TÜRKİYE CUMHURİYETİ
ADANA ALPARSLAN TÜRKEŞ SCIENCE AND TECHNOLOGY UNIVERSITY

GRADUATE SCHOOL
COMPUTER ENGINEERING DEPARTMENT

K-SALP SWARM ANOMALY DETECTION AND LINK PREDICTION BASED ANOMALY PREVENTION

Vahide Nida KILIÇ

Ph.D.

ADANA 2025

TÜRKİYE CUMHURİYETİ
ADANA ALPARSLAN TÜRKEŞ SCIENCE AND TECHNOLOGY UNIVERSITY

GRADUATE SCHOOL
COMPUTER ENGINEERING DEPARTMENT

K-SALP SWARM ANOMALY DETECTION AND LINK PREDICTION BASED ANOMALY PREVENTION

Vahide Nida KILIÇ

Ph.D.

THESIS ADVISOR
Assoc. Prof. Dr. Esra SARAÇ EŞSİZ

ADANA, 2025

CERTIFICATION OF APPROVAL

(K-SALP SWARM ANOMALY DETECTION AND LINK PREDICTON BASED ANOMALY PREVENTION)

This **Ph.D.** thesis, **completed under the conditions determined by the relevant regulations**, by Vahide Nida KILIÇ **with student number** 21801303002 in **Adana Alparslan Türkeş Science and Technology University Institute of Graduate School** Computer Engineering **Department**, has been evaluated by the following academic committee, and approved in accordance with the relevant articles of the "Adana Alparslan Türkeş Science and Technology University Graduate Education Regulations".

Assoc. Prof. Dr. Mustafa AKYOL

Institute of Graduate School Manager

Prof. Dr. Serdar YILDIRIM

Head of the Department

Assoc. Prof. Dr. Esra SARAÇ EŞSİZ

Academic Advisor

Thesis Advisory Committee Members

Assoc. Prof. Dr. Esra SARAÇ EŞSİZ

Adana Alparslan Türkeş Science and Technology University

Prof. Dr. Selma Ayşe ÖZEL

Çukurova University

Assoc. Prof. Dr. Mümine KAYA KELEŞ

Adana Alparslan Türkeş Science and Technology University

Asst. Prof. Dr. Barış ATA

Çukurova University

Asst. Prof. Dr. Mehmet SARIGÜL

Çukurova University

Thesis Defence Date

05/02/2025

DECLARATION OF CONFORMITY

In this thesis study, which was prepared following the thesis writing rules of Adana Alparslan Türkeş Science and Technology University Institute of Graduate School, I declare that I provide all the information, documents, evaluations and results in accordance with scientific ethics and moral codes without resorting to any means or assistance that would be contrary to scientific ethics and traditions. I also declare that I refer to all of the articles I used in this study with appropriate references and accept all moral and legal consequences if a situation is found contrary to my statement regarding my work.

..../..../2025

[Signature]

Vahide Nida KILIÇ

ÖZET

K-SALP SÜRÜ ANOMALİ TESPİTİ VE BAĞLANTI TAHMİNİ TABANLI ANOMALİ ÖNLEME

Vahide Nida KILIÇ
Doktora, Bilgisayar Mühendisliği Anabilim Dalı
Danışman: Doç. Dr. Esra SARAÇ EŞSİZ
Şubat 2025, 54 sayfa

Anomali tespiti ve önleme, özellikle siber güvenlik gibi alanlarda, kötü niyetli faaliyetlere karşı erken tespit ve müdahalenin kritik önem taşıdığı görevlerdir. Geleneksel çalışmalar genellikle anomali tespitine odaklanmış olsa da, risklerin önceden belirlenip azaltılmasını sağlayan anomali önleme, giderek daha fazla önem kazanmaktadır. Bu tezde, gelişmiş kümeleme ve doğadan ilham alan algoritmalar kullanarak hem anomali tespiti hem de önlemesini entegre eden kapsamlı bir çerçeve sunulmaktadır. Önerilen yaklaşım, K-medoid kümeleme yöntemi ile Salp Sürüsü Algoritması'nın (SSA) optimal eşik belirleme işlemiyle birleştirilmesini içermektedir. Bu hibrit yöntem, veri setlerindeki aykırı değerleri ve şüpheli düğümleri etkili bir şekilde tespit etmeyi sağlamaktadır. Bu yöntem Enron e-posta veri seti gibi yaygın olarak kullanılan gerçek dünya veri setlerinde uygulanmış ve hem içerik tabanlı hem de düğüm tabanlı anomali tespitindeki performansı ortaya konmuştur.

Deneysel sonuçlar önerilen yöntemin başarısını siber güvenlik alanında ortaya koymaktadır. 10 veri setinin 5'inde alternatif teknikleri geride bırakan yöntem ile Tiroid veri setinde 0.8651 AUC değeri elde edilmiştir. Ayrıca, çerçeve, verideki içerik ve yapısal özellikler arasındaki farklılıklarla belirlenen "şüpheli düğümler" kavramını tanıtmaktadır. Bu düğümler, dolandırıcılık davranışları veya kötü niyetli e-postalar gibi potansiyel zararlı eylemleri önlemek için daha fazla analiz yapılmak üzere işaretlenmektedir. Önerilen çerçeve, sadece anomali tespiti yöntemlerini geliştirmekle kalmayıp, anomali önleme konusunda da yenilikçi bir yaklaşım sunarak, risklerin gerçekleşmeden önce azaltılmasına yönelik proaktif bir çözüm sağlamaktadır.

Anahtar Kelimeler: Anomali Tespiti/Önleme, Sosyal Ağ Grafiği, Doğadan İlham Alan Algoritmalar, Bağlantı Tahmini.

ABSTRACT

K-SALP SWARM ANOMALY DETECTION AND LINK PREDICTION BASED ANOMALY PREVENTION

Vahide Nida KILIÇ
Ph.D. Department of Computer Engineering

Supervisor: Assoc. Prof. Dr. Esra SARAÇ EŞSİZ
February 2025, 54 pages

Anomaly detection and prevention are critical in various fields, particularly cybersecurity, where early identification and intervention against malicious activities are essential. While traditional studies have predominantly focused on anomaly detection, anomaly prevention has gained increasing importance in identifying and mitigating risks preemptively. This thesis presents a comprehensive framework integrating anomaly detection, prevention, advanced clustering, and nature-inspired algorithms. Our approach leverages the K-medoid clustering method and the Salp Swarm Algorithm for optimal threshold determination. This hybrid method effectively identifies outliers and suspicious nodes within datasets. The framework is applied to real-world datasets, including the widely used Enron email dataset, allowing content-based and node-based anomaly detection.

Experimental results demonstrate the success of the proposed method, particularly in the cybersecurity domain, where it outperforms alternative techniques on 5 out of 10 datasets, achieving an AUC value of 0.8651 on the Thyroid dataset. Additionally, the framework introduces a novel concept of “suspicious nodes,” identified by data discrepancies between content and structural features. These nodes are labeled for further analysis to prevent potential harmful actions, such as fraudulent behavior or malicious emails. The proposed framework enhances anomaly detection methodologies and pioneers a novel approach to anomaly prevention, offering a proactive solution for mitigating risks before they materialize.

Keywords: Anomaly Detection/Prevention, Social Network Graph, Nature Inspired Algorithms, Link Prediction.

Dedication

This thesis is dedicated to my beloved sons, Selim Alp KILIÇ & Ömer Efe KILIÇ, whose unwavering love, patience, and joy have been my constant source of strength. Your smiles have been my greatest motivation, and I hope this work inspires you to always pursue your dreams with passion and perseverance.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my invaluable advisor, Assoc. Prof. Dr. Esra SARAÇ EŞSİZ, who has always supported me with her encouraging and directing attitude and enlightened me with her knowledge and experience during my graduate education.

I would like to thank my committee members Prof. Dr. Selma Ayşe ÖZEL, Assoc. Prof. Dr. Mümine KAYA KELEŞ, Asst. Prof. Dr. Barış ATA and Asst. Prof. Dr. Mehmet SARIGÜL for their suggestions and contributions.

I would like to thank my dear husband and colleague, Res. Asst. Yasir KILIÇ for helping me to evaluate the events from a different perspective and broaden my horizons.

I would like to express my gratitude to my beloved parents and dear brother, who supported me in all matters and are always with me on my good and bad days.

Finally, I would like to thank all my teachers and friends who have contributed to me throughout my education.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

SVM	: Support Vector Machine
KNN	: K-Nearest Neighbor
NB	: Naive Bayes
PCA	: Principal Component Analysis
DLAA	: Dynamic Link Anomaly Analysis
DT	: Decision Tree
DIF	: Deep Isolation Forest
DVT	: Denoising Variational Transformer
AUC	: Area Under the Curve
BGA	: Behavioral Graph Analysis
TBS	: Trust-based Anomaly Detection
GA	: Genetic Algorithm
GS	: Greedy Search
STBM	: Stochastic Topic Block Model
DNN	: Deep Neural Network
LR	: Logistic Regression
CNN	: Convolutional Neural Network
RF	: Random Forest
RNN	: Recurrent Neural Network
GSA	: Gravitational Search Algorithm
AA	: Adamic Adar
AClCo	: Average Clustering Coefficient
ASP	: Average Shortest Path
EPAADPS	: Efficient Proactive Artificial Immune System Based Anomaly Detection
NSA	: Negative Selection Algorithm
RTM	: Repertoire Training Module
VAM	: Vulnerability Assessment Module
RM	: Response Module
SSA	: Salp Swarm Algorithm

XGBoost	: Extreme Gradient Boosting
ABC	: Artificial Bee Colony
AFS	: Artificial Fish Swarm
LR	: Logistic Regression
FCM	: Fuzzy C-means Clustering
CFS	: Correlation-Based Feature Selection
HHO	: Harris Hawks Optimization
MLSTM	: Multiplicative Long Short-Term Memory
BOA	: Bayesian Optimization Algorithm
CS	: Cuckoo Search
FLIP	: Fairness Aware Link Prediction
CLPE	: Central Node-based Link Prediction
LILP	: Local Link Prediction Model
LGLIP	: Local-Global Link Prediction Model
FSDO	: Fuzzy Spatial Distribution with Opinion Spreading
ML	: Machine Learning
HM-EIICT	: Method-Extended using Intra and Inter Community Threshold
PLP	: Potential Link Prediction
MPCP	: Metapath Computed Prediction
BP	: Balanced Precision
AUPR	: Area Under Precision-Recall Curve
MCC	: Matthews Correlation Coefficient
NDCG	: Normalized Discounted Cumulative Gain
IQR	: Interquartile Range
ED	: Euclidean Distance
LOF	: Local Outlier Factor
iForest	: Isolation Forest
CN	: Common Neighbors
JC	: Jaccard Coefficient
PA	: Preferential Attachments
TF-IDF	: Term Frequency-Inverse Document Frequency
K-SAD	: K-Medoid Salp Swarm Anomaly Detection
SNAP	: Suspicious Nodes Anomaly Prevention

TPR	: True Positive Rate
FPR	: False Positive Rate
POI	: Persons of Interest
ROC	: Receiver Operating Characteristic
CD	: Critical Difference

LIST OF SYMBOLS

D(o,c)	**:** Total Distance of Object o to Cluster c
D(o, m_i)	**:** Distance of Object o to Medoid i_{th}
Q_1	**:** First Quarter
Q_2	**:** Third Quarter
IQR	**:** Difference Between Quarters
k	**:** constant
S(i)	**:** Silhouette Score
a(i)	**:** Similarity to Other Elements in its Own Set
b(i)	**:** Similarity to its Nearest Neighbor Set
N(o)	**:** k-Nearest Neighbors of the Object
E(h(x))	**:** Average of the Length of Paths taken from Trees
N(X), N(Y)	**:** Neighbors Nodes
N(u)	**:** Set of Neighbors of a Common Neighbor u
x_j^1	**:** the Position of the Leader Salp in the j_{th} Dimension
F_j	**:** Food Source
ub_j	**:** Upper Bound
lb_j	**:** Lower Bound
c_1, c_2, c_3	**:** Random Numbers
L	**:** Total Number of Iterations
l	**:** Current Iteration Count

1. INTRODUCTION

In the era of rapid technological advancements and exponential data growth, detecting and preventing anomalies has become increasingly critical. The widespread use of the internet and the interconnected nature of modern systems have made data security a paramount concern as cybersecurity threats continue to evolve in both sophistication and scale. Anomalies—data points or behaviors that deviate from the norm—pose significant risks across many domains, including financial fraud detection, network security, medical diagnosis, and spam detection. By identifying these irregular patterns, anomaly detection systems can help mitigate various threats, such as unauthorized access, zero-day attacks, data breaches, and malicious behaviors that could compromise the integrity and confidentiality of systems and data. The challenge of anomaly detection becomes even more pronounced in dynamic and risky media like social networks, where the actions of malicious actors can have real-world consequences.

Social networks, as graph-based structures, have become ubiquitous in everyday life. The ability to reach and interact with others across the globe has brought about unprecedented convenience, but it has also provided new avenues for malicious individuals to exploit. These malicious actors often blend in with regular users, making their detection critical and complex. Anomalous behaviors in social networks can range from spreading misinformation to initiating cyberattacks or conducting fraudulent activities. Detecting these anomalies early and effectively is crucial for identifying the threats and preventing their potential harm before they escalate. In the context of anomaly detection within social networks, the task involves identifying anomalies and predicting their potential impact, requiring robust algorithms that can adapt to evolving data patterns.

Anomalies, by nature, are unpredictable and diverse. In graph-based systems like social networks, these anomalies manifest as unusual node behaviors or anomalous connections between users. However, the methods traditionally used for anomaly detection often struggle with dynamic environments where the definition of "normal" behavior constantly shifts. Thus, anomaly detection systems must be accurate and adaptable to the ever-changing strategies used by malicious actors.

In this context, clustering-based approaches have emerged as powerful tools for grouping similar data points and identifying outliers. Clustering techniques help reduce the computational load involved in processing large datasets by organizing data into meaningful groups, which in turn aids in detecting anomalies. Unlike traditional density-based, distance-based, or statistical methods, clustering algorithms can effectively differentiate between normal and anomalous behaviors by analyzing the relationships between data points. Additionally, nature-inspired algorithms, which mimic natural processes, have shown significant promise in enhancing anomaly detection. Due to their inherent randomness and robustness, these algorithms can handle complex, large-scale datasets with minimal computational overhead. By incorporating fitness functions specific to the problem, nature-inspired methods can escape local optima, making them ideal for anomaly detection tasks requiring high scalability and adaptability.

Despite the ongoing advancements in anomaly detection techniques, there is no one-size-fits-all solution. Different datasets, application contexts, and domains require tailored methods for optimal anomaly detection. While some methods analyze entire datasets effectively in batch processing, they may struggle in real-time environments that require immediate responses. Conversely, incremental methods, which update detection models iteratively as new data arrives, are more suited for real-time applications but may sacrifice accuracy for scalability. Thus, selecting the appropriate anomaly detection approach is crucial, and a comprehensive evaluation of dataset characteristics is necessary to maximize the detection system's effectiveness.

Anomaly detection alone, however, is insufficient in specific critical contexts where preventing harmful actions is more valuable than merely detecting them after the fact. Anomaly prevention, which allows for intervention before an anomalous action occurs, is a relatively underexplored area compared to anomaly detection. In particular, social networks and email communications stand out as areas where the consequences of missed anomalies can be significant. For instance, a malicious email attachment can cause irreversible damage once the sender delivers it. There is an urgent need for frameworks beyond detection to actively prevent anomalous actions.

This thesis addresses anomaly detection and prevention using a novel framework that applies to various datasets, including social networks and email communications. A method was proposed that takes advantage of the power of nature inspired algorithms, detects anomalies, predicts potential harmful actions, and prevents them through link prediction techniques. This approach is crucial in environments like email systems, where a harmful message often causes damage. For this reason, the Enron email dataset is utilized, a widely recognized real-world dataset used extensively in previous anomaly detection studies. The Enron corpus provides a unique opportunity to compare the results of the proposed approach with other studies while demonstrating the effectiveness of the proposed framework in real-world scenarios. Since the dataset includes the emails' content and metadata, such as sender, receiver, timestamps, and communication patterns, it is an ideal testbed for the proposed anomaly prevention system.

The main contributions of this thesis are as follows:

1. A novel framework that integrates anomaly detection and anomaly prevention combined with the power of nature-inspired algorithms is proposed.
2. The proposed approach combines node-based and content-based features to identify suspicious nodes in social networks and prevent their potential connections through link prediction techniques. This method offers a comprehensive solution for addressing anomalies in graph-based systems.
3. A batch method for anomaly detection that overcomes the limitations of both traditional batch and incremental methods is introduced, achieving a balance between scalability and accuracy across different datasets.

The motivation of this thesis stems from the observation that many anomaly detection studies in the literature claim to achieve the best results. However, there is no single superior method for anomaly detection, as the most effective approach depends on the dataset's characteristics. Therefore, this thesis aims to demonstrate this variability and emphasize the need for dataset-specific methodology selection. Additionally, rather than solely focusing on anomaly detection, anomaly prevention is explored, which seeks to mitigate anomalies before they occur. Since

anomaly prevention has received limited attention in the literature, this study aims to contribute to this relatively underexplored yet critical area.

This thesis is organized as follows: Section 2 provides a literature review on anomaly detection techniques and nature-inspired algorithms and their application to social networks and email communications. Section 3 outlines the background and motivation of this study, while Section 4 details the system overview and methodology. Section 5 presents the experimental results using multiple real-world datasets, including the Enron email dataset, and compares the proposed approach with existing methods. Finally, Section 6 concludes the thesis by discussing future research.

2. THEORETICAL FOUNDATIONS AND LITERATURE REVIEW

Much of the literature has traditionally focused on anomaly detection, which seeks to identify unusual or unexpected behaviors within systems to uncover vulnerabilities and potential threats. These approaches are predominantly reactive, taking action only after the threat has emerged. However, moving beyond anomaly detection to anomaly prevention is crucial, as it shifts the focus toward preventing potential threats before they occur. Few studies have focused on anomaly prevention, making it an underexplored area despite its significance in preemptively mitigating risks. This study aims to bridge this gap by proposing a proactive approach emphasizing anomaly prevention over post-event intervention. Such a strategy enhances system security and reduces the risk of costly security breaches.

The proposed study extends beyond traditional anomaly detection methods by utilizing link prediction techniques, seldom employed for anomaly prevention. The existing literature typically applies link prediction in social network analysis to predict future relationships and connections between individuals. These studies aim to model the evolution of relationships over time and forecast future interactions. However, the application of link prediction methods for anomaly prevention remains largely unexplored. This gap represents one of the innovative aspects of the proposed study. By analyzing relationships within data through link prediction, this thesis contributes to enhancing the process of anomaly prevention, particularly in social network-like datasets.

The Enron email dataset, a widely utilized real-world dataset, has been employed extensively in anomaly detection research. It offers a rich set of data capturing the email communications within the Enron Corporation, making it highly suitable for building robust anomaly detection models. The dataset's structure can also be modeled as a social network, enabling user connections and interaction analysis. Previous research has demonstrated the efficacy of the Enron dataset in detecting anomalies across various domains, including classification, clustering, and text analysis. However, existing studies mainly emphasize detection, with limited focus on preventive measures. By applying the proposed proactive anomaly prevention

framework in the context of the Enron dataset, this study aims to address this gap and explore the potential of applying preventive strategies to real-world scenarios.

A critical component of anomaly detection is the accurate setting of anomaly thresholds, which help distinguish normal from abnormal behavior. Establishing the correct threshold is crucial, as it minimizes false positive and false negative rates, leading to more effective anomaly detection. The proposed study uses nature-inspired algorithms to determine these thresholds. These algorithms have proven successful in complex optimization problems. However, their use in setting anomaly thresholds is limited in the literature, which adds to the novelty of the proposed method. The proposed method enhanced the anomaly detection and optimized the system for lower false positive rates by using nature-inspired algorithms.

While most of the literature centers on anomaly detection, this study introduces a novel framework for anomaly prevention, offering a proactive solution to potential security threats. The proposed approach, which combines link prediction and nature-inspired algorithms, goes beyond traditional detection methods and seeks to prevent threats before they occur. This study utilizes the Enron email dataset to provide a real-world application and offers insights into anomaly prevention that organizations can apply across various domains. Furthermore, using nature-inspired algorithms and link prediction methods represents a rare and innovative solution, offering the potential for more secure and efficient systems in practice.

2.1. Anomaly Detection and Prevention

Akoglu et al. (2010) proposed the OddBall algorithm. Neighborhood sub-graph rules were created and used as features for anomaly detection. New rules were created based on Near-cliques and stars, Heavy vicinities, and Dominant heavy links. Postnet, Auth2Conf, Com2Cand, Don2Com, Enron, and Oregon datasets were used. Unlike other studies, they achieved fast and good results on unlabeled data. They planned to work on dynamic datasets in future studies.

There were positive and negative data on protein interaction networks. Some data that is positive can be labeled as negative. Singh and Vig (2017) proposed an anomaly detection method for negative data. Arabidopsis thaliana, Caenorhabditis elegans, Mus musculus, and

Rattus norvegicus were used as datasets. By reducing the noise, the negative data, which are actually positive, were detected. SVM, C5.0, KNN, and NB classifiers were used. Parzen window, PCA, Nearest neighbor methods, and Gaussian process were used as anomaly detection methods. As a result, the accuracy of estimating missing links has increased.

Karakaşlı et al. (2019) proposed spam detection methods using the Twitter dataset. They applied a dynamic feature selection instead of a static one. Similar persons were grouped, and the best features were selected for each group. They achieved 7-8% higher accuracy than classical methods. They also used machine learning methods to detect spam users. SVM and KNN were used as classifiers.

Zhang and Liao (2019) performed link anomaly detection dynamically, depending on the network topology. They created a generic structure without depending on the node attributes. They proposed a framework called dynamic link anomaly analysis (DLAA). They estimated the time-dependent links created by unlink/link nodes and compare statistics such as false positives.

Huang and Zeng (2006) proposed an anomaly detection method based on link prediction. Potential scores were determined for links with Preferential Attachment, Spreading Activation, and a Generative Model. Huang and Zeng estimated the likelihood distribution for all sender-receiver tuples. Anomaly scores were determined by only considering the senders and receivers. The context was ignored. Among these three methods, the best one was the generative model, followed by spreading activation, and last, preferential attachment.

Kagan et al. (2018) developed a link prediction-based anomaly detection method using graph topology. ArXiv, Class, DBLP, Flixster, Twitter, and Yelp were used as the datasets. Fully simulated, semi-simulated, and real-world Networks were used. The method has two stages. In the first stage, the link prediction classifier was constructed, and in the second stage, meta-features were selected. They compared their results with other studies, and they achieved higher performance.

Rahman et al. (2021) used three machine learning algorithms for anomaly detection. They classified the dataset into abnormal and normal by using a Decision Tree (DT). After that, they classified abnormal data into happiest and disappointed users by using a Support Vector Machine (SVM), and finally, they classified disappointed users into suicidal and not suicidal by using Naive Bayes (NB). User profiles and content were used as features. They extracted word dictionaries for disappointed and suicidal users. They compared different machine learning algorithms, both synthetic and real datasets. Their method has the highest accuracy, according to others. In future work, they will apply online real-time detection.

Degirmenci and Karal (2022) combined iLOF and iDBSCAN for outlier detection. Furthermore, they introduced core kNN for the clustering step and used incremental Mahalanobis distance. They used 16 real-world datasets for their experiment. They compared their method with nine different methods. They had more significant results except for three datasets.

Xu et al. (2022) proposed a kNN-LOF-based outlier detection method. They renewed the reachable distance and defined a hierarchical adjacency order. They compared their method with LOF, COF, and RDOS. They used the AReM dataset from UCI. Their method was the highest with k=45 and 0.965 AUC. But, COF reached 0.935 AUC value with k=20.

Xu et al. (2023) proposed the Deep Isolation Forest (DIF) method. It was compared against iForest and its variants and several state-of-the-art deep learning-based approaches for anomaly detection across various datasets, including tabular, graph, and time series data. Experiments revealed that DIF consistently outperforms its competitors across various datasets. However, while DIF's representation diversity and optimization were less effective on certain datasets (R8, Cover, Pageblocks, Thrombin), it demonstrated superior performance in high-dimensional and complex data spaces and across diverse data types.

Min et al. (2024) introduced a denoising variational transformer (DVT) for interpretable anomaly detection in autonomous vehicles, demonstrating significant improvements over existing methods. The proposed DVT model achieves an F1 score enhancement of 8%–50%

and an area under the curve (AUC) increase of 1%–20%. Additionally, the residual interpreter improved accuracy by over 50% compared to kernel Shapley additive explanations and reduced computational time by over 99%. The model has shown substantial anomaly detection effectiveness using simulation and real-world data.

Anomaly detection methods are usually used in the studies, but Wu (2019) suggested a prevention method in his study. First, he constructed a trust-based graph. Thus, it simulated DDoS, spam, and botnet attacks. Later, he detected a collective anomaly on this graph, not an individual one. He argued that anomalies that occur collectively rather than individually are more dangerous. He improved the Spectral Clustering Algorithm. He compared his method with the Behavioral Graph Analysis (BGA) method. The Trust-based Anomaly detection method (TBS) achieved better results than the BGA method. The study's shortcomings were that it is very difficult to define the control and general stages for botnet attacks, and anomaly detection was made by considering only certain situations.

Vlasselaer et al. (2015) proposed a fraud detection method. They aimed at preventing fraud by companies. Companies that deliberately did not pay their contributions to the state were identified. For this, they used already labeled data and calculated a fraud rate for companies new to the network. They also achieved better results using a clique-based approach than traditional methods. They also tagged companies that were members of suspicious cliques as suspicious. As they calculated the future anomaly rate, they took precautionary measures and caught 22% of the fraudulent activities of large companies.

Kirchner and Gade (2011) proposed a fictional case scenario. In this scenario, banks noticed that the interest rates received were decreasing. Bank realized that the reason for this was a trick that the dealers made to agree among themselves and not pay high interest. And they created a network. Relationships in the network were continually expanded, and clusters with high fraudulent activity were identified. Key actors within each cluster were identified. The bank rejected interest requests from all dealers that fit this model. Banks confronted the big dealers who caused most scams and respected real dealer-to-dealer transactions.

Table 2.1. Anomaly Detection Literature Review Based on Clustering and Nature Inspired Algorithms

Study	Methodology	Key Techniques	Outcome
Alsaleh and Binsaeedan (2021)	SSA + NB/XGBoost	Salp Swarm for feature selection	XGBoost > NB
Pitchaimanickam (2021)	Hierarchical SSA	SSA-based Cluster Head selection	Improved network lifespan
Li et al. (2021)	ISSA-DPC	Cosine decline, chaos strategies	Outperformed DPC, KH-DPC
Degirmenci and Karal (2022)	iLOF + iDBSCAN	Incremental Mahalanobis, core kNN	Top performance
Xu et al. (2022)	Modified kNN-LOF	Reachable distance adjustments	AUC 0.965
Degirmenci and Karal (2021)	RiLOF	MoNNAD with iLOF	Superior on most datasets
Maazalahi and Hosseini (2024)	ASO-EO-FA-K-means	K-means for local search in Firefly Algorithm	Accuracy:%99; Lower Time Complexity
Mosa et al. (2024)	Meta-heuristic+RF,SVM	Feature Selection with Sail-fish Optimizer	Accuracy:%97; reduction features: %90
Choudhary et al. (2024)	ABCO + CNN	Hyperparameter optimization of CNN using ABCO	Accuracy:%100
Alzaqebah et al. (2023)	Harris Hawk + ELM	Best feature subset and optimized ELM weights	Best results in G-mean metric
This study	K-medoid+SSA	Threshold with SSA, reduce computation with K-medoid	Highest results in 5 of the 10 datasets

Table 2.1 provides an overview of previous studies in the literature. While prior research has predominantly focused on using nature-inspired algorithms for feature selection, this thesis employs these algorithms to determine thresholds, achieving significant results across ten different datasets. Additionally, although many existing methods claim superiority, no single approach can be universally regarded as the best for anomaly detection. The performance of a method is highly dependent on the dataset and parameter configurations, which can vary considerably. The proposed clustering-based approach highlights the effectiveness of clustering techniques in handling anomaly detection tasks.

2.2. Enron Email Dataset

Cormack and Lynam (2005) created a spam filtering tool. Firstly, they applied their tool to artificial collections and tried to reach the gold standard. When it reached G5, eight messages were converted from spam to ham, and 421 messages were converted from ham to spam. Then, they applied their method with the Enron corpus. When it reached G8, 632 messages were converted from spam to ham, and 383 messages were converted from ham to spam.

Wilson and Banzhaf (2009) proposed a genetic algorithm (GA) to detect email relationships between sender and receiver. They used degree, density, and proximity measures. They compared their method with Greedy Search (GS) and showed their superiority.

Corneli et al. (2019) proposed a dynamic graph analysis model. They used interactions between nodes and text data as features. They developed the Stochastic Topic Block Model (STBM) and extended a dynamic graph from the Enron corpus.

Zhang et al. (2016) created three networks such as user, subject, and keyword. This allowed them to analyze user behavior. They then proposed a keyword selection algorithm based on the Greedy Approach. Thus, they examined the effect of e-mail subjects on keywords. This has been a preliminary study for subject recommendation and character recognition. Future studies will explore methods to extract meaningless adjectives and broad-sense keywords that will not be understood in speech analysis.

Apoorva and Sangeetha (2021) proposed cluster-based classification and deep neural networks (DNN). Firstly, they extracted the email body. After those features were extracted and vectorized. Finally, with their methods, authors were predicted based on e-mail content. They used two different methods. In the first method, 94% accuracy values were reached on the five authors, 75% on the entire dataset. In the second method, they used model-bed clustering. Therefore, they reached 86% accuracy in the entire dataset.

Martino et al. (2023) highlighted that traditional machine learning algorithms remain prevalent in spam filtering compared to deep learning models. Experiments conducted with five distinct

email datasets revealed that dataset shift and spammer strategies negatively impact the performance of anti-spam filters. Specifically, tests using the Bruce Guenter datasets demonstrated deterioration in the models' predicted generalization accuracy. Discrepancies between spam sources and spammer strategies were identified as key factors contributing to declines in filter performance. In this context, the study underscored the necessity of addressing dataset shift issues in evaluating spam filters. They suggested that an in-depth analysis of spammer strategies could enhance the robustness of filters.

Bountakas and Xenakis (2023) introduced HELPHED, a novel phishing email detection system that combines Ensemble Learning with hybrid features to enhance detection accuracy. Unlike traditional methods that focus solely on content or text, HELPHED integrated both aspects, utilizing two distinct Ensemble Learning approaches: Stacking and Soft Voting. These methods employed different Machine Learning algorithms to process hybrid features separately, improving overall model performance. The evaluation of a diverse dataset, which reflects the evolution of phishing tactics, showed that HELPHED, especially with Soft Voting, significantly surpasses other algorithms with a remarkable F1-score of 0.9942.

Poobalan et al. (2025) proposed a highly efficient classifier aimed at identifying spam within email datasets, addressing the unique challenges posed by the complex and varied nature of email data. Their model, based on Bidirectional Long Short-Term Memory (BiLSTM), was evaluated across seven Enron datasets and achieved a superior accuracy rate of 98.69%. To optimize classification, they utilized TF-IDF for feature selection, ensuring the identification of relevant features. The research also have included a comparison of different models such as Logistic Regression (LR), Convolutional Neural Networks (CNN), Random Forest (RF), and Recurrent Neural Networks (RNN). Additionally, they explored solutions for cloud data security, introducing a hybrid AES-Rabbit encryption technique to enhance privacy and security in data management.

2.3. Nature Inspired Algorithms

Bastami et al. (2019) proposed a multi-level link prediction. Community detection, subgraph optimization, and local link prediction were applied. First, communities were detected. Then,

Subgraphs were optimized with a Gravitational Search Algorithm (GSA). Finally, the Adamic Adar (AA) link prediction method was applied. Subgraphs were optimized with a high Average Clustering Coefficient (AClCo) and low Average Shortest Path (ASP). They used astro-ph, Blogs, grid, protein, Internet topology, us-air, NetFlix, CiteSeer, Cora, WebKb, and Gnutella peer-to-peer as datasets.

Saurabh and Verma (2016) proposed An Efficient Proactive Artificial Immune System based Anomaly Detection and Prevention System (EPAADPS). Firstly, they have generated detectors and selected detectors above the power threshold. They have added self-tuning and detector power features to their work, which were missing in the classic Negative Selection Algorithm (NSA). Results were compared with the Real-Valued Negative Selection Algorithm (RNS). As a result, it was seen that the detection rate increased, and the false alarm rate decreased according to RNS.

Alsaleh and Binsaeedan (2021) proposed a Network Anomaly Detection based on the Salp Swarm Algorithm (SSA). UNSW-NB15 and NSL-KDD datasets were used. SSA was used for feature selection. After feature selection, Naive Bayes (NB) and extreme gradient boosting (XGBoost) algorithms were used, and their performances were compared.

Hajisalem and Babaie (2018) proposed a hybrid anomaly detection. Firstly, they have used an Artificial Bee Colony (ABC) and an Artificial Fish Swarm (AFS) together. Datasets were divided into subsets using Fuzzy C-Means Clustering (FCM). Then Correlation-based Feature Selection (CFS) was used to remove irrelevant features. The CART technique was used to separate normal behaviors from abnormal ones. UNSW-NB15 and NSL-KDD datasets were used. The proposed method and state-of-the-art methods were compared. Comparison results have shown that they were better than 5 of the 6 studies they compared.

Aswani et al. (2017) proposed a hybrid outlier detection. They used the Twitter dataset and extracted 11 attributes, like author interaction and count. K-nearest neighbors and Artificial Bee Colony (ABC) were used together. When only k-nearest neighbors were used, the model was

getting stuck in a local optimum, but with ABC, global optimization was guaranteed. The accuracy of the study was 98,37%.

Alzaqebah et al. (2023) proposed an intrusion detection system that utilizes a bio-inspired optimization algorithm for feature selection, specifically the Harris Hawks Optimization (HHO) algorithm. Their work focused on improving the accuracy and efficiency of intrusion detection by optimizing the selection of relevant features, which helped to reduce the dimensionality of the dataset and eliminate redundant or irrelevant information. Their method was tested on benchmark intrusion detection datasets, showing enhanced detection rates and lower false positives compared to traditional machine learning approaches. Their research emphasized the importance of feature selection and optimization in building robust IDSs.

Khayyat (2023) presented an Improved Bacterial Foraging Optimization with Deep Learning-Based Anomaly Detection (IBFO-ODLAD) method for anomaly detection in IoT environments, aimed at enhancing operational efficiency by identifying system failures. Their method used Z-score normalization for data preprocessing, IBFO for optimal feature selection, and Multiplicative Long Short-Term Memory (MLSTM) for classification. The model's hyperparameters were optimized through the Bayesian Optimization Algorithm (BOA). Validated on the UNSW NB-15 and UCI SECOM datasets, the IBFO-ODLAD method achieved impressive accuracy rates of 98.89% and 98.66%, respectively. Compared to existing methods like SVM, ANN, and NB, IBFO-ODLAD outperformed them in precision, recall, and F-score.

Shao et al. (2022) developed a novel outlier detection method combining image processing and the Cuckoo Search (CS) algorithm for dam monitoring data. The approach transformed monitoring data into binary scatter plot images, using Gaussian blur and Otsu binarization to filter isolated outliers. The CS algorithm identified clustered outliers by connecting pixel aggregations, with the process repeated until convergence. Comparative results have shown that the method outperformed five other techniques in detection accuracy and improved the predictive performance of monitoring models. The method was highly applicable to time series data in dam engineering and potentially other fields like road and bridge monitoring.

2.4. Link Prediction

Link prediction can be unfair. For example, a system used to recruit people may biasedly suggest that the company should always hire men since there are more male software engineers. Therefore, women cannot be employed. This problem is called a filter bubble. Masrour et al. (2020) proposed the FLIP (Fairness Aware Link Prediction) framework. They used modularity measures for postprocessing of the current link prediction and combined it with the adversarial network for solving the filter bubble problem. Dutch School, Facebook, and Google+ datasets were used.

The bipartite network is converted into a unimodal network for predicting links. But after the conversion, information loss occurred. Generally, a threshold value is used for the conversion. Choosing a threshold is so important. To solve this problem, Aslan and Kaya (2018) proposed a strengthening weighted method. They used four different threshold values according to the most frequently observed edges. Therefore, the number of link predictions and waiting time would be reduced, but more accurate links would be established. The dataset, which was an author-topic bipartite network, was collected with IEEE XPlorer service. Consequently, the potential connections have increased when the determined threshold value decreased and vice versa. With this method, high-quality links were obtained.

Jiang et al. (2020) proposed a community detection method using link prediction. The network can be ambiguous or clear. Community detection is easy for clear networks but difficult for ambiguous networks. Their perspective was that if the central node is used for detection, a node will not accidentally join the other community. Therefore, community detection would be more accurate. They compared their Central Node-based Link Prediction (CLPE) method with five state-of-the-art methods and showed the superiority of the CLPE method.

Users' perspectives can be different in different situations. For example, two people love the same style of drama movies but different styles of romantic movies. Classical methods said that high and low similarities can be concluded as middle similarities. For this reason, Su et al. (2020) proposed a vector similarity method on MovieLens and Netflix datasets with more

accurate results. Also, they used different similarity methods for different types of items, such as global, local, and meta similarities.

Tuan et al. (2019) proposed a fuzzy similarity-based link prediction in co-authorship network datasets. Neutrosophicfuzzy Adamic–Adar (FAA), Common Neighbours (FCN), and Jaccard Coefficient (FJC) were the similarity methods. A comparison with other studies was made with SVM and Gboost. Their performance was better than others.

Girdhar et al. (2019) proposed a local (LILP) and local-global link prediction model (LGLIP). Firstly, they used fuzzy for calculating trust and distrust between users. In a signed network, without fuzzy, +1 and -1 are used for trust and distrust. These are different from each other. One person cannot trust or distrust totally. After applying trust and distrust based on fuzzy, missing links were predicted in their proposed LILP and LGILP methods. They used both synthetic and real datasets. Synthetic ones were friends and foes. The real ones were Epinions and Slashdots. They reached high prediction accuracy with a larger dataset.

The similarity is calculated based on neighbors' ratings in the network. If there is not enough information, the calculation will be hard. This is called a “cold start” problem. Ai et al. (2019) have selected neighbors by using distance according to the positions of nodes. Firstly, the connection weights of items were calculated. Fuzzy Spatial Distribution with Opinion Spreading (FSDO), Tags (FSDT), Movie attribute (FSDM), and Pearson correlation coefficient (FSDP) were applied to remove unimportant links. After a complex network model was constructed and the spatial distribution layout of the items was determined, neighbors were selected according to the distance between items. Finally, users' rating values on items were calculated from previously selected neighbors. As a result, FSDP was better than some other methods, and FSDM avoided the item cold start problem and reduced computation complexity with an acceptable accuracy loss.

Samad et al. (2021) proposed structural importance-based link prediction. They added centralization to similarity. If two nodes have the same centrality and high similarity, then links can be constructed between these nodes. They used four different datasets and state-of-the-art

similarity metrics. They deleted some links in the dataset, then they tried to predict them with their methods and state-of-the-art methods. With their methods, all state-of-the-art similarity metrics gave better results according to the actual version of them. The best results were obtained with their structural importance and SAM. They tried different threshold values: 0.1, 0.3, 0.5, and 0.7. They predicted 95% links with a 0.1 threshold.

Ott et al. (2021) improved an interpretable rule-based model for link prediction. Their method used AnyBURL for extracting rules. Their method removed redundant rules by using thresholds. They compared their method with Machine Learning (ML) models. They outperformed ML models because ML models have unwanted bias and intransparency.

Saxena et al. (2022) proposed a heuristic Method-Extended using Intra and Inter Community Threshold (HM-EIICT). Firstly, they calculated similarity scores for node pairs. Then, node pairs were labeled as intra and inter-community. Finally, Thresholds were determined according to their community. They tried thresholds 0.1 to 0.9. They said that selecting thresholds was an open research question. For evaluating their model, they used accuracy and modularity reduction methods. Four different heuristic methods were tried with five different datasets. All heuristic methods with EIICT were better than those without EIICT versions.

Assouli et al. (2021) proposed a Latent Link Prediction based on Internal and Local Links for bipartite networks. They received the advantages of two other works that used the Potential Link Prediction (PLP) method, and they handled the disadvantages of these. They did not use a threshold. Instead, they used reliability. Their purpose is to block crime activities. Their methods were compared with other methods and PLP methods. They outperformed all methods.

Lande et al. (2020) proposed a metapath computed prediction (MPCP) for cooperation among scientists. The random walk was used to predict the probability of cooperation between authors. They used an empirical threshold between [0,1]. If the link weight was more than the threshold, it would be removed. When the threshold increases, the number of restored links decreases. When the threshold was equal to 1, 50% of the links remained. As a result, their method was a feasible method for link prediction.

Zhou (2023) proposed a threshold-free link prediction method. He tried to discriminate the link with three metrics. These were Area Under the Receiver Operating Characteristic Curve (AUC), Balanced Precision (BP), and Area Under the Precision-Recall Curve (AUPR). As a result, AUC and AUPR were better than BP. The author suggested that AUC and AUPR can be used together.

Jiao et al. (2024) proposed a study to find discriminative metrics in 2024 by comparing metrics such as Precision, Recall, F1-measure, Matthews Correlation Coefficient (MCC), BP, AUC, AUPR, Normalized Discounted Cumulative Gain (NDCG), and AUC-mROC. As a result of the comparison, they suggested using AUC, AUPR, and NDCG and not using BP as a metric.

3. MATERIAL AND METHOD

This section outlines the materials and methods employed in the thesis, detailing the datasets, tools, and methodologies used to conduct the research and achieve the proposed objectives.

3.1. K-Medoid

K-medoids, introduced by Kaufman and Rousseeuw (2009), is a clustering algorithm designed to partition a dataset into a predefined number of clusters. It operates by identifying medoids, which are the data points most representative of their respective clusters.

The algorithm begins by selecting an initial set of *k* medoids, which are then iteratively refined to improve cluster accuracy. The process continues until a predefined stopping criterion is met, such as reaching a maximum number of iterations or satisfying a specified similarity threshold.

$$D(o,c) = \sum_{i=1}^{n} d(o,m_i) \tag{3.1}$$

In Equation 3.1, *D(o,c)* represents the total distance of object o to cluster *c*. *d(o,* m_i*)* represents the distance of object o to medoid i_{th}.

The K-medoids algorithm is especially effective for clustering tasks, as it relies on medoids, which provide greater robustness against outliers compared to mean-based approaches.

3.2. Inter Quartile Range (IQR)

The interquartile range (IQR), defined by Wan et al. (2014), is a statistical measure of central tendency that quantifies the dispersion within a dataset. It represents the range between the second and third quartiles, encompassing the middle 50% of the data. By focusing on this central portion, IQR effectively measures variability while mitigating the influence of outliers.

IQR Formula:

$$IQR = Q_3 - Q_1 \tag{3.2}$$

In the given formula, Q_1 represents the first quartile of the dataset, while Q_3 denotes the third quartile. IQR is calculated as the difference between these two quartiles.

Tukey's Fences is an anomaly detection method based on the IQR. The upper and lower anomaly thresholds are determined using the following formulas:

$$\text{Lower} = Q_1 - \text{k} * \text{IQR} \tag{3.3}$$

$$\text{Upper} = \text{Q}_3 - k * IQR \tag{3.4}$$

In these formulas, k is a constant, typically set to 1.5. Values outside these bounds are considered outliers and may be subject to special labeling or removal.

3.3. Silhouette Score

The Silhouette Score, introduced by Rousseeuw (1987), is a widely used metric for objectively assessing the performance of clustering algorithms. It quantifies clustering effectiveness by comparing the similarity of each data point to others within the same cluster against its similarity to points in different clusters. A positive Silhouette Score indicates that a data point is well-assigned to its cluster, while a negative score suggests misclassification. This metric is valuable for comparing clustering algorithms and selecting the most suitable approach for a given dataset.

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3.5}$$

In Equation 3.5, *i* refers to the i_{th} element in the data. *S(i)* is the silhouette score, *a(i)* is the similarity to other elements in its own set, and *b(i)* is the similarity to its nearest neighbor set.

3.4. Euclidean Distance (ED)

Euclidean distance, referred to as the Euclidean metric (Danielsson, 1980), quantifies the straight-line distance between two points in Euclidean space. It is one of the most widely used distance measures across various disciplines, including geometry, physics, machine learning, and data analysis. The Euclidean distance between two points P=(p_1, p_2..., p_n) and Q=(q_1, q_2, ..., q_n) in an n-dimensional space is calculated using the following formula:

$$\text{ED} = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} \tag{3.6}$$

This metric is symmetric, meaning the distance from A to B is the same as from B to A. It satisfies the triangle inequality, a key property in distance-based algorithms.

In data analysis and machine learning, Euclidean distance is commonly employed to assess the similarity or dissimilarity between data points. It is a fundamental metric in clustering, classification, and nearest-neighbor algorithms.

3.5. Local Outlier Factor (LOF)

The Local Outlier Factor (LOF), proposed by Breunig et al. (2000), is a local anomaly detection method that evaluates an object's deviation relative to its neighbors. The process begins by determining the density of each object within its kk-nearest neighborhood. Subsequently, each object's local reachability density (LRD) is computed. Finally, the LOF value is obtained by calculating the ratio between an object's LRD and the LRD of its neighbors, allowing for the identification of anomalies based on local density variations.

$$LRD(o) = \frac{1}{\frac{\sum_{p \in N(o)} reach-dis\ (o,p)}{|N(o)|}} \tag{3.7}$$

$$LOF(o) = \frac{\sum_{p \in N(o)} \frac{LRD(p)}{LRD(o)}}{|N(o)|} \tag{3.8}$$

In Equations 3.7 and 3.8, *N(o)* represents the k-nearest neighbors of the object. *reach-dist(o,p)* is the distance measure between object *o* and neighbor *p*.

The Local Outlier Factor (LOF) is widely used for outlier detection due to its flexibility, effectiveness in handling multidimensional datasets, and ability to detect global and local anomalies.

3.6. Isolation Forest (iForest)

Isolation Forest, introduced by Liu et al. (2008), is a tree-based anomaly detection algorithm known for its speed and efficiency, mainly when applied to large and complex datasets.

The algorithm randomly selects a feature and defines a threshold within the range of that feature's minimum and maximum values. This threshold is then used to construct a tree based on the chosen feature. The process continues iteratively until a predefined tree height or number of nodes is reached. Multiple isolation trees (*n*) are generated and combined to form an isolation forest. The anomaly score for each data point is determined by averaging the lengths of the paths it follows across the trees. Data points with shorter average path lengths are identified as outliers.

$$\mathrm{s(x, n)} = 2^{-\frac{E(h(x))}{c(n)}} \tag{3.9}$$

In Equation 3.9, *E(h(x))* represents the average path length of a sample *x* across the trees in the isolation forest, while *c(n)* is a correction factor that accounts for the expected average path length of *n* samples.

iForest is widely utilized in various applications due to its speed, scalability, and adaptability to different data distributions. Its reliance on independence and randomness principles allows for efficient anomaly detection while maintaining low sensitivity to parameter selection.

3.7. Link Prediction Methods

Link prediction methods are used to forecast future connections between entities, such as people or objects. These methods predict potential connections for nodes that are currently not directly connected, helping to prevent unwanted connections if necessary.

3.7.1. Common Neighbours (CN)

The CN method (Liben-Nowell & Kleinberg, 2003) is a simple and effective technique used for link prediction in networks. The idea behind this method is that the more common connections two nodes have, the higher the probability they will connect directly. This method calculates the number of neighbors that two nodes, *X* and *Y*, share, which can be mathematically expressed as:

$$\mathrm{CN}(\mathrm{X}, \mathrm{Y}) = |N(X) \cap N(Y)| \tag{3.10}$$

Where *N(X)* and *N(Y)* denote the sets of neighbors of nodes *X* and *Y*, respectively. This metric is widely applied in social and biological networks for its simplicity and interpretability, though it struggles to capture more complex network structures.

3.7.2. Jaccard Coefficient (JC)

The Jaccard Coefficient (Jackson et al., 1989) measures the similarity between two sets by calculating the ratio of the number of common neighbors to the total number of unique neighbors. It is defined as the size of the intersection of the sets divided by the size of their union:

$$\mathrm{JC}(\mathrm{X}, \mathrm{Y}) = \left|\frac{N(X) \cap N(Y)}{N(X) \cup N(Y)}\right| \tag{3.11}$$

Where *N(X)* and *N(Y)* denote the sets of neighbors of nodes *X* and *Y*, respectively.

3.7.3. Adamic Adar (AA)

The Adamic-Adar index (Adamic & Adar, 2003) predicts the likelihood of a missing link between two nodes based on the number of common neighbors they share. It calculates the sum of the inverse logarithm of the number of neighbors for each common neighbor:

$$\mathrm{AA(X,Y)} = \sum_{u \in N(X) \cap N(Y)} \frac{1}{\log(|N(u)|)} \tag{3.12}$$

Where *N(X)* and *N(Y)* are the sets of neighbors of nodes *X* and *Y*, and *N(u)* is the set of neighbors of a common neighbor *u*.

3.7.4. Preferential Attachment (PA)

Preferential Attachment (Newman, 2001) suggests that nodes with a higher degree (i.e., more connections) are more likely to acquire new links. It is calculated as the product of the degrees of the two nodes:

$$\mathrm{PA(X,Y)} = |N(X)| \cdot |N(Y)| \tag{3.13}$$

Where *N(X)* and *N(Y)* are the degrees of nodes *X* and *Y*, respectively.

3.8. Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF (Sparck Jones, 1972) is a widely used weighting method to determine the importance of a term (or feature) within a document. It combines two components: TF and IDF. TF measures how often a term appears in a specific document by calculating the frequency of all terms in each document. To account for differences in document lengths, the term frequency is normalized by dividing it by the total number of terms in the document. The TF is calculated as follows:

$$\mathrm{TF(t)} = \frac{\#\ of\ times\ term\ t\ appears\ in\ Document\ D}{Total\ occurrences\ of\ all\ terms\ in\ Document\ D} \tag{3.14}$$

Inverse Document Frequency (IDF) measures how important a term is across the entire set of documents. A term that appears in fewer documents will have a higher IDF value, indicating it is more distinctive. Conversely, a term that appears in many documents will have a lower IDF, suggesting it is less significant. The IDF is computed as follows:

$$IDF(t) = \log\left(\frac{total\ \#\ of\ documents}{\#\ of\ documents\ with\ term\ t\ in\ it}\right) \quad (3.15)$$

By multiplying TF and IDF, the TF-IDF score is obtained, representing the significance of a term within a document relative to the entire document set.

4. SYSTEM OVERVIEW

This section will discuss parameter determination, the Salp Swarm Algorithm, and the K-Medoid Salp Swarm Anomaly Detection (K-SAD) approach, providing an overview of their roles and significance in the proposed framework.

4.1. Parameters Determination

All parameters are default values except the *k* value and fitness of the SSA. The *k* values are determined according to Degirmenci and Karal's study (Degirmenci & Karal, 2022). The SSA's fitness function is determined empirically. Table 4.1 shows these empirical experiments results.

Table 4.1. Results of Different Fitness Functions for IKS-LOF (|)

FORMULA	*a_dist*/*diff*	*a_dist*	*a_diff*/ *diff*	| *a_diff* – *diff* |
MUSK	0.9465	0.9341	**0.9494**	**0.9494**
IONOSPHERE	0.6367	0.6702	0.6367	**0.7733**
SATIMAGE-2	**0.9400**	0.9396	**0.9400**	0.9396
CARDIO	0.6752	**0.7620**	0.6755	0.6755
THYROID	0.8651	0.8651	**0.8669**	**0.8669**
ANNTHYROID	0.6811	0.6811	**0.6827**	0.6811
MAMMOGRAPHY	0.7603	0.7490	**0.7681**	0.7490
PIMA	**0.5457**	**0.5457**	0.5326	0.5326
GLASS	0.6705	0.6680	**0.7119**	0.7095
VOWELS	0.8663	0.8663	**0.8680**	**0.8680**

The variable *a_dist* represents the average distance of anomaly points, while *diff* denotes the average deviation of normal points from the overall mean. Similarly, *a_diff* indicates the average deviation of anomaly points from the total mean.

These formulas are based on the principle that anomaly points should be positioned as far from the center as possible, whereas normal points should remain close to the center. Consequently,

the fitness function is designed to be maximized, ensuring that the normal point deviation decreases while the anomaly point distance increases. The division operation highlights the inverse relationship between these two factors. This specific formula was selected over others due to its ability to detect a greater number of anomalies, despite minimal differences among alternative approaches. While other formulas may be more suitable for IKS-LOF (|), this one was chosen due to its potential applicability in other methods. For a more detailed discussion on IKS-LOF (|), refer to the Performance Comparison section of the thesis.

4.2. Salp Swarm Algorithm

Salps, as described by Mirjalili et al. (2017), inhabit the depths of the ocean. They exhibit a swarm-like behavior due to their chain formation. The Salp Swarm Algorithm includes two types of salps: the leader salp and the follower salps. The leader salp adjusts its position based on the formula given in Equation 4.1.

$$x_j^1 = \begin{cases} F_j + c_1((ub_j - lb_j)c_2 + lb_j), & c_3 \geq 0 \\ F_j - c_1((ub_j - lb_j)c_2 + lb_j), & c_3 < 0 \end{cases} \tag{4.1}$$

In Equation 4.1, x_j^1 denotes the position of the leader salp in the j-th dimension. F_j represents the food source, while ub_j and lb_j are the upper and lower bounds, respectively. The variables c_1, c_2, and c_3 are random numbers.

$$c_1 = 2e^{-(4\frac{l}{L})^2} \tag{4.2}$$

As shown in Equation 4.2, c_1 plays a crucial role in balancing exploration and exploitation. L and l denote the total number of iterations and the current iteration count, respectively. The algorithm initially emphasizes exploration and shifts towards exploitation as the iterations advance.

c_2 and c_3 are random values between 0 and 1. These values determine the direction of movement, whether towards positive or negative infinity, and also influence the step size.

$$x_j^i = \frac{1}{2}\left(x_j^i + x_j^{i-1}\right) \; where \; i \geq 2 \tag{4.3}$$

The new position of the i_{th} follower salp is calculated as the average of its current position and that of the previous salp. The leader salp is represented by 1, and the follower salps start numbering from 2.

Algorithm 1 illustrates the process of determining the threshold stage within the methodology. Once the threshold is identified, the fitness function assesses it to determine the optimal threshold value.

Algorithm 1 Salp Swarm Optimization Algorithm for Threshold

```
num_salps ← 30
num_iterations ← 100
min_dist ← minimum distance between salps
max_dist ← maximum distance between salps
salps ← [random(min_dist,max_dist) for _ in range(num_salps)]
for iter in num_iterations do
    best_salp_position ← objective_function(salp,distances)
    for i in salps do
        c1 ← 2 * exp(-(4*( iter / num_iterations )^2))
        c2 ← random values between [0,1]
        c3 ← random values between [0,1]
        ub ← max_dist
        lb ← min_dist
        a ← c1 * ((ub – lb) * c2 + lb)
        if c3 ≥ 0.5 then
            new_salp ← int(best_salp_position + a)
        else
            new_salp ← int(best_salp_position - a)
```

```
        end if
        if new_salp > max_dist then
            salps[i] ← max_dist
        else if new_salp < min_dist then
            salps[i] ← min_dist
        else
            salps[i] ← new_salp
        end if
        score ← objective_function(salps[i],distances)
        if score > global_best_score then
            global_best_score ← score
            global_best_threshold ← salps[i]
        end if
    end for
end for
```

4.3. K-Salp Swarm Anomaly Detection (K-SAD)

Algorithm 2 outlines the complete KSAD algorithm integrated with IKS-LOF (|). The preprocessing of the dataset is described in steps [1-3], where the data is initially scaled, and the upper and lower bounds are removed. Steps [4-9] involve determining the optimal k value based on the silhouette score. In step 11, clusters are identified using the K-medoid algorithm. Steps [12-30] then examine whether distances exceed the salp threshold and whether the LOF score is -1; such instances are classified as anomalies for each cluster. The LOF and Salp anomalies are combined using a logical OR operation. Finally, the combined labels and anomalies, which were previously adjusted using IQR, are integrated. The ROC-AUC score is subsequently computed by comparing the final labels with the true labels.

Algorithm 2 K-medoid Salp Swarm Anomaly Detection Algorithm

Input: Infinite data stream D=$\{x_1, x_2, \ldots, x_n, \ldots\}$ where $\boldsymbol{x_i}$ is a vector of more than one dimension

Output: Anomaly labeled data A={0,0,1,…} (1 indicates anomalies in the data)

```
1   D_scaled ← scale the D
2   D_IQR ← upper and lower anomalies determined using the interquartile range in D
3   D_rest ← D_scaled − D_IQR
4   for k in range[2,30] do
5       kmedoids ← KMedoids(n_clusters=k)
6       if silhouette_avg > best_silhouette_score then
7           best_silhouette_score ← silhouette_avg
8           best_k ← k
9       end if
10  end for
11  Clusters ← K_medoid(best_k, D_rest)
12  for cluster in Clusters do
13      cluster_threshold ← Salp_Swarm_Algorithm(distances)
14      if (distances > cluster_threshold) then
15          salp_label ← 1
16      else
17          salp_label ← 0
18      end if
19      lof_model ← LocalOutlierFactor()
20      lof_scores ← list(lof_model_predict(cluster))
21      if (lof_scores== -1) then
22          lof_label ← 1
23      else
24          lof_label ← 0
25      end if
26  end for
27  combined_label ← logic_or(lof_label,salp_label)
28  final_anomalies ← merge(D_IQR,combined_label)
29  roc_auc_score ← calculate_roc_auc(final_anomalies,true_labels)
```

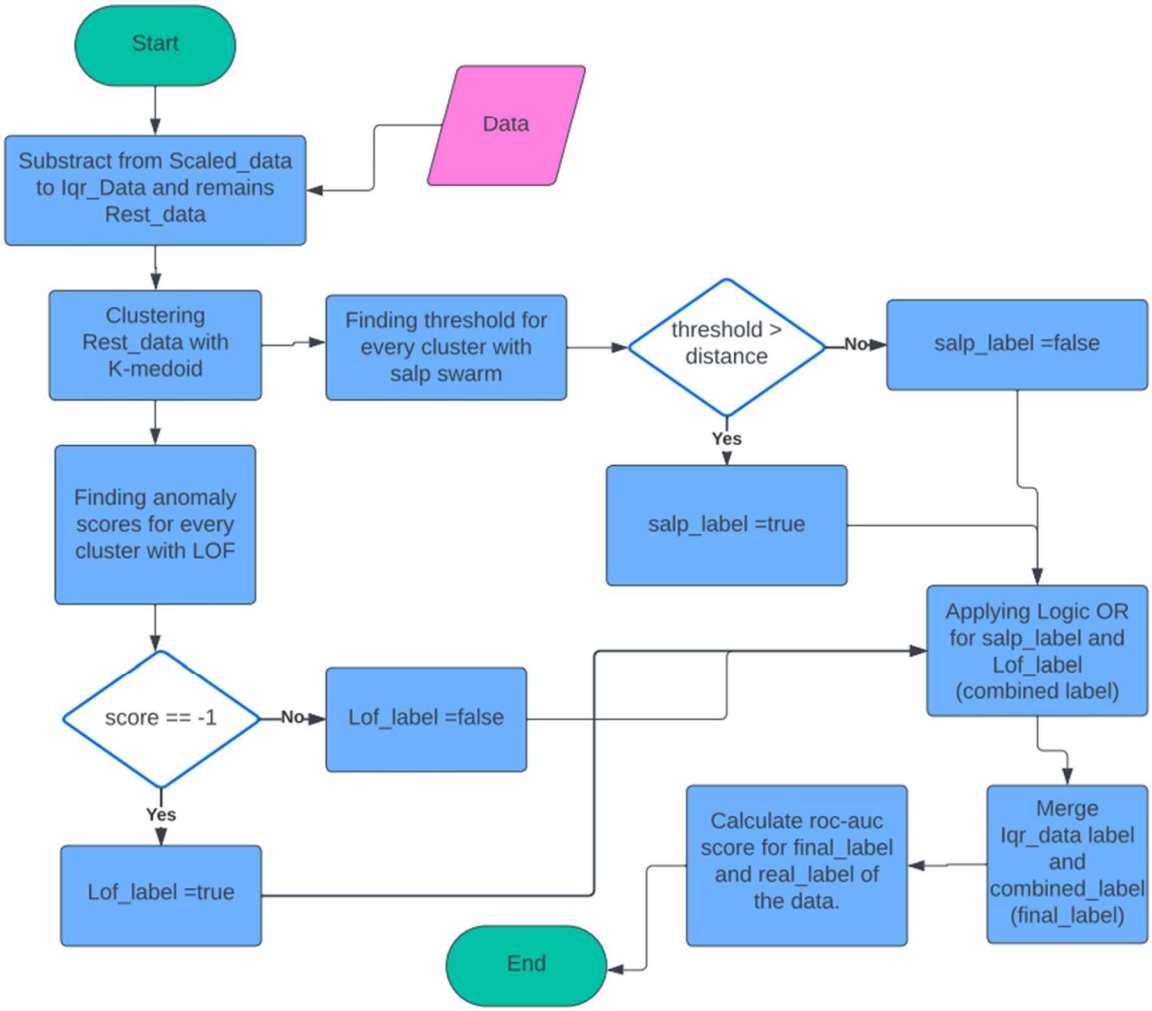

Figure 4.1. K-SAD Flowchart

5. RESULTS AND DISCUSSION

Figure 5.1 illustrates the flowchart of the Suspicious Nodes Anomaly Prevention (SNAP) Framework. Initially, node-based and content-based features are extracted from the Enron dataset. Then, the K-SAD method is applied to identify anomalous and normal nodes for both feature sets. If both methods classify a node as anomalous, it is labeled as an anomaly; if both classify it as normal, it is considered normal. However, if one method identifies the node as anomalous and the other as normal, it is classified as a suspicious node. Next, link prediction is then applied to these suspicious nodes to discover their potential future connections. Subsequently, experts examine connections that surpass a predefined threshold.

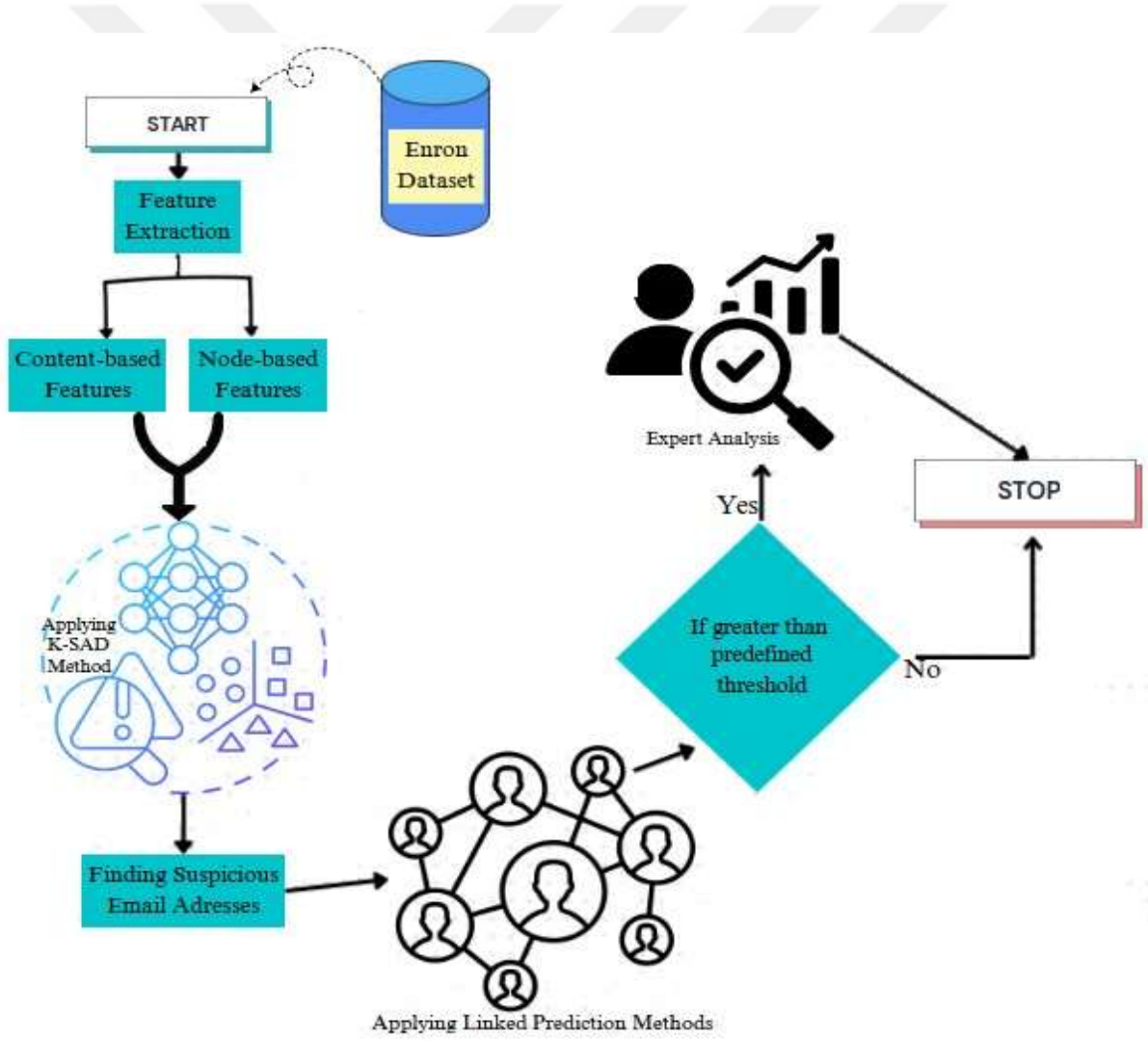

Figure 5.1. SNAP Framework

5.1. Datasets

This study utilizes ten datasets—Musk, Ionosphere, Satimage, Cardio, Thyroid, Annthyroid, Mammography, Pima, Glass, and Vowels—sourced from the UCI repository (Asuncion & Newman, 2007). These datasets are widely used in anomaly detection research. While a previous study by Degirmenci and Karal (2022) employed 16 datasets from the same repository, this study selected only 10 due to inconsistencies in the number of rows and features. Only datasets aligned precisely with the study's parameters are included to ensure data integrity and a fair comparison.

Furthermore, these datasets play a significant role in anomaly detection research, having been extensively utilized and validated within the academic community for evaluating anomaly detection methods. The use of the ODDS library (Rayana, 2016) further highlights their relevance in benchmarking anomaly detection models.

Finally, the recently developed Suspicious Nodes Anomaly Prevention (SNAP) framework utilizes the Enron dataset. The Enron corpus, introduced by Klimt and Yang (2004), contains 619,446 emails from 158 users. After preprocessing, which involved cleaning and filtering, the dataset was reduced to 200,399 messages. This corpus was selected for its large-scale, diversity, and authenticity, reflecting real-world communication patterns within a major corporation. It is particularly useful for extracting content-based and node-based features, which are crucial for anomaly detection and analysis. With 30,091 identified threads, the dataset offers a rich ground for examining communication sequences and thread dynamics, where 61.63% of the emails are part of threads, typically consisting of small-sized exchanges. This large and varied dataset provides valuable insights for modeling and understanding email interactions, making it an ideal choice for exploring advanced anomaly detection techniques.

5.2. Performance Criteria

This section outlines the performance criteria metrics, including accuracy, F1-score, and ROC-AUC score, highlighting their importance in evaluating the proposed framework's effectiveness.

5.2.1. Accuracy

Accuracy (Witten et al, 2005) is a basic evaluation metric calculated using the confusion matrix. It represents the ratio of correctly classified instances (both true positives and true negatives) to the total number of instances in the dataset. The formula for accuracy is shown in Equation 5.1.

$$Accuracy = \frac{True\ Positive\ (TP) + Tru\quad Negative\ (TN)}{TP + TN + False\ Positive\ (FP) + Fa\quad\ Negative\ (FN)} \tag{5.1}$$

5.2.2. F1-Score

F1 score (Powers, 2020) is the harmonic mean of precision and recall, providing a balance between these two metrics. Precision, also known as positive predictive value, measures the proportion of true positives out of all predicted positives, while recall, also called sensitivity, measures the proportion of true positives identified from the actual positives. F1 score is a crucial metric, especially when there is an imbalanced class distribution, as it captures the balance between Precision and Recall more effectively.

5.2.3. ROC-AUC Score

The Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) score (Metz, 1978) is utilized to assess a model's effectiveness in anomaly detection. This metric evaluates overall performance by simultaneously accounting for both sensitivity and specificity.

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{5.2}$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive} \tag{5.3}$$

The True Positive Rate (TPR) indicates the proportion of true anomalous samples correctly detected. In other words, it expresses how successfully the model identifies abnormal samples. TPR also represents sensitivity. The False Positive Rate (FPR) measures the proportion of normal samples incorrectly classified as abnormal. In other words, it refers to normal data

mistakenly flagged as abnormal. FPR is the opposite of specificity. The ROC curve illustrates the change in TPR versus FPR. AUC represents the area under the ROC curve.

A high ROC-AUC score indicates that the model successfully detects abnormal samples and has a low rate of misclassifying normal samples as abnormal.

5.3. Performance Comparison

The proposed method has been evaluated using multiple approaches, with the core methodology centered on the IKS. Each tested approach follows three main steps: (1) anomalies are identified by separating the upper and lower parts of the dataset using the IQR, (2) clustering is performed using the K-medoids algorithm, and (3) a threshold is determined through the salp swarm optimization technique.

Subsequently, after determining a threshold for each cluster and labeling anomalies, Local Outlier Factor (LOF) scores are computed for the same clusters to identify additional anomalies. When these two sets of anomaly labels are combined using a logical OR operation, the result is denoted as IKS-LOF (|), whereas a logical AND operation produces IKS-LOF (&).

To obtain the "just one" result, anomaly labels are generated using IKS, along with separate anomaly and normal labels from LOF and IF for the entire dataset. In this approach, if any one of the three methods identifies a data point as an anomaly, it is classified as an anomaly in the final label. Conversely, for the "majority" result, at least two of the three methods—IKS, LOF, and IF—must detect an anomaly for the data point to be classified as such.

The use of multiple anomaly detection techniques serves two main purposes: enhancing overall detection performance and optimizing the proposed method for improved accuracy and reliability.

Table 5.1. Without Scaling Results

Dataset	IKS	IKS-LOF (|)	IKS-LOF (&)	just one	majority
Musk	0.5309	0.4951	0.50	**0.9654**	0.5455

Ionosphere	0.5987	0.6481	0.5833	**0.6568**	0.5476
Satimage	0.9382	0.9394	0.9386	**0.9411**	0.9388
Cardio	0.6876	0.6752	**0.6879**	0.6595	0.6425
Thyroid	0.6593	0.6476	0.6783	0.6541	0.8537
Annthyroid	0.3867	0.4881	**0.5282**	0.4852	0.5183
Mammography	0.7335	**0.7603**	0.7412	0.7167	0.6521
Pima	0.5482	0.5394	0.5464	0.5623	0.5183
Glass	**0.7499**	0.7157	0.6301	0.7279	0.5312
Vowels	0.5725	**0.8683**	0.5743	0.5706	0.5529

Table 5.1 highlights that the "just one" approach, applied without scaling, demonstrates the strongest performance. In other words, only one of IKS, LOF, and IF is sufficient to identify the data as an anomaly. This means that they all find different anomalies, and since different methods evaluate the data from different angles, they can all detect different anomalies. We can conclude that combining several different anomaly detection methods with a "just one" perspective enhances performance.

Table 5.2. With Scaling Results

Dataset	IKS	IKS-LOF (|)	IKS-LOF (&)	just one	majority
Musk	0.9958	0.9465	**0.9987**	0.9648	0.9978
Ionosphere	0.5119	**0.6367**	0.5119	0.5978	0.5198
Satimage	0.9460	0.9400	0.9460	0.9346	**0.9664**
Cardio	0.6876	0.6752	**0.6879**	0.6589	0.6315
Thyroid	0.8605	**0.8651**	0.8622	0.8540	**0.8651**
Annthyroid	0.6008	**0.6811**	0.6024	0.6040	0.6127
Mammography	0.7335	**0.7603**	0.7412	0.7167	0.6225
Pima	0.5396	**0.5457**	0.5265	0.5400	0.5233
Glass	0.4775	**0.6705**	0.5190	0.4531	0.5336
Vowels	0.5725	**0.8663**	0.5743	0.5716	0.5622

Table 5.2 demonstrates that while scaling enhances performance in some cases, its effectiveness varies across datasets. Specifically, scaling negatively impacted the results for the Glass and Ionosphere datasets, whereas it had no effect on the Cardio and Mammography datasets. Additionally, after applying scaling, IKS-LOF (|) achieved superior performance compared to other approaches. Combining IKS with LOF and the 'OR' command led to better results. Again, as in Table 5.1, since the methods have different perspectives, they find different anomalies, so the 'OR' command further improved the result.

Table 5.3. Comparison with Other Methods

Dataset	K-SAD	Scaling	Comparison
Musk	**0.9987**	y	Almost same
Ionosphere	0.6568	n	below average
Satimage	**0.9664**	y	best except LODA
Cardio	0.6879	both	above average
Thyroid	**0.8651**	y	best
Annthyroid	0.6811	y	below average
Mammography	**0.7603**	both	best except LODA
Pima	0.5623	n	above average
Glass	**0.7499**	n	best
Vowels	0.8683	n	above average

In Table 5.3, a comparison with the findings of Degirmenci and Karal (2022) reveals that the best or near-best results were achieved in five datasets. In two datasets, the performance was below average among the ten methods, while in three datasets, it was above average. Overall, eight datasets produced acceptable results, whereas only two fell below average. Additionally, the best outcomes were obtained with scaling for four datasets and without scaling for another four datasets, while scaling had no noticeable impact on two datasets. These findings indicate that scaling does not consistently enhance performance across all datasets.

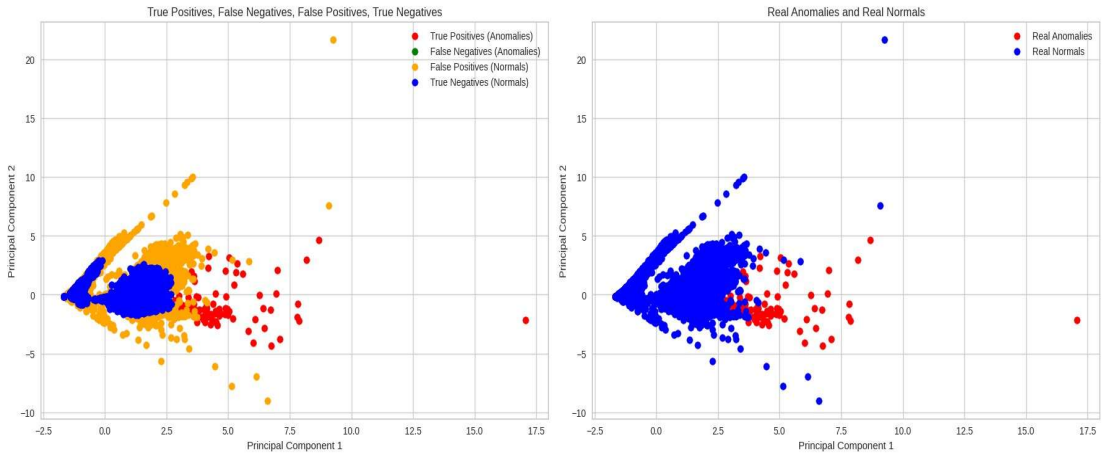

Figure 5.2. Mammography Result

On the right side of Figure 5.2, true normals are shown in blue, and true abnormals are shown in red. On the left side, blue, yellow, green, and red indicate true negatives, false positives, false negatives, and true positives, respectively. There were no false negatives in the Mammography dataset. In other words, no false anomalies were detected. However, some normal instances were detected as anomalies.

5.4. Statistical Test Analysis

Two non-parametric tests were performed, both using a significance level of 0.05. The Friedman test was first applied to determine whether there were notable differences among the methods. The test result was 2.55, with a p-value of 0.9795, indicating that no significant difference was observed, and the null hypothesis—stating that no method is significantly different from the others—was accepted.

Additionally, the Nemenyi test was conducted to analyze pairwise comparisons among the methods, leading to the results presented in Figure 5.3. The findings indicate that ILDCBOF exhibited significant differences with some methods but not with iLOF, LODA, LDOF, or K-SAD. This supports the study's argument that no single method is universally superior in anomaly detection.

In addition, the critical difference was the closest with a K-SAD p-value of 0.07 with SO-GAAL and 0.1 with MO-GAAL. An almost significant difference will occur between K-SAD & SO-GAAL and K-SAD & MO-GAAL. None of the other methods had such a significant difference.

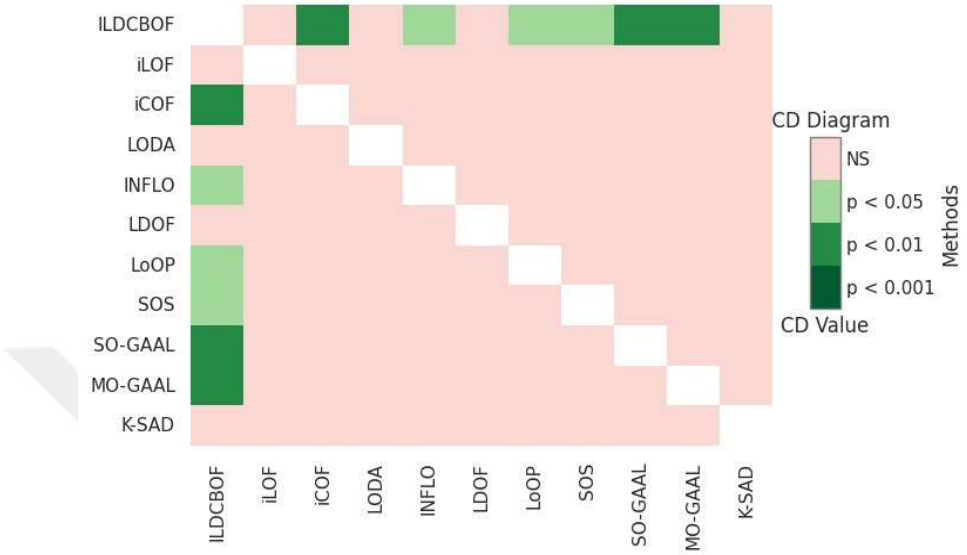

Figure 5.3. Critical Difference (CD) Diagram

5.5. Results

The Enron dataset includes multiple folders, which can lead to the repetition of the same emails across different folders. For instance, emails were found in both the sent and received folders. To simplify the analysis, we utilized only the sent folder, aligning with practices in previous studies that also focused solely on this folder (Styler, 2011).

A NetworkX graph was constructed from the "from-to" relationships in the sent emails. This graph initially contained 15,868 nodes and 32,433 edges. However, some nodes were isolated, resulting in five distinct connection graphs with sizes of 15,852, 10, 2, 2, and 2. The four smaller graphs were excluded, and the associated email addresses were removed from the analysis. The excluded email addresses were: rjbaker@ttu.edu, lcampbe, tim.derrick, patricia.hunter, tom.nemila, denise.williams, mark.fisher, andre.rast, kurt.anderson, jeff.duff, joe.chapman, julie.johnson, all.states, click.home, kid.mailing, and enron.announcements. Email addresses without a domain were assumed to have the domain @enron.com. The graph's division into five distinct components can be seen in Figure 5.4.

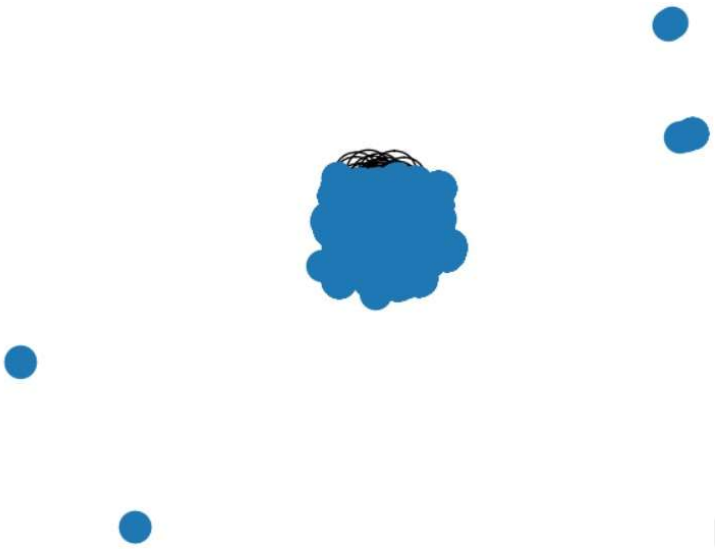

Figure 5.4. Enron Graph Visualization

Based on the NetworkX graph, several node-based features were extracted, including degree centrality, eigenvector centrality, closeness centrality, betweenness centrality, clustering coefficient, PageRank, HITS (hubs and authorities), eccentricity, community, degree, average common neighbors, average Jaccard coefficient, average Adamic-Adar score, preferential attachment, and effective size.

We also used partially labeled Enron nodes, where another work (Noever, 2020) was marked as 'persons of interest' (POIs). By comparing the labels generated through our methods with the actual POIs (representing anomaly emails), we achieved a ROC-AUC score of 0.6781. Out of 18 anomalies, our method misclassified only one instance as normal. The confusion matrix displaying these results for the node-based features can be seen in Figure 5.5.

Although our method performed well in identifying anomalies, it tended to misclassify normal instances as anomalies. This is an area for improvement in future work, focusing on reducing false positives and improving the accuracy of normal data classification.

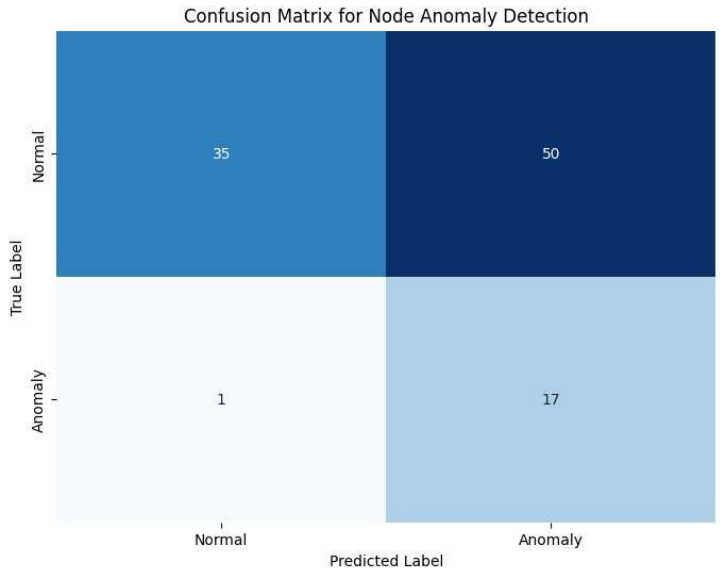

Figure 5.5. Node-based Anomaly Detection Confusion Matrix

Similarly, by applying TF-IDF on the email content and removing duplicates, the dataset was reduced to 96,148 emails. The proposed method was then applied to identify whether these connections were anomalies. As shown in the confusion matrix in Figure 5.6, while the method demonstrated a high success rate in detecting anomalies, its performance in identifying normal instances was slightly lower.

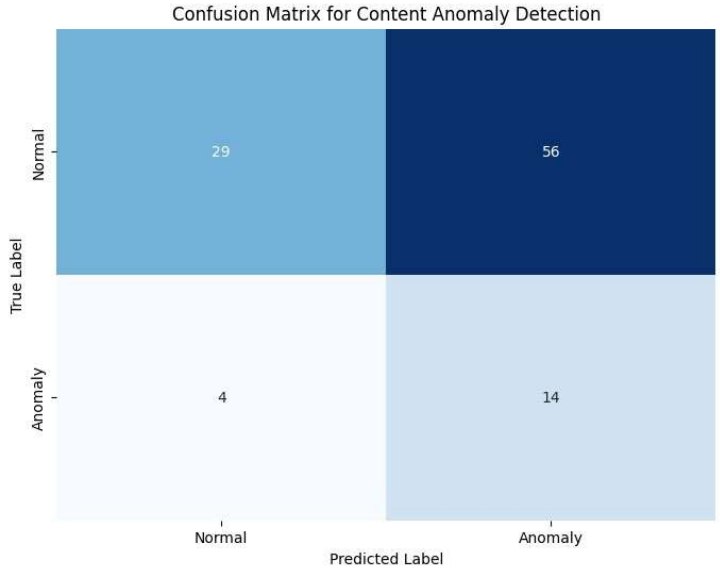

Figure 5.6. Content-based Anomaly Detection Confusion Matrix

Accuracy, F1-score, and ROC-AUC scores for both node-based and content-based features are presented in Table 5.4. Additionally, as shown in the ROC curve in Figure 5.7, it is evident that anomalies detected using node-based features yield better results compared to those detected using content-based features.

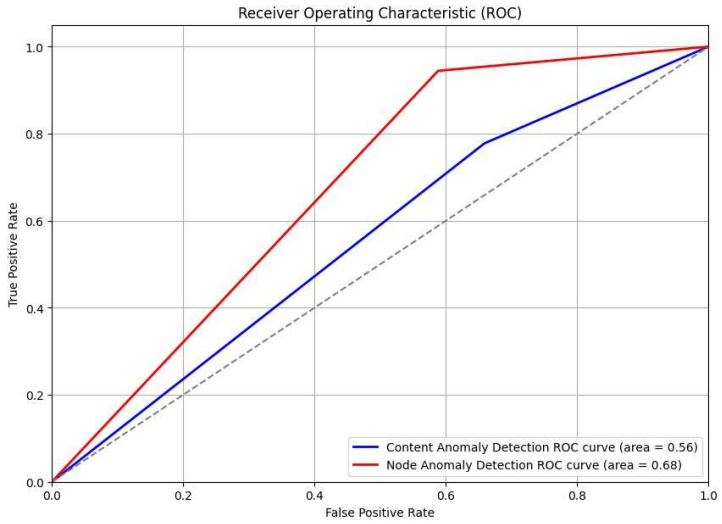

Figure 5.7. ROC Curve Results for Content and Node Anomaly Detections

Table 5.4. Content and Node Results

Method	Accuracy	F1-score	ROC-AUC
CAD	0.4175	0.4612	0.5595
NAD	0.5049	0.5473	0.6781

In the analysis of labeled data, 56 instances were classified as anomalies by both node-based and content-based features. Among the POI labels, there were 18 anomalies, with the remaining being normal email addresses. Both node-based and content-based methods identified 38 additional anomalies compared to the POI labels. In both methods, 22 nodes were classified as normal, while in the POI labels, only one was considered anomalous, with the rest being normal. Instances classified as anomalies by one method but normal by the other, referred to as suspicious email addresses, are detailed in Table 5.5.

Table 5.5. Suspicious Nodes Content, Node, and Real Labeled Data

Email Address	contentLabel	nodeLabel	realLabel
kenneth.lay	0	1	1
richard.shapiro	0	1	0
andrew.fastow	0	1	1
bill.cordes	1	0	0
david.haug	0	1	0
diomedes.christodoulou	1	0	0
gene.humphrey	1	0	0
james.bannantine	1	0	0
joe.hirko	0	1	1
joe.kishkill	0	1	0
john.buchanan	1	0	0
ken.powers	1	0	0
lou.pai	0	1	0
mark.pickering	1	0	0
marty.sunde	0	1	0
matthew.scrimshaw	1	0	0
michael.moran	1	0	0
mitchell.taylor	1	0	0
rebecca.mcdonald	0	1	0
richard.lewis	1	0	0
rick.bergsieker	1	0	0
robert.hayes	1	0	0
tod.lindholm	0	1	0
tracy.foy	1	0	0
w.duran	0	1	0

Table 5.6 shows the contacts of suspicious nodes with all nodes, only suspicious nodes, and anomaly nodes. The summary title provides an overview of the email content, and the details of the risky instances are shown in Table 5.8.

Table 5.6. Suspicious Nodes Contacts with All, Only, and Anomalies

From	To	Summary Title	Email Chain	Between
john. buchanan	terry. kowalke	TMS Small Talk Upgrade Testing: Timing and Assistance Request	Yes	All
diomedes. christodoulou	thomas. white	Private Flight Request and Scheduling Communication	No	All
diomedes. christodoulou	j.metts	Project Summer Transfer Updates and Staff Changes	Yes	All
diomedes. christodoulou	sanjay. bhatnagar	Preparation for Media Inquiries on Potential Dabhol Sale	No	All
rebecca. mcdonald	andrew. fastow	Delay in DASH signing; Need Fastow's signature	Yes	Only
james. bannantine	diomedes. christodoulou	Government reaction and regulatory considerations	Yes	Only
kevin. hannon	rebecca. mcdonald	System Testing Update and Gas Logistics Organization	No	Anomalies
ken.rice	david.w. delainey	Building Lease Opportunity for Car Storage	Yes	Anomalies
andrew. fastow	rebecca. mcdonald	Harrah's DASH Approval Issue	Yes	Anomalies
andrew. fastow	david. haug	Harrah's DASH Approval Issue	Yes	Anomalies
david. haug	richard. causey	Internal Feedback Forms Completed	No	Anomalies
andrew. fastow	mark. koenig	Mid-Day Info on EOL Trades and Counterparty Recap	No	Anomalies
andrew. fastow	kevin. hannon	System Testing Update, Global Focus, and Gas Logistics	No	Anomalies

Table 5.7. shows the risky connections identified through link prediction for suspicious nodes. The 'Between' column presents connections between suspicious nodes and all nodes in the dataset, while the 'Only' column shows connections exclusively between suspicious nodes. The 'Anomalies' column highlights connections between suspicious nodes and anomalies. The calculations were performed using four link prediction methods: CN, JC, AA, and PA, with specific thresholds for each method. A threshold of 0.7 was applied for 'All' connections, while the other cases used a threshold of 0.5. The higher threshold for 'All' connections was necessary,

as no connections were identified below this level. By setting these thresholds, the link prediction process ensured that the connections predicted by all four methods were reliable, thereby improving the accuracy of the results.

Table 5.7. Potential Contacts of Suspicious Nodes

From	To	Threshold	Email Chain	Between	Num
diomedes. christodoulou	sanjay. bhatnagar	0.7	No	All	1
james. bannantine	diomedes. christodoulou	0.5	Yes	Only	2
ken.rice	david.w. delaniey	0.5	Yes	Anomalies	3
andrew.fastow	mark.koenig	0.5	No	Anomalies	4

Table 5.8 presents the email contents. Initially, more connections were detected: four for 'All,' three for 'Only,' and nine for 'Anomalies.' However, upon detailed review, specific emails were labeled as anomalies based on their content. Given the context of corruption in the Enron dataset, these emails were labeled due to their content related to financial and governmental matters. For example, the first email discusses discounts and media but subtly suggests moving to phone conversations to avoid leaving a record. The second email appears to involve governmental approval processes, hinting at efforts to expedite acceptance through influential individuals. The third email extends congratulations for a workplace role while proposing an offer that could be interpreted as a bribe. The fourth email raises concerns about low credit ratings and suggests the likelihood of engaging in anomalous activities.

Table 5.8. Contents of Suspicious Nodes Potential Contacts

Num	Suspicious Mail Content
1	Joe asked me to get in touch with you regarding Indian media reports/inquiries on possible sale or sell down of Dabhol and to discuss a possible statement in the event we get additional inquiries. Give me a call - 713-853-1586.
2	James emphasizes the need for early action to avoid delays in the governmental approval stage of Project California. He notes that Jose, with his expertise and local knowledge, is best suited to handle government-related assessments and regulatory approvals. James recommends including Jose in the process to effectively manage governmental interactions and prevent potential delays.

3	Ken Rice congratulates David Delainey on his new role and offers a building in Bellaire for lease or purchase. The space is suggested for car storage, with a rental cost of $12-$15 per square foot annually. Leasing one-third of the building would cost about $1,600 per month, potentially less than a climate-controlled storage unit. Ken invites David to discuss further if interested.
4	A mid-day summary compares the Enron Online (EOL) trades on October 22, 2001, with the full-day transaction counts from September 21, 2001. A detailed recap can be provided at the end of the day if needed. The counterparties listed have higher credit ratings than Enron. It is suggested to contact a relevant person for any credit-related questions.

5.6. Discussions

This study proposed and evaluated a framework for anomaly prevention using a real-world dataset. A novel nature-inspired anomaly detection method was developed, utilizing node and content features. While nature-inspired algorithms are typically employed for clustering tasks in the literature, their use here for threshold estimation represents a key innovation. Anomalies, by definition, are data points that do not fit into any predefined group. Therefore, by first clustering the data and then accurately identifying an optimal threshold, it becomes possible to label data points falling outside this threshold as anomalies. However, the threshold used in our current implementation might be too strict, leading to more false positives, where non-anomalous data points are mistakenly classified as anomalies. Future research will focus on refining the clustering process and optimizing the threshold to enhance the accuracy of anomaly detection.

Table 5.9. Characteristics of Datasets

Dataset	Samples	Features	Anomaly Ratio (%)	PCA Explained Variance (%)
Mammography	11183	6	2,32	61,32
Annthyroid	7200	6	7,42	97,11
Thyroid	3772	6	2,47	82,96
Glass	214	9	4,21	73,94
Pima	768	8	34,9	95,01
Vowels	1456	12	3,43	54,12

Cardio	1831	21	9,61	44,56
Ionosphere	351	33	35,9	43,62
Satimage-2	5803	36	1,22	84,83
Musk	3062	166	3,17	57,06

Additionally, the effectiveness of the newly proposed method, K-SAD, was discussed across different datasets. The characteristics of the datasets are presented in Table 5.9, while the results of our method's performance across these datasets are detailed in Table 5.3. Based on the analysis of the dataset characteristics and K-SAD performance, our method achieved the best results on the Musk, Satimage-2, Thyroid, Mammography, and Glass datasets. These datasets generally have lower anomaly ratios (ranging from approximately 1.22% to 4.21%) and, in some cases, higher PCA-explained variance percentages (e.g., Mammography at 61.32% and Satimage-2 at 84.83%), which appear to facilitate more effective anomaly detection. In contrast, the Ionosphere and Annthyroid datasets produced the poorest outcomes, likely due to their higher anomaly ratios (35.9% for Ionosphere and 7.42% for Annthyroid) combined with lower PCA-explained variance in Ionosphere (43.62%), indicating that a high proportion of anomalies or less distinctive structural features may challenge the method's effectiveness. Datasets such as Vowels, Pima, and Cardio yielded moderate results, suggesting that while these datasets possess some favorable characteristics, such as moderate anomaly ratios and acceptable variance retention, they do not align optimally with the strengths of our approach. Overall, these findings underscore the sensitivity of the K-SAD method to inherent dataset properties, highlighting its robustness in environments with low anomaly prevalence and well-defined data structures and its limitations when confronted with high anomaly rates or less discriminative feature representations.

The classical TF-IDF method was employed for its efficiency and widespread acceptance as one of the best traditional methods for content analysis. However, with the advent of more semantically powerful methods like BERT, TF-IDF may appear limited. Nonetheless, the speed of TF-IDF was critical due to the large volume of emails in the dataset. While newer methods like BERT offer superior semantic understanding, they are often slower in processing large datasets. By applying the same anomaly detection approach to both node and content features,

instances labeled as anomalies by both methods were categorized as confirmed anomalies. In contrast, those marked as normal by both methods were considered normal. The main focus of our study, however, was on those instances where one method identified an anomaly while the other did not. These were labeled as suspicious nodes.

Further investigation was conducted into the potential future connections of these suspicious nodes. Since their potential connections may carry anomaly risks, these were flagged for further analysis. This approach allows for preventative measures to be taken before suspicious activities manifest into significant anomalies. For instance, financial-related emails marked as suspicious within the Enron dataset could be presented to experts for further scrutiny, helping identify individuals involved in unethical activities more clearly. While some high-level executives have already been implicated, there may still be hidden offenders whose activities could be uncovered through this framework.

6. CONCLUSION

This study presents a significant advancement in anomaly detection and prevention. The proposed K-SAD method exhibits robust efficacy in anomaly detection, outperforming many existing techniques across various datasets. Despite its success, the method's performance varies depending on the dataset, emphasizing anomaly detection tasks' inherent complexity and variability. This variation highlights the impracticality of achieving universal optimization with a single approach, underscoring the need for customized methodologies that address the specific characteristics of each dataset. Anomaly detection remains intricately linked to data specifics, necessitating the development of bespoke solutions for each unique dataset.

Nature-inspired algorithms, commonly used in clustering and feature selection, have been effectively adapted in this study for threshold detection, representing a novel approach to anomaly detection. By employing a fitness function aligned with the characteristics of anomaly data, the study successfully advances the field of anomaly detection methodologies. However, the study encountered limitations in optimizing results independent of dataset characteristics. Future research will refine parameter selection, such as center and threshold values, to optimize results for specific datasets.

In addition, the study introduces a new framework for anomaly prevention, utilizing a novel anomaly detection technique based on clustering and nature-inspired algorithms. This framework employs email data as a social network dataset, enabling node- and content-based analyses. We classify instances identified as anomalies by one method but normal by another as suspicious nodes and analyze their connections to detect potential anomalies. This proactive approach prevents anomalous activities by identifying risks before they materialize. Unlike previous studies, this research emphasizes anomaly prevention rather than mere detection.

Future work should explore advanced clustering techniques to improve anomaly identification accuracy and incorporate temporal data for link prediction. Such advancements could enhance the framework's ability to prevent anomalous activities and provide a more proactive approach to anomaly prevention.

REFERENCES

Akoglu, L., McGlohon, M., & Faloutsos, C. (2010, June). Oddball: Spotting anomalies in weighted graphs. In Pacific-Asia conference on knowledge discovery and data mining (pp. 410-421). Springer, Berlin, Heidelberg.

Singh, K. V., & Vig, L. (2017). Improved prediction of missing protein interactome links via anomaly detection. Applied Network Science, 2(1), 1-20.

Karakaşlı, M. S., Aydin, M. A., Yarkan, S., & Boyaci, A. (2019). Dynamic feature selection for spam detection in Twitter. In International Telecommunications Conference (pp. 239-250). Springer, Singapore.

Zhang, T., & Liao, Q. (2019). Dynamic link anomaly analysis for network security management. Journal of Network and Systems Management, 27(3), 600-624.

Huang, Z., & Zeng, D. D. (2006, October). A link prediction approach to anomalous email detection. In 2006 IEEE international conference on systems, man and cybernetics (Vol. 2, pp. 1131-1136). IEEE.

Kagan, D., Elovichi, Y., & Fire, M. (2018). Generic anomalous vertices detection utilizing a link prediction algorithm. Social Network Analysis and Mining, 8(1), 1-13.

Rahman, M. S., Halder, S., Uddin, M. A., & Acharjee, U. K. (2021). An efficient hybrid system for anomaly detection in social networks. Cybersecurity, 4(1), 1-11.

Degirmenci, A., & Karal, O. (2022). Efficient density and cluster based incremental outlier detection in data streams. Information Sciences, 607, 901-920.

Xu, H., Zhang, L., Li, P., & Zhu, F. (2022). Outlier detection algorithm based on k-nearest neighbors-local outlier factor. Journal of Algorithms & Computational Technology, 16, 17483026221078111.

Xu, H., Pang, G., Wang, Y., & Wang, Y. (2023). Deep isolation forest for anomaly detection. IEEE Transactions on Knowledge and Data Engineering, 35(12), 12591-12604.

Min, H., Lei, X., Wu, X., Fang, Y., Chen, S., Wang, W., & Zhao, X. (2024). Toward interpretable anomaly detection for autonomous vehicles with denoising variational transformer. Engineering Applications of Artificial Intelligence, 129, 107601.

Wu, X. (2019). A trust-based detection scheme to explore anomaly prevention in social networks. Knowledge and Information Systems, 60(3), 1565-1586.

Van Vlasselaer, V., Akoglu, L., Eliassi-Rad, T., Snoeck, M., & Baesens, B. (2015, January). Guilt-by-constellation: Fraud detection by suspicious clique memberships. In 2015 48th Hawaii International Conference on System Sciences (pp. 918-927). IEEE.

Kirchner, C., & Gade, J. (2011). Implementing social network analysis for fraud prevention. CGI Gr. Ind.

Cormack, G. V., & Lynam, T. R. (2005, July). Spam Corpus Creation for TREC. In CEAS.

Wilson, G., & Banzhaf, W. (2009, May). Discovery of email communication networks from the enron corpus with a genetic algorithm using social network analysis. In 2009 IEEE Congress on Evolutionary Computation (pp. 3256-3263). IEEE.

Corneli, M., Bouveyron, C., Latouche, P., & Rossi, F. (2019). The dynamic stochastic topic block model for dynamic networks with textual edges. Statistics and Computing, 29(4), 677-695.

Zhang, L., Zhou, T., Zhixin, Q., Guo, L., & Xu, L. (2016). The research on e-mail Users' behavior of participating in Subjects based on social network analysis. China Communications, 13(4), 70-80.

Apoorva, K. A., & Sangeetha, S. (2021). Deep neural network and model-based clustering technique for forensic electronic mail author attribution. SN Applied Sciences, 3(3), 1-12.

Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., Fidalgo, E., & Alegre, E. (2023). A review of spam email detection: analysis of spammer strategies and the dataset shift problem. Artificial Intelligence Review, 56(2), 1145-1173.

Bountakas, P., & Xenakis, C. (2023). Helphed: Hybrid ensemble learning phishing email detection. Journal of network and computer applications, 210, 103545.

Poobalan, A., Ganapriya, K., Kalaivani, K., & Parthiban, K. (2025). A novel and secured email classification using deep neural network with bidirectional long short-term memory. Computer Speech & Language, 89, 101667.

Bastami, E., Mahabadi, A., & Taghizadeh, E. (2019). A gravitation-based link prediction approach in social networks. Swarm and evolutionary computation, 44, 176-186.

Saurabh, P., & Verma, B. (2016). An efficient proactive artificial immune system based anomaly detection and prevention system. Expert Systems with Applications, 60, 311-320.

Alsaleh, A., & Binsaeedan, W. (2021). The influence of salp swarm algorithm-based feature selection on network anomaly intrusion detection. IEEE Access, 9, 112466-112477.

Hajisalem, V., & Babaie, S. (2018). A hybrid intrusion detection system based on ABC-AFS algorithm for misuse and anomaly detection. Computer Networks, 136, 37-50.

Aswani, R., Ghrera, S. P., Kar, A. K., & Chandra, S. (2017). Identifying buzz in social media: a hybrid approach using artificial bee colony and k-nearest neighbors for outlier detection. Social Network Analysis and Mining, 7(1), 1-10.

Alzaqebah, A., Aljarah, I., & Al-Kadi, O. (2023). A hierarchical intrusion detection system based on extreme learning machine and nature-inspired optimization. Computers & Security, 124, 102957.

Khayyat, M. M. (2023). Improved bacterial foraging optimization with deep learning based anomaly detection in smart cities. Alexandria Engineering Journal, 75, 407-417.

Shao, C., Zheng, S., Gu, C., Hu, Y., & Qin, X. (2022). A novel outlier detection method for monitoring data in dam engineering. Expert Systems with Applications, 193, 116476.

Masrour, F., Wilson, T., Yan, H., Tan, P. N., & Esfahanian, A. (2020, April). Bursting the filter bubble: Fairness-aware network link prediction. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 01, pp. 841-848).

Aslan, S., & Kaya, M. (2018). Topic recommendation for authors as a link prediction problem. Future Generation Computer Systems, 89, 249-264.

Jiang, H., Liu, Z., Liu, C., Su, Y., & Zhang, X. (2020). Community detection in complex networks with an ambiguous structure using central node based link prediction. Knowledge-Based Systems, 195, 105626.

Su, Z., Zheng, X., Ai, J., Shen, Y., & Zhang, X. (2020). Link prediction in recommender systems based on vector similarity. Physica A: Statistical Mechanics and its Applications, 560, 125154.

Tuan, T. M., Chuan, P. M., Ali, M., Ngan, T. T., Mittal, M., & Son, L. H. (2019). Fuzzy and neutrosophic modeling for link prediction in social networks. Evolving Systems, 10(4), 629-634.

Girdhar, N., Minz, S., & Bharadwaj, K. K. (2019). Link prediction in signed social networks based on fuzzy computational model of trust and distrust. Soft Computing, 23(22), 12123-12138.

Ai, J., Su, Z., Li, Y., & Wu, C. (2019). Link prediction based on a spatial distribution model with fuzzy link importance. Physica A: Statistical Mechanics and its Applications, 527, 121155.

Samad, A., Azam, M., & Qadir, M. (2021). Structural Importance-based Link Prediction Techniques in Social Network. EAI Endorsed Transactions on Industrial Networks and Intelligent Systems, 7(25).

Ott, S., Meilicke, C., & Samwald, M. SAFRAN: An interpretable, rule-based link prediction method outperforming embedding models. arXiv 2021. arXiv preprint arXiv:2109.08002.

Saxena, A., Fletcher, G., & Pechenizkiy, M. (2022). HM-EIICT: Fairness-aware link prediction in complex networks using community information. Journal of Combinatorial Optimization, 44(4), 2853-2870.

Assouli, N., Benahmed, K., & Gasbaoui, B. (2021). How to predict crime—informatics-inspired approach from link prediction. Physica A: Statistical Mechanics and its Applications, 570, 125795.

Lande, D., Fu, M., Guo, W., Balagura, I., Gorbov, I., & Yang, H. (2020). Link prediction of scientific collaboration networks based on information retrieval. World Wide Web, 23, 2239-2257.

Zhou, T. (2023). Discriminating abilities of threshold-free evaluation metrics in link prediction. Physica A: Statistical Mechanics and its Applications, 615, 128529.

Jiao, X., Wan, S., Liu, Q., Bi, Y., Lee, Y. L., Xu, E., ... & Zhou, T. (2024). Comparing discriminating abilities of evaluation metrics in link prediction. Journal of Physics: Complexity, 5(2), 025014.

Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis. John Wiley & Sons.

Wan, X., Wang, W., Liu, J., & Tong, T. (2014). Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. BMC medical research methodology, 14, 1-13.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

Danielsson, P. E. (1980). Euclidean distance mapping. Computer Graphics and image processing, 14(3), 227-248.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (pp. 93-104).

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In 2008 eighth ieee international conference on data mining (pp. 413-422). IEEE.

Liben-Nowell, D., & Kleinberg, J. (2003, November). The link prediction problem for social networks. In Proceedings of the twelfth international conference on Information and knowledge management (pp. 556-559).

Jackson, D. A., Somers, K. M., & Harvey, H. H. (1989). Similarity coefficients: measures of co-occurrence and association or simply measures of occurrence?. The American Naturalist, 133(3), 436-453.

Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. Social networks, 25(3), 211-230.

Newman, M. E. (2001). Clustering and preferential attachment in growing networks. Physical review E, 64(2), 025102.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 28(1), 11-21.

Mirjalili, S., Gandomi, A. H., Mirjalili, S. Z., Saremi, S., Faris, H., & Mirjalili, S. M. (2017). Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems. Advances in engineering software, 114, 163-191.

Asuncion, A., & Newman, D. (2007). UCI machine learning repository.

Rayana, S. (2016). ODDS library. Stony Brook University, Department of Computer Sciences.

Klimt, B., & Yang, Y. (2004, July). Introducing the Enron corpus. In CEAS.

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & Data, M. (2005, June). Practical machine learning tools and techniques. In Data mining (Vol. 2, No. 4, pp. 403-413). Amsterdam, The Netherlands: Elsevier.

Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061.

Metz, C. E. (1978, October). Basic principles of ROC analysis. In Seminars in nuclear medicine (Vol. 8, No. 4, pp. 283-298). WB Saunders.

Styler, W. (2011). The enronsent corpus. University of Colorado-Boulder, 1-7.

Noever, D. (2020). The Enron Corpus: Where the Email Bodies are Buried?. arXiv preprint arXiv:2001.10374.

ATTACHMENTS

Attachment A: CD of Computer Program