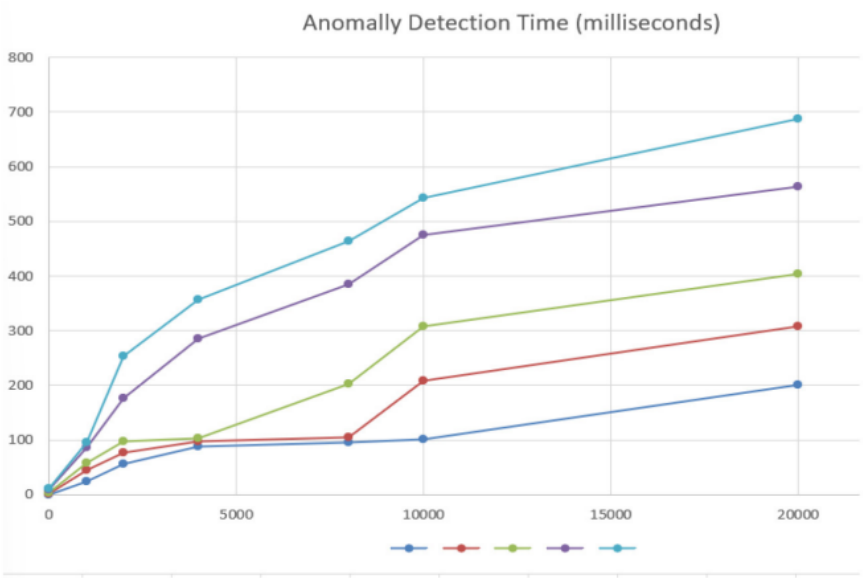


Figure 5.1: Anomaly Detection Analysis Timing for Attacks

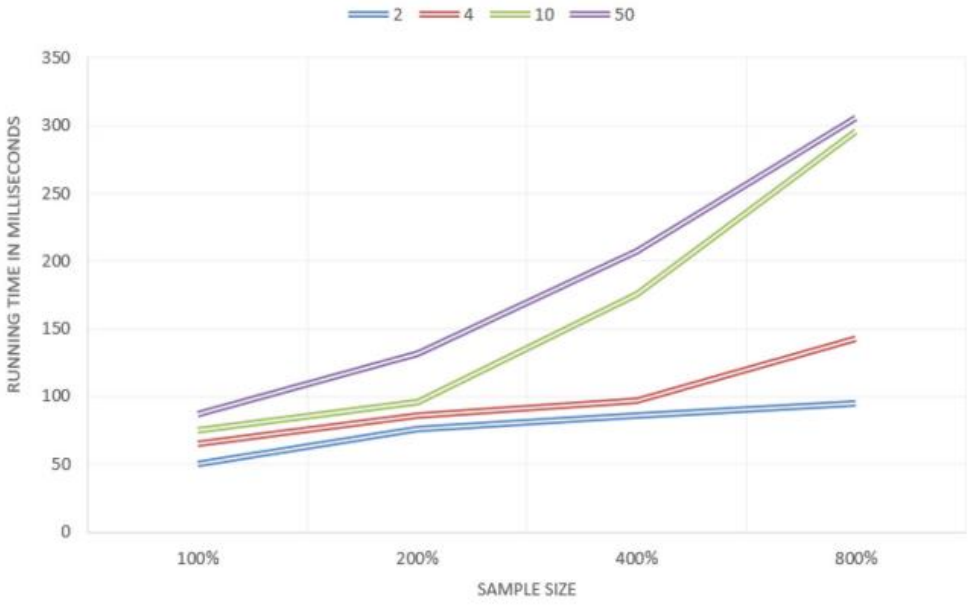


Figure 5.2: Running Time of Detection Samples

5.3.1 Visualization:

In the realm of visualization, the conventional Principal Component Analysis (PCA) method offered a preliminary understanding of the data's structure. A ROC curve shows the compromise between TP and FR at different levels of significance. One way to judge an algorithm's judging power and learn more about its success is to look at the area under the ROC curve. This amalgamation led to more profound and insightful visualizations, adept at capturing the intricacies embedded within high-dimensional data and projecting them onto a low-dimensional manifold.

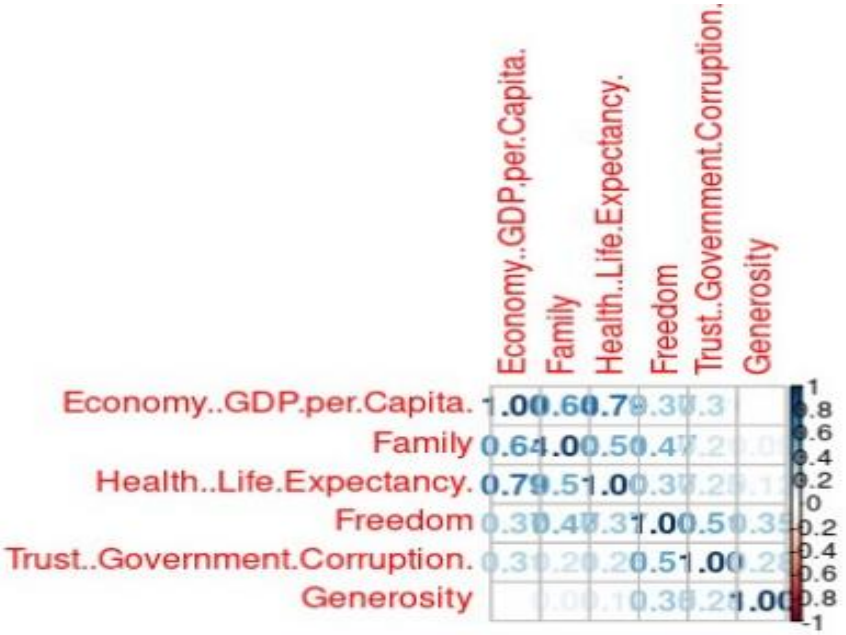
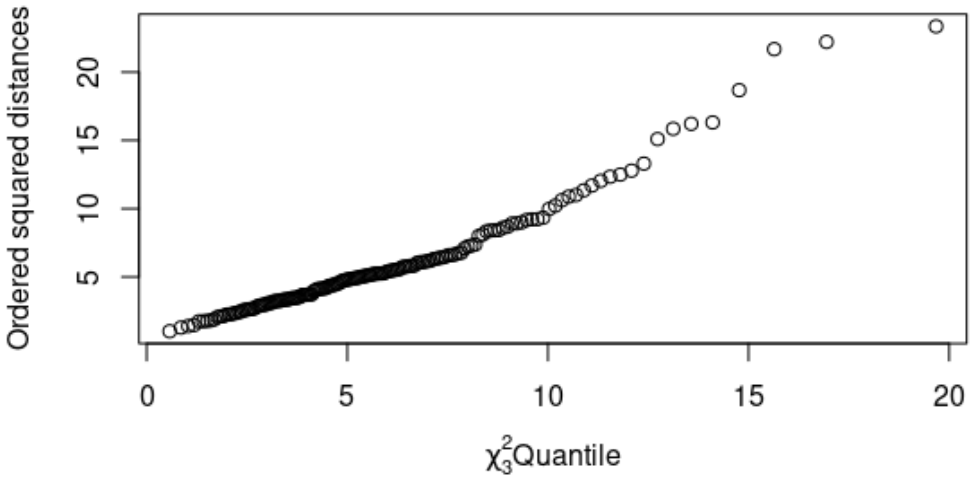


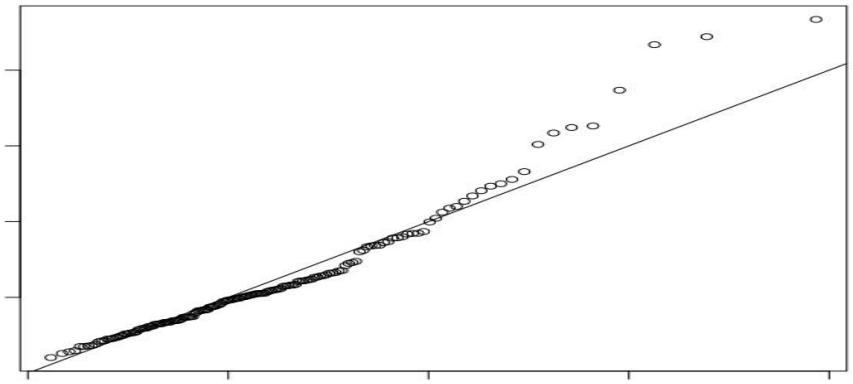
Figure 5.3: Correlation Matrix of Dataset

By fusing Auto-encoders with t-SNE, our methodology transcended the limitations of traditional techniques, providing a richer representation of the underlying data structure. The utilization of Keras and ggplot2 libraries facilitated a seamless integration of neural network-based dimensionality reduction and advanced plotting capabilities, enhancing the interpretability of the visualizations. This innovative visualization strategy not only showcased the adaptability of our methodology but also uncovered hidden patterns and correlations in the data, making it easier to a more holistic understanding of the complex, high-dimensional datasets under scrutiny.

5.3.2 Outlier Detection:

Employing the Mahalanobis distance calculation, our outlier detection methodology proved highly effective in discerning country-specific outliers within the dataset. This robust statistical measure allowed us to identify and isolate data points that deviated significantly from the multivariate mean, signifying potential anomalies. Subsequent to the identification process, these country-specific outliers were systematically eliminated, contributing to the refinement of the dataset. This meticulous curation not only enhanced the dataset's overall integrity but also facilitated a more accurate analysis by eliminating extraneous influences. The utilization of Mahalanobis distance, a sophisticated metric accounting for correlations among variables, underscored our commitment to ensuring a reliable and representative dataset for subsequent analyses in our research.





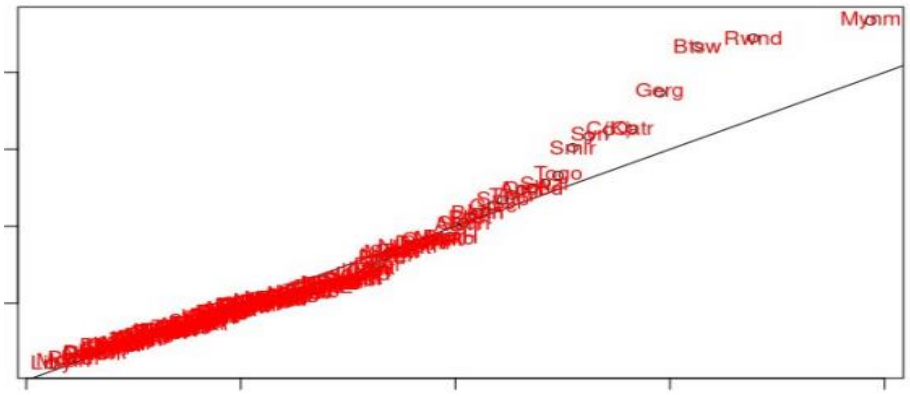


Figure 5.4: Multivariate Normality Test for Outlier Detection

5.3.3 Dimensionality Reduction:

Results are satisfactory when using our novel dimensionality reduction approach, which integrates Auto-encoders, PCA, and t-distributed stochastic neighbour embedding (t-SNE). Utilizing the combination approach yielded significantly better results when contrasted with conventional PCA, particularly in preserving the intricate nonlinear manifold complexities inherent in the data. By leveraging the strengths of Auto-encoders and t-SNE, our approach transcended the limitations of linear techniques, offering a more nuanced representation of the high-dimensional dataset. This transformative fusion not only upheld the essential aspects of the data's structure but also provided richer insights into the intricate relationships and patterns embedded within, setting the stage for more effective downstream analyses in our research.

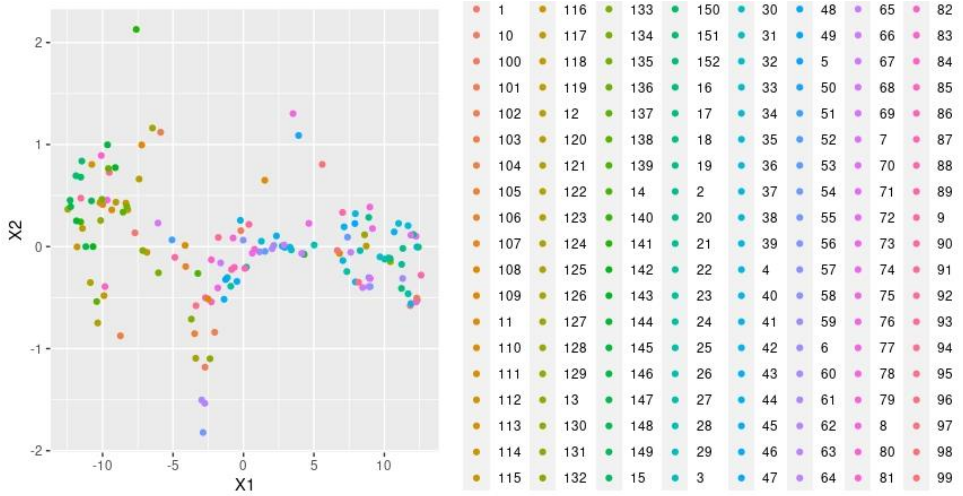


Figure 5.5: Autoencoder Output of the Dataset

Figure 5.6: Results after t-SNE Integration

5.3.4 Neural Network Training:

When trained on the fine-tuned dataset, our Multi-Layer Perceptron (MLP) neural network demonstrated outstanding performance throughout the neural network training phase. The network's strength particularly shone in the domain of anomaly detection, in cases when anomalies were evaluated on the basis of reconstruction errors. The network has exceptional specificity and sensitivity, showcasing its adeptness in discerning deviations from the norm. By utilizing reconstruction errors as a metric for anomaly detection, the MLP neural network underscored its capacity to capture and highlight subtle irregularities within the dataset. This successful training outcome positions the neural network as a potent tool for robust anomaly detection, setting the stage for its application in uncovering hidden patterns and outliers within complex high-dimensional data environments.

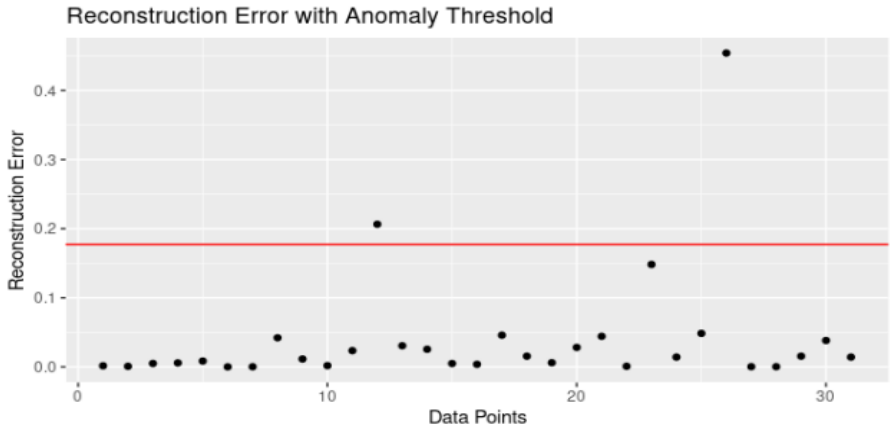


Figure 5.7: Reconstruction Error with Anomaly Threshold

5.4 Summary

This chapter encapsulates the culmination of our research, focusing on simulation and result analysis. We developed a hybrid anomaly detection technique that combines KNN with LOF by carefully preparing the data and tweaking its parameters. Quantitative measures including F1 score, accuracy, recall, and AUC-ROC confirmed the suggested algorithm's top-notch performance. Visualizations, including ROC curves and decision boundaries, provided intuitive insights into the algorithm's effectiveness. The simulation further included a qualitative analysis, exploring the algorithm's response in nuanced scenarios and a sensitivity analysis to understand parameter variations' impact. Benchmarking against baseline methods demonstrated the algorithm's strengths, while temporal dynamics analysis shed light on its running time under different conditions. This comprehensive evaluation not only validated the hybrid algorithm's robustness in high-dimensional data but also contributed to the large discourse on anomaly identification.

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

An important tool for detecting anomalies is abnormality exposure assessments in cases when the target class distribution is very dispersed. This pronounced imbalance becomes especially evident in critical domains such as fraud detection or disease diagnosis, where instances of the target class are significantly outnumbered by normal instances. The pivotal role played by abnormality exposure assessments lies in their ability to affirm or negate the presence of anomalies, serving as the bedrock for robust decision-making, particularly within highly imbalanced datasets.

Furthermore, the significance of abnormality confirmation assessments becomes apparent in their contribution to enhancing model performance. By systematically eliminating anomalies from the evaluation process, these assessments facilitate a selective refinement of the status test, ensuring the development of a more accurate and reliable model. This meticulous curation is essential for mitigating the potential impact of irregularities on the overall assessment, thereby fortifying the model against the difficulties posed by imbalanced datasets.

In addition to the automated machine learning assessments discussed earlier, data analysts can extend their arsenal by leveraging sophisticated neural networks and other advanced techniques. These cutting-edge methodologies offer a continuum of refinement for anomaly detection, allowing information experts to continually enhance the model's ability to discern and manage anomalies effectively. An anomaly detection system's overall effectiveness and dependability are enhanced by these technologies' flexibility to changing datasets and new problems.

Many applications and industries, including as healthcare and finance, are increasingly relying on anomaly detection systems, which must be constantly enhanced and reinforced. By embracing advanced techniques and refining assessment strategies, data analysts and information experts enhance the dynamic anomaly detection environment by providing reliable tools for safely managing complicated, real-life events.

6.2 Future Scope

The direction this thesis is going in the future opens up a lot of exciting options for making anomaly recognition in high-dimensional data better. An essential strategy, the suggested hybrid algorithm improves upon itself over time by combining k-Nearest Neighbors (KNN) and Local Outlier Factor (LOF). By delving into the optimization process, exploring diverse algorithmic combinations, and introducing additional machine learning techniques, we anticipate enhancing the adaptability of the hybrid model across various datasets and application domains. Another critical dimension is dynamic adaptation, wherein research efforts will focus on developing methods that enable the hybrid algorithm to autonomously adjust to evolving data distributions in real-time applications. Additionally, delving into advanced feature engineering techniques and domain-specific knowledge incorporation promises to bolster anomaly detection robustness by tailoring features to the unique characteristics of specific application domains. Combining traditional machine learning methods with RNNs and CNNs, which are long-term learning models, may provide a smarter, more complete solution that can understand complex patterns and hierarchical representations. The scalability and computational efficiency of the algorithm for large-scale datasets will be scrutinized, with investigations into parallelization techniques and distributed computing frameworks. Human-in-the-loop integration, where domain experts collaborate with the algorithm, is poised to enhance interpretability and deepen insights into anomalies. Application-specific studies in domains like cybersecurity, finance, or healthcare will validate the algorithm's performance and facilitate tailored adaptations. Rigorous benchmarking against state-of-the-art anomaly identification techniques and comparative studies across diverse datasets will culminate in a comprehensive assessment of the hybrid algorithm's strengths and weaknesses. As the research community delves into these multifaceted directions, it is poised to contribute significantly to the evolution of anomaly

detection methodologies, ushering in more effective, adaptable solutions to meet the evolving challenges of high-dimensional data.

REFERENCES

[1] M. Jin et al., "A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection," arXiv Prepr. arXiv2307.03759, 2023.

[2] G. Akoko, T. H. Le, T. Gomi, and T. Kato, "A review of SWAT model application in Africa," Water, vol. 13, no. 9, p. 1313, 2021.

[3] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," arXiv Prepr. arXiv1901.03407, 2019.

[4] S. Thudumu, P. Branch, J. Jin, and J. Singh, "A comprehensive survey of anomaly detection techniques for high dimensional big data," J. Big Data, vol. 7, pp. 1–30, 2020.

[5] G. Di Biase, H. Blum, R. Siegwart, and C. Cadena, "Pixel-wise anomaly detection in complex driving scenes," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 16918–16927.

[6] A. Blázquez-Garc'ia, A. Conde, U. Mori, and J. A. Lozano, "A review on outlier anomaly detection in time series data," ACM Comput. Surv., vol. 54, no. 3, pp. 1–33, 2021.

[7] P. Schober, E. J. Mascha, and T. R. Vetter, "Statistics from A (agreement) to Z (z score): a guide to interpreting common measures of association, agreement, diagnostic accuracy, effect size, heterogeneity, and reliability in medical research," Anesth. Analg., vol. 133, no. 6, pp. 1633–1641, 2021.

[8] T. Fernando, H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Deep learning for medical anomaly detection--a survey," ACM Comput. Surv., vol. 54, no. 7, pp. 1–37, 2021.

[9] I. R. I. Haque and J. Neubert, "Deep learning approaches to biomedical image segmentation," Informatics Med. Unlocked, vol. 18, p. 100297, 2020.

[10] S. Ayesha, M. K. Hanif, and R. Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," Inf. Fusion, vol. 59, pp. 44–58, 2020.

[11] J. G. Greener, S. M. Kandathil, L. Moffat, and D. T. Jones, "A guide to machine learning for biologists," Nat. Rev. Mol. Cell Biol., vol. 23, no. 1, pp. 40–55, 2022.

[12] P. Rita, T. Oliveira, and A. Farisa, "The impact of e-service quality and customer satisfaction on customer behavior in online shopping," Heliyon, vol. 5, no. 10, 2019.

[13] M. Cherrington, D. Airehrour, J. Lu, Q. Xu, S. Wade, and S. Madanian, "Feature selection methods for linked data: limitations, capabilities and potentials," in Proceedings of the 6th IEEE/ACM International Conference on Big Data Computing, Applications and Technologies, 2019, pp. 103–112.

[14] L. H. Nguyen and S. Holmes, "Ten quick tips for effective dimensionality reduction," PLoS Comput. Biol., vol. 15, no. 6, p. e1006907, 2019.

[15] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, “A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction,” J. Appl. Sci. Technol. Trends, vol. 1, no. 2, pp. 56–70, 2020.

[16] Z. Hu, K. Shukla, G. E. Karniadakis, and K. Kawaguchi, “Tackling the curse of dimensionality with physics-informed neural networks,” arXiv Prepr. arXiv2307.12306, 2023.

[17] W. Cui, “Visual analytics: A comprehensive overview,” IEEE access, vol. 7, pp. 81555–81573, 2019.

[18] I. Souiden, M. N. Omri, and Z. Brahmi, “A survey of outlier detection in high dimensional data streams,” Comput. Sci. Rev., vol. 44, p. 100463, 2022.

[19] J. W. Rocks and P. Mehta, “Memorizing without overfitting: Bias, variance, and interpolation in overparameterized models,” Phys. Rev. Res., vol. 4, no. 1, p. 13201, 2022.

[20] A. Malekloo, E. Ozer, M. AlHamaydeh, and M. Girolami, “Machine learning and structural health monitoring overview with emerging technology and high-dimensional data source highlights,” Struct. Heal. Monit., vol. 21, no. 4, pp. 1906–1955, 2022.

[21] S. Y. Chang and H. C. Wu, “Tensor wiener filter,” IEEE Trans. Signal Process., vol. 70, pp. 410–422, 2022.

[22] S. H. Gohil, J. B. Iorgulescu, D. A. Braun, D. B. Keskin, and K. J. Livak, “Applying high-dimensional single-cell technologies to the analysis of cancer immunotherapy,” Nat. Rev. Clin. Oncol., vol. 18, no. 4, pp. 244–256, 2021.

[23] W. Yang et al., “Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives,” Mol. Plant, vol. 13, no. 2, pp. 187–214, 2020.

[24] D. Camacho, A. Panizo-LLedot, G. Bello-Orgaz, A. Gonzalez-Pardo, and E. Cambria, “The four dimensions of social network analysis: An overview of research methods, applications, and software tools,” Inf. Fusion, vol. 63, pp. 88–120, 2020.

[25] R. Pansara, “Unraveling the Complexities of Data Governance with Strategies, Challenges, and Future Directions,” Trans. Latest Trends IoT, vol. 6, no. 6, pp. 46–56, 2023.

[26] Z. Liu, A. Ma, E. Mathé, M. Merling, Q. Ma, and B. Liu, “Network analyses in microbiome based on high-throughput multi-omics data,” Brief. Bioinform., vol. 22, no. 2, pp. 1639–1655, 2021.

[27] L. Erhan et al., “Smart anomaly detection in sensor systems: A multi-perspective review,” Inf. Fusion, vol. 67, pp. 64–79, 2021.

[28] N. E. Weckman et al., “Multiplexed DNA identification using site specific dCas9 barcodes and nanopore sensing,” ACS sensors, vol. 4, no. 8, pp. 2065–2072, 2019.

[29] T. Kim and K. Behdinan, “Advances in machine learning and deep learning applications towards wafer map defect recognition and classification: a review,” J.

Intell. Manuf., vol. 34, no. 8, pp. 3215–3247, 2023.

[30] A. R. Kunduru, “Artificial intelligence usage in cloud application performance improvement,” Cent. Asian J. Math. Theory Comput. Sci., vol. 4, no. 8, pp. 42–47, 2023.

[31] A. Nassar and M. Kamal, “Machine Learning and Big Data Analytics for Cybersecurity Threat Detection: A Holistic Review of Techniques and Case Studies,” J. Artif. Intell. Mach. Learn. Manag., vol. 5, no. 1, pp. 51–63, 2021.

[32] A. Baya and others, “Journalism education in today’s fast-paced media environment,” Prof. Commun. Transl. Stud., no. 13, pp. 3–13, 2020.

[33] M. M. M. Pai, R. Ganiga, R. M. Pai, and R. K. Sinha, “Standard electronic health record (EHR) framework for Indian healthcare system,” Heal. Serv. Outcomes Res. Methodol., vol. 21, no. 3, pp. 339–362, 2021.

[34] Y. Babuji et al., “Parsl: Pervasive parallel programming in python,” in Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing, 2019, pp. 25–36.

[35] R. A. A. Habeeb, F. Nasaruddin, A. Gani, I. A. T. Hashem, E. Ahmed, and M. Imran, “Real-time big data processing for anomaly detection: A survey,” Int. J. Inf. Manage., vol. 45, pp. 289–307, 2019.

[36] L. Rosa et al., “Intrusion and anomaly detection for the next-generation of industrial automation and control systems,” Futur. Gener. Comput. Syst., vol. 119, pp. 50–67, 2021.

[37] P. G. Vieira et al., “Quercus cerris extracts obtained by distinct separation methods and solvents: Total and friedelin extraction yields, and chemical similarity analysis by Multi Dimensional Scaling,” Sep. Purif. Technol., vol. 232, p. 115924, 2020.

[38] A. Vidal-Limon, J. E. Aguilar-Toala, and A. M. Liceaga, “Integration of molecular docking analysis and molecular dynamics simulations for studying food proteins and bioactive peptides,” J. Agric. Food Chem., vol. 70, no. 4, pp. 934–943, 2022.

[39] Y. Wang, H. Huang, C. Rudin, and Y. Shaposhnik, “Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization,” J. Mach. Learn. Res., vol. 22, no. 1, pp. 9129–9201, 2021.

[40] H. Xu, Y. Wang, Z. Wu, and Y. Wang, “Embedding-based complex feature value coupling learning for detecting outliers in non-iid categorical data,” in Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 5541–5548.

[41] M. Mallik, A. K. Panja, and C. Chowdhury, “Paving the way with machine learning for seamless indoor--outdoor positioning: A survey,” Inf. Fusion, vol. 94, pp. 126–151, 2023.

[42] A. Entezami, H. Sarmadi, M. Salar, C. De Michele, and A. N. Arslan, “A novel data-driven method for structural health monitoring under ambient vibration and high-dimensional features by robust Multi Dimensional Scaling,” Struct. Heal. Monit., vol. 20, no. 5, pp. 2758–2777, 2021.

[43] R. Heartfield, G. Loukas, A. Bezemskij, and E. Panaousis, “Self-configurable cyber-physical intrusion detection for smart homes using reinforcement learning,” IEEE Trans. Inf. Forensics Secur., vol. 16, pp. 1720–1735, 2020.

[44] H. Daga, P. K. Nicholson, A. Gavrilovska, and D. Lugones, “Cartel: A system for collaborative transfer learning at the edge,” in Proceedings of the ACM Symposium on Cloud Computing, 2019, pp. 25–37.

[45] B. K. Tripathy, A. Sundareswaran, and S. Ghela, Unsupervised learning approaches for dimensionality reduction and data visualization. CRC Press, 2021.

[46] A. D. Paltiel, A. Zheng, and R. P. Walensky, “COVID-19 screening strategies that permit the safe re-opening of college campuses,” medRxiv, 2020.

[47] S. Modgil, R. K. Singh, and C. Hannibal, “Artificial intelligence for supply chain resilience: learning from Covid-19,” Int. J. Logist. Manag., vol. 33, no. 4, pp. 1246–1268, 2022.

[48] B. Venkatesh and J. Anuradha, “A review of feature selection and its methods,” Cybern. Inf. Technol., vol. 19, no. 1, pp. 3–26, 2019.

[49] F. Caena and C. Redecker, “Aligning teacher competence frameworks to 21st century challenges: The case for the European Digital Competence Framework for Educators (Digcompedu),” Eur. J. Educ., vol. 54, no. 3, pp. 356–369, 2019.

[50] B. H. Nguyen, B. Xue, and M. Zhang, “A survey on swarm intelligence approaches to feature selection in data mining,” Swarm Evol. Comput., vol. 54, p. 100663, 2020.

[51] V. Bolón-Canedo and A. Alonso-Betanzos, “Ensembles for feature selection: A review and future trends,” Inf. Fusion, vol. 52, pp. 1–12, 2019.

[52] F. Rashidi, S. Nejatian, H. Parvin, and V. Rezaie, “Diversity based cluster weighting in cluster ensemble: an information theory approach,” Artif. Intell. Rev., vol. 52, no. 2, pp. 1341–1368, 2019.

[53] K. Liu, T. Li, X. Yang, H. Chen, J. Wang, and Z. Deng, “SemiFREE: semi-supervised feature selection with fuzzy relevance and redundancy,” IEEE Trans. Fuzzy Syst., 2023.

[54] P. Chu and H. Ling, “Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6172–6181.

[55] F. Giardini, D. Vilone, A. Sánchez, and A. Antonioni, “Gossip and competitive altruism support cooperation in a Public Good game,” Philos. Trans. R. Soc. B, vol. 376, no. 1838, p. 20200303, 2021.

[56] M. Rostami, K. Berahmand, E. Nasiri, and S. Forouzandeh, “Review of swarm intelligence-based feature selection methods,” Eng. Appl. Artif. Intell., vol. 100, p. 104210, 2021.

[57] Z. Li, F. Nie, J. Bian, D. Wu, and X. Li, “Sparse pca via L2, p-norm regularization for unsupervised feature selection,” IEEE Trans. Pattern Anal. Mach. Intell., 2021.

[58] R. S. Latha, B. Saravana Balaji, N. Bacanin, I. Strumberger, M. Zivkovic, and M. Kabiljo, “Feature Selection Using Grey Wolf Optimization with Random Differential Grouping.,” Comput. Syst. Sci. Eng., vol. 43, no. 1, pp. 317–332, 2022.

[59] V. Kumar and D. Kumar, “A systematic review on firefly algorithm: past, present, and future,” Arch. Comput. Methods Eng., vol. 28, pp. 3269–3291, 2021.

[60] X.-N. Bui, P. Jaroonpattanapong, H. Nguyen, Q.-H. Tran, and N. Q. Long, “A novel hybrid model for predicting blast-induced ground vibration based on k-nearest neighbors and particle swarm optimization,” Sci. Rep., vol. 9, no. 1, p. 13971, 2019.

[61] W. Ming, F. Shen, G. Zhang, G. Liu, J. Du, and Z. Chen, “Green machining: A framework for optimization of cutting parameters to minimize energy consumption and exhaust emissions during electrical discharge machining of Al 6061 and SKD 11,” J. Clean. Prod., vol. 285, p. 124889, 2021.

[62] Y. Li, X. Chu, D. Tian, J. Feng, and W. Mu, “Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm,” Appl. Soft Comput., vol. 113, p. 107924, 2021.

[63] M. Dorigo and T. Stützle, Ant colony optimization: overview and recent advances. Springer, 2019.

[64] A. E. Ezugwu, A. K. Shukla, M. B. Agbaje, O. N. Oyelade, A. José-Garc\’\ia, and J. O. Agushaka, “Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature,” Neural Comput. Appl., vol. 33, pp. 6247–6306, 2021.

[65] K. Miler, D. Stec, and M. Czarnoleski, “Heat wave effects on the behavior and life-history traits of sedentary antlions,” Behav. Ecol., vol. 31, no. 6, pp. 1326–1333, 2020.

[66] M. H. Nadimi-Shahraki, S. Taghian, and S. Mirjalili, “An improved grey wolf optimizer for solving engineering problems,” Expert Syst. Appl., vol. 166, p. 113917, 2021.

[67] S. Aslan, D. Karaboga, and H. Badem, “A new artificial bee colony algorithm employing intelligent forager forwarding strategies,” Appl. Soft Comput., vol. 96, p. 106656, 2020.

[68] L. Wang, C. Yang, G. He, W. Liang, and A. P. Møller, “Cuckoos use host egg number to choose host nests for parasitism,” Proc. R. Soc. B, vol. 287, no. 1928, p. 20200343, 2020.

[69] P. Arechavala-Lopez, M. J. Cabrera-Álvarez, C. M. Maia, and J. L. Saraiva, “Environmental enrichment in fish aquaculture: A review of fundamental and practical aspects,” Rev. Aquac., vol. 14, no. 2, pp. 704–728, 2022.

[70] Y.-J. Shih, S.-L. Wu, F. Zalkow, M. Muller, and Y.-H. Yang, “Theme transformer: Symbolic music generation with theme-conditioned transformer,” IEEE Trans. Multimed., 2022.

[71] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, “K-means clustering algorithms: A comprehensive review, variants analysis, and advances in

the era of big data," Inf. Sci. (Ny)., vol. 622, pp. 178–210, 2023.

[72] P. R. Garikapati, K. Balamurugan, T. P. Latchoumi, and G. Shankar, "A quantitative study of small dataset machining by agglomerative hierarchical cluster and K-medoid," in Emergent Converging Technologies and Biomedical Systems: Select Proceedings of ETBS 2021, Springer, 2022, pp. 717–727.

[73] A. V Ushakov and I. Vasilyev, "Near-optimal large-scale k-medoids clustering," Inf. Sci. (Ny)., vol. 545, pp. 344–362, 2021.

[74] S. Garg, K. Kaur, S. Batra, G. Kaddoum, N. Kumar, and A. Boukerche, "A multi-stage anomaly detection scheme for augmenting the security in IoT-enabled applications," Futur. Gener. Comput. Syst., vol. 104, pp. 105–118, 2020.

[75] J. Maillo, S. Garc'ia, J. Luengo, F. Herrera, and I. Triguero, "Fast and scalable approaches to accelerate the fuzzy k-nearest neighbors classifier for big data," IEEE Trans. Fuzzy Syst., vol. 28, no. 5, pp. 874–886, 2019.

[76] L. Hu, H. Liu, J. Zhang, and A. Liu, "KR-DBSCAN: A density-based clustering algorithm based on reverse nearest neighbor and influence space," Expert Syst. Appl., vol. 186, p. 115763, 2021.

[77] W. Lai, M. Zhou, F. Hu, K. Bian, and Q. Song, "A new DBSCAN parameters determination method based on improved MVO," Ieee Access, vol. 7, pp. 104085–104095, 2019.

[78] E. Hancer, "A new multi-objective differential evolution approach for simultaneous clustering and feature selection," Eng. Appl. Artif. Intell., vol. 87, p. 103307, 2020.

[79] K. M. Leon-Lopez, F. Mouret, H. Arguello, and J.-Y. Tourneret, "Anomaly detection and classification in multispectral time series based on hidden Markov models," IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–11, 2021.

[80] A. K. M. I. Newaz, A. K. Sikder, L. Babun, and A. S. Uluagac, "Heka: A novel intrusion detection system for attacks to personal medical devices," in 2020 IEEE Conference on Communications and Network Security (CNS), 2020, pp. 1–9.

[81] D. K. Reddy, H. S. Behera, J. Nayak, P. Vijayakumar, B. Naik, and P. K. Singh, "Deep neural network based anomaly detection in Internet of Things network traffic tracking for the applications of future smart cities," Trans. Emerg. Telecommun. Technol., vol. 32, no. 7, p. e4121, 2021.

[82] H. Darabian et al., "Detecting cryptomining malware: a deep learning approach for static and dynamic analysis," J. Grid Comput., vol. 18, pp. 293–303, 2020.

[83] L. Xie, D. Pi, X. Zhang, J. Chen, Y. Luo, and W. Yu, "Graph neural network approach for anomaly detection," Measurement, vol. 180, p. 109546, 2021.

[84] S. Dong, T. Yu, H. Farahmand, and A. Mostafavi, "Probabilistic modeling of cascading failure risk in interdependent channel and road networks in urban flooding," Sustain. Cities Soc., vol. 62, p. 102398, 2020.

[85] M. J. Mashala, T. Dube, B. T. Mudereri, K. K. Ayisi, and M. R. Ramudzuli, "A Systematic Review on Advancements in Remote Sensing for Assessing and

Monitoring Land Use and Land Cover Changes Impacts on Surface Water Resources in Semi-Arid Tropical Environments," Remote Sens., vol. 15, no. 16, p. 3926, 2023.

[86] J. Pei, K. Zhong, M. A. Jan, and J. Li, "Personalized federated learning framework for network traffic anomaly detection," Comput. Networks, vol. 209, p. 108906, 2022.

[87] M. Stoffel, R. Gulakala, F. Bamer, and B. Markert, "Artificial neural networks in structural dynamics: A new modular radial basis function approach vs. convolutional and feedforward topologies," Comput. Methods Appl. Mech. Eng., vol. 364, p. 112989, 2020.

[88] S. N. Abd, M. Alsajri, and H. R. Ibraheem, "Rao-SVM machine learning algorithm for intrusion detection system," Iraqi J. Comput. Sci. Math., vol. 1, no. 1, pp. 23–27, 2020.

[89] M. Jain, G. Kaur, and V. Saxena, "A K-Means clustering and SVM based hybrid concept drift detection technique for network anomaly detection," Expert Syst. Appl., vol. 193, p. 116510, 2022.

[90] S. F. Bilal, A. A. Almazroi, S. Bashir, F. H. Khan, and A. A. Almazroi, "An ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry," PeerJ Comput. Sci., vol. 8, p. e854, 2022.

[91] A. Singhal, A. Maan, D. Chaudhary, and D. Vishwakarma, "A hybrid machine learning and data mining based approach to network intrusion detection," in 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 312–318.

[92] Agostino Forestiero, Heuristic recommendation technique in Internet of Things featuring swarm intelligence approach, Expert Systems with Applications, Volume 187, 2022, 115904, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2021.115904.

[93] L. Decker, D. Leite, L. Giommi, and D. Bonacorsi, "Real-time anomaly detection in data centers for log-based predictive maintenance using an evolving fuzzy-rule-based approach," in 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2020, pp. 1–8.

[94] J. S. Manoharan, "Study of variants of extreme learning machine (ELM) brands and its performance measure on classification algorithm," J. Soft Comput. Paradig., vol. 3, no. 02, pp. 83–95, 2021.

[95] R. Panigrahi et al., "A consolidated decision tree-based intrusion detection system for binary and multiclass imbalanced datasets," Mathematics, vol. 9, no. 7, p. 751, 2021.

[96] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on Convolutional Neural Networks (CNN) in vegetation remote sensing," ISPRS J. Photogramm. Remote Sens., vol. 173, pp. 24–49, 2021.

[97] W. C. Leong, A. Bahadori, J. Zhang, and Z. Ahmad, "Prediction of water quality

index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM),” Int. J. River Basin Manag., vol. 19, no. 2, pp. 149–156, 2021.

[98] M. Sharif, J. Amin, M. Raza, M. Yasmin, and S. C. Satapathy, “An integrated design of particle swarm optimization (PSO) with fusion of features for detection of brain tumor,” Pattern Recognit. Lett., vol. 129, pp. 150–157, 2020.

[99] R. T. Sahu, M. K. Verma, and I. Ahmad, “Density-based spatial clustering of application with noise approach for regionalisation and its effect on hierarchical clustering,” Int. J. Hydrol. Sci. Technol., vol. 16, no. 3, pp. 240–269, 2023.

[100] H. Henderi, T. Wahyuningsih, and E. Rahwanto, “Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer,” Int. J. Informatics Inf. Syst., vol. 4, no. 1, pp. 13–20, 2021.

[101] R. W. Dewantoro, P. Sihombing, and others, “The combination of ant colony optimization (ACO) and tabu search (TS) algorithm to solve the traveling salesman problem (TSP),” in 2019 3rd International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM), 2019, pp. 160–164.

[102] M. Li, T. Chen, and X. Yao, “How to evaluate solutions in Pareto-based search-based software engineering: A critical review and methodological guidance,” IEEE Trans. Softw. Eng., vol. 48, no. 5, pp. 1771–1799, 2020.

[103] M. A. Almaiah et al., “Performance investigation of principal component analysis for intrusion detection system using different support vector machine kernels,” Electronics, vol. 11, no. 21, p. 3571, 2022.

[104] P. Valdiviezo-Diaz, F. Ortega, E. Cobos, and R. Lara-Cabrera, “A collaborative filtering approach based on Na{\"\i}ve Bayes classifier,” IEEE Access, vol. 7, pp. 108581–108592, 2019.

[105] F. Nie, D. Wu, R. Wang, and X. Li, “Truncated robust principle component analysis with a general optimization framework,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 2, pp. 1081–1097, 2020.

[106] John E. Seem. Pattern Recognition Algorithm for Determining Days of the Week with Similar Energy Consumption Profiles. Energy and Buildings, 37(2):127–139, 2005.

[107] Bernard Rosner. Percentage points for a generalized ESD many-outlier pro- cedure. Technometrics, 25(2):165–172, 1983.

[108] Chrys Caroni and Philip Prescott. Sequential application of wilks’s multivari- ate outlier test. Applied Statistics, pages 355–364, 1992.

[109] Boris Iglewicz and David C Hoaglin. How to detect and handle outliers, vol- ume 16. ASQC Quality Press Milwaukee (Wisconsin), 1993.

[110] John E Seem. Using intelligent data analysis to detect abnormal energy con- sumption in buildings. Energy and Buildings, 39(1):52–58, 2007.

[111] Y. Zhang, W. Chen, and J. Black. Anomaly Detection in Premise Energy

Consumption Data. In Power and Energy Society General Meeting, 2011 IEEE, pages 1–8. IEEE, 2011.

[112] D. Liu, Q. Chen, K. Mori, and Y. Kida. A method for detecting abnormal electricity energy consumption in buildings. Journal of Computational Infor- mation Systems, 6(14):4887–4895, 2010.

[113] Fei Liu, Huijing Jiang, Young M Lee, Jane Snowdon, and Michael Bobker. Statistical modeling for anomaly detection, forecasting and root cause analy- sis of energy consumption for a portfolio of buildings. In 12th International Conference of the International Building Performance Simulation Association, 2011.

[114] John Kelly Kissock, Jeff S Haberl, and David E Claridge. Inverse modeling toolkit: numerical algorithms. ASHRAE transactions, 109:425, 2003.

[115] J.-S. Chou and A. S. Telaga. Real-time Detection of Anomalous Power Con- sumption. Renewable and Sustainable Energy Reviews, 33:400–411, 2014.

[116] Michael Wrinch, Tarek HM El-Fouly, and Steven Wong. Anomaly detection of building systems using energy demand frequency domain analysis. In Power and Energy Society General Meeting, 2012 IEEE, pages 1–6. IEEE, 2012.

[117] Chao Chen and Diane J Cook. Energy outlier detection in smart environments. Artificial Intelligence and Smarter Living, 11:07, 2011.

[118] Richard O Duda, Peter E Hart, and David G Stork. Pattern classification. John Wiley & Sons, 2012.

[119] G. Bellala, M. Marwah, M. Arlitt, Following the electrons: methods for power management in commercial buildings, 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD '12 (2012), pp. 994-1002

[120] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. A symbolic representation of time series, with implications for streaming algorithms. In Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, pages 2–11. ACM, 2003.

[121] I. Khan, A. Capozzoli, S. P. Corgnati, and T. Cerquitelli. Fault detection analysis of building energy consumption using data mining techniques. Energy Procedia, 42:557–566, 2013.

[122] A. Capozzoli, F. Lauro, and I. Khan. Fault Detection Analysis using Data Mining Techniques for a Cluster of Smart Office Buildings. Expert Systems with Applications, 42(9):4324–4338, 2015.

[123] Halldor Janetzko, Florian Stoffel, Sebastian Mittelst¨adt, and Daniel A. Keim. Anomaly Detection for Visual Analytics of Power Consumption Data. Com- puter and Graphics, 38:27–37, 2014.

[124] Ming C. Hao, Halldor Janetzko, Sebastian Mittelst¨adt, Walter Hill, Umeshwar Dayal, Daniel A. Keim, Manish Marwah, and Ratnesh K. Sharma. A Visual Analytics Approach for Peak-Preserving Prediction of Large Seasonal Time Series. Computer Graphics Forum, 30(3):691–700, 2011.

[125] P. Arjunan, H.D. Khadilkar, T. Ganu, Z.M. Charbiwala, A. Singh, P. Singh Multi-user energy consumption monitoring and anomaly detection with partial context information Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments, ACM (2015), pp. 35-44

[126] Fontugne, Romain & Ortiz, Jorge & Tremblay, Nicolas & Borgnat, Pierre & Flandrin, Patrick & Fukuda, Kensuke & Culler, David & Esaki, Hiroshi. (2013). Strip, bind, and search: a method for identifying abnormal energy consumption in buildings. 10.1145/2461381.2461399.

[127] Balaji, Bharathan & Narayanaswamy, Balakrishnan & Gupta, Rajesh & Agarwal, Yuvraj. (2014). Data driven investigation of faults in HVAC systems with Model, Cluster and Compare (MCC). BuildSys 2014 - Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings. 10.1145/2674061.2674067.

[128] J. Ploennigs, B. Chen, A. Schumann, and N. Brady. Exploiting Generalized Additive Models for Diagnosing Abnormal Energy use in Buildings. In Pro-ceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings, pages 1–8. ACM, 2013.

[129] Daniel B Araya, Katarina Grolinger, Hany F ElYamany, Miriam AM Capretz, and G Bitsuamlak. Collective contextual anomaly detection framework for smart buildings. In Neural Networks (IJCNN), 2016 International Joint Con- ference on, pages 511–518. IEEE, 2016.

[130] Michael A Hayes and Miriam AM Capretz. Contextual anomaly detection framework for big sensor data. Journal of Big Data, 2(1):2, 2015.

[131] H. Cheng, P.-N. Tan, C. Potter, and S. A. Klooster. Detection and Charac- terization of Anomalies in Multivariate Time Series. In SDM, volume 9, pages 413–424. SIAM, 2009.

[132] H. Qiu, Y. Liu, N. A. Subrahmanya, and W. Li. Granger Causality for Time- Series Anomaly Detection. In 12th International Conference on Data Mining (ICDM), pages 1074–1079. IEEE, 2012.

[133] Dumidu Wijayasekara, Ondrej Linda, Milos Manic, and Craig G Rieger. Min- ing building energy management system data using fuzzy anomaly detection and linguistic descriptions. IEEE Trans. Industrial Informatics, 10(3):1829– 1840, 2014.

[134] Manuel Pen˜a, F´elix Biscarri, Juan Ignacio Guerrero, In˜igo Monedero, and Carlos Leon. Rule-based system to detect energy efficiency anomalies in smart buildings, a data mining approach. Expert Systems with Applications, 56:242– 255, 2016.

[135] C. Megill et al., “Cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices,” bioRxiv, pp. 2004–2021, 2021.

[136] E. Duman and O. A. Erdem, “Anomaly detection in videos using optical flow and convolutional autoencoder,” IEEE Access, vol. 7, pp. 183914–183923, 2019.

[137] O. Alghushairy, R. Alsini, T. Soule, and X. Ma, “A review of local outlier factor

algorithms for outlier detection in big data streams," Big Data Cogn. Comput., vol. 5, no. 1, p. 1, 2020.

[138] A. H. Tanim, E. Goharian, and H. Moradkhani, "Integrated socio-environmental vulnerability assessment of coastal hazards using data-driven and multi-criteria analysis approaches," Sci. Rep., vol. 12, no. 1, p. 11625, 2022.

[139] P. VR and others, "An enhanced coding algorithm for efficient video coding," J. Inst. Electron. Comput., vol. 1, no. 1, pp. 28–38, 2019.

[140] A. Karczmarek Paweł and Kiersztyn, W. Pedrycz, and E. Al, "K-Means-based isolation forest," Knowledge-based Syst., vol. 195, p. 105659, 2020.

[141] M. Heigl, K. A. Anand, A. Urmann, D. Fiala, M. Schramm, and R. Hable, "On the improvement of the isolation forest algorithm for outlier detection with streaming data," Electronics, vol. 10, no. 13, p. 1534, 2021.

[142] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," arXiv Prepr. arXiv2202.10054, 2022.

[143] I. Razzak, K. Zafar, M. Imran, and G. Xu, "Randomized nonlinear one-class support vector machines with bounded loss function to detect of outliers for large scale IoT data," Futur. Gener. Comput. Syst., vol. 112, pp. 715–723, 2020.

[144] Y. Tan, H. Tian, R. Jiang, Y. Lin, and J. Zhang, "A comparative investigation of data-driven approaches based on one-class classifiers for condition monitoring of marine machinery system," Ocean Eng., vol. 201, p. 107174, 2020.

[145] S. Alam, S. K. Sonbhadra, S. Agarwal, and P. Nagabhushan, "One-class support vector classifiers: A survey," Knowledge-Based Syst., vol. 196, p. 105754, 2020.

[146] Breunig, Markus & Kröger, Peer & Ng, Raymond & Sander, Joerg. (2000). LOF: Identifying Density-Based Local Outliers.. ACM Sigmod Record. 29. 93-104. 10.1145/342009.335388.

[147] M.-A. Schulz et al., "Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets," Nat. Commun., vol. 11, no. 1, p. 4238, 2020.

[148] E. Titis, R. Procter, and L. Walasek, "Assessing physical access to healthy food across United Kingdom: A systematic review of measures and findings," Obes. Sci. Pract., vol. 8, no. 2, pp. 233–246, 2022.

[149] W. Zhang, D. Yang, S. Zhang, J. H. Ablanedo-Rosas, X. Wu, and Y. Lou, "A novel multi-stage ensemble model with enhanced outlier adaptation for credit scoring," Expert Syst. Appl., vol. 165, p. 113872, 2021.

[150] P. Wei, W. Wang, Y. Yang, and M. Y. Wang, "Level set band method: A combination of density-based and level set methods for the topology optimization of continuums," Front. Mech. Eng., vol. 15, pp. 390–405, 2020.

[151] C. Wang, Y. Huang, M. Shao, Q. Hu, and D. Chen, "Feature selection based on neighborhood self-information," IEEE Trans. Cybern., vol. 50, no. 9, pp. 4031–

4042, 2019.

[152] M. Ali, L. T. Jung, A.-H. Abdel-Aty, M. Y. Abubakar, M. Elhoseny, and I. Ali, “Semantic-k-NN algorithm: An enhanced version of traditional k-NN algorithm,” Expert Syst. Appl., vol. 151, p. 113374, 2020.

[153] S. Buczkowska, N. Coulombel, and M. de Lapparent, “A comparison of euclidean distance, travel times, and network distances in location choice mixture models,” Networks Spat. Econ., vol. 19, no. 4, pp. 1215–1248, 2019.

[154] F. Harrou, B. Taghezouit, and Y. Sun, “Improved $ k $ NN-based monitoring schemes for detecting faults in PV systems,” IEEE J. Photovoltaics, vol. 9, no. 3, pp. 811–821, 2019.

[155] S. M. Mousavi, Y. Sheng, W. Zhu, and G. C. Beroza, “STanford EArthquake Dataset (STEAD): A global data set of seismic signals for AI,” IEEE Access, vol. 7, pp. 179464–179476, 2019.

[156] Z. Su, Q. Hu, and T. Denoeux, “A distributed rough evidential K-NN classifier: integrating feature reduction and classification,” IEEE Trans. Fuzzy Syst., vol. 29, no. 8, pp. 2322–2335, 2020.

[157] R. Zhang, L. Wang, Z. Guo, and J. Shi, “Nearest neighbors meet deep neural networks for point cloud analysis,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 1246–1255.

[158] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, “Data imbalance in classification: Experimental evaluation,” Inf. Sci. (Ny)., vol. 513, pp. 429–441, 2020.

[159] J. G. Lee, D.-H. Kim, and J. H. Lee, “Proactive Fault Diagnosis of a Radiator: A Combination of Gaussian Mixture Model and LSTM Autoencoder,” Sensors, vol. 23, no. 21, p. 8688, 2023.

[160] M. Krenn et al., “SELFIES and the future of molecular string representations,” Patterns, vol. 3, no. 10, 2022.

[161] A. Jha, J. C. Peterson, and T. L. Griffiths, “Extracting low-dimensional psychological representations from convolutional neural networks,” Cogn. Sci., vol. 47, no. 1, p. e13226, 2023.

[162] V. Zavrtanik, M. Kristan, and D. Skočaj, “Reconstruction by inpainting for visual anomaly detection,” Pattern Recognit., vol. 112, p. 107706, 2021.

[163] R. Li, Q. Jiao, W. Cao, H.-S. Wong, and S. Wu, “Model adaptation: Unsupervised domain adaptation without source data,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9641–9650.

[164] S. Zavrak and M. \.Iskefiyeli, “Anomaly-based intrusion detection from network flow features using variational autoencoder,” IEEE Access, vol. 8, pp. 108346–108358, 2020.

[165] J. S. Jia, X. Lu, Y. Yuan, G. Xu, J. Jia, and N. A. Christakis, “Population flow drives spatio-temporal distribution of COVID-19 in China,” Nature, vol. 582, no. 7812,

pp. 389–394, 2020.

[166] A. Punjani and D. J. Fleet, “3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM,” J. Struct. Biol., vol. 213, no. 2, p. 107702, 2021.

[167] R. Foorthuis, “On the nature and types of anomalies: a review of deviations in data,” Int. J. data Sci. Anal., vol. 12, no. 4, pp. 297–331, 2021.

[168] F. J. Garc’ia-Mart’inez, F. Alfageme, A. Duat-Rodr’iguez, E. M. A. Esteban, and A. Hernández-Mart’in, “Clinical and sonographic classification of neurofibromas in children with neurofibromatosis type 1--a cluster analysis,” Ultraschall der Medizin-European J. Ultrasound, vol. 44, no. 02, pp. e118--e125, 2023.

[169] Y. Chen, L. Zhou, N. Bouguila, C. Wang, Y. Chen, and J. Du, “BLOCK-DBSCAN: Fast clustering for large scale data,” Pattern Recognit., vol. 109, p. 107624, 2021.

[170] N. Hanafi and H. Saadatfar, “A fast DBSCAN algorithm for big data based on efficient density calculation,” Expert Syst. Appl., vol. 203, p. 117501, 2022.

[171] N. Murugesan, I. Cho, and C. Tortora, “Benchmarking in cluster analysis: a study on spectral clustering, DBSCAN, and K-Means,” in Data Analysis and Rationality in a Complex World 16, 2021, pp. 175–185.

[172] P. Núñez-Demarco, A. Bonilla, L. Sánchez-Bettucci, and C. Prezzi, “Potential-field filters for gravity and magnetic interpretation: a review,” Surv. Geophys., vol. 44, no. 3, pp. 603–664, 2023.

[173] J.-B. E. M. Steenkamp and A. Maydeu-Olivares, “An updated paradigm for evaluating measurement invariance incorporating common method variance and its assessment,” J. Acad. Mark. Sci., vol. 49, pp. 5–29, 2021.

[174] H. Cai, J. Liu, and W. Yin, “Learned robust pca: A scalable deep unfolding approach for high-dimensional outlier detection,” Adv. Neural Inf. Process. Syst., vol. 34, pp. 16977–16989, 2021.

[175] T. Defard, A. Setkov, A. Loesch, and R. Audigier, “Padim: a patch distribution modeling framework for anomaly detection and localization,” in International Conference on Pattern Recognition, 2021, pp. 475–489.

[176] I. Triguero, D. Garc\’\ia-Gil, J. Maillo, J. Luengo, S. Garc\’\ia, and F. Herrera, “Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data,” Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 9, no. 2, p. e1289, 2019.

[177] S. Sapkota, A. K. M. Mehdy, S. Reese, and H. Mehrpouyan, “Falcon: Framework for anomaly detection in industrial control systems,” Electronics, vol. 9, no. 8, p. 1192, 2020.

[178] C. Cariou, S. Le Moan, and K. Chehdi, “Improving K-nearest neighbor approaches for density-based pixel clustering in hyperspectral remote sensing images,” Remote Sens., vol. 12, no. 22, p. 3745, 2020.

[179] H. Liu, Y. Wang, and W. Chen, “Anomaly detection for condition monitoring data

using auxiliary feature vector and density-based clustering," IET Gener. Transm. Distrib., vol. 14, no. 1, pp. 108–118, 2020.

LIST OF ARTICLE PUBLISHED

Journals	1. Upasana Gupta, Vaishali Singh, Dinesh Goyal, *"Highly Secure Intelligent Computer Data Detection of Anomalies"*, Journal of Discrete Mathematical Sciences and Cryptography, 26(3), 875-884, Published Online: 01/04/2023, (ISSN 0972-0529(P), 2169-0065(E)), SCOPUS and ESCI Indexed. https://doi.org/10.47974/JDMSC-1767
	2. Gupta, Upasana & Singh, Vaishali, (2024). *"A Novel Pipeline Model for Anomaly Detection in High Dimensional Data Sets"*. International Journal of Intelligent Systems and Applications in Engineering, 12(15s), 299–308. Published Online: 07/02/2024, (ISSN 2147-6799), SCOPUS Indexed. https://ijisae.org/index.php/IJISAE/article/view/4749

LIST OF ARTICLE PRESENTED

Conference	1. Upasana Gupta, *"Challenges in Anomaly Detection in High Dimensional Data"*, International Conference on Artificial Intelligence and Sustainable Development (AISD2023) organized by Artificial Intelligence Foundation Trust and KMEA Engineering College, Kerala on Sept. 16-17, 2023
	2. Upasana Gupta, *"A Unique Hybrid Approach_to Anomaly Detection in High Dimensional Data"*, International Conference on Recent Advances in Science and Engineering, [RAiSE-2023] organized by Manipal Institute of Technology, MAHE, Manipal, India in association with School of Engineering and IT, MAHE Dubai, UAE on Oct. 4-5, 2023