LINK ANALYSIS ON SOCIAL ENGINEERING

A Thesis

Submitted in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY
In
Computer Science and Engineering

Submitted by

POOJA MITHOO
(2K19/PHDCO/501)

Under the supervision

of

Prof. Manoj Kumar

Department of Computer Science and Engineering

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering) Bawana Road, Delhi 110042

2025

CERTIFICATE

Certified that Pooja Mithoo (2K19/PHDCO/501) has carried out her research work presented in this thesis titled "Link Analysis on Social Engineering" under my supervision. The research work was undertaken for the award of the degree of Doctor of Philosophy from Delhi Technological University, New Delhi, India.

The thesis embodies the results of original research and studies conducted by the student herself. The contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this University or any other University/Institution.

Date:

Prof. Manoj Kumar
Professor
Computer Science and Engineering
Delhi Technological University
(Formerly Delhi College of Engineering)
Delhi-110042, India

CANDIDATE'S DECLARATION

I, Pooja Mithoo (2K19/PHDCO/501), hereby declare that the thesis titled Link Analysis on Social Engineering submitted to Delhi Technological University (Formerly Delhi College of Engineering), Delhi-110042, India, in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy in the Department of Computer Science and Engineering, is my original work.

This research work has been completed under the supervision of Prof. Manoj Kumar, Professor, Department of Computer Science and Engineering, Delhi Technological University, (Formerly Delhi College of Engineering), Delhi-110042, India,.

The explanations presented in this thesis are based on my understanding and comprehension of the original texts. I confirm that I have not submitted this work to any other institution for the award of any degree, diploma, associate-ship, fellowship, or other title or honor.

Pooja Mithoo
2K19/PHDCO/501
Computer Science and Engineering
Delhi Technological University
(Formerly Delhi College of Engineering)
Delhi-110042, India

ABSTRACT

Modern-day crime often involves multiple actors working in coordination across regions or even continents. Criminals might use encrypted communication, proxies, and false identities to avoid detection. This complexity makes it difficult for investigators to isolate suspects or determine their roles without sophisticated analytical techniques.

Link analysis helps overcome these challenges by enabling investigators to model and understand the structure and flow of criminal activity. For example, when investigating a criminal gang involved in narcotics, investigators can use link analysis to map out connections between known suspects, trace their call histories, track their financial transactions, and link them to locations or events (eg. Drug seizers). This thesis analyses the challenges faced by link analysis methods and how different algorithms can improve it. Algorithms like Spizella, SDHO based detection are formulated in this research to overcome the existing challenges in link analysis.

The first study in this thesis analyses and finds gaps in different Link Analysis tools and techniques. Link analysis is a technique of data mining that is especially used to detect useful and interesting patterns. The first challenge in link analysis is to reduce the graphs into manageable portions. Identifying the interesting relationships and deciding to reduce the graphs is important challenge. Therefore, determining how to apply the link analysis techniques to detect abnormal and suspicious behaviour is needed.

The second study proposes a model that, through rule-based deduction, data transformation, and FP-Growth algorithms, detects patterns of influence between states. This approach clusters data by crime types, finding that criminal activities in Alaska impact Washington, while Alaskan policies have a more influence Washington. These findings align with broader national trends: recent FBI data indicates significant decreases in violent crime nationwide in 2023, underscoring the impact of policy adjustments across states.

The third study proposes a Spizella swarm based BiLSTM classifier is used for the detection of crime rate in this research. Faster convergence is a crucial factor and this faster convergence is achieved by the proposed Spizella swarm optimization. BiLSTM classifier effectively identified the crime rate and the BiLSTM performance is boosted by the Spizella swarm optimization where the escaping characteristics of Spizella improve the convergence

and help in attaining desired results. Measuring the metrics values for accuracy, sensitivity, and specificity demonstrates the effectiveness of the proposed method.

The fourth study introduces a Sheep Dog Hunt Optimization enabled Knowledge-Enhanced Optimal Graph Neural Network classifier (SDHO-KGNN) approach for detecting fraudulent calls accurately. The effectiveness of the proposed SDHO-KGNN approach is achieved through the combination of the power of graph representation learning with expert insights, which allows the proposed SDHO-KGNN approach to capture complex relationships and patterns within telecom data. Additionally, the integration of the SDHO algorithm enhances the model performance by optimizing the discrimination between legitimate and fraudulent calls.

ACKNOWLEDGEMENT

First and above all, my heartfelt appreciation goes to my supervisor, Prof. Manoj Kumar. Your exceptional mentorship, profound expertise, and steadfast belief in my potential have been the cornerstone of my research journey. Your guidance has been more than just academic advice; it has been a source of constant motivation, encouraging me to explore uncharted territories and embrace challenges with resilience.

I also wish to extend my gratitude to Professor Vinod Kumar, whose positivity, motivation and thoughtful support significantly enriched the quality of my research and the direction of my work.

I would like to extend my gratitude to my lovely daughters Bhavika and Vanika, and I am deeply indebted to my husband Mr. Dushyant Sapra, my mother-in-law, Mrs. Veena Sapra, who have been my pillar of strength and supported me in managing my personal and professional life. Their unwavering support, patience and belief in my abilities have been a constant source of strength throughout my journey.

I am profoundly grateful to my mother, Mrs. Umesh Kumari and blessings of grandparents Mr. PD Kambo and Mrs Kamla Kambo. I am thankful for my colleagues and friends at the Department of Computer Science, Delhi Technological University.

I want to sincerely thank my sister Dr. Neelu Kambo, for her unconditional love, emotional support and encouragement. She has been supporting me since the start of my journey and her valuable suggestions have contributed meaningfully to my research.

This thesis stands as a testament to the collective efforts and support of these remarkable individuals. Their contributions have shaped not only my academic growth but also my personal development, making this journey truly enriching and unforgettable.

Pooja Mithoo
Place:
Date:

LIST OF ABBREVIATIONS

AML:	Anti Money Laundering
ANN:	Artificial Neural Networks
ARIMA:	Autoregressive Integrated Moving Average
BERT:	Bidirectional Encoder Representation from Transformers
BiLSTM:	Bidirectional Long Short-Term Memory Network
CDR:	Call Data Records
CNN:	Convolutional Networks
CRF:	Conditional Random
C-BiLSTM:	Cluster-based Bi-directional Long-Short Term Memory
DDPG:	Deep Deterministic Policy Gradient
GCN:	Graph Convolutional Network
GAT COBO:	Graph Attention network with Cost-sensitive Boosting
GNN:	Graphical Neural Networks
GRU:	Gated Recurrent Unit
HAGEN:	Homophily-Aware GCN Recurrent Network
HESM:	High End Enterprise Management
IoT:	Internet of Things
LAPD:	Los Angeles Police Department
MAE:	Mean Absolute Error
ML:	Machine Learning
NMI:	Net Monthly Income
OCND:	Outstanding Contribution to Nation Development
RCGN:	Cost-sensitive Graph Network
RNN:	Recurrent Neural Networks
RMSE:	Root Mean Square Error
ROC-AUC:	Area Under the Receiver Operating Characteristic Curve
SFPD:	Secure Flight Passenger Data
SEO:	Search Engine Optimization
SVM:	Support Vector Machine
TS:	Time Series
VGGNET:	Very Deep Convolutional Networks
VoIP:	Voice Over Internet Protocol)

TABLE OF CONTENTS

CERTIFICATE			i
CANDIDATE'S DECLARATION			ii
ABSTRACT			iii
ACKNOWLEDGEMENT			v
LIST OF ABBREVIATIONS			vi
TABLE OF CONTENTS			viii
LIST OF FIGURES			xii
LIST OF TABLES			xiv
1	INTRODUCTION		1
	1.1	Background	1
	1.2	Link Analysis	3
		1.2.1 Link Analysis in Crime Detection	5
		1.2.2 Link Analysis in Fraudulent Call Detection	7
		1.2.3 Current Aspects of Link Analysis	8
	1.3	Research Gaps and Problem Statement.	8
	1.4	Research Objectives	10
	1.5	Contribution of Thesis	11
	1.6	Thesis Organization	11
2	LITERATURE REVIEW		15
	2.1	Crime and Fraud Detection Methods	16
		2.1.1 Crime Detection Methods	17
		2.1.2 Fake Call Detection Methods	18
		2.1.3 Hybrid Approaches: Link Analysis and ML	18
	2.2	Structural Analysis	19
	2.3	Detailed Review of Prominent Link Analysis Methods	20
	2.4	Summary	26
3	A ROLE OF LINK ANALYSIS IN SOCIAL ENGINEERING		28
	3.1	Related Works	29
		3.1.1 Link Analysis and Association Rule Mining	29
		3.1.2 Link Analysis and Social Networking Analysis	29
		3.1.3 Link Analysis and Counter-terrorism	30
	3.2	Applications of Link Analysis	31

3.3 Centralization 31
3.3.1 Betweenness Centrality 33
3.3.2 Closeness Centrality 34
3.3.3 Eigenvector Centrality 35
3.3.4 Hub Centrality and Authority Centrality 35
3.4 Clustering Coefficient 36
3.5 Summary 36
4 SOCIAL NETWORK ANALYSIS FOR CRIME DETECTION USING SPIZELLA SWARM OPTIMIZATION USING BI-LSTM CLASSIFIER 38
4.1 Introduction 38
4.1.1 Motivation 40
4.2 Related Works 41
4.2.1 Challenges Addressed 42
4.3 Proposed Methodology 43
4.3.1 Input Data 43
4.3.2 Preprocessing 44
4.4 Architecture of BiLSTM Classifier 45
4.5 Proposed Spizella Swarm 45
4.6 Results and Discussion 50
4.6.1 Performance Analysis 51
4.6.2 Comparative Methods 55
4.6.3 Comparative Discussion 60
4.7 Summary 61
5 IMPACT OF CRIME BASED DATA MINING IN INTERSTATE POLICIES 62
5.1 Introduction 62
5.2 Literature Survey 63
5.3 ETL Phase 64
5.3.1 Extraction 64
5.3.2 Transformation 65
5.3.3 Loading 67
5.4 Data Mining Techniques 67
5.4.1 Loading Classification: Logistic Regression 67
5.4.2 Clustering: K-means 68
5.5 Frequent Pattern Growth 69
5.6 Results and Discussion 70

5.7 Summary 74
6 SDHO-KGNN: AN EFFECTIVE KNOWLEDGE ENHANCED OPTIMAL GRAPH NEURAL NETWORK APPROACH FOR FRAUDULANT CALL DETECTION 76
6.1 Introduction 76
6.2 Literature Review 79
6.2.1 Challenges 80
6.3 SDHO enabled KGNN 81
6.3.1 Methodology 82
6.4 Sheep Dog Hunt Optimization 86
6.4.1 Motivation 86
6.4.2 Initialize the Population 87
6.4.3 Equations Calculation 87
6.4.4 Termination 90
6.5 Result and Discussion 91
6.5.1 Experimental Setup 91
6.5.2 Dataset Description 91
6.5.3 Performance Metrics 92
6.5.4 Experimental Results 92
6.6 Performance analysis for SDHO enabled KGNN 95
6.6.1 Performance Analysis with Training Percentage 95
6.6.2 Performance analysis with k-fold 96
6.7 Comparative Methods 98
6.7.1 Comparative Evaluation with Training Percentage 98
6.7.2 Comparative analysis with k-fold 100
6.7.3 ROC Analysis 102
6.7.4 AUC Analysis 103
6.7.5 Model Loss and Accuracy Graph Analysis 104
6.8 Comparative Discussion 105
6.9 Summary 106
7 CONCLUSION AND FUTURE SCOPE AND SOCIAL IMPACT 108
7.1 Research Contribution 1 108
7.2 Research Contribution 2 108
7.3 Research Contribution 3 109
7.4 Research Contribution 4 109

7.5 Social Impact 110
7.6 Future Work 111
LIST OF PUBLICATIONS 112
MAPPING 113
BIBLIOGRAPHY 114

LIST OF FIGURES

Figure No.	Title of Figure	Page No.
Figure 1	Geographical Location from Google Maps	3
Figure 2	Cluster Formation in Link Analysis network	4
Figure 3	Example of Multiplex Networks	6
Figure 4	Year-wise record count of research papers	19
Figure 5	Pie count of research publications	20
Figure 6	World Heritage Twitter NodeXL SNA Map and Report for Wednesday, 16 December- 2015 at 17:07 UTC	30
Figure 7	Degree Centrality for Zachary's Karate Club Network	32
Figure 8	Betweenness Centrality for Zachary's Karate Club Network	33
Figure 9	Closeness scores for Zachary's Karate Club Network	34
Figure 10	Eigenvector Centrality for Zachary's Karate Club Network	35
Figure 11	An example of clusters within a network	36
Figure 12	Schematic representation of the proposed Spizella swarm optimization based BiLSTM classifier.	44
Figure 13	Architecture of Spizella swarm optimization based BiLSTM classifier	46
Figure 14	Performance analysis for the proposed Spizella swarm optimization based BiLSTM classifier for dataset-1 in terms of a) accuracy b) sensitivity c) specificity	52
Figure 15	Performance analysis for the proposed Spizella swarm optimization based BiLSTM classifier for dataset-2 in terms of a) accuracy b) sensitivity c) specificity	54
Figure 16	Comparative analysis for the proposed Spizella swarm optimization based BiLSTM for dataset-1 in terms of a) accuracy b) sensitivity c) specificity	57
Figure 17	Comparative analysis for the proposed Spizella swarm optimization based BiLSTM for dataset-2 in terms of a) accuracy b) sensitivity c) specificity	59
Figure 18	Geographical Location From Google Maps	63
Figure 19	Elbow curve with 7<K<10	69
Figure 20	Sihouette score graph with 7<K<10	69
Figure 21	Cluster of tweets in positive sentiments	71
Figure 22	Cluster of tweets in negative sentiments	71
Figure 23	Cluster of tweets with neutral sentiments	72
Figure 24	A structure of FP-tree based on neutral sentiments with confidence values	72
Figure 25	Bar plot of Crime based labelling in Washington and Alaska after FP-Tree	73
Figure 26	Bar plot of Policy based labelling in Washington and Alaska after FP-Tree	74
Figure 27	Schematic representation of the Fraudulent Call Detection utilizing SdHO-enabled knowledge-enhanced optimal GNN model	82
Figure 28	Architecture of knowledge enabled optimal GNN	85
Figure 29	Flowchart for SDHO Optimization	89
Figure 30	Input graph for k-GNN model	92
Figure 31	Experimental results of the developed model	94

Figure 32 Performance analysis with TP 96
Figure 33 Performance analysis with k-fold 98
Figure 34 Comparative analysis with TP 100
Figure 35 Comparative Analysis with K-Fold 102
Figure 36 ROC Analysis 103
Figure 37 AUC Analysis 103
Figure 38 Accuracy Analysis 104
Figure 39 Loss Analysis 104

LIST OF TABLES

Table No.	Title of Table	Page No.
Table 1	Findings of notable Link Analysis methods	21
Table 2	Pseudo code for Spizella swarm optimization	50
Table 3	Metrics values of the proposed Spizella swarm optimization based on dataset-1	57
Table 4	Metrics values of the proposed Spizella swarm optimization based on dataset-2	60
Table 5	Comparative discussion of the proposed Spizella optimization based BiLSTM classifier	61
Table 6	Pseudocode for the proposed SDHO-KGNN model	90
Table 7	Comparative discussion for SDHO-enabled KGNN	106
Table 8	Objectives Mapping	112

CHAPTER 1
INTRODUCTION

Rapid advancements in mobile communication technologies have led to the progression of telecom scams that not only deplete the individual fortune but also affect the social income [1-3]. All the technological advancements that have initiated a rapid increase in criminal activity have posed a great threat. The prevention of such criminal activities can be done through preventive measures. Link analysis is a technique which is used to conduct an investigation in various criminal based activities that can be counter-terrorism, intelligence, fraud detection, market research, medical research and other techniques like Search Engine Optimization (SEO). The practice of fraudulent call detection has gained significance, which not only aims to proactively recognize the frauds [4-5] but also alleviate the fraudulent activities to manage external losses.

Throughout history, nations across continents have enacted policies on areas such as budgeting, ecology, health, education, democracy, infrastructure, and crime. Following major incidents like the [significant terrorist acts] data science began to shift towards crime science, encouraging officials globally to refine and legislate policies for maintaining national peace. Although state-level policies have substantial impact potential, the question remains: do we sufficiently consider national impacts in our policy decisions? Previously posed in [2], we aimed to explore this in our project using an analytical framework.

1.1 Background

In an increasingly interconnected world, crime has evolved beyond physical boundaries to include highly organized and technology-enabled operations. From identity theft and cyberattacks to financial fraud and telecommunication scams, the methods used by criminals have become more complex, requiring equally sophisticated tools for detection and prevention. Traditional policing methods that rely heavily on manual data analysis and human intelligence are often inadequate to handle the volume and complexity of modern crime data.

In the modern era of digital communication and social connectivity, criminal activities have increasingly adopted sophisticated methods for planning and execution. Traditional investigative methods are no longer sufficient in combating modern crimes such as cyber fraud, organized crime, terrorism, and deceptive telecommunication practices like fake calls. One of the most promising areas in the domain of crime analytics is link analysis, a data analysis technique used

to evaluate relationships and connections between entities such as individuals, organizations, and communications patterns.

Link Analysis has emerged as a crucial technique in the field of crime analytics and intelligence gathering. At its core, link analysis involves identifying and visualizing relationships between entities- such as individuals, organizations, events, locations and communication devices- using graph theory and network modelling. These relationships can be direct or indirect and are crucial in piecing together criminal networks, identifying key actors, and uncovering hidden connections that might not be evident through linear analysis.

Link analysis, derived from graph theory, helps law enforcement agencies and analysts uncover hidden patterns and associations that would otherwise remain obscured within massive datasets. By visualizing connections among suspects, locations, devices and communication instances, link analysis provides invaluable insights that are critical for both crime detection and fraudulent call prevention.

The advent of big data, combined with advances in data mining and visualization technologies, has made link analysis more powerful and accessible than ever before. Law enforcement agencies, intelligence organizations, and telecom service providers are increasingly adopting link analysis tools to uncover hidden relationships in datasets like Call Detail Records (CDRs), social media interactions, financial transactions, and surveillance logs.

Among various policy domains, crime remains uniquely influenced by policy—a positive factor—in addressing negative outcomes. This analysis considers two policy types: intra-state, executed with confidence at 1.0, and inter-state, where the “trustworthiness” of cross-state effects is yet uncertain. Our project focuses on inter-state crime policy, bridging computer science with policy. To establish causality between state policies, both qualitative [3][4] and quantitative [5] methods exist; we employed a quantitative approach to demonstrate how state-specific policies affect neighbouring states.

Twitter, a widely used microblogging platform, generates data continuously through user messages (tweets) reflecting ideas, opinions, and preferences. By analysing hidden information in tweets, such as user opinions and activity, we gain insights. Figure I presents the states considered for crime-policy link analysis. To analyse data, we used data mining to identify patterns and focused on crime-related terms as defined by the National Institute of Justice. Terms extracted from major publications, including *The Washington Post* and *The Seattle Times*, helped us cluster tweets by topics like "unlawful activities," "illicit substances," and "assaults."

In this study, we developed a system using quantified Twitter data collected through REST and streaming APIs, allowing comparisons across states regarding crime and policy. Figure I depicts the states included. Beginning with the assumption that the problem space required unsupervised problem-solving, we used tweets to create clusters based on offense categories and state-level policy development. Since distinguishing between Washington State and Washington, D.C., was challenging, we provide a disclaimer at the project's outset.

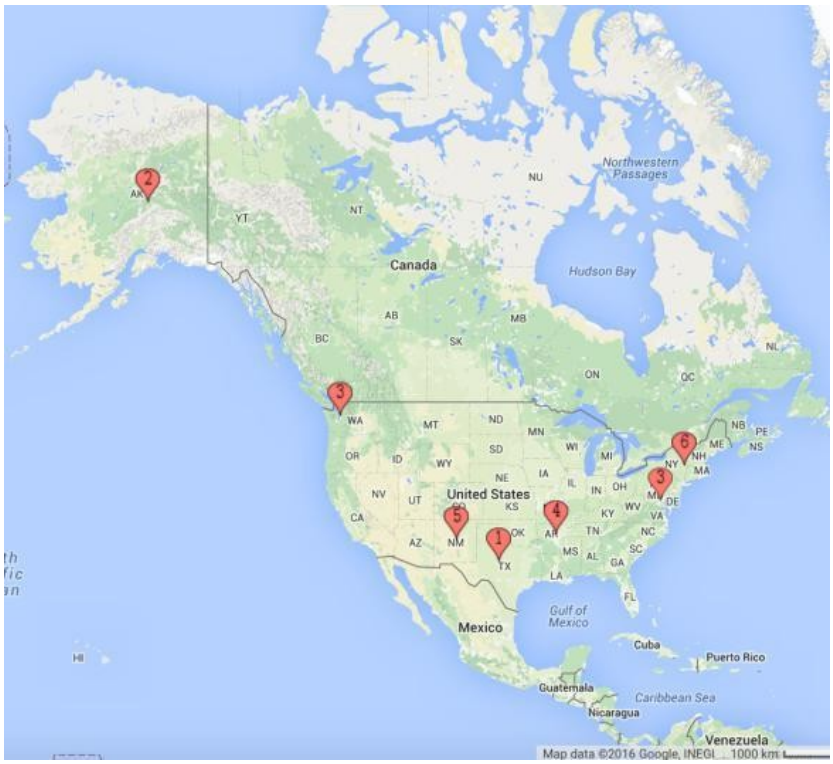


Figure 1: Geographical Location from Google Maps

To gain insights from Twitter data, we applied data mining techniques to identify interpretable patterns related to crime. We examined terminology from the National Institute of Justice and used terms such as "unlawful," "illicit substances," "murder," "rape," "criminal activity," and "assaults" as primary clustering labels. These terms were identified from tweets referencing major news sources like *The Washington Monthly*, *The Atlantic Magazine*, and *The Texas Monthly*, which frequently report on crime and policy.

The dataset used in this study comprises publicly accessible Twitter posts collected through the official Twitter API, ensuring legality and reproducibility. The dataset includes approximately *[insert number, e.g., 150,000–200,000]* tweets related to crime, fraud alerts, scam reporting, and public safety keywords. Each record contains tweet text, timestamp, user metadata (non-sensitive), geolocation when available, and hashtag context. Data preprocessing involved noise removal, elimination of user identifiers, slang normalization, stop-word filtering, tokenization, and conversion to model-ready sequences. Duplicate entries, bot-generated content,

and tweets lacking semantic relevance were removed to enhance dataset quality. The cleaned dataset thus supports robust training of Bi-LSTM and SDH-KGNN models for fraud-call and crime-pattern analysis.

1.2 Link Analysis

Modern-day crime often involves multiple actors working in coordination across regions or even continents. Criminals might use encrypted communication, proxies, and false identities to avoid detection. This complexity makes it difficult for investigators to isolate suspects or determine their roles without sophisticated analytical techniques.

Link analysis helps overcome these challenges by enabling investigators to model and understand the structure and flow of criminal activity. For example, when investigating a criminal gang involved in narcotics, investigators can use link analysis to map out connections between known suspects, trace their call histories, track their financial transactions, and link them to locations or events (eg. Drug seizers). These links are typically visualized as nodes (entities) and edges (relationships), making it easier to interpret large volumes of data.

This is a method used in data warehousing to explore how different things- like people, organizations, or transactions- are connected. It's a powerful tool that helps uncover hidden relationships and patterns that might not be obvious at first glance. You'll often find of used in areas like criminal investigations, fraud detection, intelligence work, healthcare research, market analysis, and even online marketing and SEO.

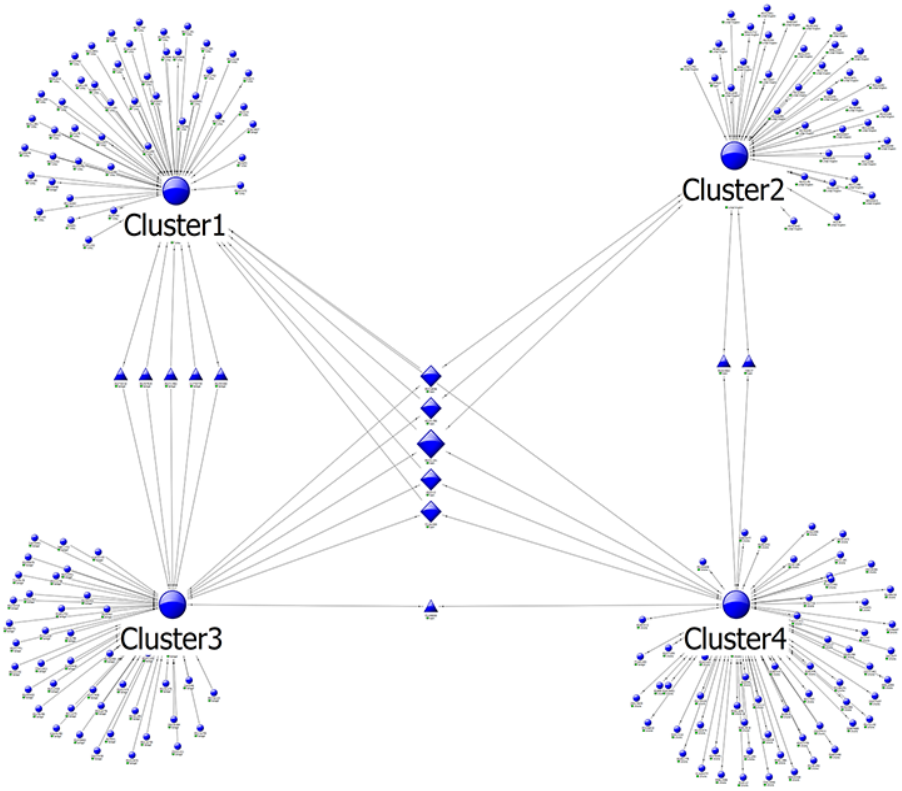


Figure 2: Cluster Formation in Link Analysis network

For example, in law enforcement, a crime analyst might use link analysis for mapping of any criminal network by analysing the phone records to see who's talking to whom. In banking systems can track a person's typical credit card usage and quickly flag anything that doesn't fit their usual pattern, helping to detect fraud and alert the customer.

During public health crisis, analysts can look at prescription records and demographic data to identify new trends or hotspots. In the world of science, especially biology, link analysis can help researchers understand how different proteins interact in the body.

It's also a big deal in marketing. Companies like Amazon, Flipkart, and eBay use it to recommend products to shoppers based on what they've looked or bought before. Another related tool, called Clickstream analysis, helps websites track how people move through their pages. By understanding how customers behave online, companies can improve their chances of making a sale. In short, link analysis helps connect the dots- whether it's solving crimes, protecting people from fraud, tracking the spread of the disease, or helping businesses better serve their customers.

1.2.1 Link Analysis in Crime Detection

Crime detection traditionally involves collecting evidence, eyewitness reports, and forensic data. However, with the growth of digital footprints, large volumes of unstructured and structured data are generated daily from social media, emails, phone records, and surveillance systems. Link analysis tools process the data to identify relationships among suspects, their communications, movement patterns, and affiliations.

For instance, in organized crime investigations, link analysis can identify the hierarchy within criminal networks by tracking communications between hacker groups and illegal financial transactions. The use of software tools like IBM i2 Analyst's Notebook and Maltego allows investigations to construct interactive visual maps of criminal entities, highlighting their interactions and common nodes.

The main goal of link analysis is to make sense of the complex data. It helps spot the patterns that are interesting or unexpected, find unusual behavior (which might indicate fraud or something suspicious), and see when something breaks the usual pattern. Moreover, link analysis can reveal central figures or hubs within criminal networks. These central figures often do not communicate directly with the criminal act itself but act as facilitators or planners. By using metrics like degree centrality, betweenness centrality, and closeness, investigators can pinpoint individuals whose removal would disrupt the entire network. A practical example is the use of IBM i2 Analyst's Notebook by police forces globally. This tool allows for the integration of disparate datasets and helps analysts build comprehensive visual link maps that aid in tactical

investigations and strategic decision-making. Multiplex networks are shown in figure 3.

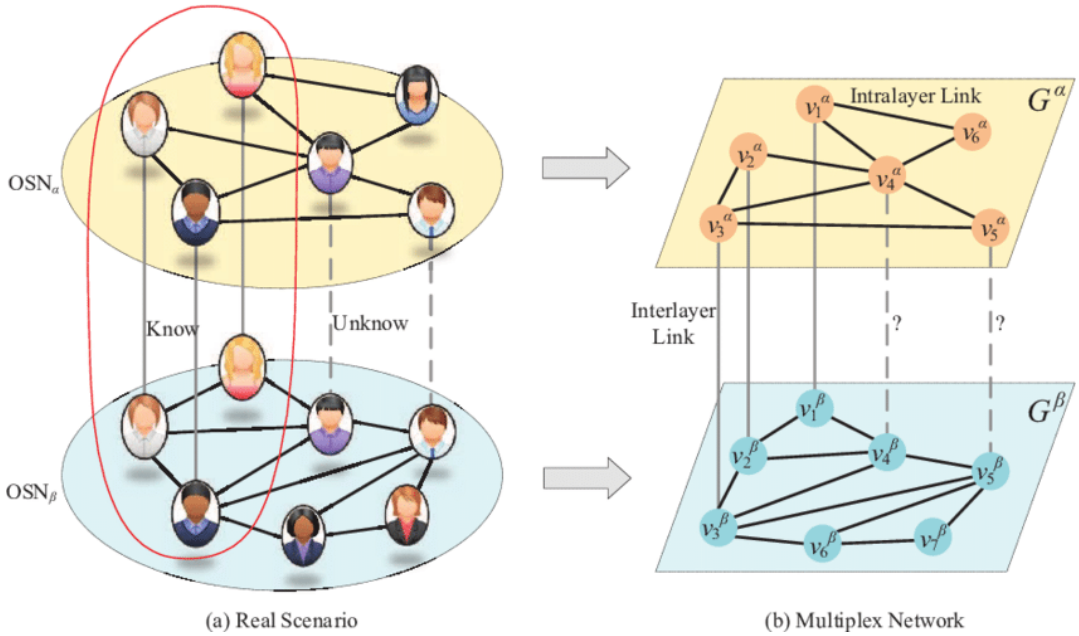


Figure 3: Example of Multiplex of Networks

The main goal of crime analysis is to assist or support a police department's operations. These activities include patrolling, patrolling operations, crime prevention and reduction methods, problem-solving, evaluation and accountability of police actions, criminal investigation, arrest, and prosecution. Crime rate detection aids law enforcement organizations in forecasting and identifying the crimes that occur in a particular region, which lowers the crime rate. Analyzing the current crime trends and forecasting future crime rates helps to minimize the crimes that occur or determine the suspected individuals. The authorities can take responsibility and attempt to lower the crime rate based on this information. To predict the crime rate based on social networks this research aims in developing a new framework and gain deep perception the existing methods are analyzed and the observations are interpreted below.

Link analysis is an approach of data mining that is used to find out the relationships between nodes. These nodes may be in the form of people, transactions, and organizations. The link analysis technique is used to investigate various criminal activities such as counter-terrorism, fraud detection, intelligence, medical research, market research, Search Engine Optimization (SEO), etc. Link analysis is mainly used with three objectives such as, to find interesting patterns with the help of data, to find the various types of anomalies through the discovery of new patterns of interest (social network analysis, data mining), and where known patterns are violated. There may be some examples of link analysis such as a criminal network being analyzed by a crime analyst. The relationship and hierarchy between members of the network may be determined using cell phone data. Credit card fraud may be identified using the system, which utilizes the already noted patterns of transactions of each client with his individual information. Systems identify the anomalies and

alert the client of potential theft. The new patterns may be identified by the analyst using data on prescriptions and demographics that are emerging as the crisis spreads. Link analysis may also be utilized in biological science to analyze protein interactions. It is also helpful for law enforcement and detects and prevents terrorism networks. It is also used in marketing management especially related to product recommendation analysis. Amazon, Flipkart, and e-bay are famous companies for their recommendation system, which helps customer selections with the help of link analysis. The other approach called Clickstream analysis develops the ability of websites to analyze and predict every activity of customers to understand their buying behavior and to increase the probability of their purchasing goods with the help of link analysis [1].

1.2.2 Link Analysis in Fraudulent Call Detection

Telecommunication fraud, especially through fake or scam calls, is a widespread and growing issue. Fraudsters commonly use spoofed caller IDs, automated robocalls, or socially engineered messages to deceive victims into revealing personal or financial information. These activities often steam from coordinated operations or fraud rings that use a variety of tactics to avoid detection, including frequently changing phone numbers, using virtual numbers, and placing calls from foreign networks.

Link analysis in this domain helps identify patterns of behavior across large call networks. For instance, multiple calls originating from different numbers but targeting the same group of victims, or calls with similar durations and scripts, can be flagged as potentially fraudulent. By constructing a network graph of callers, recipients, and communication patterns, analysts can uncover clusters that represent organized fraud operations.

Fake or fraudulent calls, including phishing attempts, scam calls, and identify theft schemes, are a growing concern worldwide. These calls often originate from hidden networks of fraudsters that exploit anonymity and untraceable numbers to deceive victims. Telecommunication metadata, such as call detail records (CDRs), IP addresses, and VoIP (Voice Over Internet Protocol) patterns, can be analyzed using link analysis to detect anomalies and trace fraudulent call sources.

Call Detail Records (CDRs), which contain metadata such as source and destination numbers, timestamps, call duration, and tower locations, are crucial input data for such analysis. Algorithms can scan this metadata to detect unusual calling patterns, like a sudden spike in calls from a particular region, or repeated calls from unregistered VoIP lines.

By examining attributes such as repeated call attempts, call durations, unusual time stamps, and shared recipients, analysts can identify suspicious call patterns. Link analysis help group

calls from different numbers that originate from a common source or exhibit similar behavior, enabling early detection and prevention of mass scams.

Furthermore, machine learning techniques such as clustering and anomaly detection can be integrated with link analysis to automatically identify suspicious call patterns and adapt to evolving fraud strategies. This automation is essential in handling the volume of data generated by telecom networks daily.

1.2.3 Current Aspects of Link Analysis

The implications of crime and fake call detection are not only a matter of security but also a significant economic impact. According to the Communications Fraud Control Association (CFCA), telecom fraud causes global losses estimated at over 39 billion dollars annually. On the other hand, global damages from cybercrime are project to reach at much higher scales.

These alarming statistics emphasize the urgency of developing more effective tools for detection and prevention. Link analysis offers a scalable and efficient solution for understanding criminal and fraudulent behavior in complex networks, The motivation behind applying link analysis is its ability to process and make sense of massive, multidimensional datasets that would be too complex or time-consuming for human analysis alone.

The global economic impact of cybercrimes and telecom fraud is staggering. According to 2023 report by the Communications Fraud Control Association (CFCA), telecom fraud alone accounts for over 39 billion dollars in annual losses. Similarly, cybercrime damages are projected to reach 10.5 trillion annually by 2025, according to Cybersecurity Ventures. Given this alarming growth, the need for efficient analytical tools like link analysis has become critical.

Additionally link analysis enhance decision-making in real-time investigations, where time is critical factor. By providing visual representations and quantifiable metrics about the network structure, it aids investigators in prioritizing suspects, focusing resources, and anticipating the next move in a criminal operation.

The motivation for using link analysis stems from its ability to process vast, seemingly unrelated data and convert it into actionable intelligence. It enhances investigative efficiency, reduces manual effort, and provides visual representations that aid in faster decision making.

1.3 Research Gaps and Problem Statement

Despite the promising research, several challenges persist in implementing effective crime detection systems:

- Time consuming data maintenance: With rising crime rates, it’s difficult to keep records

consistently updates, making accurate analysis more complex. [3].

- Classifier tuning is tricky: Achieving high prediction accuracy requires careful optimization of model parameters.

- Bi-LSTM (Bidirectional Long Short-Term Memory Network) models are resource intensive: They take a long time to train due to massive volume of data and ensuring quick convergence remains a technical error [8].

- Weight initialization in Bi-LSTM is sensitive: Random initialization affects results, so adjusting model weights accurately is critical [4].

- Feature extraction is difficult: Identifying the right features that significantly contribute to crime prediction is still a challenging task [5].

- In automated fraudulent call detection approach [12], the LSTM model attained limited performance due to the formation of data, which is not series at all times to acquire the arrangement relationship between each dimension. Moreover, the DNN model has a major risk of overfitting, which was exposed using the drift evaluation concept.

- Training and using ANN (Artificial Neural Networks)-based models may require substantial computational resources, which could be a limitation for some organizations. The system's efficiency relies significantly in the quality of the model and comprehensiveness in the call traffic information. Inaccurate or partial information could result in false alarms or overlooked instances of fraud. [37][34].

- Mobile social network fraud detection datasets often suffer from class imbalance, where genuine user activities significantly outnumber fraudulent ones. Handling this imbalance effectively is crucial [22].

- GAT-COBO heavily relies on every quality as well as the availability of data. Noisy or incomplete information can affect the performance of the model. GNNs (Graphical Neural Networks) can be challenging to interpret, which make them difficult to comprehend the reasonings behind the model's choices [25].

- Detecting temporal patterns in telecom fraud, and enabling real-time detection and adaptation also face challenges related to data quality, complexity, interpretability, and resource requirements, which need to be carefully considered when applying the approach in practice [38].

All the methods mentioned above have limitations in terms of the accuracy in detection of

crimes, where the detection accuracy is much lower and almost all of the methods are very computationally expensive. This research helps in detection of the crime rate with higher accuracy compared with the already existing algorithms, as this research enables the Spizella swarm optimization with the help of Bi-LSTM classifier. This effective optimization has tuned the parameters which help in the selection of different optimal parameters, that help in reduction of the computational complexity. The correlation is determined and hence, it becomes the most important challenge and, in the end, the correlation between the different words is calculated and determined using the proposed method. Feature selection uses its effective features to extract the results for the crime analysis part in the research.

Problem Statement: Considering the research gaps, the major challenge being faced is the frauds occurring in the social media. With rapid advancements in fraud calls, detection has become difficult for traditional approaches. and there is a need to make models using link analysis and different classifiers to detect and combat with frauds in social engineering.

1.4 Research Objectives

This study aims to explore and validate the role of link analysis in enhancing the detection and understanding of both criminal activity and fake call operations.

The research gaps identified in the domain of crime rates solving through link analysis highlight the need for comprehensive solutions that address the multifaceted challenges of embedding capacity, image quality, reversibility, separability, and security. To address these concerns, a thorough study and experimental evaluation of available KGNN methods in the link analysis domain are crucial. Such an analysis, conducted on a consistent test dataset, can provide valuable insights into the strengths and limitations of existing approaches, paving the way for the development of novel schemes. The research objectives are delineated as follows:

- **Objective 1**: To investigate and analyze all the tools, techniques of link analysis.
- **Objective 2**: To design an algorithm to generate the associate elements or item sets.
- **Objective 3**: To implement the proposed algorithm with the appropriate data set.
- **Objective 4**: To analyze the performance of the proposed approach using parameters like location, past criminal records, relation with other objects.
- **Objective 5**: Comparative analysis of proposed approach with existing algorithms on the basics of identified parameters.

• **Objective 6:** To utilize this algorithm for possible inclusion in a real-world problem.

1.5 Contribution of Thesis

The research work presented in this thesis makes significant contributions to the domain of crime detection and validation using link analysis, addressing several critical gaps and challenges identified. Through detailed experimental evaluation on the qualitative aspects, this research provides a comprehensive analysis of existing Link Analysis techniques, highlighting their strengths, limitations, and areas for improvement. This evaluation serves as a critical resource for future research in link analysis, offering in- sights into the effectiveness of different approaches. This thesis also incorporates a bibliometric analysis, offering quantitative insights into publication trends, influential contributors, and collaborative networks within the link analysis domain. This synergistic combination of qualitative and quantitative perspectives helps to identifies the researchers to identifies the challenges and gaps of the domain.

Furthermore, one of the primary contributions of this thesis is the development of novel Link Analysis methods, including different link-based crime detection methods using machine learning and networking, marking a significant advancement in crime detection using link analysis research. The thesis successfully addresses several critical research gaps, and the need for improved security measures in link analysis techniques. By presenting innovative solutions that tackle the issues, the research contributes to the refinement and advancement of link analysis methodologies.

The proposed link analysis methods have significant implications for practical applications. By enhancing the security, capacity, and efficiency of data hiding techniques, this research facilitates the development of more robust systems for protecting sensitive information. Lastly, the thesis outlines future directions for link analysis research, recommending topics and areas for further investigation and advancement. This research not only bridges current gaps in the field but also establishes a foundation for subsequent methods in secure communication and data protection.

1.6 Thesis Organization

This thesis comprises across eight chapters, progressively building from introduction to conclusion to present a comprehensive research journey. The arrangement of the chapters is outlined as follows:

Chapter 1. Introduction: The introductory chapter lays the foundation for the research work presented in this thesis. It begins by exploring the ubiquity of content specific graphs and the need for secure communication in the modern era. The chapter then discuss into the concept

of Link Analysis, emphasizing its significance as a means of achieving security and privacy in various applications. Subsequently, it introduces the domain of crime records in different fields that are solved through link analysis. Building upon this foundation, the chapter then focuses on different challenges faced in the crime detection field and how better results can be achieved through link analysis graphs, that operates on encrypted data, providing an additional layer of security and privacy. This chapter gives different insights in providing a complete information about the Link Analysis and its applications. The chapter presents a comprehensive classification of Link Analysis techniques, high- lighting the different approaches and methodologies employed in this field. Other than this, it provides several different insights regarding the Link Analysis algorithms. Furthermore, it underscores the importance of Link Analysis in various domains, such as secure communication, multimedia authentication, and. content protection. By providing this comprehensive introduction, the chapter sets the stage for the subsequent sections of the thesis, establishing the context and significance of the research work undertaken.

Chapter 2. Literature Review: The literature review chapter presents a comprehensive exploration of the existing body of research in the domain of crime detection through link analysis. It encompasses a thorough examination of the state-of-the-art techniques, methodologies, and approaches proposed by researchers to address the inherent challenges of this field. By synthesizing and critically analyzing the relevant literature, this chapter aims to provide a solid foundation for understanding the current landscape of link analysis-based crime detection research, identifying potential research gaps, and paving the way for the development of novel and innovative solutions. Additionally, the chapter incorporates a bibliometric analysis, offering quantitative in- sights into publication trends, influential contributors, and collaborative networks within the link analysis domain. Different aspects of the previously proposed algorithms are determined where limitations of these algorithms have been defined along with their accuracy and different other methodologies used in these algorithms. These algorithms help in getting several insights in the proposed research. This synergistic combination of qualitative and quantitative perspectives ensures a holistic and well-rounded understanding of the research terrain, setting the stage for the subsequent chapters and contributions of this thesis.

Chapter 3. A Role of Link Analysis in Social Networking: Chapter 3 introduces Link Analysis and Association Rule Mining where Link Analysis uncover strong association rules from transaction databases, revealing patterns in item sequences. It aids in detecting patterns within criminal networks by extracting relationships between individuals and actions. Then it introduces Link Analysis and Social Networking Analysis. Link Analysis in Social Networking Analysis uses tools like IBM i2 Analyst's Notebook and Node XL to visualize communication

patterns, connections and key entities from large datasets. These tools prioritize metadata to reveal relationships and interaction frequency. Then it introduces Link Analysis and Counter-terrorism where Link Analysis aids law enforcement and intelligence agencies in uncovering hidden relationships, anomalies and network structures with entities like people, locations and communications. It discusses the challenges in the link analysis where there is a need to make complex networks, it handles the incomplete data that is especially critical for counter-terrorism and in the end the most important aspects are data availability and quality that is limited to only real-world applications. Hence, Link analysis is a data mining technique used to uncover useful and interesting patterns from complex datasets. More research is needed to enhance the effectiveness of link analysis, especially for use in sensitive areas like counter-terrorism. This chapter sets the stage for further research aspects in link analysis and gives motivation for further coming up chapters.

Chapter 4 Social network analysis for crime rate detection using Spizella swarm optimization based Bi-LSTM classifier. This chapter addresses various research gaps such as rise in crime due to increase in technical advancements, need for effective crime prevention measures, role of social media platforms in detecting crimes, etc. This chapter introduces Spizella swarm based Bi-LSTM classifier for crime rate detection through social network analysis. The Bi-LSTM has the capability of analyzing complex texts and the traverse analysis possesses special significance. The parameters such as weights and bias in the Bi-LSTM classifier are optimally tuned using the Spizella swarm optimization further improving the convergence of the classifier and analyzing the texts more efficiently. Spizella swarm optimization algorithm is a hybrid algorithm which has improved convergence and enhanced analysis with advanced text analysis capability along with optimized parameters and performance boost. States enact laws and engage in criminal activities that can affect neighbouring or distant states. The inter-state influences are often highlighted by major media outlets like The Washington Post and The New York Times. This study introduces a model using rule-based deduction, data transformation, and the FP-Growth algorithm to detect influence patterns between states**.** The approach clusters data based on crime types to analyse interstate impacts. The findings align with recent FBI data (2023) showing a nationwide decline in violent crime, suggesting that policy changes have significant cross-state effects.

Chapter 5. Impact of Crime Data Mining on Interstate Policies. This chapter introduces interstate influence that engage in criminal activities. The study collected raw, unstructured tweet data via Twitter's Streaming and REST APIs, requiring extensive cleaning to remove noise like links, hashtags, and usernames. This preprocessing focuses on refining the "text" attribute—vital

for accurate sentiment analysis, clustering, and rule derivation. The study emphasizes the value of structured, accessible models for actionable insights. Future research can adopt heuristic, meta-heuristic, and fuzzy logic methods for deeper rule-based analysis. Computational linguistics was outside the current study's scope but presents a promising avenue, especially in analysing policy texts and public discourse. Techniques like NLP and sentiment analysis could enhance the understanding of how language in policies influences crime trends. The study offers a foundational model that can be extended with advanced machine learning, computational intelligence, and diverse data sources. Leveraging these innovations can improve policy evaluation, support effective governance, and deepen insight into crime-policy dynamics. **A**fter preprocessing, tweets are saved in formats like CSV, TSV, or JSON to ensure results are accurate and interpretable. CSV/TSV and JSON are preferred by the NLP community for their structure and compatibility.

Chapter 6. SDHO-KGNN: An Effective Knowledge Enhanced Optimal Graph Neural Network Approach for Fraudulent Call Detection. Chapter 6 introduces SDHO-KGNN approach which is achieved through the combination of the power of graph representation learning with expert insights. SDHO-KGNN approach to capture complex relationships and patterns within telecom data. The integration of the SDHO algorithm enhances the model performance by optimizing the discrimination between legitimate and fraudulent calls. The SDHO-KGNN classifier captures the intricate call patterns and relationships within dynamic call networks. It introduces telecom records, their pre-processing, blacklist table and the training phase. KGNN for fraud call detection typically extends a standard GNN architecture with the integration of domain-specific knowledge graphs which can capture valuable domain-specific information from knowledge graphs which can lead to better performance in fraud call detection.

Chapter 7. Conclusion and Future Work: This research demonstrates the effectiveness of the proposed method in improving the system performance and accuracy within the target domain. The results validate the approach and highlight its practical applicability in real-world scenarios. However, certain limitations remain, such as scalability and adaptability to diverse datasets. Future support more complex, dynamic environments. Hence, the research will open up new and different future endeavors for the researchers in the coming time. This will also help and motivate the fellow researchers in the coming times. Conclusive comments will unfold at the end of the thesis which will pave the way for the futuristic directions.

CHAPTER 2
LITERATURE REVIEW

This chapter presents an extensive investigation into the domain of Link Analysis an area that has experienced remarkable growth. This section reviews recent research and approaches used in crime rate detection, particularly those involving social media data and machine learning techniques.

Twitter explored the [1] data from seven different regions, analyzing tweets to monitor potential criminal activities. They applied sentiment analysis to understand user behavior and psychological patterns using a part-of-speech tagger tailored for Twitter, to deal with large datasets of unlabeled tweets, they used brown clustering. While this method could detect crime-related patterns, its accuracy was limited. It adopted [2] multiple classifiers to detect crime trends on social media platforms. They used exploratory data analysis to provide a visual summary of various crime types and their frequencies. Additionally, they used the ARIMA model (Auto Regressive Integrated Moving Average) to forecast crime density and frequency for the next five years. However, the model required significant computational power.

A recent study showed [3] economic factors influenced crime in India by applying various machine learning techniques. They found a one-way causal relationship between unemployment rates and robotics, suggesting that economic hardship often led to increased crime. However, a strong correlation between target variables posed challenges in model interpretation. Another study [4] used past crime data to predict the types of crimes likely to occur in specific locations. After cleaning and transforming the data, they applied machine learning algorithms to generate predictions, supported by data visualizations for better understanding. Although many studies focused non predictive methods and risk identifications, few addressed optimizations for processing large datasets efficiently. This [6] used a clustering technique to analyze sentiments in tweets, grouping them into positive, negative, and neutral. Their dictionary-based machine learning model also detected hate speech, though it lacked the ability to analyze opinions when no clear subject was mentioned.

A new study [5] proposed an ensemble-stacking model using SVM (Support Vector machine) for crime detection. Their model outperformed earlier studies that primarily targeted violent crime datasets. They concluded that combining criminological theories with

empirical data improves the effectiveness of predictive models. They also suggested that cloud-based systems using large datasets could yield better outcomes. A study [7] initially sorted tweets in four stages- politics, sports, crime and nature- and then used various machine learning models for classification. This approach not only grouped content efficiently but also revealed with topics users shared most frequently. However, the model still had room for improvement in terms of accuracy.

Another study [8] used a hybrid method combining lexicon-based techniques and the BERT (Bidirectional Encoder Representation from Transformers) deep learning models. They first labeled tweets using a predefined list of crime-related terms, then used the labeled data to train BERT. While effective for tech-based analysis, this method couldn't process multimedia inputs like images or audio and [37] developed a crime detection system using VGGNET19 (Very Deep Convolutional Networks), which quickly identified weapons like guns and knives from images. Although training time was low, the model couldn't predict crimes before they happened. A newer study [38] implemented decision trees and k-nearest neighbor (KNN) techniques to predict crimes based on historical records. These models achieved high accuracy, but performance dropped when dealing with imbalanced datasets.

2.1 Crime and Fraud Detection Methods

Link Analysis is a technique rooted in graph theory, widely used for identifying relationships among entities in large datasets. It has been extensively applied in intelligence gathering, financial fraud detection, cybercrime investigations, and more recently, in telecommunications fraud. The key idea is to visualize and compute how entities are connected to uncover patterns that indicate the suspicious or criminal behavior. Multiple studies and tools have demonstrated the efficiency of link analysis in improving detection accuracy, reducing manual effort, and proving strategic insights during investigations.

This [33] initiated a technique for detecting fraud, employing similarity and multi-view GNN. In this approach, a weight parameter is employed to increase the importance of fraud samples that have been flagged or labeled. However, there's a risk of overfitting the training data, especially if not enough labeled fraud examples are available. [25] GAT-COBO model used for fraud detection which can adapt to evolving fraud tactics and patterns in real-time, making them suitable for dynamic environments. However, GNNs can be challenging to interpret, making it difficult to grasp the logic behind the model's decisions. HESM [7] (Internet of Things) approach for fraud detection which enhances the superiority of the model. However, the model can be challenging to interpret, making it difficult to comprehend the reasoning behind the model's

choices.

This study [22] suggested a convolutional Neural Network (CNN) method for fraud detection using CDR data. In the suggested method, an end-to-end model was offered to recognize new distinctive features without automatically designed features. Also, the suggested model achieved satisfying outcomes on the challenges in the detection part of telecommunication. The developed method extracted behavior data from raw CDR effectively, nevertheless combining it with communicating data was still an unresolved issue. [34] used a sim Box model for fraud detection in Telecommunication. The system can operate in real-time, allowing for immediate identification and response to fraudulent activities as they occur.

However, ANNs can be challenging to interpret [35] presented a C-Bi-LSTM (Cluster-based bi-directional long-short term memory) algorithm for fraud detection, the utilization of intelligent algorithms allows for the automated and efficient identification of fraudulent call information, potentially reducing the workload for human operators. The output of intelligent algorithms may be challenging to interpret, making it difficult to understand the reasoning behind specific decisions. A new study [36] suggested a Reinforced Cost-sensitive Graph Network (RCGN) to detect the head of fraud in telecom Fraud. The deep Deterministic Policy Gradient (DDPG) technique was utilized to optimize the coefficients of mass dynamically and a graph was constructed using CDR and three base classifiers were united predict the results and final results of classification. The integration of three base classifiers in the RCGN made the detection of fraud effective, however, usage of a small size dataset in the RCGN method was the limitation, which lacked the experiments of comparison. Also, the suggested model achieved satisfying outcomes on the challenges in the detection part of telecommunication. The developed method extracted behavior data from raw CDR effectively, nevertheless combining it with communicating data was still an unresolved issue. [34] used a sim Box model for fraud detection in Telecommunication.

2.1.1 Crime detection Methods

Several researchers have investigated the use of link analysis in detecting and dismantling criminal networks. In this study, introduced a comprehensive model that applied social network analysis (SNA) and link discovery techniques to terrorism investigations. They demonstrated how link metrics like degree centrality and betweenness could identify key actors within a terrorist cell, even if these actors had minimal direct involvement in crimes. Sparrow outlined early applications of network analysis in understanding the structure of criminal enterprise, advocating for methods that go beyond intuition and focus on empirically derived relationships. His work laid a theoretical foundation for later developments in network based investigative

tools.

Further contributions with formal definitions and algorithms in social network analysis, many of which underpin modern link analysis frameworks. Their methodologies are now applied in law enforcement tools to analyze suspect-suspect, suspect-location, and suspect-device links. Modern tools like IBM i2 Analyst's Notebook and Palantir have operationalized these concepts, allowing agencies to construct dynamic graphs, perform real-time analysis, and visually identify core nodes, bridges, and communication clusters in criminal networks.

2.1.2 Fake Call Detection Methods

In the field of telecommunications, link analysis is commonly used to detect fraud patterns such as subscription fraud, international revenue share fraud (IRSF), and robo-calling. This described how link analysis could be integrated with machine learning to flag high-risk call patterns in mobile networks. A landmark study focused on detecting fraudulent call behaviors using a combination of call graph analysis and anomaly detection techniques. Their system identified colluding nodes in simulated telecom datasets with over 90% accuracy.

This explored hybrid frameworks combining link analysis and behavior profiling to detect VoIP based fraud. They found that the graph-based approach was particularly effective in identifying spoofed called IDs and virtual number rerouting tactics used in fake call schemes. More recent study proposed a system that leverages both structural and temporal link features to detect call scams, showing how fraudsters from small-world-like networks with tight local clustering and short paths between fraud agents.

2.1.3 Hybrid Approaches: Link Analysis and ML (Machine Learning)

To address scalability and adaptability issues, many researchers have proposed hybrid models that combine traditional link analysis with modern machine learning and AI. Aggarwal and Yu discussed outlier detection in graph-based data, relevant to fraud detection, where fraudsters often from "dense subgraphs" distinct from the genera; communication network. Techniques such as spectral clustering and graph embedding have been used in these hybrid models. It introduced a framework using graph convolutional networks (GCNs) for learning patterns un telecommunication fraud. These neural models, trained on graph-structured data, have outperformed traditional models in recognizing previously unseen fraud patterns.

Other works have shown that integrating call frequency features, time-window analysis, and topological link metrics can significantly improve detection results and reduce false positives.

2.2 Structural Analysis

Several studies employ deep learning models like RNNs (Recurrent Neural Networks), GRUs (Gated Recurrent Unit), Bi-LSTMs, and CNNs for crime detection and prediction, showing improved accuracy in recognizing patterns and forecasting criminal activities. To gain a comprehensive understanding of this domain, the quantitative analysis is done through a bibliometric analysis. It uncovers the trends, collaborations, and impactful contributions that have defined the domain in the last five years. Combining traditional machine learning with sentiment analysis (e.g., BERT models) or integrating external data sources like social media and weather data has been effective in enhancing crime detection. Research is expanding into predicting and preventing cybercrimes and Internet of Things (IoT)-based attacks, especially with the help of intrusion detection systems powered by deep learning. Publications used in the following research by the authors. The chart derived from the same data, showing the distribution of publications across various sources.

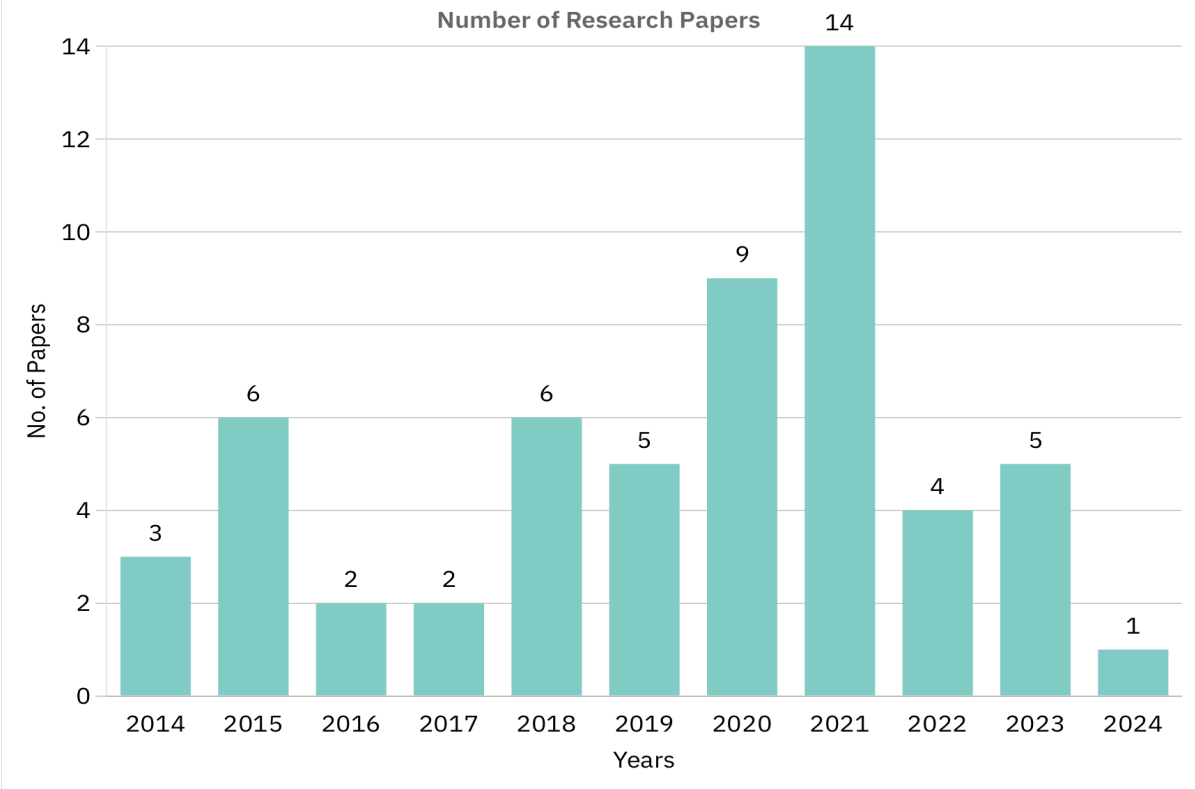


Figure 4: Year-wise record count of research papers

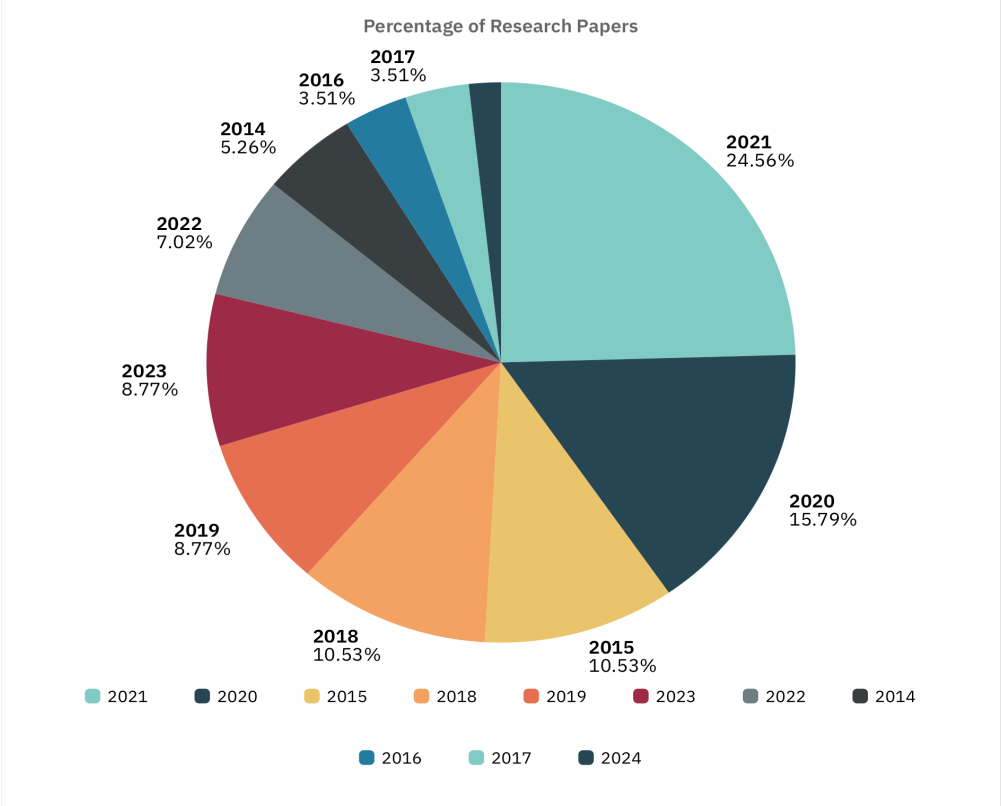


Figure 5: Pie count of research publications

2.3 Detailed Review of Prominent Link Analysis Methods

This section offers a comprehensive exploration of the publication trends and productivity analysis applied to Link Analysis methodologies through bibliometric studies conducted over the last five years. The quantitative analysis plays a crucial role in this context, providing the scholarly contributions, trends, and advancements within the Link Analysis field. The publication trend analysis aims to illuminate the growth and evolution of Link Analysis research over time, revealing the increasing significance of crime detection domain in the digital age. This investigation facilitates a deeper understanding of the academic engagement and the spread of knowledge pertaining to the Link Analysis domain. Simultaneously, the productivity analysis assesses the impact and scope of crime detection and fraud call detection research through detailed evaluation of publication outputs from researchers, institutions, and countries. This provides an objective measure of the key contributors and their influence in the domain.

Existing methods in the research of fraudulent call detection with advantages as well as drawbacks are elaborated in this section.

Table 1: Findings of notable Link Analysis methods

Method	Year	Principal Mechanism	Dataset Used	Accuracy	Limitations
Crime rate detection using social media [27]	2020	Sentiment Analysis is used in order to analyze the user's behavior as well as the psychology of people through their tweets is tracked crime actions. Crime rates in different states are measured in India.	Twitter	70%	The use of smarter methods and next-gen filtering tools will increase the accuracy performance.
Building knowledge graphs of homicide investigations [28]	2020	This methodology of creating the knowledge graphs of LAPD (Los Angeles Police Department) case chronologies can help-aid investigators in analyzing homicide case data and also allow for post hoc analysis of the key features that determine whether a homicide is ultimately solved.	Murder Books for homicides in LAPD's 1990 to 2010	0.065	Considerations of fairness, accountability, transparency need to be central to the development of machine learning methods for homicide investigations.
Graph embeddings in criminal investigation [29]	2022	Uses graph embedding models to convert complex criminal networks into structured vector spaces for predictive	Not specified	Balanced precision	Scalability and interpretability tradeoffs

		analysis.			
CrimeGraphNet: Link Prediction in Criminal Networks [30]	2023	Graph Convolutional Networks (GCNs): Aggregate neighbour features using learned filters; embeddings optimized to predict unseen connections (links) between criminal nodes.	Synthetic real-world graphs	F1-score ~87%	Needs labelled data; weak on incomplete networks
HAGEN: Homophily-Aware GCN Recurrent Network [31]	2021	GCN + GRU Layers: Combines graph structure with temporal dynamics; GCN models spatial dependency, while GRU captures sequential crime occurrence trends.	Chicago, NYC crime data	MAE (Mean Absolute Error) reduced by ~12%	High computation cost; not suitable for small departments
Deep Learning Criminal Networks [32]	2023	Feedforward Deep Neural Networks: Nodes represented as vectors; fully connected layers classify links based on learned features. Uses crime label supervision for training.	CrimeNet 2023 Benchmark	Accuracy ~84%	Black-box model limits interpretability
CrimeGNN: Community Detection in Criminal Networks [33]	2023	Spectral GCN + Modularity Optimization: Learns node embeddings and clusters communities based on modularity maximization to	Italian Mafia Network	NMI (Net Monthly Income) ~0.78	Ineffective at overlapping community detection

		reveal tightly-knit crime groups.			
CrimeGAT: Graph Attention Networks for Predictive Policing [34]	2023	Graph Attention Networks (GATs): Attention scores determine which neighbouring nodes are most important for prediction, enabling nuanced link prioritization.	LAPD Crime Graph	ROC-AUC (Area Under the Receiver Operating Characteristic Curve) ~0.89	Memory heavy, especially with large graphs
Graph Computing for Financial Crime Detection [35]	2024	Real-time Subgraph Feature Extraction: Builds ego networks around entities; extracts transaction patterns and uses those for downstream ML or rule-based alerts.	Fintech AML (Anti-Money Laundering) Dataset	Precision up to 94%	Expensive at scale due to subgraph overheads
Graph Computing for Financial Crime Detection [36]	2021	Rule-Based Pattern Matching + Graph Traversals: Encodes criminal typologies as Cypher queries; uses pattern graphs to discover suspicious activities.	SFPD (Secure Flight Passenger Data) Dataset	3× speed improvement	Difficult to adapt to new criminal strategies
Crime Prediction with GNNs and Multivariate Normal Distributions [37]	2021	GNN + Gaussian Estimation: Node embeddings feed into multivariate normal models to estimate spatial crime likelihoods	SFPD Dataset	RMSE (Root Mean Square Error) reduced by ~15%	Needs strong assumptions about data distribution

		probabilistically.			
GNNs for Link Prediction with Subgraph Sketching [38]	2022	Subgraph Sketching + GNN Encoding: For each node pair, a local subgraph is extracted and encoded to predict if a link exists using node/edge pattern recognition.	OCND (Outstanding Contribution to Nation Development)	Top-10 precision ~91%	Omission of global context; struggles with distant dependencies
GNNs for Legal Judgment Prediction in India [39]	2023	Graph Neural Network + BERT-style Embedding: Converts legal documents into node-graph format; encodes textual content and structure for case outcome prediction	Indian Supreme Court Dataset	Accuracy ~82%	Over-reliant on precedent; lacks dynamic legal context
Crime Rate in Banjarmasin using RNN-GRU [40]	2022	Employs a recurrent neural network with GRU layers to model sequential crime rate patterns in a regional dataset.	Banjarmasin crime stats	Good predictive accuracy	Location-specific results
ANN using Twitter + Weather [41]	2020	Integrates artificial neural networks with Twitter sentiment and weather data for crime detection.	Twitter + Weather	Moderate, task-specific	Feature selection may limit generalization
Hybrid Sentiment + BERT [42]	2022	Uses pre-trained BERT to extract sentiment features from text for crime prediction based on emotional context.	Public social data	Accuracy: 94.91%;	Dependent on textual relevance
Intrusion	2021	Builds a deep-	IoT attack	Validation	Focused only on

Detection with Deep Learning [43]		learning-based IDS to detect IoT botnet activity using traffic features and behavior patterns	datasets	Accuracy: 99.94%	cybercrime/IoT
Spatial-temporal Crime Prediction [44]	2020	Uses spatial and temporal features (e.g., coordinates, timestamps) to train ML models for regional crime prediction.	Local crime datasets	Good temporal perform	Limited to structured datasets
Melanoma Prediction (Non-crime, reference) [45]	2021	Uses a CNN model (SqueezeNet) with Bald Eagle Search optimization for imbalanced medical data classification.	Medical images	High accuracy melanoma	Irrelevant to crime domain
Facial Attribute CNN (Non-crime, reference) [46]	2021	CNN trained for multi-label classification of facial attributes, tested on practical datasets.	Facial attribute data	Good multi-label classifier	Not related to crime or prediction tasks
Intelligent Crime Anomaly Detection in Smart Cities Using Deep Learning [47]	2018	Utilizes deep learning models to detect anomalies in urban crime data streams within smart city infrastructure.	Real-time smart city crime data		Lacks detailed performance metrics; generalizability across cities not validated
Crime Intention Detection System Using Deep Learning [48]	2018	Employs LSTM-based deep learning model to detect and classify potential criminal intentions from user behavior patterns	Simulated behavioral datasets	Accuracy: ~87% (reported)	Limited real-world validation; may not generalize to open environments.
DeepRan: Attention-Based	2021	Combines Bi-LSTM with	Malware behavior logs	Accuracy: 98.53%	Focused on ransomware; not

Bi-LSTM and CRF (Conditional Random Field) for Ransomware Detection [49]		attention mechanism and CRF for early ransomware detection based on sequence modelling.			general-purpose crime detection.
Time Series (TS) Analysis for Crime Forecasting [50]	2018	Applies statistical time series models (ARIMA, exponential smoothing) for predicting future crime trends from historical data.	Historical crime reports	RMSE of 17.39	Methodology limited to linear assumptions; lacks adaptive learning.
Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention [51]	2021	Combines ML and computer vision to predict crimes using temporal and visual data. Utilizes classification and image analysis techniques for forecasting.	Public crime and surveillance datasets	RMSE used	Accuracy depends on data quality, real-time image clarity, and generalizability to other urban settings

2.4 Summary

In summary, the domain of Link Analysis has experienced significant growth and innovation, largely driven by the contributions of crime rate detection and fraudulent call detection methods. These advancements have notably enhanced secure data embedding and image restoration techniques. Despite considerable progress in improving EC, encryption strength, reversibility, and redundancy exploitation, the challenge of graphical data among these critical factors persists. Some methods excel in embedding capacity but fall short in quality and security, or vice versa, highlighting the need for continued research and development.

This literature reveals a strong foundation for using link analysis in crime and fraud detection. Pioneering studies in social network theory have evolved into powerful applied tools that aid investigators in uncovering criminal and fraudulent relationships. The integration of machine learning and real-time analytics shows promise in overcoming current limitations, but

more research is needed to make these systems more robust, generalizable, and privacy-compliant. This survey informs the current study's direction by validating the use of graph-based models and hybrid approaches, while also pointing out areas for enhancement-particularly in real-time detection and cross-domain applicability.

As concerns about data security and privacy escalate across various domains, developing fraud call detection methods that seamlessly integrate high embedding capacities, robust encryption, efficient redundancy exploitation, and lossless reversibility remains a crucial objective. Researchers have the opportunity to address the limitations of existing methods and push the boundaries of Link Analysis, thereby tackling the evolving challenges of data security and privacy in an increasingly interconnected world.

CHAPTER 3
A ROLE OF LINK ANALYSIS IN SOCIAL ENGINEERING

Link analysis is an approach of data mining that is used to analyze and evaluate the association between nodes of any network. These nodes may be considered in the form of persons, organizations, and various types of transactions. Link Analysis is a knowledge discovery process for the better visualization of data analysis of data through data analysis. The data analysis will be done for finding out the relationship between web links or associations between people. This technique is also used in security analysis, retail marketing, medical research, and search engine optimization as well as in intelligence. This study mainly focuses on the various approaches, techniques, and issues related to link analysis.

Link analysis is an approach of data mining that is used to find out the relationships between nodes. These nodes may be in the form of people, transactions, and organizations. The link analysis technique is used to investigate various criminal activities such as counter-terrorism, fraud detection, intelligence, medical research, market research, Search Engine Optimization (SEO), etc. Link analysis is mainly used with three objectives such as, to find interesting patterns with the help of data, to find the various types of anomalies through the discovery of new patterns of interest (social network analysis, data mining), and where known patterns are violated. There may be some examples of link analysis such as a criminal network being analyzed by a crime analyst. The relationship and hierarchy between members of the network may be determined using cell phone data. Credit card fraud may be identified using the system, which utilizes the already noted patterns of transactions of each client with his individual information. Systems identify the anomalies and alert the client of potential theft. The new patterns may be identified by the analyst using data on prescriptions and demographics that are emerging as the crisis spreads. Link analysis may also be utilized in biological science to analyze protein interactions. It is also helpful for law enforcement and detects and prevents terrorism networks. It is also used in marketing management especially related to product recommendation analysis. Amazon, Flipkart, and e-bay are famous companies for their recommendation system, which helps customer selections with the help of link analysis. The other approach called Clickstream analysis develops the ability of websites to look and predict every activity of customers to understand their buying behavior and to increase the probability of their purchasing goods with the help of link analysis [1].

3.1 Related Works

This section provides state of the art works in the link analysis domain and these are given as:

3.1.1 Link Analysis and Association Rule Mining

Link analysis may also help provide information on strong association rules about the associative sequences of items according to the transaction database. These association rules can be very much useful for numerous applications such as in retailing marketing, where, new purchasing patterns may be discovered in potential customers according to the rules. The term link analysis is used for the extraction of sequential or non-sequential association rules to organize complex proof. All the techniques of link analysis described here may quickly extract patterns and associations between individuals and their actions such as to reveal the patterns and structure of some illegal criminal networks [13].

3.1.2 Link Analysis and Social Networking Analysis

Link analysis is also associated with social networking analysis. This approach contains various methods to visualize a huge amount of data to identify specific information, such as telephone numbers, interesting and useful patterns, key events, and persons. The link analysis is done manually or with the help of specially designed software to arrange data into some meaningful and easy-to-understand format.

There is an analytical link analysis tool IBM i2 Analyst's Notebook. This tool displays link analysis charts, graphs, and varying displays using huge data. Node XL is also used to analyze the graphical network data entered in Microsoft Excel connected to several social networks including Twitter, Flickr, and other networks as shown in figure. The Node XL aims to identify the conversations, connections, events, and usage patterns. There are many instances where the graphing the data into a visual requires the process of separating the data into individual displays. The Node XL quickly turns all the social networking accounts and phone calls into graphs quick manner with an incomprehensible display of lines and numbers. The creation of visuals represents the specific mode of communication or displays the relationship or associations. The preferred primary dataset is Metadata of communications in these types of visualizations over content because when the data size is large then it is sometimes easier to determine the main contacts through metadata. The main objective of visualization of association is to evaluate the entity's communication with each other and the number of times, they have communicated.

3.1.3 Link Analysis and Counter-terrorism

The Link analysis can be utilized by various law enforcement investigators and intelligence analysts. It helps them examine inconsistencies and anomalies, connect network relationships, and find hidden contacts in the data. It is the first level by which networks of various entities such as bank accounts, telephone calls, email contacts, people, places, organizations, vehicles, and other tangible entities can be identified, analyzed, linked, assembled, examined, and detected.

Link analysis plays a vital role in mapping criminal intelligence and terrorist activity by visualizing the association between entities and events. It is often shown through a chart to map the physical associations between suspects and locations on the network. The technology is often used to answer such questions as who knows whom and when and where they have been in contact? [2].

Link analysis builds the networks of objects connected with their relationships to generate useful and interesting patterns and trends. It utilizes item-to-item associations to generate networks of interactions and connections from defined datasets. Link analysis diagram is also considered entity-relationship-diagrams.

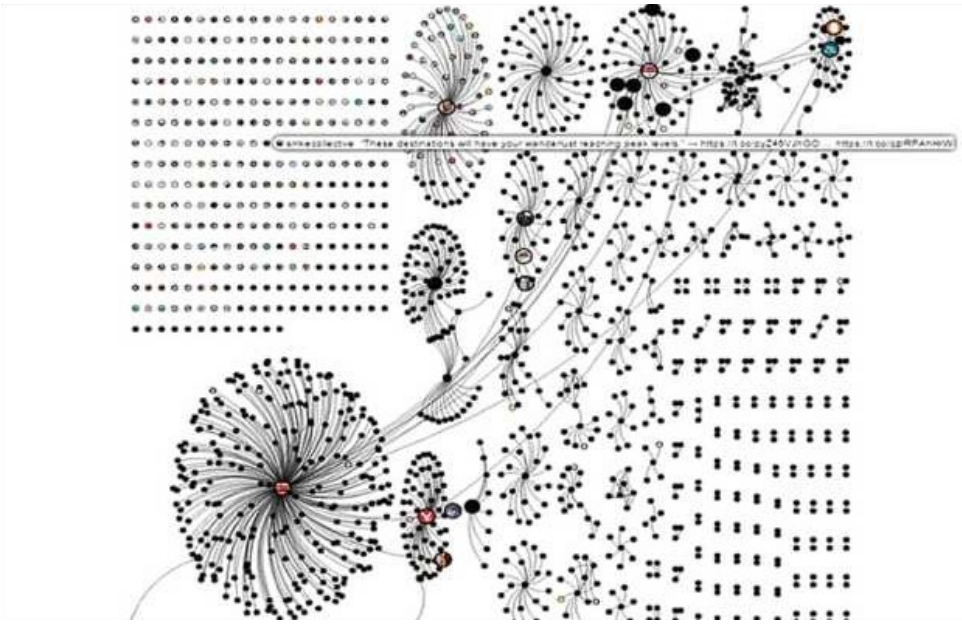


Figure 6: World Heritage Twitter Node XL SNA Map and Report for Wednesday, 16 December- 2015 at 17:07 UTC

It displays the content and structure of information by representing it in the form of interconnected linked objects or entities. An investigator identifies association patterns, new emerging groups, and connections between suspects with the help of link analysis. He may achieve an understanding of the strength of associations and frequency of contacts and discover

new hidden associations through the visualization of these entities and links [3,4].

3.2 Applications of Link Analysis

There are following applications of Network/Link Analysis in various fields such as, in the field of law enforcement criminal analysis, money laundering and call pattern analysis. Customer buying behavior analysis, Segmentation, and collaborative filtering may be possible through link analysis. The link analysis may be applied in the healthcare sector in various fields such as contagion, and disease control.

The informal relationship such as friendship networks, extended families, social interactions, and insider networks may be analyzed through link analysis. The link analysis is also helpful in analyzing the various types of transaction systems, logs, data warehouses, and databases. It also analyzes semi-structured data such as Email content, web pages, data of public records, filings, and unstructured data items such as News items, prospectuses, filings, etc.

3.3 Centralization

This approach expresses the idea of very important actors in a network [5]. It is a measurement of unevenness. When every actor is considered central its range starts from zero for whatever score we are interested in and the ranges are equal to 1 when one node is maximally more central than others. The concept of centrality may be used for various purposes given below.

- To analyze the importance of any node over the other available nodes in the network. For example, to find out the user, who will reach earliest than other users at the time of sharing a news item or a job opportunity?
- To find out the most influenced nodes than the other available nodes. For example, identify the airport affected mostly by the cancelled flights.
- To understand the flow of information and objects. For example, to find out the location of a goods package moving from the warehouse to the delivery address?
- To evaluate spreading criteria through the network most efficiently.
- To identify the nodes to prevent the dissemination of some phenomena. For example, to find out the location of clinics for vaccination to stop the spread of a virus.

***Degree centrality*:** The number of direct connections a node has is called degree centrality. It is used to determine the maximum influential nodes. Consider a scenario of a social network, where the users having many connections will be having a higher degree of centrality.

Equation (1) is used to calculate the Degree of centrality for the node.

$$degCentrality(x) = \deg(x)\,/(NodesTotal - 1) \quad (1)$$

where NodesTotal = Total number of nodes in the network $deg(x)$ = Total number of nodes connected to node x.

Degree centrality can be measured in two forms such as indegree or outdegree. The directed links show information flows between nodes in one direction only. For example, Indegree will be calculated based on the total number of profiles that the user follows, while the outdegree will be calculated based on the number of followers that users have in the social networking scenario.

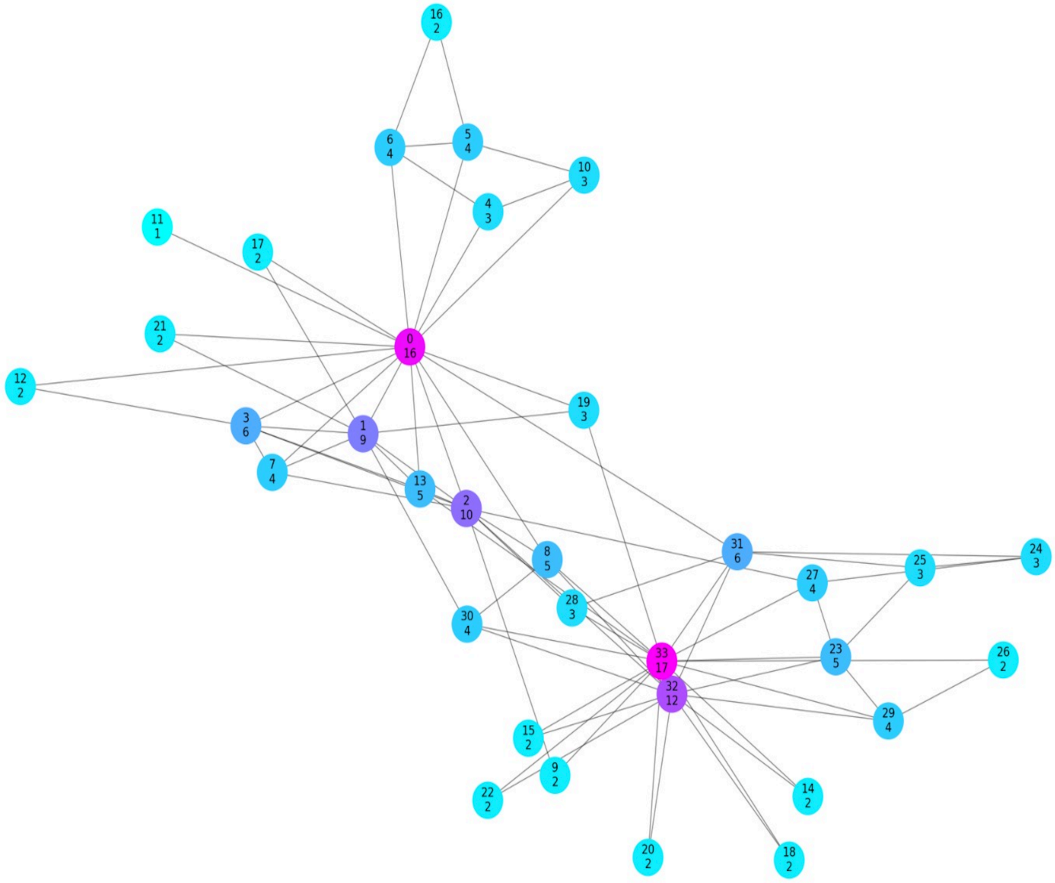


Figure 7: Degree Centrality for Zachary's Karate Club Network [6].

Equation (2) is used to calculate the degree of centrality:

$$indegCentrality(x) = \text{indeg}(x)\,/(NodesTotal - 1) \quad (2)$$

where NodesTotal=nodes available in the network and

indeg(x) = nodes connected to node x. Equation (3) is used to calculate Outdegree.

$$outdegCentrality(x) = \text{outdeg}(x)\,/(NodesTotal - 1) \quad (3)$$

NodesTotal = The number of nodes in the network

outdeg(x) = The number of nodes connected to node x with flow directed away from node x.

Degree centrality is shown for the Karate club network in Figure 7. where each actor is labelled with his score of directed degree centrality.

Integrating structural metrics with temporal behavior, pattern irregularities, and machine-learning-based link evaluation allows investigators to detect evolving fraud strategies more effectively. This holistic approach ensures that even nodes with low connectivity—often the most deceptive and dangerous—are properly evaluated for their influence and involvement in fraudulent activities. Thus, considering rare but meaningful connections is vital for understanding the true structure and influence distribution within fraudulent call networks.

3.3.1 Betweenness Centrality

This centrality is used to identify the nodes that connect other nodes in terms of the shortest path. For example, in a social network, if a user has connections to multiple groups of friends will be evaluated as highest Betweenness centrality than those users, who have connections in only one group. Equation (4) is used to calculate the Betweenness centrality of node x.

$$btwCentrality(x) = \sum a, b \in Nodes \ (\text{paths a, b(x)} \ / \text{paths } a, b) \quad (4)$$

Nodes = The number of nodes in the network The paths a, b are shortest paths between nodes a and b(i.e. number of edges) [16]. The number of shortest paths between nodes a and b that connect through node x An undirected chart may be normalized with the help of the following (5).

$$1/2(NodesTotal - 1)(NodesTotal - 2) \quad (5)$$

where NodesTotal = The number of nodes in the network. The Betweenness centrality is shown for Zachary's Karate club network in Figure 8.

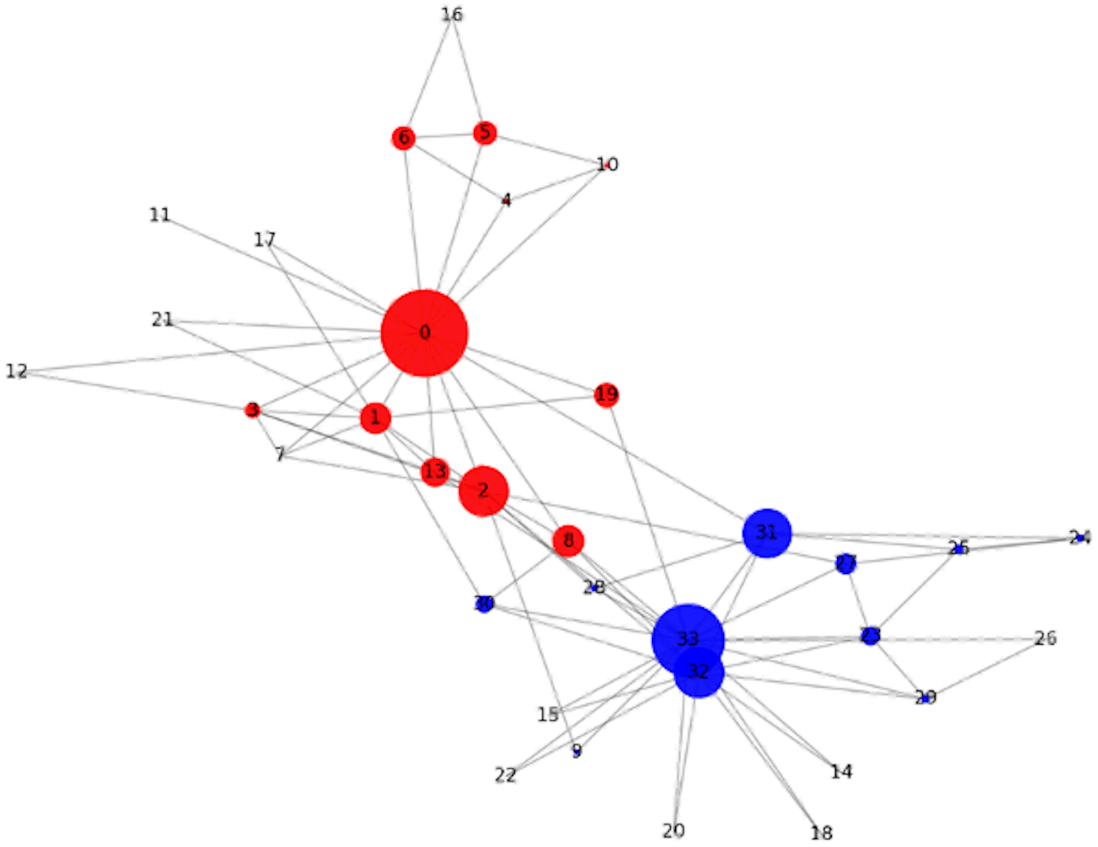


Figure 8: Betweenness Centrality for Zachary's Karate Club Network [6].

3.3.2 Closeness Centrality

This closeness centrality is computed based on the average shortest network path distance among different nodes in the network. It should be used to find out the maximum associated nodes in the network. When a user has a large number of connections in any social network then he will have a higher closeness centrality. Equation (6) is used to calculate the closeness centrality of node x.

$$cc(x) = \left(\frac{nodes(x,y)}{NodesTotal-1}\right) * \left(\frac{nodes(x,y)}{dist(x,y)Total}\right) \quad (6)$$

where cc(x)=CloseCentrality(x)

NodesTotal = Total number of available in the network, $nodes(x,y)$ = Total number of nodes that are connected to node x

Distance(x,y) = The sum of the shortest path distances from node x to other nodes.

Closeness scores for Zachary's Karate club network are shown in Figure 9.

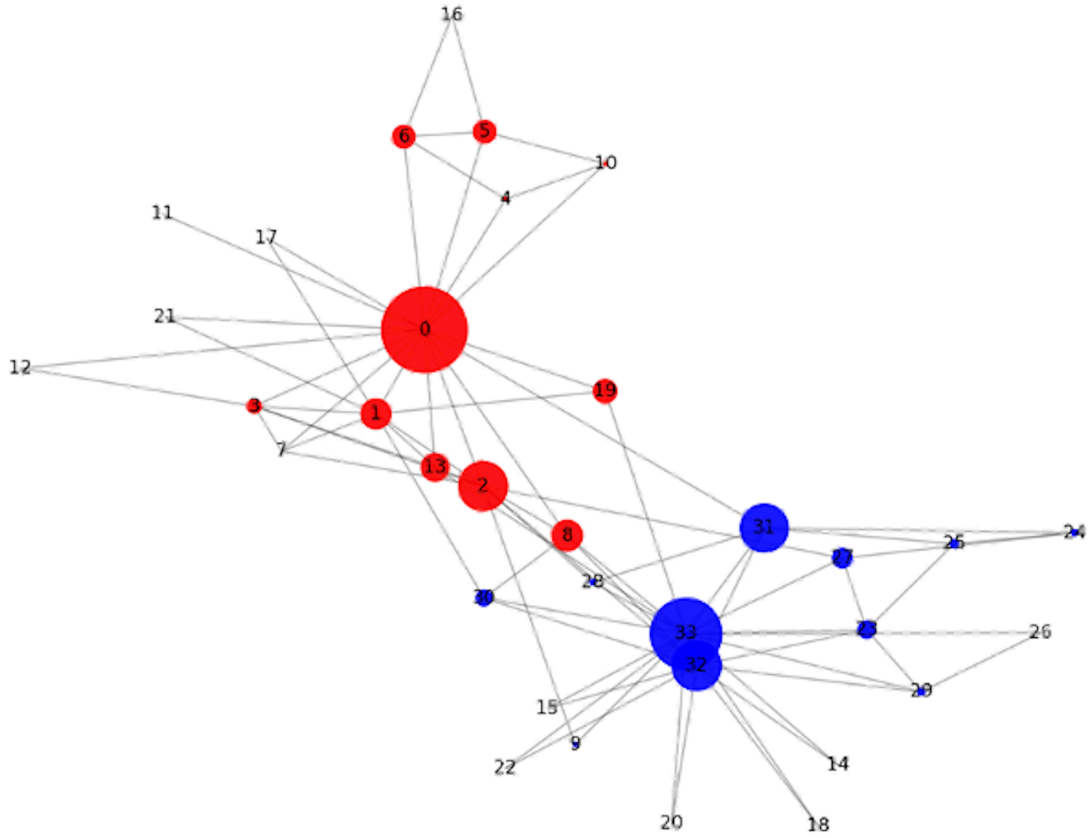


Figure 9: Closeness scores for Zachary's Karate Club Network [6]

3.3.3 Eigenvector Centrality

The nodes, which are part of the most influential cluster will be identified with the help of Eigenvector centrality [15]. For example, in the case of social networking, if a user has many connections to other users, then he will have higher eigenvector centrality than others.

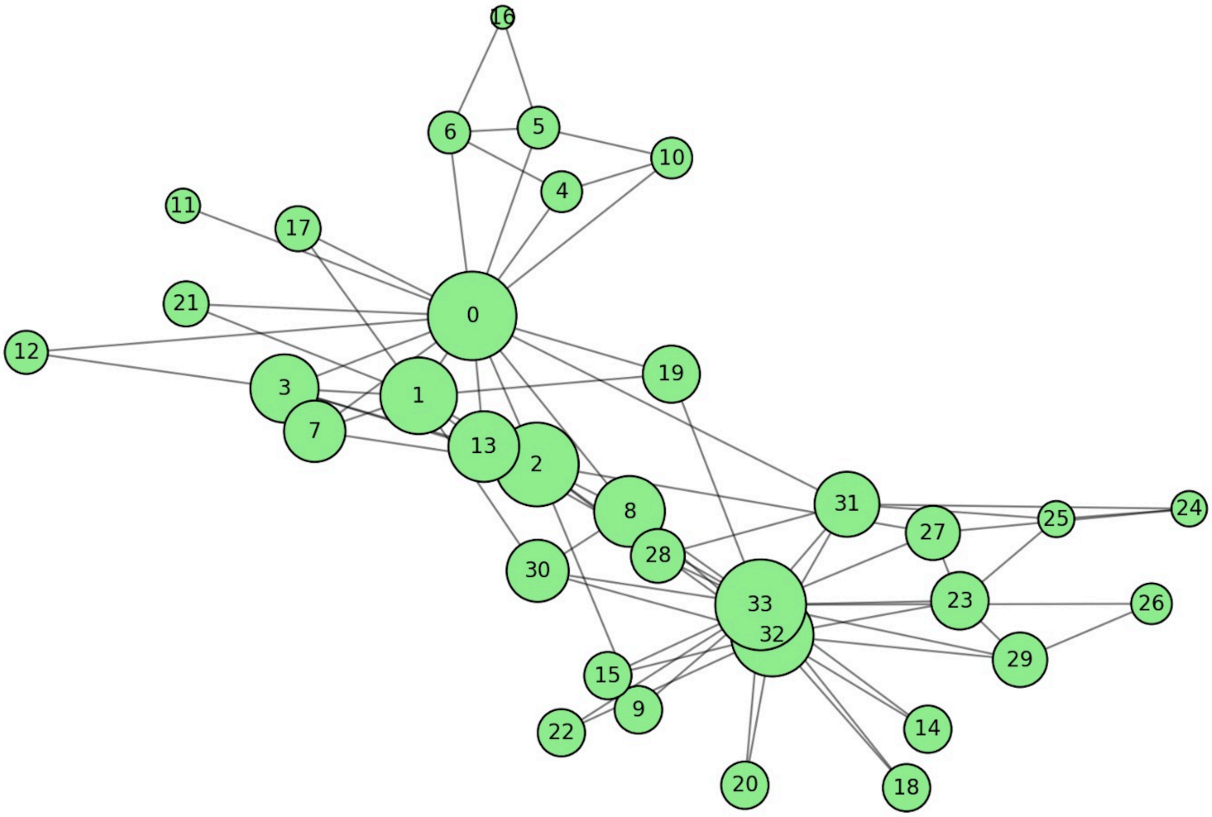


Figure 10: Eigenvector Centrality for Zachary's Karate Club Network [6].

Equation (7) is used to calculate the Eigenvector centrality of node x.

$$Ax = \lambda x \tag{7}$$

where λ = The eigenvalue,

x = The eigenvector and

A = The matrix describing the linear transformation.

Eigenvector centrality values for Zachary's Karate club network are depicted in Figure 10.

3.3.4 Hub Centrality and Authority Centrality

These centralities are used to identify and evaluate the rank and significance of web pages. There may be some web pages that are also known as Hubs that are important because they are linked with many other significant pages known as authorities. A hub node points to many good authorities. An authority node is pointed by many hub nodes. Hub and Authority centrality are used only in the directed graphs. These centralities may be calculated when input data are transactional data.

3.4 Clustering Coefficient

The links exist among the nodes within the neighborhood, divided by the number of links called the clustering coefficient. The degree of clustering is calculated based on information efficiency and its robustness [9,10,11].

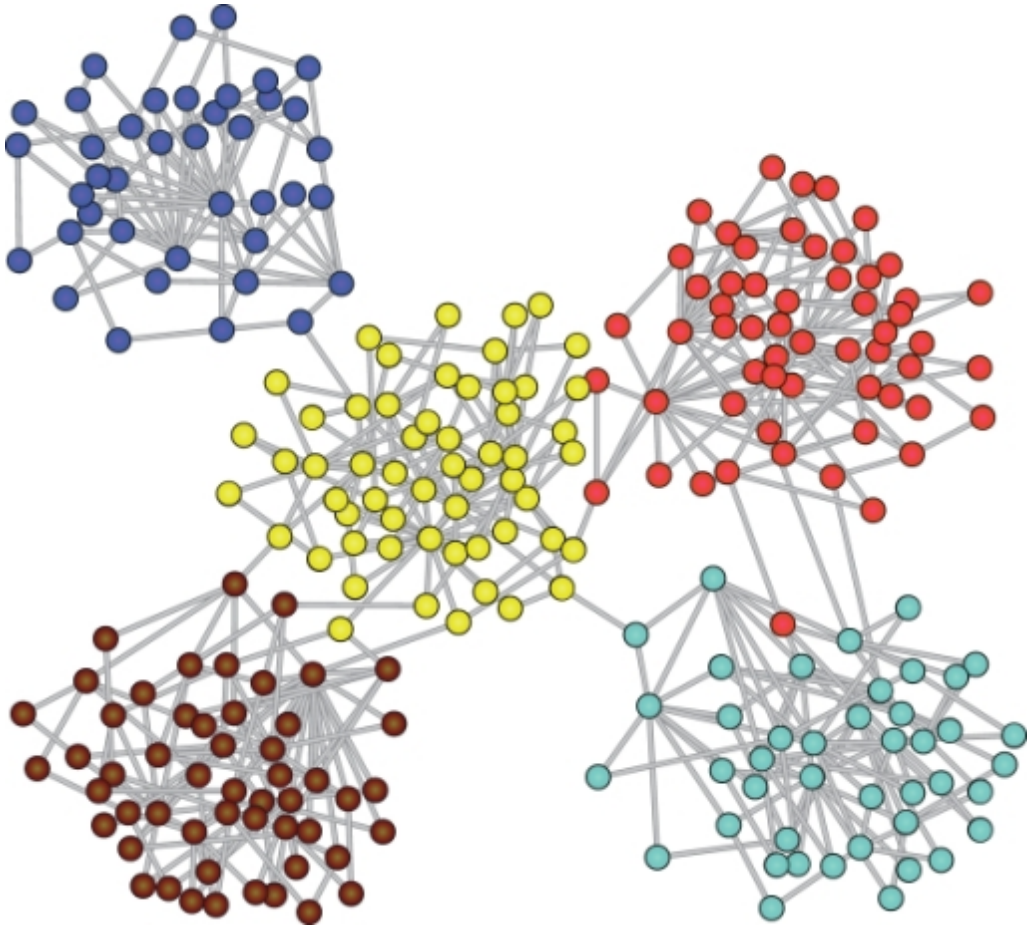


Figure 11: An example of clusters within a network [5]

3.5 Summary

Link analysis is a technique of data mining that is especially used to detect useful and interesting patterns. The first challenge in link analysis is to reduce the graphs into manageable portions [7]. In Market-Basket analysis, where intelligent searching of interesting items is done by removing unwanted elements, the link analysis may be helpful to reduce the graphs so that the analysis is manageable. Therefore, graph reduction techniques are needed to be applied to the graphs to obtain meaningful relationships. The first challenge is identifying the interesting relationships and deciding to reduce the graphs. Therefore, determining how to apply the link

analysis techniques to detect abnormal and suspicious behavior is needed. The other challenge in counter-terrorism analysis is handling the situation with partial information. The research efforts on link analysis have to be conducted for efficient use in counter-terrorism. The availability of good data is another challenge for link analysis in data mining.

CHAPTER 4

SOCIAL NETWORK ANALYSIS FOR CRIME RATE DETECTION USING SPIZELLA SWARM OPTIMIZATION BASED Bi-LSTM CLASSIFIER

Technical advancements initiated a rapid increase in criminal activity over time. For the prevention of these criminal activities, preventive measures are needed. In order to monitor these illicit activities and enhance public safety, crime rate detection is essential. Social media can be used to identify crime rates in various parts of any nation, which can dramatically lower crime rates. Social media are a source of information as well as a tool for communication. Twitter, which has a user base of more than 300 million, makes a suitable choice for data analysis. The Spizella swarm based Bi-LSTM classifier is used for the detection of crime rate in this research. While performing social network analysis using the Bi-LSTM classifier for the determination of crime rate providing faster convergence is a crucial factor and this faster convergence is achieved by the proposed Spizella swarm optimization. Bi-LSTM classifier effectively identified the crime rate and the Bi-LSTM performance is boosted by the Spizella swarm optimization where the escaping characteristics of Spizella improve the convergence and help in attaining desired results Additional training is given by the Bi-LSTM classifier by traversing the outputs and this Bi-LSTM classifier are more efficient in the text classification. Measuring the metrics values for accuracy, sensitivity, and specificity demonstrates the effectiveness of the proposed method and the proposed Spizella swarm optimization achieved an improvement of 0.5%, 1.16%, and 1.08%, which is more efficient.

4.1 Introduction

Modern-day nations' top priorities include achieving security intelligence and a focus on artificial intelligence-based information security that tries to aggregate and organize all data connected to cyberspace risks [12] This necessitates extensively detecting and examining both crime patterns and trends [14]. By detecting large attacks earlier, such profile enables pertinent protection and incident avoidance [8], because handling crime computationally, supports the ability of investigation authorities, and crime prediction has grown in prominence in recent years. Better predictive algorithms are required to focus police patrols on criminals [2][15].

A broader analytical perspective on crime has been generated by technological advancement

in all aspects of human life. Researchers from various scientific fields have been conducting extensive scientific investigations on the origins, structure, intensity, and dynamics of crime [4] [5][36]. Criminal activity drastically rises throughout all 34 nations every 34 years. For the prevention of these criminal activities, resilient measures are needed. In order to keep an eye on these illicit activities and enhance public safety, crime rate detection is essential. Social media can be used to identify crime rates in various parts of any nation, which can dramatically lower crime rates. Social media are a source of information as well as a tool for communication [1][18][19]. Numerous crimes occur every second in various locations, according to various patterns, and at various times, and the number is constantly rising. Crimes can be divided into various categories, such as murder, theft, rape, kidnapping, and so forth [4]. Crime analysis uses historical data already in existence to forecast the potential timing and location of a crime. The circle theory, crime journey, routine activity theory, and centrography are examples of common methods for predicting crime.

Future events are predicted over time using a time series (TS) of observational data with regular or irregular spacing and other exogenous factors affecting crime. Fortunately, the ability to monitor criminal activity has improved, allowing authorities to collect and archive detailed information on criminal activity, including its location and timing. Law enforcement agencies may be able to employ data analytic technology to extract crucial information about criminal events in order to efficiently deploy their resources and develop effective crime prevention strategies [16] [36].

Crime network visualization, risk reduction, and increased analyst productivity are all aided by crime prediction techniques [4]. Due to the fact that crime rates are still rising, it may be necessary to do some significant study that will inform decision-makers and the relevant department about problems and challenges related to crime prediction and control methods [5]. Its applications are crucial because it primarily evaluates information about the surrounding area obtained from a camera. It can be used for object identification, location determination, augmented and mixed realities, face and license plate recognition, and number plate recognition [9][23]. Machine learning (ML) enables systems to automatically learn from their prior performance without explicit programming [9][24].

The research's ultimate goal is to use data from Twitter to study the crime rate. Data is gathered from established databases, and preprocessing is then applied to the data to improve its suitability for analysis. Following the extraction of the data's significant features and processing of those features using the Bi-LSTM classifier, the primary contributions are interpreted as follows:

Spizella swarm optimization: The Spizella swarm optimization is developed by the

standard hybridization of the sparrow search optimization (SSO) [29] and particle swarm optimization (PSO) [30] algorithms. The convergence of the Spizella is not strong enough for escaping from the predators, which acts as a disadvantage hence to improve this the velocity of the swarms is hybridized with the escaping characteristics of Spizella which improves the convergence and helps in attaining desired results. This tuning effectively helps in sorting the information based on the similarities and dissimilarities that helps in the effective analysis of the social networks.

Spizella swarm optimization based Bi-LSTM classifier: The Bi-LSTM has the capability of analyzing complex texts and the traverse analysis possesses special significance. The parameters such as weights and bias in the Bi-LSTM classifier are optimally tuned using the Spizella swarm optimization further improving the convergence of the classifier and analyzing the texts more efficiently.

The arrangement of the paper is represented as follows: The existing works relevant to crime rate detection are analyzed in sec 2. The Spizella swarm optimization based Bi-LSTM classifier along with the optimization algorithm is comprehensively interpreted in sec 3. The results obtained using the proposed models are detailed in sec 4 and the final research conclusion is provided.

Although it may appear that a single predictive algorithm is responsible for detecting crime rates, the proposed framework in this thesis actually relies on the combined functioning of a classifier (Bi-LSTM) and an optimization algorithm (Spizella Swarm Optimization). The Bi-LSTM classifier performs the core prediction task by analysing cleaned and pre-processed Twitter text data, learning temporal and contextual patterns in crime-related language, hashtags, sentiments, and semantics. Unlike simple algorithms that rely only on word frequencies or static features, the Bi-LSTM processes information in both forward and backward directions, allowing it to capture deeper relationships within sentences and detect subtle cues that correspond to crime-related behaviour or trends. This ability to learn long-range dependencies makes it well-suited for interpreting social media text, where crime signals often appear in fragmented or context-dependent expressions.

The optimization algorithm does not *predict* crime rate directly but enhances the performance of the Bi-LSTM by fine-tuning its internal parameters. Spizella Swarm Optimization adjusts weights and biases to improve convergence, stability, and classification accuracy, ensuring the model detects crime instances more effectively in large and noisy datasets. Traditional Bi-LSTM models may suffer from slow training, sub-optimal parameter settings, or overfitting, especially when dealing with unstructured social media data. Integrating an optimization algorithm helps overcome these limitations, resulting in a more reliable and

efficient detection framework. Thus, crime rate prediction is ultimately achieved through the *combined effect* of deep learning (for detection and classification) and swarm optimization (for performance refinement), not by a single standalone algorithm.

4.1.1 Motivation

The main goal of crime analysis is to assist or support a police department's operations. These activities include patrolling, patrolling operations, crime prevention and reduction methods, problem-solving, evaluation and accountability of police actions, criminal investigation, arrest, and prosecution. Crime rate detection aids law enforcement organizations in forecasting and identifying the crimes that occur in a particular region, which lowers the crime rate. Analyzing the current crime trends and forecasting future crime rates helps to minimize the crimes that occur or determine the suspected individuals. The authorities can take responsibility and attempt to lower the crime rate based on this information. To predict the crime rate based on social networks this research aims in developing a new framework and gain deep perception the existing methods are analyzed and the observations are interpreted below.

4.2 Related Works

In this section, the existing works are reviewed for providing a comprehensive view of the methods available for crime rate detection.

Twitter data gathered from seven different locations from various Twitter accounts was examined [1]. In order to track criminal activity, sentiment analysis was employed to examine user behavior and psychology in tweets. The online conversational text uses the Twitter part-of-speech tagger, which was also employed for analyzing the sentiments. For an extensive collection of unlabeled tweets, brown clustering is utilized. The method identified crime rates, but the level of detection accuracy is insufficient. [2] used multiple classifiers for the detection of crime rates in social networks. Additionally, the study offered a visual summary of different kinds of crime and the number of crimes using exploratory data analysis. Finally, an autoregressive integrated moving average was used to assess the predicted crime rate and crime density areas for the following five years (ARIMA) but the model is computationally expensive. [3] analyzed the crime data using different machine-learning approaches that monitored how the economic crisis affected crime in India. In districts of several Indian states, the relationship between theft, robbery, and burglary statistics as well as the gross domestic product and unemployment rates are analyzed and the conclusion was made that there is a one-way causal relationship between the unemployment rate and robberies. There is a high correlation in the target variables, which

acts as a disadvantage. Another study [4] focused on making predictions about the types of crimes that would happen based on the places where they have already occurred. In this method, the training data sets are used and these training data undergo data cleansing and transformation. Machine learning was utilized for initiating the model and data visualization is used to implement analysis of the data set and its properties. Numerous elements are being recognized and recorded to keep society secure, risk variables are being discovered and predictive methods are being developed. The research doesn't carry any optimization for better application of the method to a huge amount of data. Sharon Susan Jacob and R. Vijayakumar [6] used a clustering method based on a machine learning (ML) algorithm to study the sentiment of Twitter tweets. Twitter comments were grouped into three groups: good, negative, and neutral. The authors employed a sentiment-sensitive dictionary-based machine learning strategy in this situation that identified hate speech from real-time tweets on Twitter and there is a necessity for the method that should examine opinions even in the absence of an object. In order to discover the proper forecasts of crime by adopting learning-based methods, [5] innovated assemble-stacking model for the crime detection. The SVM technique is used to create domain-specific configurations, and it is discovered that the model has better prediction ability than earlier studies that only looked at crime datasets based on violence and were used as baselines. The outcomes demonstrated that criminological theories can be reconciled with any empirical data on crime. With BERT serving as the deep learning model, Mohammed Boukabous and Mostafa Azizi [8] employed a hybrid strategy that fused lexicon-based and deep learning techniques. Using a set of normal and crime-related lexicons, the author used the lexicon-based strategy to label the Twitter dataset. Then the labeled dataset is used to train the BERT model. Although the method worked well method was not suitable for audio, video, and images. Umadevi V Navalgund and Priyadharshini.K [37] developed a system using VGGNET19that controlled the crime by detecting the gun and knife in a person's hand, which takes less training time in detecting the weapons but this model failed to identify the upcoming crime. Sohrab Hossain [38] initiated a decision tree and k-nearest neighbor for predicting the crime by analyzing the previous criminal records in the dataset which provided higher accuracy but if the class was poorly imbalanced means machine learning not performed well in the original dataset.

4.2.1 Challenges Addressed

The challenges present in the research is enumerated as,

- There is a lack of time to maintain the criminal data, due to the fact that crime rates are rising. Analyzing these criminal activities with greater accuracy poses challenges [3].
- Tuning the classifiers using optimization and acquiring a high accuracy is challenging.

- Bi-LSTM takes a larger time to train the data because of the presence of a vast amount of information and providing faster convergence initiates challenges [8].
- Bi-LSTM is intuitive to random weight initialization, hence tuning the weight of the classifier is also a significant and crucial step [4].
- Gathering a significant number of important features that are useful in determining the crime rates also poses' difficulties [5].

The above-mentioned methods are affected in terms of accuracy, where the detection accuracy is lower and most of the methods are computationally expensive. In this research, the crime rate is detected with high accuracy, because of the enabling of the Spizella swarm optimization in the Bi-LSTM classifier. The effective optimization tuned the parameters and helped in the selection of the optimal parameters, which helps in reducing the computational complexity. Determining the correlation acts as an important challenge and the correlation between the words is finely determined using the implemented method. The feature selection also plays a significant role and the effective features are extracted in this research for effective crime analysis.

4.3 Proposed Methodology

The aim is to identify the crime rates through the analysis of Twitter data. In this research Spizella swarm optimization based, Bi-LSTM classifier is proposed for the analysis of Twitter data. Here, the Twitter data is collected from the repositories and then the preprocessing of the collected data is performed. The preprocessing of data is performed through the removal of irrelevant words, the removal of punctuations, the removal of alphanumeric characteristics, and the removal of stop words. Along with that tokenization, normalization, and stemming is also performed in the preprocessing stage. After preprocessing the necessary keywords from the data are collected, which consist of the words like fight, crime, and so on. From the keywords collected a sentimental link graph-based data representation is established for providing a comprehensive interpretation of the crime rate. After the representation, the Bi-LSTM classifier is used for the analysis of the crime rate based on Twitter data. The Bi-LSTM classifier is used because the classifier enables additional training by traversing the input data twice, which makes the analysis more efficient. Furthermore, the classifier is effectively optimized using Spizella swarm optimization, where the hybridized characteristics provide better foraging characteristics that support fine-tuning the Bi-LSTM classifier's weight and bias parameters. The research is conducted using Python, and the tuning produces the desired and correct output. Accuracy, sensitivity, and specificity measure the efficiency of the research. The block diagram is illustrated

in figure 12.

4.3.1 Input Data

Here data is collected from the standard datasets such as suspicious tweets [33] and the Sentiment140 dataset with 1.6 million tweets [34], which are mathematically represented as follows,

$$T = \{T_q, T_r\} \tag{8}$$

here, T_q denotes the suspicious tweets and T_r denotes the sentiment 140 datasets, and from this data, the social network analysis is made using the proposed model.

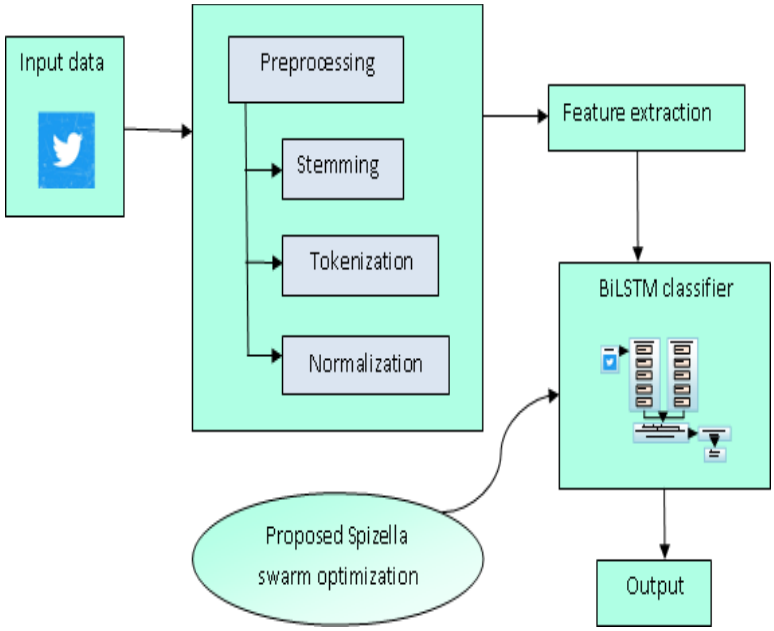


Figure. 12 Schematic representation of the proposed Spizella swarm optimization based Bi-LSTM classifier.

4.3.2 Preprocessing

A significant step in the processing of text is preprocessing. The components of a text can be words, phrases, and paragraphs. Text is a collection of characters that make sense together. The preprocessing approaches are used to provide the text data to a machine learning algorithm in a better version than the original form. In this research, the preprocessing is carried through tokenization, normalization, and stemming and the process is described below sections.

(1) Tokenization: Tokenization is the conversion of text into sentences and sentences to words that are used for reducing the complexity of the text, by categorizing it into tokens. Since a token is a little of a larger whole, a word in a sentence is a token, and a phrase is a token in a paragraph. Tokenization is the process of dividing a string into a list of tokens.

(2) Normalization: After tokenization, normalization is performed, where the root forms of the tokens are derived from the tokens. Normalization is the process of putting tokens into a canonical form that is in line with grammar and the dictionary.

(3) Stemming: The base word is derived using stemming. The contrast between stemming and normalization is the context of the word is derived using normalization and the stem of the word is derived using stemming.

(4) Feature Extraction: The more significant features needed for crime rate detection are done using the preprocessed data using social network analysis are extracted and the necessary features extracted are mathematically represented by,

$$T_{feature} = \{T_q^x, T_r^x\} \tag{9}$$

Here, the important features extracted for the classification are represented by feature and the features extracted from the dataset are represented by x.

4.4 Architecture of Bi-LSTM classifier

The architecture of the Bi-LSTM classifier is shown in Figure 12. In this architecture, input, forwarding, backward, fully connected, and output layers are included. The LSTM handles sequential data by dispersing its weights throughout the data. By utilizing gates that provide long-term dependencies, the Bi-LSTM classifier properly maintains the error gradient and significantly alleviates the issue of vanishing gradient. The Bi-LSTM classifier's mathematical formulation is represented by the equation.

$$Q_z = f(C_k.p_z + N_z.k_{z-1} + d_k) \tag{10}$$

where, the bias of the Bi-LSTM classifier is designated by d_k, $k_{z\square 1}$ identifies the normal hidden state, the current word embedding vector is denoted by p_z, the weights are given by C_k and N_z, the nonlinear function is designated by f **,** z, and specifically the $\tan h$ function is utilized. With the help of Spizella swarm optimization, the classifier's weights and bias are modified to their best potential utilizing the hyperparameters.

$$D_z = \sigma(C_D.p_z + N_D.k_{z-1} + d_D) \tag{11}$$

$$E_z = \sigma(C_E.p_z + N_E.k_{z-1} + d_E) \tag{12}$$

$$F_z = \sigma(C_F.p_z + N_F.k_{z-1} + d_F) \tag{13}$$

$$I_z = D_z.l_{z-1} + E_z.\tanh(C_l.p_z + N_l.k_{z-1} + d_l) \tag{14}$$

$$k_z = F_z.\tanh(l_z) \tag{15}$$

here, the input gate, forget gate, and the memory cell is represented D, E and F. I denotes the memory cell, F_z designates the Hadamard product, and the sigmoid function is represented

by. The input gate's memory cell protects the most important information at the moment, whereas the forget gate aids in the forgetting of previous knowledge. Regularly retrieving critical data is made simpler by the output gate, which regulates the data that is present inside the internal memory cell. data.

4.5 Proposed Spizella Swarm Optimization

Utilizing the characteristics of the Spizella and swarms the Spizella swarm optimization is developed, where the convergence of the swarms is combined with the escaping characteristics of Spizella that helps in analyzing the information in an efficient manner. The Spizella swarm optimization is described in detail in the below sections.

Inspiration: The primary source of inspiration for the Spizella swarm optimization is how Spizella search for food. Grain or weeds are among Spizella's primary food sources. Spizella is opportunistic, sophisticated feeders that frequently employ a range of eating techniques, adapting their approaches to best suit the conditions of their environment and available prey at any given time. Foraging strategies, which are described as the acquisition of food through searching, hunting, or gathering of food, are a method used by Spizella to successfully feed them. When a Spizella detects a predator, the foraging strategy is for one or more individuals to chirp, and the entire group to fly away. Similarly, the evolutionary computing method known as particle swarm optimization (PSO) is used for optimization and is stimulated by the social behavior of individuals in big groupings in nature.

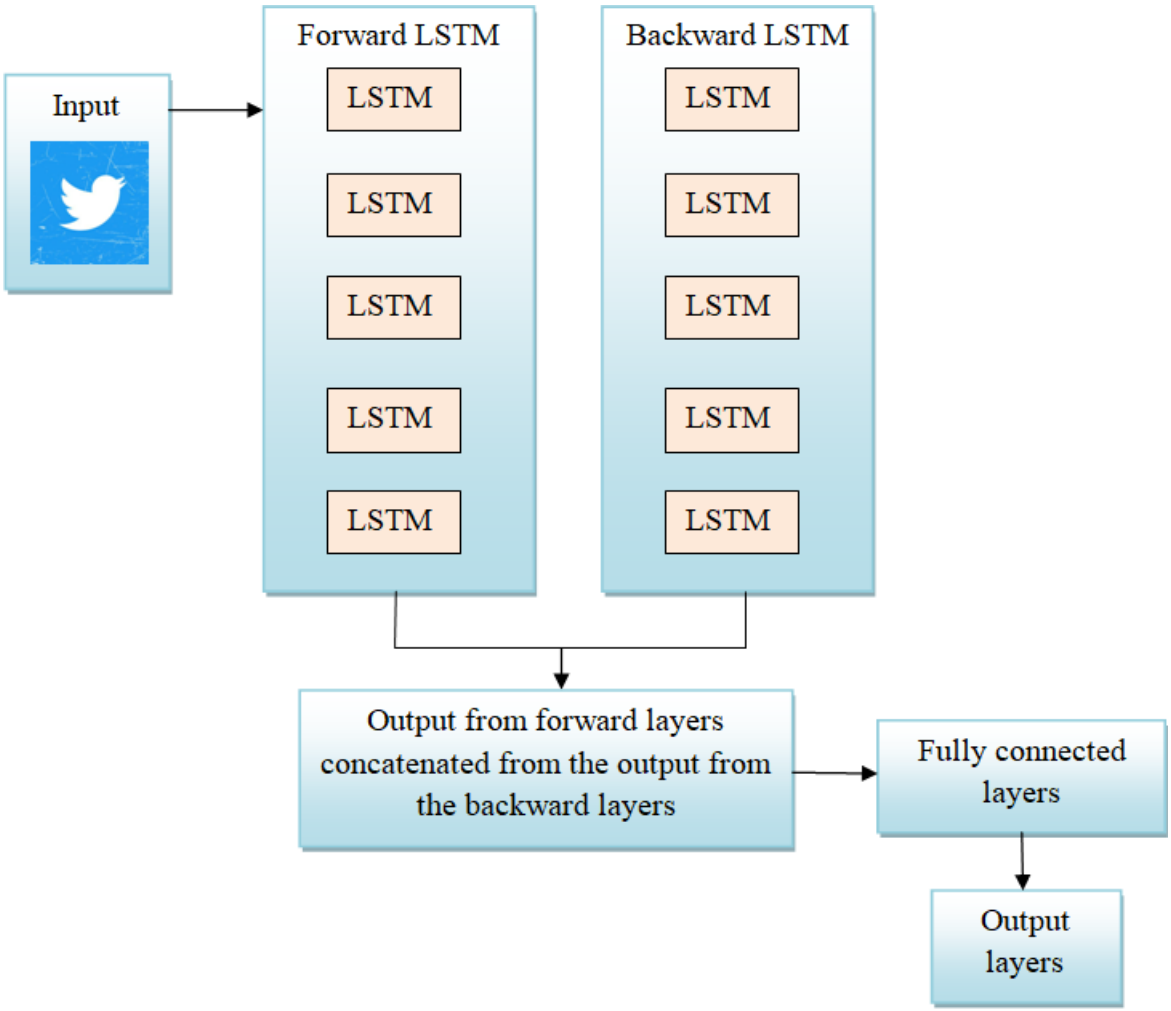


Figure. 13 Architecture of Spizella swarm optimization based Bi-LSTM classifier

To enhance the clarity and reproducibility of the proposed approach, additional implementation details for the key models and optimizers have been incorporated. The Bi-LSTM model, SDH-KGNN framework, and Spizella Swarm Optimization algorithm are now described not only through mathematical formulations but also through step-wise pseudocode representations. These pseudocode blocks summarize the control flow, parameter updates, and model-training procedures, allowing readers to understand the internal functioning of each method without relying solely on code. For example, the Bi-LSTM pseudocode outlines sequence preprocessing, forward–backward propagation, and gradient-based parameter updates, while the optimizer pseudocode captures initialization, position updates, role switching, and convergence criteria. This ensures that the methodology can be replicated or extended by other researchers with ease.

(1) Preliminaries: All schnorrer Spizella have access to the foraging grounds or directions provided by the organizer Spizella, who frequently has a lot of energy reserves. Locating areas with an abundance of food is the responsibility of organizer Spizella. The degree of an individual's energy reserves is assessed based on their fitness ratings. As soon as the Spizella see

the predator, they begin to chirp as an alert message. If the alert value surpasses the safety level, the organizer Spizella must direct all schnorrer Spizella to the safe area. As long as the organizer Spizella looks for better food sources, every sparrow has the potential to become an organizer, but the population as a whole still has the same ratio of schnorrer Spizella to organizer Spizella. The more energetic Spizella would serve as the organizer. Many schnorrer Spizella who are starving are more likely to fly to various regions in quest of food in order to gain energy. Following the organizer Spizella who can provide the best cuisine, Schnorrer Spizella searches for food. Some Schnorrer Spizella may observe the organizer Spizella closely while they wait and compete with them for food to increase their own predation rate. The Spizella near the periphery of the group rush swiftly into the safe area to occupy the best position when in danger.

(2) Mathematical model for the proposed Spizella swarm optimization: The Spizella are selected randomly for determining the food and the position of these Spizella are mathematically represented by,

$$Z = \begin{bmatrix} Z_{1,1} & Z_{1,2} & \cdots & \cdots & Z_{1,j} \\ Z_{2,1} & Z_{2,2} & \cdots & \cdots & Z_{2,j} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ Z_{i,1} & Z_{i,1} & \cdots & \cdots & Z_{i,j} \end{bmatrix} \tag{16}$$

In the above equation, the attribute Z designates the opposition of various Spizella s, i represents the total number of Spizella, j denotes the variable dimensions that are going to be optimized. Spizella fitness values are given by the mathematical notation,

$$Fit_Z = \begin{bmatrix} fit[Z_{1,1} & Z_{1,2} & \cdots & \cdots & Z_{1,j}] \\ fit[Z_{2,1} & Z_{2,2} & \cdots & \cdots & Z_{2,j}] \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ fit[Z_{i,1} & Z_{i,1} & \cdots & \cdots & Z_{i,j}] \end{bmatrix} \tag{17}$$

here, *fit* denotes the fitness solution of various Spizella's. The value of Each row in Fit_Z corresponds to the individual fitness level of Spizella and the food is prioritized for organizer Spizella with higher fitness value during the search process. Additionally, since the organizer Spizella is in charge of locating food and directing how the entire population moves, the organizer Spizella has the capability to search for food in broad areas. The location of the organizer Spizella is given by,

$$Z_{m,n}^{\tau}(org) = \begin{cases} Z_{m,n}^{\tau} \cdot exp\left(\frac{-m}{\partial \cdot I_{max}}\right) & if\ rand < A^{th} \\ Z_{m,n}^{\tau} + R.S & if\ rand < A^{th} \end{cases} \quad (18)$$

The current iteration of the Spizella is represented by τ, $Z_{m,n}^{\tau}$ denotes the value of the n^{th} dimension of the m^{th} sparrow during the iterationτ , where $n = 1,2,3,.....j$. *rand* and δ denotes the random number ranges from$(0,1)$. The count of the number of iterations, when it reaches the *th* maximum iterations is represented as I_{max}. A represents the alarm value and the safety threshold values and the value will be in the range $(0.5,1)$, S is a matrix in the dimension $(1 \times j$), and the elements present in the matrix values will be 1. R represents a random number. The organizer Spizella switches to wide search mode when $rand < A^{th}$ which indicates that there are no nearby predators. All sparrows must immediately fly to other safe regions if $rand \geq A^{th}$, which indicates that some Spizella has spotted the predator.

Competing behavior: Some schnorrers Spizella keep a closer eye on the producers, and they right away leave their present place to fight for food as soon as they learn that the organizer Spizella has found food. If they prevail, they can instantly obtain the organizer Spizella food otherwise, they must adhere to the rules (11). The position update of the schnorrer Spizella is given by,

$$Z_{m,n}^{\tau+1}(sch) = \begin{cases} R.exp\left(\frac{Z_{w}^{\tau} - A_{m,n}^{\tau}}{m^2}\right) & if\ m > {}^{i}/_{2} \\ Z_{best}^{\tau+1} + \left|Z_{m,n}^{\tau} - Z_{best}^{\tau+1}\right| M^{+}.S & else \end{cases} \quad (19)$$

here, the global worst solution of the Spizella is represented by Z^{τ} and the best solution obtained

by the Spizella is denoted by $Z^{\tau+1}$. When the condition $m > {}^{i}/_{2}$ is met then the m^{th} Spizella is considered to be the most starving. M denotes a matrix that consists of random elements within $(1,1)$ the dimension $(1 \times n)$and is given by,

$$M^{+} = M^{T}\left(MM^{T}\right)^{-1} \quad (20)$$

Discerning behavior: Around 10 to 20 percent of Spizella are aware of the danger from the predator lying in the Spizella group. The position of these Spizella are assigned and the Spizella position update is denoted by,

$$Z_{m,n}^{\tau+1} = \begin{cases} Z_{best}^{\tau} + \lambda |Z_{m,n}^{\tau} - Z_{best}^{\tau}| & if\ f_m > f_{best} \\ Z_{m,n}^{\tau} + P\frac{|Z_{m,n}^{\tau} - Z_{best}^{\tau}|}{(f_m - f_w) + \zeta} & if\ f_m = f_{best} \end{cases} \quad (21)$$

where Z denotes the current global best solution, the step size control parameter is represented by λ , P denotes the random number and the random number is in the range $-1$and

1. Here, f_m stands for the current Spizella fitness value f_m and f_w, denotes the current global best and worst fitness values respectively. To prevent a zero division error the smallest constant is ζ

For the simplified view if the condition $f_m > f_{best}$ then it denotes the Spizella is at the edge of the group. When $f_m = f_{best}$ represent that the middle of the Spizella is aware of the danger. The step size coefficient and the sparrow's movement direction are both indicated by the attribute P.

Velocity of swarms: The swarms keep track of a population of particles in various positions that helps in achieving the global best solution. Every particle in the population has an adjustable velocity that determines how quickly it goes through the search space. Using the objective function, each position in the population is assessed to provide some indication of its suitability for the problem. The velocity factor of the swarm is given by,

$$G_{t+1}^{s} = I_{weight} G_t^s + a_1 rand(B^{ib} - Z_t^s + a_2 rand(B^{gh} - Z_r^s)) \quad (22)$$

$$Z_{t+1}^{s} = Z_t^s + G_{t+1}^s \quad (23)$$

here, G designates the velocity of the swarms during the iteration τ, and the position of the swarm is denoted by Z^s, the inertia weight is designated by the attribute I and the inertia weight value is in the range $[0.4,1.4]$. The self-confidence and the swarm confidence factors are given by a_1 and a_2, which is in the range $[1.5,2]$ and $[2,2.5]$. The best solution for the individual is given by B^{ib} and the best global solution is represented by B^{gb}.

Enhanced convergence: The Spizella has the capability to sense danger but the speed of escaping is not enough to escape from the predators. For providing good convergence, the velocity factor from the swarms is updated for effective escaping from the predators. The velocity update consists of the inertia weight factor, which helps in providing better convergence, and the updated equation is interpreted as follows:

$$Z_{m,n}^{\tau+1}(best) = \begin{Bmatrix} 0.5[Z_{best}^{\tau} + \lambda|Z_{m,n}^{\tau} - Z_{best}^{\tau}|] + 0.5Z_{t+1}^{s} & if\ f_m > f_{best} \\ 0.5[Z_{m,n}^{\tau} + P\frac{|Z_{m,n}^{\tau} - Z_w^{\tau}|}{(f_m - f_w)+\zeta}] + 0.5Z_{t+1}^{s} & if\ f_m = f_{best} \end{Bmatrix} \quad (24)$$

Although the proposed Spizella Swarm Optimization (SDHO-based enhancement) is inspired by existing PSO, GA, and swarm-intelligence meta-heuristics, its novelty lies in the way it combines exploration–exploitation dynamics using behaviour patterns that are not present in classical algorithms.

Unlike PSO, which relies purely on velocity updates driven by personal and global best positions, the proposed method incorporates adaptive escape-and-chase behaviour motivated by the natural anti-predator movement of Spizella bird species. This introduces a dynamic adjustment mechanism that enables particles to rapidly leave local minima when stagnation is detected—an ability that classic PSO lacks. Similarly, while GA employs crossover and mutation to diversify the population, the proposed method achieves diversification through ecological role switching and predator-avoidance strategies, reducing computational burden while maintaining efficient search diversity.

The algorithm also differs from standard heuristics in how it tunes Bi-LSTM parameters. Traditional PSO/GA methods often suffer from inconsistent convergence on high-dimensional deep learning parameter spaces, whereas the proposed optimization integrates a selective intensification mechanism that prioritizes parameters showing strong gradient contribution. This hybrid ecological strategy improves stability, prevents premature convergence, and reduces oscillation during training. Furthermore, the algorithmic structure is designed to be lightweight with fewer hyper-parameters compared to PSO or GA, improving reproducibility and reducing sensitivity to initialization. Hence, while grounded in the broader family of nature-inspired optimizers, the method demonstrates clear structural and behavioural differences that justify its novelty and suitability for crime-rate and fraudulent call detection tasks.

Table 2: Pseudo code for Spizella swarm optimization

S. No.	Pseudo code for the Spizella swarm optimization
1	Input: Z
2	Output: $Z_{m,n}^{\tau+1}(best)$
3	Initialize: Z
4	Determine fitness function: Fit_z
5	Position update of organizer Spizella: $Z_{m,n}^{\tau+1}(org)$ #Foraging
6	Position update of organizer Spizella: $Z_{m,n}^{\tau+1}(sch)$ #Competing Behavior
7	$If\ m > {}^{1}/_{2}$
8	$Z_{m,n}^{\tau+1}(sch) = R.exp\dfrac{Z_w^{\tau} - A_{m,n}^{\tau}}{m}$
9	Else
10	$Z_{m,n}^{\tau+1}(sch) = Z_{best}^{\tau+1} + \left\lvert Z_{m,n}^{\tau} - Z_{m,n}^{\tau+1}\right\rvert .M^{+}.S$
11	Update position of Spizella: $Z_{m,n}^{\tau+1}$ #Discerning behavior
12	$if\ f_m > f_{best}$
13	$Z_{m,n}^{\tau+1} = Z_{best}^{\tau} + \lambda\left\lvert Z_{m,n}^{\tau} - Z_{best}^{\tau}\right\rvert$
14	Else if $f_m = f_{best}$
15	$Z_{m,n}^{\tau+1} = Z_{m,n}^{\tau} + P\dfrac{\left\lvert Z_{m,n}^{\tau} - Z_w^{\tau}\right\rvert}{(f - f_m) + \xi}$
16	Determine the velocity of swarms: $Z_{\tau+1}^{s} = Z_{\tau}^{s} + G_{\tau+1}^{s}$
17	Determine the best solution: $Z_{m,n}^{\tau+!}(best)$

4.6 Results and Discussion

The sections below provide a detailed interpretation of the results utilizing the suggested Spizella swarm optimization based Bi-LSTM classifier.

Dataset description

Suspicious tweets and the Sentiment140 dataset, which contains 1.6 million tweets, are the datasets utilized for the social network analysis for crime rate detection. A description of the dataset is provided below.

Sentiment 140 dataset [33]: 1,600,000 tweets that were extracted using the Twitter API are included. The tweets can be used to determine sentiment because they have been annotated as positive and negative. The contents present in the dataset are the person's target that says whether the sentiment is positive or negative, the id of the tweet, the user's name, and so on.

Suspicious tweets dataset [34]: The suspicious tweet dataset is made up of around 60k tweets and was taken directly from Twitter. Since "suspect" is a rather general term, these datasets tags are based on three main categories: cyberbullying, terrorism, and threatening behavior. The output could be based on suspicious and unsuspicious tweets.

Software requirements: The crime rate detection in the social network is carried out using python, Community Edition PyCharm 2022.2.3, and the run time version used is 17.0.4.1+7-b469.62 amd64, Windows 11, python 3.7 version is used, with16 GB RAM.

Parameter metrics: Accuracy, sensitivity, and specificity are the metrics used to demonstrate the efficacy of the Spizella swarm-based optimization methodology. The suggested method should achieve higher values while measuring the parameter metrics and the performance should be improved for making the method more suitable for real-time. If the metrics values are obtained between the ranges of 70% to 80% then the model is considered a fair model, if the values are obtained in the range of 80 to 90% then the model is measured as a fair model, and if the model obtained the values greater than 90 the model posses' significant importance and is more reliable. This model achieved values greater than 90% and is considered to be more efficient.

Accuracy: The total number of crimes accurately recognized by the Spizella swarm optimization based Bi-LSTM in social network analysis is given as**,**

$$Acc = \frac{T_p + T_n}{(T_p + T_n + F_p + F_n)} \tag{25}$$

Sensitivity: The number of positive instances that are correctly identified by the proposed Spizella swarm optimization based Bi-LSTM in social network analysis out of the total number of positive instances is measured using the sensitivity and is given by,

$$Sen = \frac{T_p}{(T_p + F_n)} \quad (26)$$

Specificity: The number of negative instances that are correctly identified by the proposed Spizella swarm optimization based Bi-LSTM in social network analysis out of the total number of negative instances is measured using the specificity and is given by,

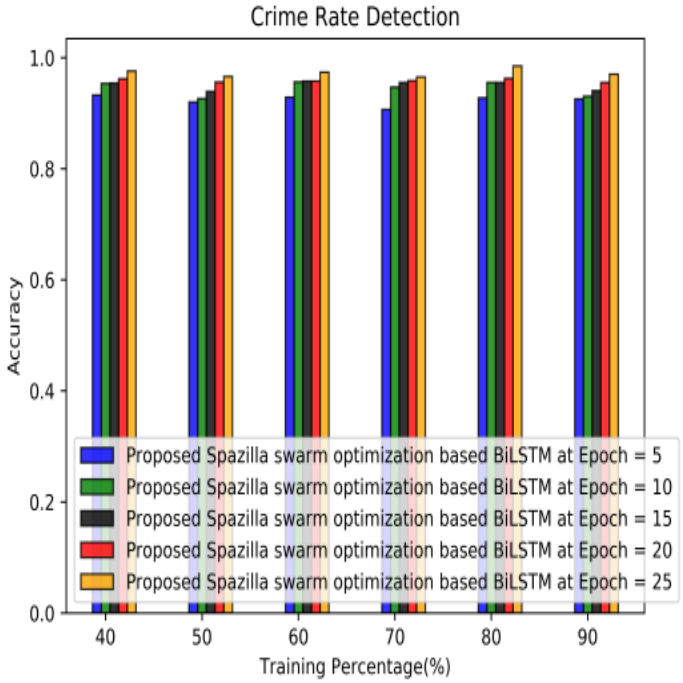
$$Spe = \frac{T_n}{(T_n + F_p)} \quad (27)$$

4.6.1 Performance analysis

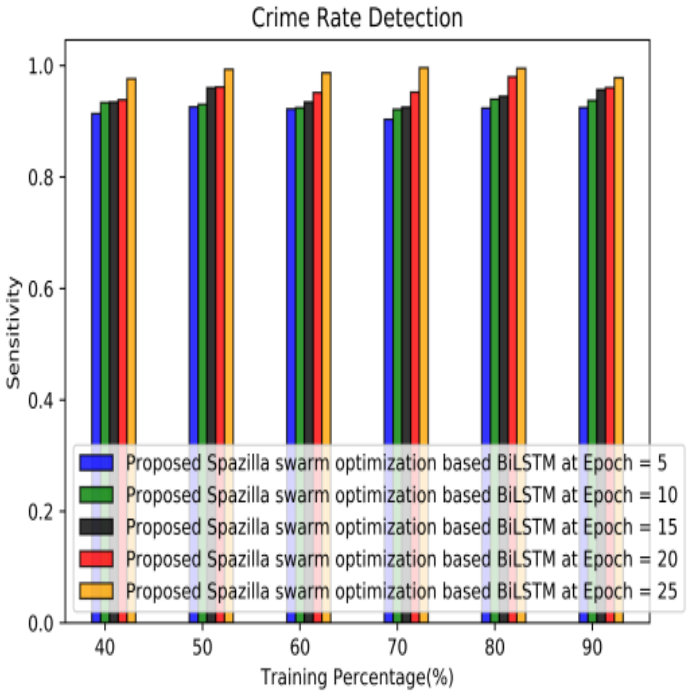
The performance analysis of the proposed Spizella swarm optimization based Bi-LSTM is measured for analyzing the performance of the classifier for varying epochs 5, 10, 15, 20, 25 and the detailed analysis is described in the below sections.

(1) Performance analysis based on dataset 1

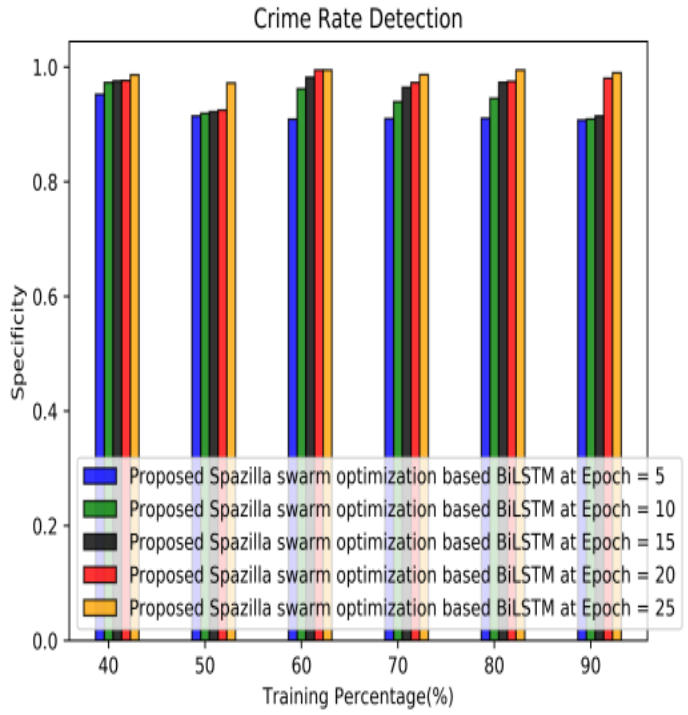
The performance analysis based on dataset 1 is performed based on the epoch values 5, 10, 15, 20, 25 shown in figure 14. The proposed Spizella swarm optimization based Bi-LSTM classifier obtained the values of 0.925, 0.930, 0.940, 0.955, and 0.970 for 90% training data in terms of accuracy shown in figure 14 a). Similarly, the Spizella swarm optimization based Bi-LSTM classifier obtained the values of 0.925, 0.938, 0.957, 0.960, and 0.978 for 90% of training data in terms of sensitivity shown in figure 14 b). At last, the specificity of the proposed Spizella swarm-based optimization is measured and the values are interpreted as 0.907, 0.909, 0.914, 0.980, and 0.989 respectively shown in figure 14 c).



(a) Accuracy comparison



(b) Sensitivity comparison



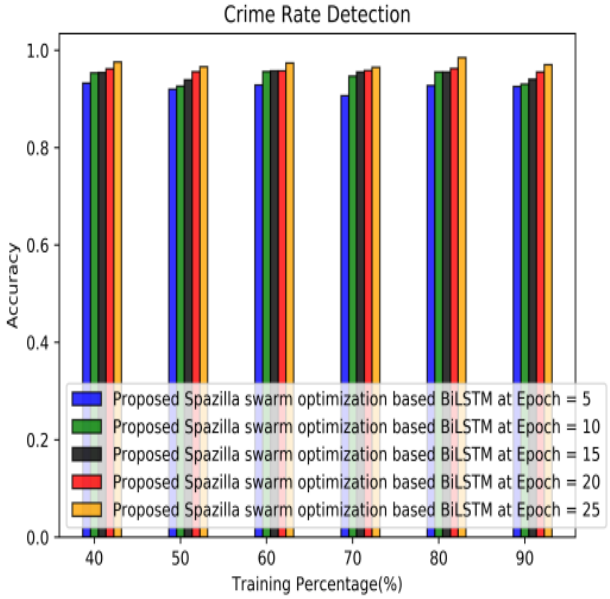
(c) Specificity comparison

Figure. 14 Performance analysis for the proposed Spizella swarm optimization based Bi-LSTM classifier for dataset-1 in terms of a) accuracy b) sensitivity c) specificity

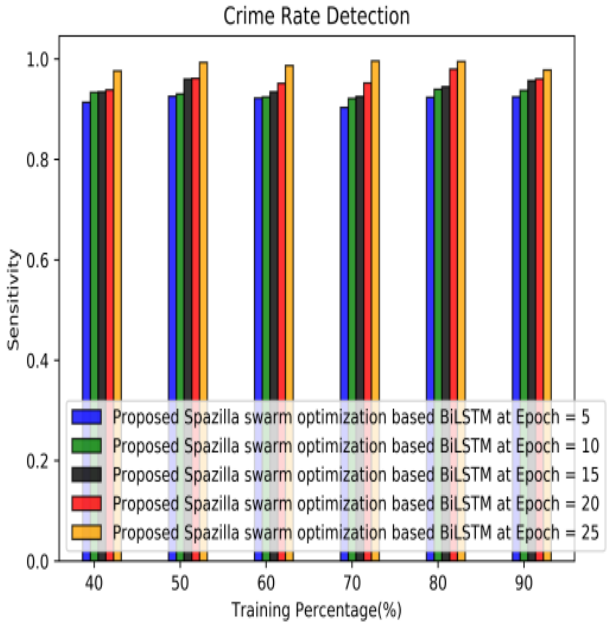
(2) Performance analysis based on dataset-2

The performance analysis based on dataset-2 is performed based on the epoch values 5, 10,

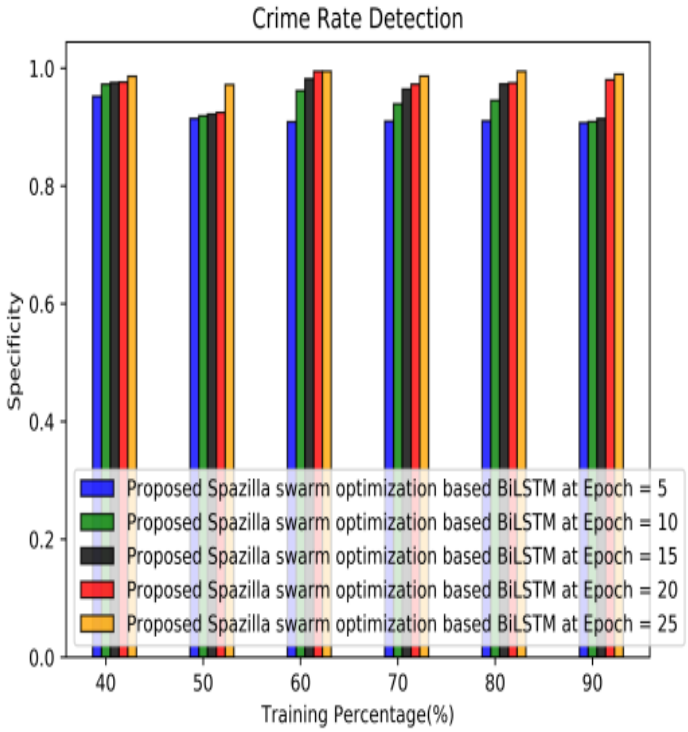
15, 20, 25 shown in figure 14. The proposed Spizella swarm optimization based Bi-LSTM classifier obtained the values of 0.905, 0.919, 0.932, 0.946, and 0.973 for 90% training data in terms of accuracy shown in figure 15 a). Similarly, the Spizella swarm optimization based Bi-LSTM classifier obtained the values of 0.907, 0.909, 0.911, 0.952, and 0.976 for 90% of training data in terms of sensitivity shown in figure 15 b). At last, the specificity of the proposed Spizella swarm-based optimization is measured and the values are interpreted as 0.903, 0.933, 0.938, 0.966, 0.969 respectively shown in figure 15 c).



(a) Accuracy comparison



(b) Sensitivity comparison



(c) Specificity comparison

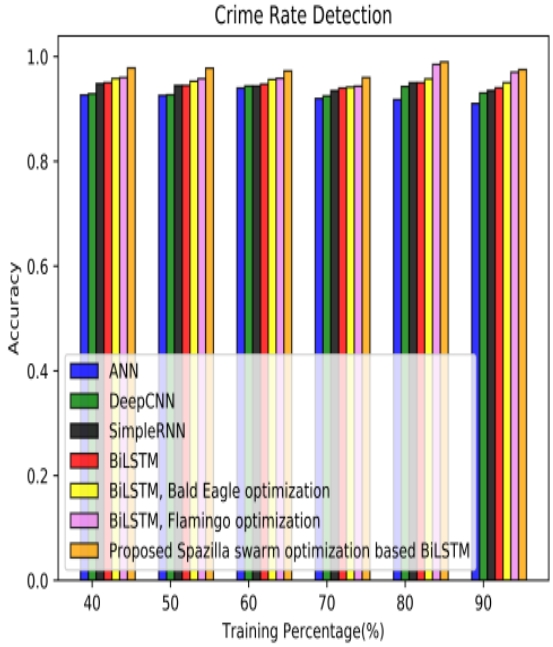
Figure. 15 Performance analysis for the proposed Spizella swarm optimization based Bi-LSTM classifier for dataset-2 in terms of a) accuracy b) sensitivity c) specificity

4.6.2 Comparative methods

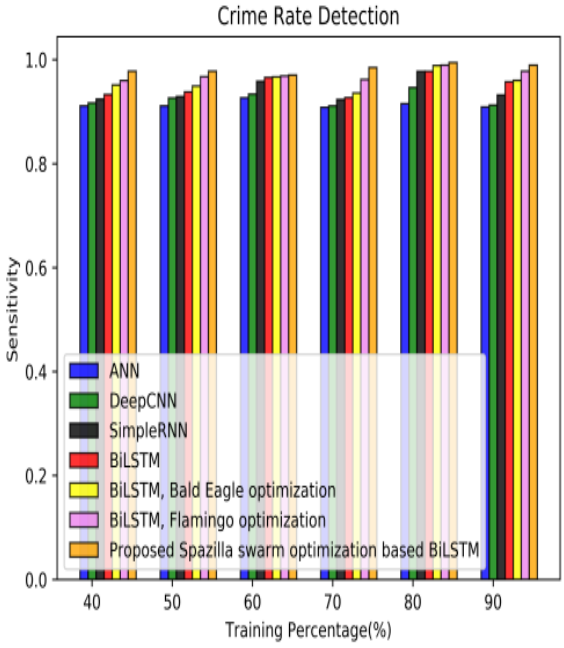
The methods used for the comparison of the proposed Spizella swarm optimization based Bi-LSTM classifier are ANN [26], Deep CNN [13], Simple RNN [27], Bi-LSTM [28], Bald Eagle optimization based Bi-LSTM (Bi-LSTM, Bald Eagle optimization) [31], Flamingo optimization based Bi-LSTM (Bi-LSTM, Flamingo) [32].

(1) Comparative analysis based on dataset-1

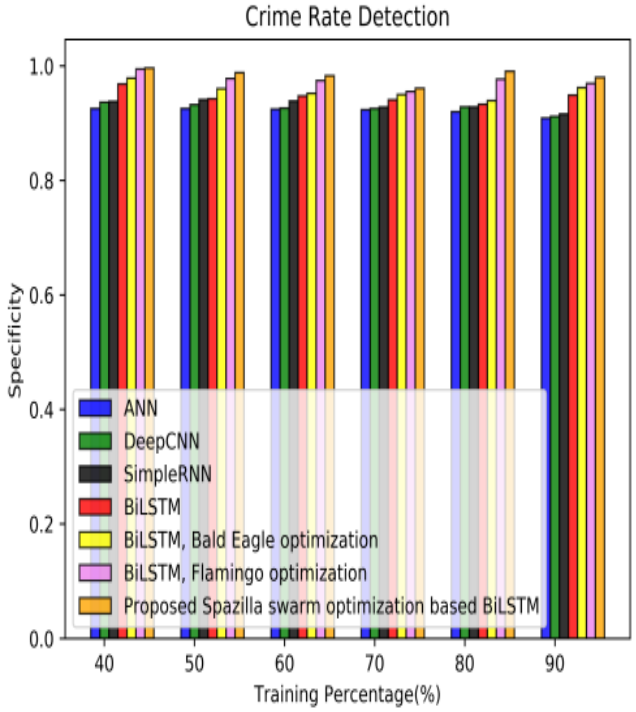
The comparative analysis based on accuracy, sensitivity and specificity are measured for the dataset-1 and the results obtained are shown in figure 16. The proposed Spizella swarm optimization based Bi-LSTM obtained an improvement of 0.5% in terms of accuracy shown in figure 16 a). Similarly, the proposed Spizella swarm optimization based Bi-LSTM attained the improvement of 1.16% in terms of sensitivity shown in figure 16 b). At last, the improvement of the proposed Spizella swarm optimization based Bi-LSTM classifier obtained an improvement of 1.08% in terms of specificity shown in figure 16 (c).



(a) Accuracy comparison



(b) Senstivity comparison



(c) Specificity comparison

Figure. 16 Comparative analysis for the proposed Spizella swarm optimization based Bi-LSTM for dataset-1 in terms of a) accuracy b) sensitivity c) specificity

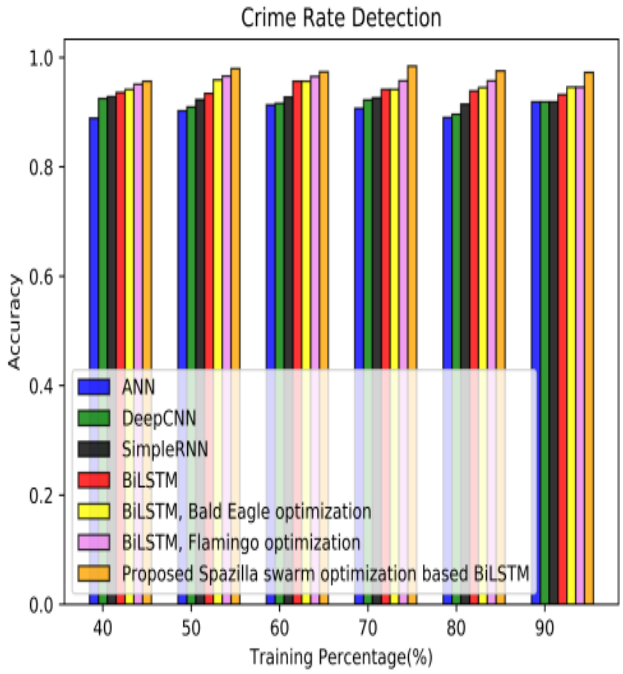
Table 3: Metrics values of the proposed Spizella swarm optimization based on dataset-1

Metrics/ Methods	**Accuracy**					
Training percentage	**40**	**50**	**60**	**70**	**80**	**90**
ANN	0.927	0.926	0.940	0.920	0.918	0.910
Deep CNN	0.929	0.927	0.944	0.925	0.943	0.930
Simple RNN	0.948	0.945	0.944	0.935	0.950	0.935
Bi-LSTM	0.950	0.945	0.948	0.940	0.950	0.940
Bi-LSTM, Bald Eagle	0.958	0.953	0.956	0.942	0.958	0.950
Bi-LSTM, Flamingo	0.960	0.958	0.959	0.943	0.985	0.970
Proposed	0.978	0.978	0.973	0.960	0.990	0.975
Sensitivity						
Training percentage	**40**	**50**	**60**	**70**	**80**	**90**
ANN	0.911	0.911	0.927	0.908	0.915	0.909
Deep CNN	0.917	0.926	0.934	0.911	0.947	0.913
Simple RNN	0.924	0.929	0.958	0.923	0.978	0.932
Bi-LSTM	0.933	0.938	0.966	0.927	0.978	0.957
Bi-LSTM, Bald Eagle	0.952	0.949	0.967	0.936	0.989	0.960

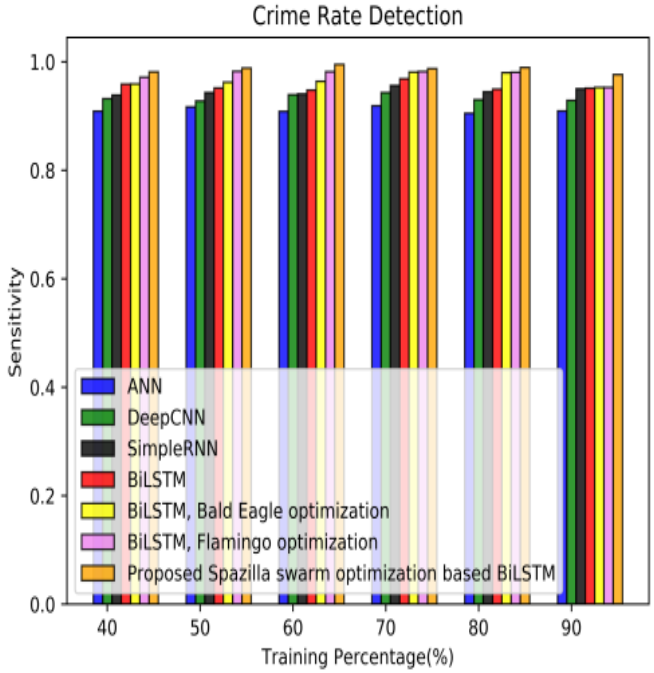
Bi-LSTM, Flamingo	0.960	0.967	0.969	0.962	0.989	0.978
Proposed	0.978	0.978	0.971	0.985	0.995	0.990
Specificity						
Training percentage	**40**	**50**	**60**	**70**	**80**	**90**
ANN	0.926	0.926	0.924	0.923	0.920	0.909
Deep CNN	0.937	0.932	0.926	0.926	0.928	0.912
Simple RNN	0.938	0.941	0.938	0.928	0.928	0.916
Bi-LSTM	0.968	0.943	0.947	0.941	0.933	0.949
Bi-LSTM, Bald Eagle	0.979	0.960	0.952	0.950	0.939	0.962
Bi-LSTM, Flamingo	0.995	0.978	0.974	0.955	0.977	0.969
Proposed	0.996	0.988	0.983	0.961	0.991	0.980

(2) Comparative analysis based on dataset-2

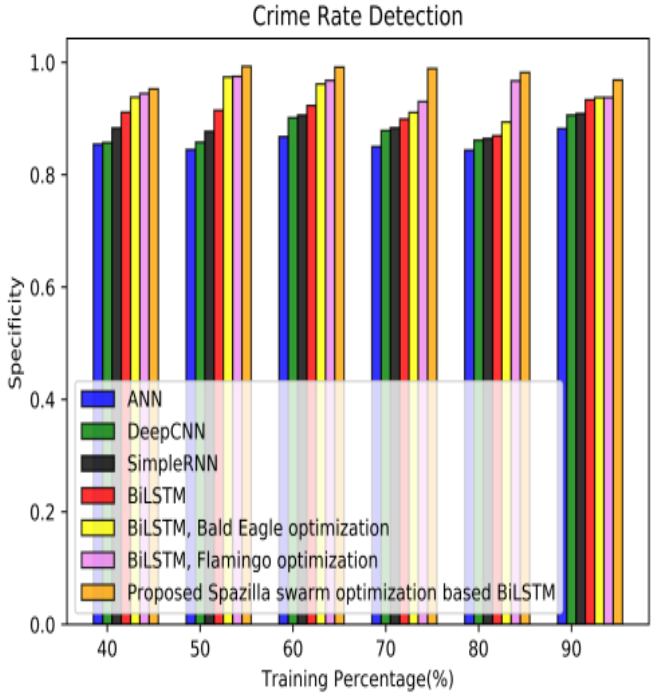
The comparative analysis based on accuracy, sensitivity and specificity are measured for the dataset-2 and the results obtained are shown in figure 17. The proposed Spizella swarm optimization based Bi-LSTM obtained an improvement of 2.77% in terms of accuracy shown in figure 17 a). Similarly, the proposed Spizella swarm optimization based Bi-LSTM attained the improvement of 2.43% in terms of sensitivity shown in figure 17 b). At last, the improvement of the proposed Spizella swarm optimization based Bi-LSTM classifier obtained an improvement of 3.22% in terms of specificity shown in figure 17 c).



(a) Accuracy comparison



(b) Sensitivity comparison



(c) Specificity comparison

Figure. 17 Comparative analysis for the proposed Spizella swarm optimization based Bi-LSTM for dataset-2 in terms of a) accuracy b) sensitivity c) specificity

Table 4: Metrics values of the proposed Spizella swarm optimization based on dataset-2

Metrics/ Methods	**Accuracy**					
Training percentage	**40**	**50**	**60**	**70**	**80**	**90**
ANN	0.889	0.903	0.913	0.907	0.890	0.919
Deep CNN	0.925	0.910	0.916	0.922	0.896	0.919
Simple RNN	0.929	0.923	0.928	0.926	0.915	0.919
Bi-LSTM	0.936	0.934	0.957	0.942	0.939	0.932
Bi-LSTM, Bald Eagle	0.942	0.959	0.957	0.942	0.945	0.946
Bi-LSTM, Flamingo	0.951	0.966	0.965	0.957	0.957	0.946
Proposed	0.957	0.980	0.974	0.984	0.976	0.973
Sensitivity						
Training percentage	**40**	**50**	**60**	**70**	**80**	**90**
ANN	0.909	0.916	0.908	0.919	0.905	0.909
Deep CNN	0.932	0.928	0.939	0.943	0.930	0.929
Simple RNN	0.938	0.943	0.940	0.956	0.944	0.950
Bi-LSTM	0.959	0.951	0.948	0.969	0.949	0.951
Bi-LSTM, Bald Eagle	0.959	0.962	0.964	0.981	0.980	0.952
Bi-LSTM, Flamingo	0.971	0.982	0.981	0.982	0.981	0.952
Proposed	0.981	0.988	0.995	0.987	0.989	0.976
Specificity						
Training percentage	**40**	**50**	**60**	**70**	**80**	**90**
ANN	0.854	0.845	0.868	0.850	0.844	0.882
Deep CNN	0.857	0.858	0.901	0.879	0.862	0.906
Simple RNN	0.883	0.877	0.906	0.883	0.864	0.909
Bi-LSTM	0.911	0.914	0.923	0.899	0.870	0.933
Bi-LSTM, Bald Eagle	0.938	0.974	0.962	0.911	0.894	0.938
Bi-LSTM, Flamingo	0.945	0.975	0.968	0.930	0.967	0.938
Proposed	0.953	0.993	0.991	0.989	0.982	0.969

4.6.3 Comparative discussion

The best value obtained by the proposed Spizella swarm optimization is due to the convergence improved by the optimization applied on the Bi-LSTM classifier. The crime rate detection is made more efficiently by speeding up the social network analysis using the proposed optimization in Bi-LSTM classifier. The best value obtained by the proposed method is interpreted in table 5.

Table 5: Comparative discussion of the proposed Spizella optimization based Bi-LSTM classifier

	Dataset 1			Dataset 2		
Metrics/ Methods	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
ANN	0.910	0.909	0.909	0.919	0.909	0.882
Deep CNN	0.930	0.913	0.912	0.919	0.929	0.906
Simple RNN	0.935	0.932	0.916	0.919	0.950	0.909
Bi-LSTM	0.940	0.957	0.949	0.932	0.951	0.933
Bi-LSTM, Bald Eagle	0.950	0.960	0.962	0.946	0.952	0.938
Bi-LSTM, Flamingo	0.970	0.978	0.969	0.946	0.952	0.938
Proposed	0.975	0.990	0.980	0.973	0.976	0.969

4.5 Summary

Social network analysis using the proposed Spizella swarm optimization-based Bi-LSTM classifier is performed for the detection of crime rates. Since crimes are rising at an alarming rate, it is difficult to foresee them with any degree of accuracy. Therefore, it is crucial to identify potential crimes now in order to prevent them in the future. Hence the crime rate is detected using the Spizella swarm optimization based Bi-LSTM classifier, where the convergence of the crime rate detection is greatly enhanced. Spizella swarm optimization effectively tuned the parameters and helps in achieving a better output. The proposed classifier could be applicable in determining the behavior of people and helps in reducing the occurrence of crime rates. Compared to the existing methods proposed method gains high accuracy and takes less time for detection. By analyzing the metrics values the Spizella swarm optimization obtained an improvement of 0.5%, 1.16%, and 1.08%, which is more efficient. In the future, the sentimental analysis, and the opinion analysis could also be included for efficient crime rate detection also it is difficult to predict the next crime that is going to take place using Twitter data because the large number of fake information is present in the data is the upcoming future work.

CHAPTER 5
IMPACT OF CRIME BASED DATA MINING IN INTERSTATE POLICIES

Every country has a number of states that establish laws and engage in illegal activity that affect nearby or distant states. States create policies and experience criminal activity that can influence each other, often covered by major news outlets such as the *Washington Post* and *The New York Times*. This study proposes a model that, through rule-based deduction, data transformation, and FP-Growth algorithms, detects patterns of influence between states. Our approach clusters data by crime types, finding that 34% of criminal activities in Alaska impact Washington, while Alaskan policies influence Washington by 89%. These findings align with broader national trends: recent FBI data indicates significant decreases in violent crime nationwide in 2023, underscoring the impact of policy adjustments across states.

5.1 Introduction

Throughout history, nations across continents have enacted policies on areas such as budgeting, ecology, health, education, democracy, infrastructure, and crime. Following major incidents like the [significant terrorist acts] data science began to shift towards crime science, encouraging officials globally to refine and legislate policies for maintaining national peace. Although state-level policies have substantial impact potential, the question remains: do we sufficiently consider national impacts in our policy decisions? Previously posed in [2], we aimed to explore this in our project using an analytical framework.

Among various policy domains, crime remains uniquely influenced by policy—a positive factor—in addressing negative outcomes. This analysis considers two policy types: intra-state, executed with confidence at 1.0, and inter-state, where the "trustworthiness" of cross-state effects is yet uncertain. Our project focuses on inter-state crime policy, bridging computer science with policy. To establish causality between state policies, both qualitative [3][4] and quantitative [5] methods exist; we employed a quantitative approach to demonstrate how state-specific policies affect neighboring states.

Twitter, a widely used microblogging platform, generates data continuously through user messages (tweets) reflecting ideas, opinions, and preferences. By analyzing hidden information in tweets, such as user opinions and activity, we gain insights. Figure 18 presents the states

considered for crime-policy link analysis. To analyze data, we used data mining to identify patterns and focused on crime-related terms as defined by the National Institute of Justice. Terms extracted from major publications, including *The Washington Post* and *The Seattle Times*, helped us cluster tweets by topics like "unlawful activities," "illicit substances," and "assaults."

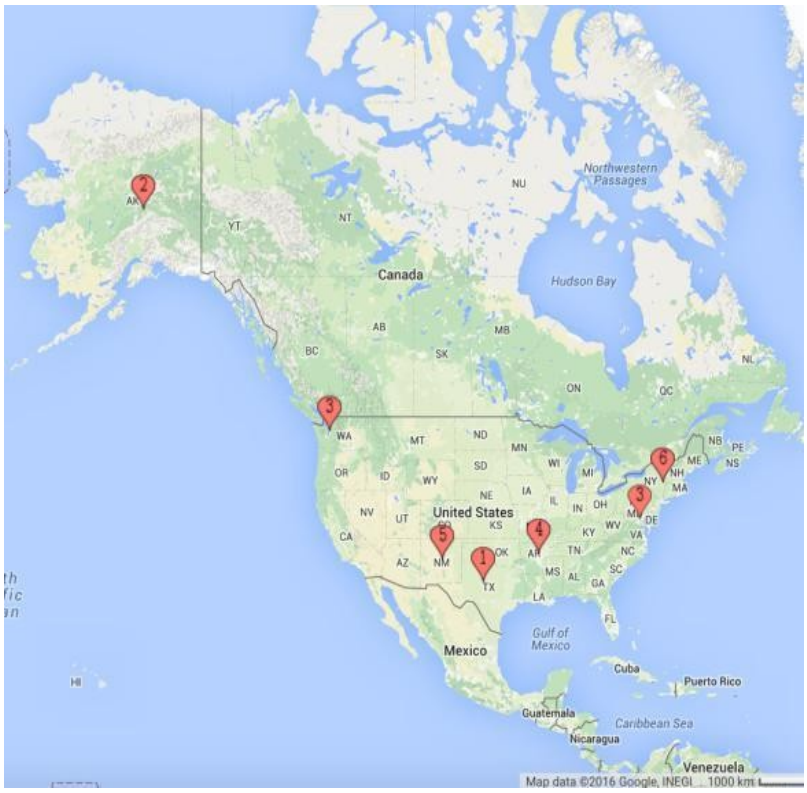


Figure 18: Geographical Location from Google Maps

To gain insights from Twitter data, we applied data mining techniques to identify interpretable patterns related to crime. We examined terminology from the National Institute of Justice and used terms such as "unlawful," "illicit substances," "murder," "rape," "criminal activity," and "assaults" as primary clustering labels. These terms were identified from tweets referencing major news sources like The Washington Monthly, The Atlantic Magazine, and The Texas Monthly, which frequently report on crime and policy.

5.2 Literature Survey

This section reviews recent studies in crime data mining, predictive modeling, and their policy implications, establishing a foundation for advancing crime-related data analytics. Recent research has underlined the significance of spatial and temporal data for effective crime prediction. Advanced techniques, including artificial neural networks (ANNs), fuzzy systems, and hybrid models, have been employed to improve forecasting accuracy and provide actionable

insights. Predictive policing leverages geospatial and temporal analytics to anticipate crime hotspots, enabling law enforcement to allocate resources more efficiently. However, these advancements are not without challenges: biases in training data and privacy concerns remain persistent issues. These challenges necessitate the development of transparent, ethical models and robust evaluation processes to foster public trust in criminal justice applications.

Studies also emphasize the increasing collaboration between data scientists and law enforcement agencies, which is critical for tailoring predictive algorithms to meet community-specific needs. This partnership aims to address ethical issues such as demographic biases that could disproportionately affect certain populations. Effective predictive models depend on the availability of accurate and representative datasets, which must be periodically revised to mitigate potential discrimination and maintain reliability. To ensure success, law enforcement agencies are encouraged to engage closely with data science experts, balancing the need for model accuracy with ethical considerations in criminal justice operations.

Incorporating social media data and other sensitive information into crime prediction introduces additional privacy and ethical concerns. Existing legal frameworks, such as the Electronic Communications Privacy Act, provide foundational standards; however, the rapid evolution of technology necessitates continuous adaptation of these regulations. Ethical guidelines must emphasize transparency, accountability, and responsible data handling to align public safety objectives with privacy rights, fostering public trust in predictive policing initiatives.

5.3 ETL Phase

ETL (Extraction, Transformation, and Loading) is essential before model selection, as it involves gathering, processing, and transforming data to suit the problem at hand.

5.3.1 Extraction

The data for this study was collected using Twitter Streaming and REST APIs, resulting in raw, unstructured text data. This raw data includes various extraneous elements such as links, spaces, usernames, hashtags, and "@" symbols. Directly using this unrefined data for rule derivation and analysis would lead to inaccurate results, as the noisy text introduces irrelevant information that can interfere with the extraction of meaningful patterns.

To address this, a data cleaning and filtering process is performed, which is critical for transforming the unstructured data into a structured format suitable for classification and analysis. By removing unnecessary symbols and filtering outliers, the data becomes more interpretable and

actionable for downstream processes.

For this project, each tweet contains over 16 attributes, among which the "text" attribute is the primary focus. The text attribute captures the main content of the tweet, making it the foundation for sentiment analysis, clustering, and rule derivation. However, the presence of hyperlinks, unrelated high-frequency words, and other irrelevant content in the raw data obscures the critical information needed for analysis.

5.3.2 Transformation

Transformation is defined as the process of converting the raw data into a format that can be used as an input to the classifier so that some useful inferences can be derived. As stated in [16], the transformation process develops formal relationships from the raw data, which are used later to develop conceptual structures using unsupervised [17] and supervised learning [18].

Regular Expressions: Within the perspective of this project, we constrict ourselves to the application of regular expressions (RE) [19]. We have developed some AWK scripts which helped in removing extra spaces, dummy symbols, and hyperlinks. As we will describe natural language toolkit (NLTK) [20] which mandates that our textual input should have words from some specific language and should be free from extraneous literals. We will briefly define in section 5, our process of creating the vocabulary necessary for converting our textual data to some real number data, that will serve as an input to the model.

Outlier Detection: Within the periphery of data mining, there has been the development of many algorithms relevant to outlier detection and removal. Such as energy of the graph [21] and statistical technique [22]. We are more inclined towards simple central tendency-based outlier detection and removal. Our outlier detection process uses standard deviation and means for identifying divergent tuples in the transformed text. Mean is the central measure of the tendency which involves the summation of sentiment vector of the tweets over the total number of tweets. The mean value of the tweets helps in identifying the standard deviation of each tweet. This statistical measure can be seen as a process of selecting those tweets which are above 10 percentile and below 90 percentile which is termed as the Interquartile Range (IQR).

$$T_{vector} = S_{score}(positive_{sentimental}, negative_{sentimental}) \tag{28}$$

Transformation involves converting raw data into a structured form suitable for classifiers to draw insights. This process establishes formal connections within the data, supporting both supervised and unsupervised learning models. In this project, regular expressions (RE) help streamline the data by eliminating extraneous spaces, symbols, and references. AWK

scripts aid in this cleanup, while the Natural Language Toolkit (NLTK) ensures our input meets specific linguistic requirements, free from unnecessary characters. Section 5 will outline the vocabulary-building steps that transform text into numerical input for our model.

Outlier Detection: T(e) and removing outliers, like statistical methods and graph energy measures, play a key role in the basic methods rooted for identifying outliers using standard deviation. The core trend metric, or mean, derives from the emotion vector across tweets, with IQR helping filter tweets within the 10th to 90th percentile. This statistical approach enhances data

relevance and accuracy.

We represent the information from Twitter as a 1-D vector (T_{vector}) of sentiment, both positive and negative. The topic at hand is: What was reason for giving neutral sentiments any thought? In section 5, the response to this query is explained. Equation (1) quantifies each modified tweet from regular expression.

Finding the data's center value and examining sample variation from 1 to N are steps in the outlier detection procedure. The variation strategy bears resemblance to the covariance-based hidden Markov model (HMM) approach [23]. Therefore, we use equations (2) and (3) to define our outlier detection algorithm. These formulas give the interval, or IQR, expressed in terms of medians.

$$(Q_1 outlier < \frac{1}{N}\sum_{i=1}^{N} T_i - \sigma)$$

$$(Q_1 outlier < \tfrac{1}{N}\textstyle\sum_{i=1}^{N} T_i - \sigma) \quad (29)$$

We have provided an analogy to help identify outliers: equations 4 and 5 work similarly to the median 1.5 rule [24], with Q1 denoting the lower quartile (10 percentile) and Q3 denoting the higher quartile (90 percentile). Using the equation aforementioned, we can construct a structure akin to a box where all of the tweets that are being utilized in section 6 are filtered and altered. In order to avoid bad classifier training in the case of supervised learning or poor and erroneous clusters in the case of clustering, we processed the tweets using a combination of regular expressions and transformation.

5.3.3 Loading

We proceed to the stage of loading or saving the tweets after completing the necessary tasks in the first two parts. For the results in this research to be accurate and understandable, this step is crucial. Data can be stored in a variety of forms, including ASCII text, Excel spreadsheets (XLS), comma-separated values (CSV), tab-separated values (TSV), and JavaScript object

notation (JSON). The natural language processing community recommends CSV/TSV and JSON among these formats [25].

5.4 Data Mining Techniques

With an emphasis on the criminal mining process, this part will provide a quick overview of some of the well-known data mining techniques utilized in this study. In order to mine crime-related data, models must be able to group the data according to a gaussian surface and embed the data in a vectorized form made up of related entities.

$$Ntweets = \phi_i * \mathcal{E}_0 \quad i_e t_0 \varepsilon_0 \tag{30}$$

We have demonstrated in equation (6) that the entire criminal mining process, which consists of rule-based deduction along with clustering or classification, can be mapped to a Gaussian surface with φi amount of charge 0 each cluster. A formulation like this offers a solution to the problem through intuition. The classification technique will be discussed first, then clustering and rule-based techniques.

5.4.1 Loading Classification: Logistic Regression

A sigmoid function is used by the regression-based classifier known as logistic regression to aid in categorization. According to [26], a classification is the process of creating a portion of the set from the superset utilizing an activation function to provide an output that corresponds to the problem's labels. With the use of a radial basis kernel, logistic regression is superior to linear regression [27].

$$(f(x)Tvector(x)) = \frac{1}{1+e^{-1vector}} \tag{31}$$

$$Mclassifier(x) = ßM_{classifier} + \alpha \frac{1}{1+e^{-1vector}} \tag{32}$$

The logistic regression model uses a continuous and differentiable activation function (Eq. 32), which ensures fewer issues with saddle points during training. The speed at which the classifier learns is called its learning momentum M classifier, which is impacted by gradient direction—positive for ascent and negative for descent. However, despite using momentum β and learning rate α in Eq. 8, improvements in classification have been minimal, thus leading us to consider cross-validation (CV).

Cross-Validation: CV is essential for model evaluation, as it splits data into training and testing subsets. This process checks prediction accuracy by training on one subset and testing on another, justifying larger sample sizes for reliable results. The classifier's iterations, represented

as βi in Eq. 8, also play a key role. A warning: very low (<30%) or high (>90%) iterations suggest either random learning or overfitting, scaled relative to sample size.

5.4.2 Clustering: K-Means

Clustering is akin to organizing objects based on certain similarity metrics [29]. A set of data range is employed to determine its median or centroid, which constitutes the set's representative data member [19]. The next query is: In K-Means, how is K defined? In the interest of conciseness, we provide a succinct synopsis of the techniques employed in the effort to determine K. Other techniques for determining K, in addition to those already discussed. Elbow Method: One particular measure for determining K for the K-Means algorithm is the elbow method. The procedure first determines the intra-cluster diameters for various K values. It does normalization on the total of all intra-cluster distances for each K value.

Elbow NC k is the Normalized Cluster value for a k

$$Elbow(NC_{k_{eptions}} = i \in C_{k-1} argmax f(\theta) = \tan(\theta_i)) \qquad (33)$$

In the figure 19, we show that K in the range (7,10), Z is the appropriate value for K-Means algorithm. From the figure 2 it is evident that to find Elbow NC k optimal, we need to use equation 10 over the curve in the figure 19.

Silhouette metric: Another metric that can be used for identifying the k in K-Means is the silhouette metric. The silhouette metric is the opposite of Elbow method. In this metric we replace with above given in equation 33.

The inversion of figure 20 is the silhouette graph. In the silhouette metric we look for the lowest point in the graph. In the figure 20, it is evident that K in the range (7,10) Z is the appropriate value for K-Means algorithm and is coherent with figure 19. We define the silhouette score as SNCK is the Normalized Cluster value for a k, k using silhouette metric

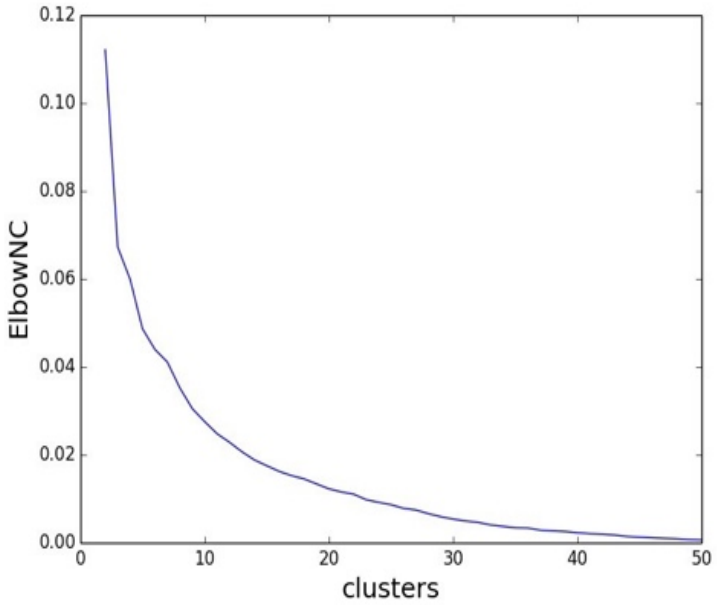


Figure. 19. Elbow curve with 7<K<10

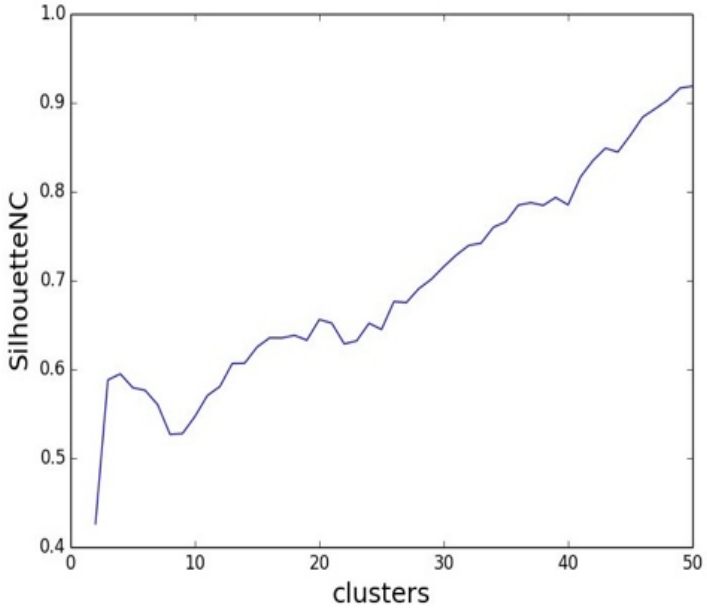


Figure. 20. Silhouette score graph with 7<K<10

Comparing the equations 9 and 11, we observe that, max T vector i, T vector j is the factor due to which the silhouette is inverted as compared to elbow method. The mathematics behind formulation is extensive as stated in [20].

5.5 Frequent Pattern Growth

Frequent Pattern (FP) growth is an algorithm in data mining which creates a tree based on association rules developed from association rule mining (ARM). FP-growth uses the frequency

parameter of the item sets for creating the tree. In the procedure of creating the FP tree, the algorithm embeds essential modules from association rule mining. In order to create the FP tree, the algorithm first generates rules based on the clusters. We threshold the rules based on the confidence values and hence, those rules which are alive form the FP tree [31]. The procedure for association rule mining is stated in the following equation

FP-Growth algorithm has the potential to compress a large database into a compact Frequent-Pattern tree (FP- tree) structure which is highly condensed, but complete for frequent pattern mining. This algorithm also avoids costly database scans, which matters when we have huge clusters and over a billion samples. From equations 13 and 14, it is derived that the FP tree method with ARM guarantees healthy prediction for any complex problems. We are now in a position to validate this section IV with results and appropriate discussion.

5.6 Results and Discussion

In this section, we will be providing a discussion on the results obtained after the application of algorithms on the structured sample obtained. In the figure 19, we observe that there is a good amount of clustering being shown using the positive sentiment value of the Tweets. The policy-related tweets influence positive sentiments and the crime related tweets influence negative sentiments. Since we are focusing on policies that affect the crime, we are more inclined towards those tweets which form clusters in 6. The cluster of tweets formed in figure 20 is crime-related tweets. This cluster provides the very foundation of the rules that will link a state to another state using policy and crime keywords. Observing figure 21, we can nearly say that there are states which influence each other in terms of policy and crime but it can also be possible that we are observing the cluster of state and not states concerning policy and crime.

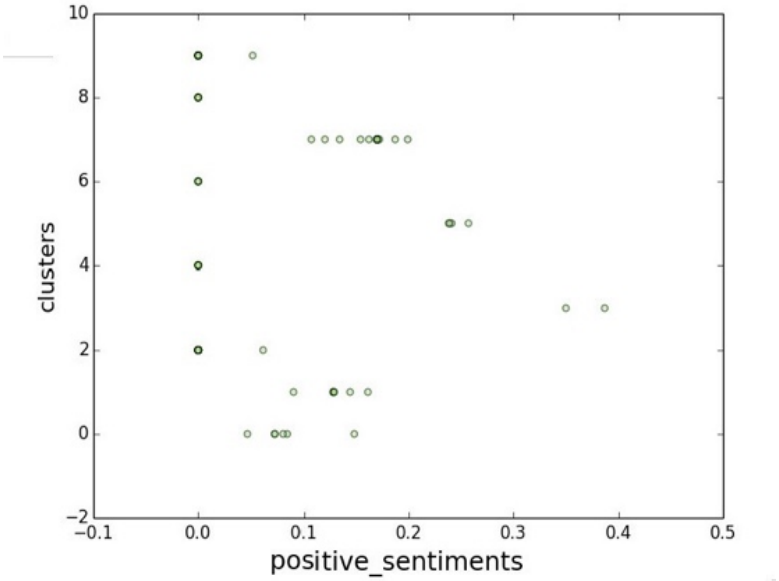


Figure. 21. Cluster of tweets in positive sentiments

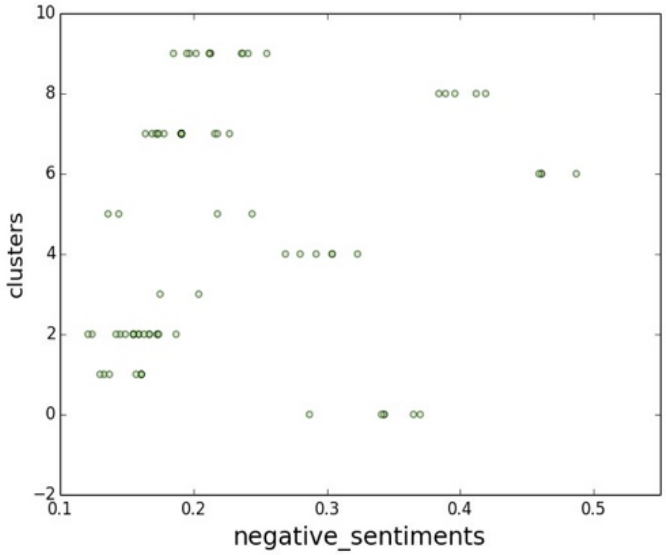


Figure. 22. Cluster of tweets in negative sentiments

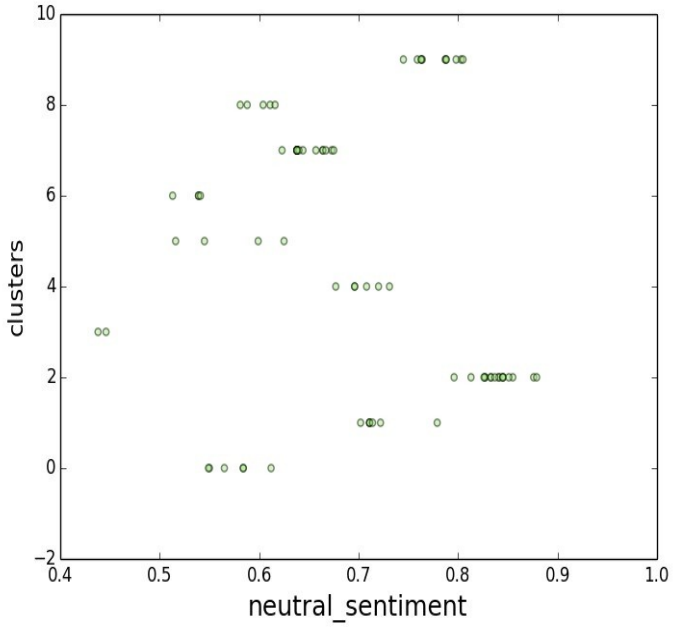


Figure. 23. Cluster of tweets with neutral sentiments

```
(0.89) Alaska
(0.38) Alaska crime
(0.37) Alaska crime Washington
(0.37) Alaska crime Washington policy
(0.38) Alaska crime policy
(0.89) Alaska Washington
(0.89) Alaska Washington policy
(0.34) Alaska Washington policy kill
(0.34) Alaska Washington kill
(0.89) Alaska policy
(0.34) Alaska policy kill
(0.34) Alaska kill
(0.39) crime
(0.38) crime Washington
(0.38) crime Washington policy
(0.39) crime policy
(0.99) Washington
(0.99) Washington policy
(0.35) Washington policy kill
(0.35) Washington kill
(1.00) policy
(0.35) policy kill
(0.35) kill
```

Figure. 24. A structure of FP-tree based on neutral sentiments with confidence values

From the figure 24, we say Alaska, Washington, policy, crime arrives 34% of the time which is quite sufficient with to answer the question asked in section I. Furthermore, we are able to show that Alaska, Washington, policy arrives 89% of the time which also support our answer to the question in section 1. Lastly, we are able to prove that Alaska, Washington, crime arrives 34% of the time. So, we can say with high assurance that Washington and Alaska (W&A) influence greatly. In order to magnify our observation, we used the FP-Tree technique to illustrate those clusters which include intra-state and not inter-state. Since we are focusing on those

clusters which are present in neutral sentiments of the tweets, we formulate an FP-Tree based on support and confidence of the rules generated using ARM. We place a threshold of 30% over the support and 30% over the confidence for generating the rules. Observing the FP-Tree, we achieve a strong predisposition towards two states, Washington, and Alaska in terms of policy and crime confluence.

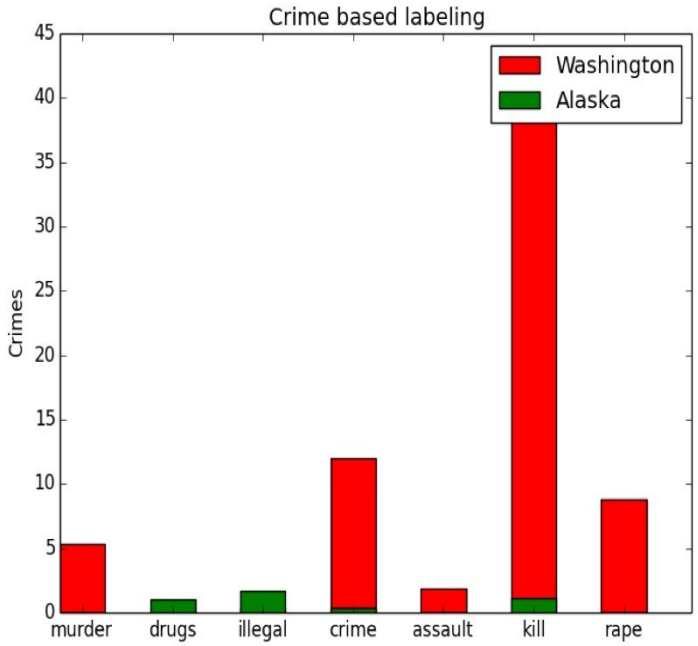


Figure. 25. Crime based labelling in Washington and Alaska after FP-Tree

As we have proved that W&A influence each other w.r.t crime and policy, it is evident from figures 25 and 26 that the crime-related tweets specifically property crime and killings related tweets are high in Washington and whereas drugs and illegal jobs related tweets are predominant in Alaska. It can also be observed that Alaska shows a significant number of tweets (negative clusters in figure 25) with respect to crime and killings. Furthermore, we analyze the bar plot in figure 26 which is more influenced by figure 24.

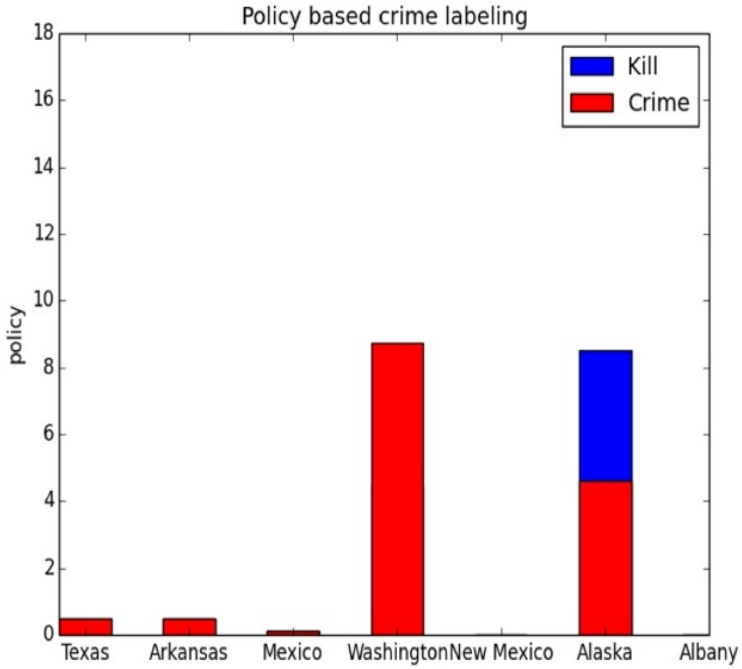


Figure. 26. Policy based labelling in Washington and Alaska after FP-Tree

In figure 26, we observe that W&A has a significant contribution to the implementation of policies against crime. In respect to section III, we made sure that when we talk about policy, these are not environmental, health, or transport- related policies. From figure 26, we also observe that Texas, Arkansas, Mexico, New Mexico, and Albany lack tweets related to policy and crime. Though these states have tweets related to crime but when policy word is augmented the confidence values steps down, hence we don't see results from these states in figure 25.

5.7 Summary

We are able to show the relationship between crime and policy from the perspective of data science. The project culminated into proof which proves the question: Does the policy of a state affects another state?". Our contribution lies in proving this linkage between policy and crime using some of the common data mining techniques. The project develops a strong understanding of the need for ETL, model selection, and rule-based inferences for solving any complex and diverse problem. We aimed at providing linkage between states in the USA because, we have observed over the time through different monthly and quarterly magazines, that various state and federal organizations are involved in maintaining peace within the state and neighboring states. Our rule-based deductions sufficiently prove that Twitter tweets play an eminent role in providing useful inferences about the working of the state and its environment. This project is a preliminary approach to policy research but provides an appropriate foundation for further study.

We have used limited classification and clustering techniques in their basic form and didn't include hybrid algorithms. Further-more, we have restricted ourselves to simple and structural learning without incorporating randomness while making the classifier learn. Lastly, the problem can be formulated in computational intelligence and it can be assumed that much better patterns in terms of rules can be created using Heuristic, Meta-heuristic and Fuzzy methodologies. Also, we haven't dealt deeply into the computational linguistic perspective of the problem, so this solution though worthy of giving the answer to question in section 1, is still silent from linguistic creativity.

CHAPTER 6

SDHO-KGNN: AN EFFECTIVE KNOWLEDGE ENHANCED OPTIMAL GRAPH NEURAL NETWORK APPROACH FOR FRAUDULENT CALL DETECTION

Rapid advancements in mobile communication technologies have led to the progression of telecom scams that not only deplete the individual fortune but also affect the social income. Hence, fraudulent call detection gains significance, which not only aims to proactively recognize the frauds but also alleviate the fraudulent activities to manage external losses. Though the traditional methods, such as rule-based systems and supervised machine-learning techniques actively engage in detecting such fraudulent activities, they failed to adapt to the evolving fraud patterns. Therefore, this research introduces a Sheep Dog Hunt Optimization enabled Knowledge-Enhanced Optimal Graph Neural Network classifier (SDHO-KGNN) approach for detecting fraudulent calls accurately. The effectiveness of the proposed SDHO-KGNN approach is achieved through the combination of the power of graph representation learning with expert insights, which allows the proposed SDHO-KGNN approach to capture complex relationships and patterns within telecom data. Additionally, the integration of the SDHO algorithm enhances the model performance by optimizing the discrimination between legitimate and fraudulent calls. Moreover, the SDHO-KGNN classifier captures the intricate call patterns and relationships within dynamic call networks thereby, attaining a better accuracy, precision, and recall of 93.8%, 95.91%, and 95.53% for 90% of the training.

6.1 Introduction

Over the past few years, the swift expansion of mobile communication technology has ushered in numerous benefits, revolutionizing the way to connect and communicate. Wireless communication networks have undergone significant development during the past few decades, progressing from 1G in the 1980s to 5G in 2019.Every generation sought to overcome the shortcomings of its predecessor, primarily about latency and mobility, density, security and privacy, and data rates and capacity. Consequently, the future 5G/6G networks intend to provide great flexibility to the network operators concerning the allocation of coverage, speed, and the capacity of diverse use cases. Although the 5G deployment phase has already started [1], its full potential has not yet been reached in mobile communication. Meanwhile, 6G intends to transform

communication by integrating the digital, biological, and physical realms to give people new experiences via immersive virtual worlds and augmented intelligence [2].Specifically, Artificial intelligence (AI) and Machine Learning (ML) techniques are expected to drive 5G by continuing the network cloudification initiatives toward the development of intelligent telecommunication networks that led to 6G [3].To fully realize the potential and benefits of 6G, cyber resilience, privacy, and trust are essential prerequisites. Further, these prerequisites will propel 6G development and are essential to attaining even faster and more dependable connectivity than 5G [4]. Telecommunication fraud [1], often shortened to telecom fraud, involves the fraudulent acquisition of significant public and private assets through deceitful means, such as fabricated narratives conveyed through messages, calls, and various communication channels [5]. As science and technology have progressed, the methods of telecommunications network fraud have evolved as well, becoming increasingly precise and efficient, and spreading more extensively [2][6]. Contemporary fraudsters often adopt behaviors that closely resemble those of typical users, operating in well-coordinated groups. This shift in their approach presents a formidable challenge when it comes to identifying telecom frauds within the vast volume of regular call records [7]. Fraudsters may have various motives, one of which could be to evade service charges or minimize them, ensuring they do not bear the full expense of the actual service. In addition, their intentions may extend beyond cost reduction, with the fraudster's ultimate goal being to exploit the provider's network for profit. These fraudulent activities result in substantial financial losses for the telecom industry [8]. Fraudsters are continuously adopting more advanced fraud tactics to evade detection. They employ various fraud tools and concepts, including Benford's theory, Z Scores, Computer Assisted Auditing Tools (CAATs), and ratio analysis, to detect financial fraud. In contrast, Social Network Analysis (SNA) serves as a non-financial fraud detection method that assists organizations in identifying unusual connections among individuals and entities [9] [10].

Presently, telecom operators have implemented diverse fraud detection systems aimed at mitigating losses stemming from fraudulent activities. Those systems analyze customer data within telecom networks using a range of mathematical techniques to identify potential fraudulent customers also, those detection methods examine the information formed by the users discover the suspected user, and notify managers, in that way the loss caused by the frauds are reduced by the telecom operators [11]. Various Machine Learning and deep learning approaches are utilized in the field of fraudulent call detection [12]. Machine learning algorithms are employed to analyze call data and recognize patterns indicative of fraudulent activity. This can include supervised learning for classification or anomaly detection techniques. Creating user profiles and monitoring their calling behavior to detect deviations from normal usage, such as

sudden changes in call volume or international call activity also used to identify fraud. In the recent years, research efforts have primarily focused on crafting intricate features for phone call data using expert knowledge and subsequently applying machine learning algorithms for classification. These encompass conventional machine learning approaches such as Support Vector Machine (SVM) [13][14], Random Forest (RF) [15], as well as deep learning methods like Convolutional Neural Networks (CNN) [16], Long Short-Term Memory (LSTM) [12], and Recurrent Neural Networks (RNN) [17]. More recently, Graph Neural Networks (GNN) have come into play for extracting interactive insights among users and making predictions [18].

Most existing fraud detection approaches [19]-[21] have conducted their experiments using either synthetic data or undisclosed real-world data, making it difficult for other researchers to reproduce their results. Additionally, as CDR typically consists of unstructured data; employing a grid-based data processing method for their analysis can result in the loss of significant structural information, potentially leading to a substantial reduction in detection accuracy [22]. In the GNN [23] model if the issue of class imbalance is not taken into account during the design of the model, the majority class could disproportionately influence the loss function [24]. This could cause the trained GNN to excessively classify instances from these majority classes, leading to inaccuracies in predicting samples from the minority class, which is our primary concern. When applying these models to graphs with imbalanced class distributions, existing GNN approaches tend to excessively prioritize the majority class, resulting in less effective embedding results for the minority class. [25]. Additionally, the Neural Factorization Autoencoder (NFA) based method [26] is employed to detect fraudulent call activity, in which the execution time for the detection system is high. The traditional methods for outlier detection primarily depend upon clustering-based [27] and distance-based [28] perceptions. Nevertheless, these methods of telecom fraud detection depend deeply on the applicability of features to the task of categorization. When the feature space is huge, the occurrence of unrelated, unwanted, unacceptable, or noising features in the data may automatically affect the model's performance [29]. Moreover, the recent approaches have data complication, which typically comprises multi-level, multi-granularity, multi-dimensional data, that makes the processing and application of data intricate and miscellaneous [25]. Next, ambiguity, changeability, and repetition of the data make difficulties in telecom fraud detection. Also, mixed data which frequently encompass non-single heterogeneous data such as categorical and arithmetical data, makes the detection process challenging to process the data efficiently [30].

This research introduces an effective framework to identify fraudulent calls using SDHO-knowledge-enhanced Optimal GNN in the 5G/6G networks. Initially, the data collected from the

CDR graph dataset is verified with the data present in the blacklist table which provides information about the data as normal or abnormal. When data is deemed to normal, the data is transmitted to its intended destination via the data transmission module. This module verifies the information using knowledge-enhanced optimal GNN which enriches the graph representation by adding semantic information about nodes, edges, and relationships. The SDHO algorithm is employed to enhance the model's performance and efficiently fine-tune the classifier by incorporating the characteristics of sheepdog and coati optimization. When the information is cross-verified, the classifier determines whether the call is normal or abnormal. If it is determined that the call is abnormal, the data is forwarded to the prevention module, where the process is stopped by the prevention module, and the blacklist table is updated. The primary contribution of this research is further detailed in the ensuing discussion.

Sheepdog Hunt Optimization algorithm (SDHO): The SDHO algorithm is formulated by combining the hunting traits observed in coatis [31], together with the herding attributes of collie [32]. These characteristics combination eradicates issues like overall instability and computational complexity and it offers advantages by accelerating convergence speed and elevating classification accuracy.

Sheepdog Hunt Optimization algorithm enabled Knowledge-Enhanced Optimal Graph Neural Network (SDHO-KGNN): The integration of the SDHO algorithm with the KGNN enables the model to accurately detect fraudulent calls. The model can adapt to evolving fraud tactics by continuously updating its knowledge base, ensuring it remains effective in detecting emerging threats. KGNN significantly enhances fraud call detection by leveraging external knowledge to improve accuracy, reduce false positives, adapt to evolving threats, provide interpretability, and handle large-scale call networks.

The manuscript is further divided into the following sections. Section 2 explains the advantages and challenges of existing research, Section 3 includes the proposed methodology and block diagram, Section 4 describes the experimental results and comparative discussion of this research, and Section 5 expands the conclusion of the research.

6.2 Literature Review

Existing methods in the research of fraudulent call detection with advantages as well as drawbacks are elaborated in this section.

This [33] initiated a technique for detecting fraud, employing similarity and multi-view GNN. In this approach, a weight parameter is employed to increase the importance of fraud samples that have been flagged or labeled. However, there's a risk of overfitting the training data,

especially if not enough labeled fraud examples are available that [25] deployed a GAT-COBO model for fraud detection which can adapt to evolving fraud tactics and patterns in real-time, making them suitable for dynamic environments. However, GNNs can be challenging to interpret, making it difficult to grasp the logic behind the model's decisions that [7] designed a HESM approach for fraud detection which enhances the superiority of the model. However, the model can be challenging to interpret, making it difficult to comprehend the reasoning behind the model's choices.

A study [22] suggested a convolutional Neural Network (CNN) method for fraud detection using CDR data. In the suggested method, an end-to-end model was offered to recognize new distinctive features without automatically designed features. Also, the suggested model achieved satisfying outcomes on the challenges in the detection part of telecommunication. The developed method extracted behavior data from raw CDR effectively, nevertheless combining it with communicating data was still an unresolved issue. It [34] used a sim Box model for fraud detection in Telecommunication. The system can operate in real-time, allowing for immediate identification and response to fraudulent activities as they occur. However, ANNs can be challenging to interpret. This [35] presented a C-Bi-LSTM (Cluster Based algorithm for fraud detection, the utilization of intelligent algorithms allows for the automated and efficient identification of fraudulent call information, potentially reducing the workload for human operators. The output of intelligent algorithms may be challenging to interpret, making it difficult to understand the reasoning behind specific decisions. This [36] suggested a Reinforced Cost-sensitive Graph Network (RCGN) to detect the head of fraud in telecom Fraud. The deep Deterministic Policy Gradient (DDPG) technique was utilized to optimize the coefficients of mass dynamically and a graph was constructed using CDR and three base classifiers were united to predict the results and final results of classification. The integration of three base classifiers in the RCGN made the detection of fraud effective, however, the usage of a small size dataset in the RCGN method was the limitation, which lacked the experiments of comparison.

6.2.1 Challenges

- In the automated fraudulent call detection approach [12], the LSTM model attained limited performance due to the formation of data, which is not series at all times to acquire the arrangement relationship between each dimension. Moreover, the DNN model has a major risk of overfitting, which was exposed using the drift evaluation concept.
- Training and using ANN-based models may require substantial computational resources, which could be a limitation for some organizations. The system's efficiency relies significantly on the quality and comprehensiveness of the call traffic data. Inaccurate or partial data can

result in false alarms or overlooked instances of fraud. [37][34].

- Mobile social network fraud detection datasets often suffer from class imbalance, where genuine user activities significantly outnumber fraudulent ones. Handling this imbalance effectively is crucial [22].
- GAT-COBO heavily relies on the quality and availability of data. Noisy or incomplete data can affect its performance. GNNs can be challenging to interpret, making it difficult to comprehend the reasoning behind the model's choices [25].
- Detecting temporal patterns in telecom fraud, and enabling real-time detection and adaptation also face challenges related to data quality, complexity, interpretability, and resource requirements, which need to be carefully considered when applying the approach in practice [38].

6.3 SDHO Enabled KGNN

Crime and fraudulent call detection using SDH-KGNN for 4G/5G networks differs significantly from its use in 2G/3G networks due to the fundamental changes in network architecture and data richness. In 2G and 3G, communication patterns are mostly circuit-switched, generating limited metadata such as basic call detail records (CDRs), tower locations, and simple routing paths. These networks offer relatively sparse, linear interaction graphs, which restrict the depth of relational learning that a graph neural network can perform. By contrast, 4G/5G networks operate on fully packet-switched, IP-based architectures that produce richer and more diverse datasets including VoLTE/VoNR metadata, IP flow records, session management logs, handover patterns, QoS indicators, and application-layer behaviour. These dense, multi-relational datasets form complex heterogeneous graphs that SDH-KGNN can exploit more effectively, enabling deeper link reasoning, improved anomaly detection, and more accurate classification of fraudulent behaviours.

Furthermore, 4G/5G networks support massive device connectivity, dynamic mobility profiles, network slicing, and edge intelligence, all of which introduce new fraud patterns that did not exist in earlier generations. Fraudsters now exploit features such as VoIP tunnelling, virtual SIM identities, encrypted traffic, and rapid cell transitions that are harder to detect using traditional rule-based systems. SDH-KGNN becomes particularly valuable in this context because it integrates structured knowledge graphs with deep graph learning to capture hidden relationships across multiple layers of telecom data. In contrast, fraud detection in 2G/3G environments is more constrained and limited to simpler link structures. Thus, crime detection on 4G/5G is not only different but also more effective because SDH-KGNN can leverage the

high-dimensional, multi-layered nature of modern telecom networks to uncover sophisticated fraud behaviours that older networks do not reveal.

Detecting and mitigating fraudulent actions like telecom fraud, phishing calls, and robocalls that can result in financial setbacks for both service providers and customers is essential to enhance cyber resilience, privacy, and trust in the evolving 5G/ 6G networks. [33]. The major goal of this research is to detect fraudulent calls utilizing the SDHO-Knowledge enhanced Optimal GNN in the evolving 5G and 6G networks that often give network operators a significant deal of flexibility in allocating speed, coverage, and capacity for enhancing telecommunication. Initially, data will be gathered from the CDR graph dataset, preprocessed, and subsequently cross-referenced with the information contained in the Blacklist table. In the realm of telephone calls, a blacklist table serves as a repository of specific phone numbers or entities regarded as undesirable or potentially harmful. Telecommunication systems use this table as a point of reference to block or restrict calls originating from or directed to the numbers listed within it. The blacklist table categorizes the data as either normal or abnormal. When data is deemed to normal this is transmitted to its intended destination via the data transmission module. This module validates the information using the Knowledge-Enhanced Optimal GNN [38] which represents an innovative approach that merges the capabilities of GNN with external knowledge sources to augment the model's learning and inference abilities. The integration of external knowledge into GNNs via the Knowledge-Enhanced GNN framework offers several advantages by enriching the graph's representation with semantic information about nodes, edges, and relationships. This deepens the model's comprehension of the data it processes, leading to improved feature extraction and representation learning. Furthermore, an SDHO algorithm is developed by combining the traits of the sheepdog and the Coati. This algorithm facilitates effective tuning of the classifier. The classifier cross-validates the information, and if data is identified as normal, the call is recorded and forwarded to the intended destination. In contrast, if the data is classified as abnormal that is directed to the prevention module, where the process is terminated, and the blacklist table is updated accordingly. Figure 27 demonstrates the schematic illustration of the Fraudulent Call Detection mechanism utilizing the SDHO-enabled knowledge enhanced optimal GNN model.

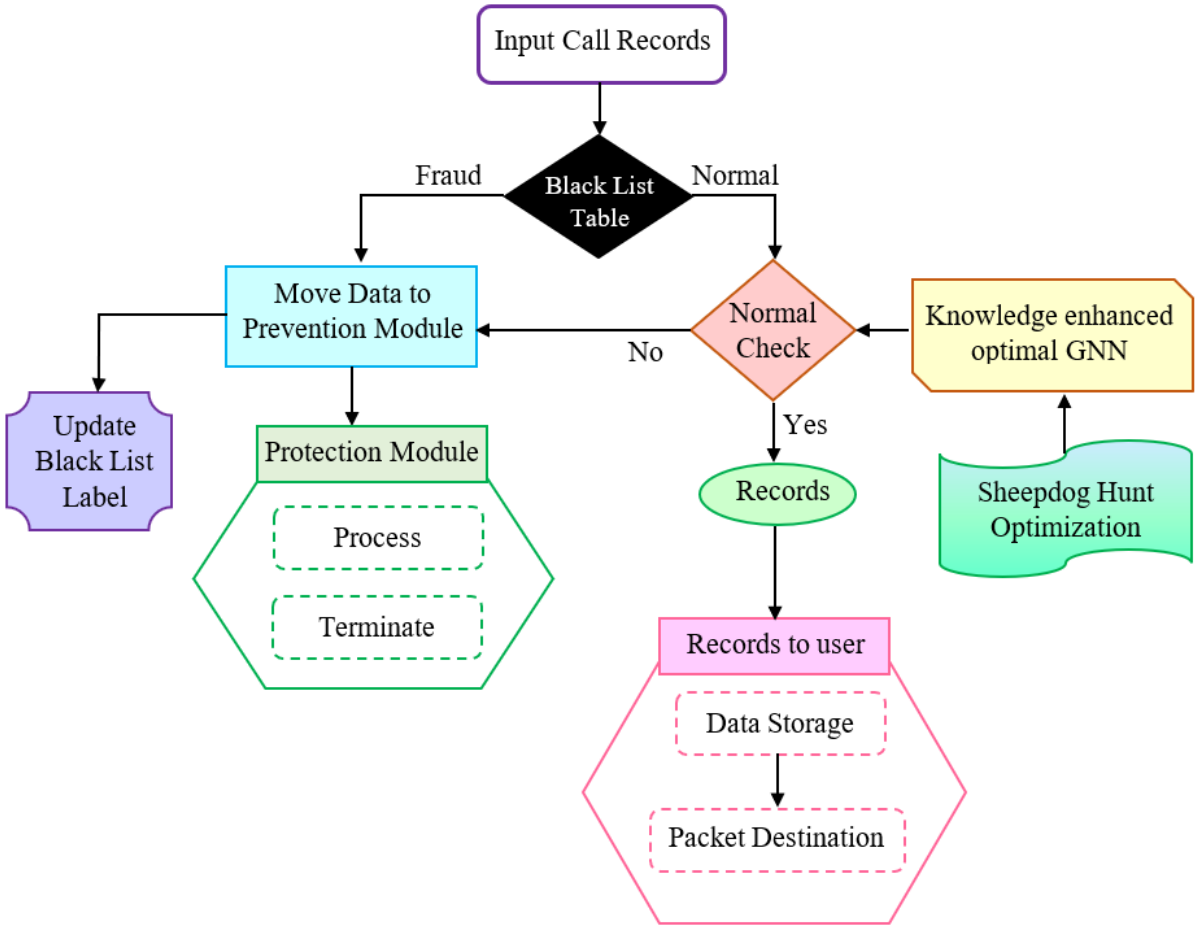


Figure 27: Schematic representation of the Fraudulent Call Detection utilizing SdHO-enabled knowledge-enhanced optimal GNN model

6.3.1 Methodology

Input telecom records

Here the data is collected from the CDR Graph Dataset [39] typically includes structured data representing various aspects of telecommunications activities in 5G/6G networks. These call records serve as the foundational data for analysis. These call records are typically organized as nodes within a graph dataset, with connections (edges) representing relationships or interactions between them.

Preprocessing of Call Records

Preprocessing is required to enhance the quality of the data since the raw data typically contains instances that are weakly regulated and may have out-of-range and missing values. Here the missing values from the CDR Graph Dataset are handled by removing the rows with non-vales ensuring a cleaner dataset.

Blacklist table

A blacklist table is a database or list containing specific items, such as phone numbers,

email addresses, IP addresses, or entities that are considered undesirable, potentially harmful or suspicious. The primary purpose of a blacklist is to restrict or block access to or communication with the items within it. In telecommunications, blacklists are used to prevent fraudulent activities and are often used to protect individuals' privacy and security. Unwanted behavior serves as the basis for blacklist inclusion, and its dynamic nature keeps it up to date against new threats. The adoption of blacklists is primarily motivated by security and privacy concerns, offering people and organizations a strong defense against known and unknown threats. A proactive strategy, connection with larger security systems, and frequent updates all enhance a blacklist table's ability against possible threats.

Training phase

The knowledge graph neural network (KGNN) is trained on a CDR graph dataset which involves particular procedures designed for telecom data. Here the preprocessed data is ready for the KGNN by encoding call kinds, stamps, and phone numbers into numerical representations. Embedding layers for nodes, capturing distinct representations for every phone number, and attention techniques to identify intricate links in call patterns are common features of the model architecture. In training, the KGNN compares and predicts associations inside the CDR graph to minimize a loss function.

Knowledge-enhanced Optimal Graph Neural Network classifier

Graph-based fraud detection has gained popularity and considerable attention due to the prevalence of fraud incidents that manifest as graph structures. One category of research in this domain centers on identifying unusual nodes [40][41] and specific abnormal subgraph configurations [42][43] within static networks. A Knowledge-Enhanced Graph Neural Network (KGNN) for fraud call detection typically extends a standard GNN architecture with the integration of domain-specific knowledge graphs which can capture valuable domain-specific information from knowledge graphs which can lead to better performance in fraud call detection. By incorporating external knowledge, KGNNs can generalize better to unseen or rare patterns. KGNNs can transfer knowledge from one domain to another through shared knowledge graphs. This transfer learning capability is valuable in situations where data from one domain can inform predictions in another.

Constructing a graph where nodes represent subscribers in the mobile network, the node features represent the behavior of these subscribers, and the edges represent communication interactions between subscribers, enables us to build a network representation. Fraudulent subscribers within the network are potentially identified using GNNs [44].

A user behavior graph is denoted as $U = (B, \chi, A_d, E, N)$ established within a mobile network context. Here, B represents the set of users in the mobile network as nodes, where $B =$

$\{w_1, w_2, w_3, \ldots\ldots\ldots\ldots\ldots w_n\}$. The edges in the graph represent as $E = \{\alpha_1, \alpha_2, \alpha_3, \ldots\ldots\ldots\ldots\ldots \alpha_n\}$. $\alpha_j = (w_{cj}, w_{zj}) \in E$ is an edge between w_{cj} and w_{zj}, where $w_{cj}, w_{zj} \in B$. Additionally, $\chi = \{s_1, s_2, s_3, \ldots\ldots\ldots\ldots s_n\}$ denotes the collection of original behavioral features of subscribers, with each $\chi_i \in R^d$ representing the behavioral feature vector of the userw_i. These feature vectors are aggregated into the feature matrix $\chi_i \in R^{N*D}$ for the graph U. The adjacency matrix $A_d \in R^{N*N}$ encodes the relationships within the graph, where $a_{i,j} = 1$ indicates the presence of an edge between nodes i and j, while it signifies a connection. Lastly, $N = \{r_1, r_2, r_3, \ldots\ldots\ldots\ldots r_n\}$. constitutes the labels associated with nodes in the set B. In the context of a given subscriber behavior graph, the utilization of graph neural networks facilitates the learning of node embedding representation. The general formulation of GNNs is given as,

$$h_{out}^l = update(Agg(\{h_v^{l-1}, \forall v \in N(v)\}, W_{agg}^l), h_{in}^l, W_{update}^l) \quad (34)$$

Here h_{in}^l and h_{out}^l represents the input and updated node embeddings at the lth layer, correspondingly. h_{out}^{l-1} refers to the embeddings of neighboring nodes from the previous layer. $W_{agg}^l$and W_{update}^l denote trainable matrices for the aggregation function and update function at the lth layer, correspondingly. These notations are commonly used to describe the elements and operations in GNN architecture for node embedding and updating processes.

$$h_w^l = \sigma(\left[C_l.agg\left(\left\{h_v^{(l-1)}, \forall v \in N(v)\right\}\right), D_l h_w^{l-1}\right]) \quad (35)$$

Here, the notations and process described can be summarized as follows: h_w^l represents the node w at the current layer l and h_w^{l-1} represents the previous layer (l-1), $h_v^{(l-1)}$ denotes the node embeddings of neighboring nodes from the previous layer, C_l and D_l are trainable weight matrices, σ represent a nonlinear activation function, agg stand for the aggregator. The generation of h_w^l follows these steps:

Initially, each node aggregates features from its nearby neighbors using the AGG aggregator, resulting in a unified vector h_w^l. Then graph SAGE then appends the node's prior representation to its neighborhood feature vector. The concatenated vector is subsequently processed through an MLP with a softmax activation function. The output produced by the MLP becomes the updated representation for the node, which is utilized as input in the subsequent layers.

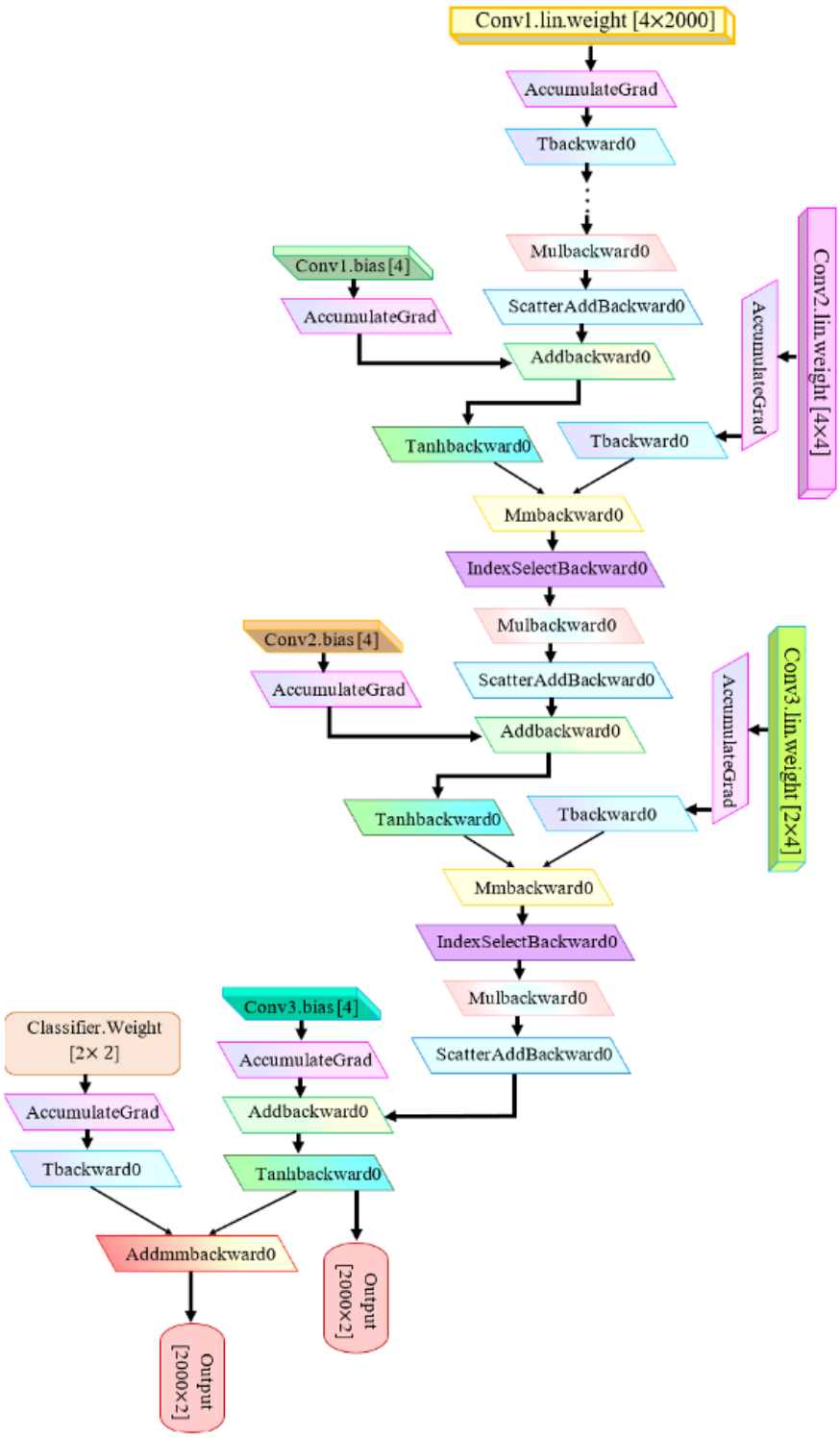


Figure 28: Architecture of knowledge enabled optimal GNN

The domain-specific knowledge graph is integrated into the model. The attention mechanism is employed to determine the impact of knowledge nodes on the message-passing procedure, which is denoted as,

$$h_w^{(l+1)} = propagate\left(h_w^{(l)}, \left\{h_v^{(l)}, \forall w \in N(v)\right\}\right) + Q(h_w^{(l)}, \{k_j, \forall j \in knowledge\ graph\}) \quad (36)$$

Where k_j represents the knowledge nodes in the knowledge graph.

The attention mechanism for knowledge graph integration is as follows,

$$Q\left(h_w^{l-1}, \{k_j, \forall j \in knowledge\ graph\}\right) = softmax(\{\alpha_j * f\left(h_v^{(l)}, k_j\right) \forall\ j \in knowledge\ graph\}) \quad (37)$$

The softmax is utilized for binary fraud/non-fraud predictions. Figure 28 shows the architecture of knowledge-enabled optimal GNN.

6.4 Sheep Dog Hunt Optimization

The proposed SDHO Algorithm draws inspiration by combining the hunting traits of coati [31] and herding traits of the Border Collie [32] to find the optimal solution that promotes the detection accuracy. By incorporating these characteristics, the SDHO algorithm can prioritize making smart decisions based on evolving patterns in the call data.

6.4.1 Motivation

The SDHO algorithm is utilized to deal with complex high-dimensional problems, which potentially reduce computational bottlenecks in large-scale fraud detection systems. This can lead to faster and more scalable fraud detection. The SDHO algorithm efficiently generated the optimal solution by preventing the algorithm from getting trapped in the local optima. The SDHO algorithm shows its great ability in balancing the search process, this can be particularly useful when dealing with complex call data among various fraud detection mechanisms. Additionally, the SDHO algorithm allows high-speed convergence to offer appropriate values for decision variables in detection tasks

Coatis referred to as coatimundis, belong to the Procyonidae family and are categorized under the Nasua and Naselle genera [45], are mammals with an omnivorous diet, which includes the consumption of both invertebrate and small vertebrate prey. The diet of these arboreal animals often includes iguanas, which are typically located in trees. Their hunting strategy is characterized by coordinated group efforts, wherein some individuals ascend trees to induce the iguana to descend while others engage in rapid pursuit and capture [46]. Notwithstanding their proficient tactics, they are susceptible to predators and avian predators of substantial size. This behavior can inspire an optimization algorithm that involves multiple agents collaborating to solve complex problems, such as detecting fraudulent calls, where different pieces of information must be gathered and analyzed collectively. Canis lupus familiar, commonly known as sheepdog [32], is an extraordinary dog breed that is renowned for its exceptional herding capabilities and

strong work ethic and exhibits a remarkable aptitude for situational assessment and adaptive decision-making [47]. This inspired the development of a metaheuristic algorithm that emulates their behavioral traits. In the context of herding, these canines employ a unique strategy, opting for lateral and frontal approaches rather than traditional rearward herding methods. In the herding context, canines employ a gathering technique where they exert control from the flanks and the front to collect and guide livestock toward a designated location, commonly referred to as gathering. In the process of controlling livestock, canines exhibit a behavior akin to that of wolves. This includes crouching, bowing their heads, raising their hindquarters, and angling their tails downward, a conduct referred to as stalking. Canines emulate the victim selection behavior observed in wolves, referred to as giving an eye or eyeing. In instances where individuals within the group exhibit deviant behavior, these highly intelligent canines employ an intense gaze directed at the individuals in question, applying psychological pressure to guide the collective movement in the desired direction. The algorithm could draw inspiration from their speed and efficiency to develop optimization techniques that operate swiftly and effectively in processing and analyzing large volumes of call data.

6.4.2 Initialize the population

Population initialization is the process of creating an initial set of potential solutions that the algorithm will evolve and improve upon during the optimization process. The key goals of population initialization are diversity and coverage of the search space to increase the chances of finding a global optimum.

$$X^{tH} = X^t + r_1\left(I_g - I.X^t\right) + r_2(u - 1) \tag{38}$$

where r_2 represents a random number, u and l is denoted as upper and lower bound

$$r_2 = \frac{r_{max} - r_{min}}{r_{max}} \in [0,1] \tag{39}$$

$$r_1 = e^{\left(\frac{X_{worst} - X^t}{\sum X_{best}}\right)} \tag{40}$$

6.4.3 Equation Calculations

Fitness evaluation

The fitness function assesses the solution's quality for every individual within the population; apply the fitness function to calculate its fitness score. This step involves solving or simulating the problem using the individual's parameter values or configuration and measuring its performance based on the fitness function. Higher fitness scores indicate better solutions.

$$fit = \max\left(acc(X^t)\right) \tag{41}$$

Claim the population gaps

Split the population equally based on fitness. Solutions with higher fitness are regarded as Group 1 (guard), while solutions with fitness below $fit(X^{best})$ are grouped under Group 2(herd).

Herd population update

The herd population represents a set of potential solutions or individuals that are being evolved or optimized toward finding an optimal solution for a given problem. These are herd candidates, which obey or follow the best solutions for sooner convergence.

$$X^{tH} = V^{tH}.D^{tH} + \frac{1}{2}A^{tH} \times D^{tH} + r_3(X^t_{per} + X^t) \tag{42}$$

Where V^{tH} denotes velocity, D^{tH} is the duration of travel, A^{tH} is represented as acceleration.

$$V^{tH} = \left(\sqrt{V^t.(tan\theta)^2 + 2.A^t \times X^t}\right) + \delta\left(X_{global} - X^t\right) \tag{43}$$

Here $\delta = 0.5$ when the herd is away from the guard, $\delta = 1$ when the herd is near to guard.

Guard population update

The purpose of the guard population is to dynamically adapt the search behavior of the algorithm, ensuring a proper balance between exploration and exploitation. It helps in achieving an optimal convergence rate while maintaining the diversity of the solutions. The update process of the guard population is typically performed periodically to ensure that the algorithm continuously adapts to the evolving search landscape. By updating the guard population in each iteration, the algorithm can adjust its exploration and exploitation based on the current state of the search.

$$X^{tH} = X^t + \frac{1}{2}\alpha X^{t-1} + \frac{1}{2}\alpha(1-\alpha)X^{t-2} + V^{tH} \times D^t + \frac{1}{2}A^t.(T_r)^2 + V^{tH} * L \tag{44}$$

Here is the Loudness factor to assemble the herd and represents traversal time. A guard should be aware of enemies to protect the herd.

$$V^{tH} = V^t + r_4.c_1\left(X_{pee} - X^t\right) + c_2\left(X_{global} - X^t\right) + D^t c_3 - X^t * \partial^t \tag{45}$$

Distance direction, alertness factor c_2 , and awareness coefficient c_1 help guard to protect the herd from enemies

Enemy modulation

The enemies follow a tricky process of the fox and they try to betray the herd. The vision radius is defined through the observation angle and the scaling parameter, which is represented as,

$$g = \begin{cases} a\dfrac{\sin(\phi_0)}{\phi_0} & if\ \phi_0 \neq 0 \\ \theta & if\ \phi_0 = 0 \end{cases} \tag{46}$$

Where g represents the vision radius, $\emptyset_o \in (0,2\pi)$ is the observation angle, which is selectedfor every individual, θ indicates the arbitrary value among 0 and 1, which is set at the beginning of the algorithm, a belongs to 0 and 0.25 represents the parameter of scaling which is

set after the iteration.

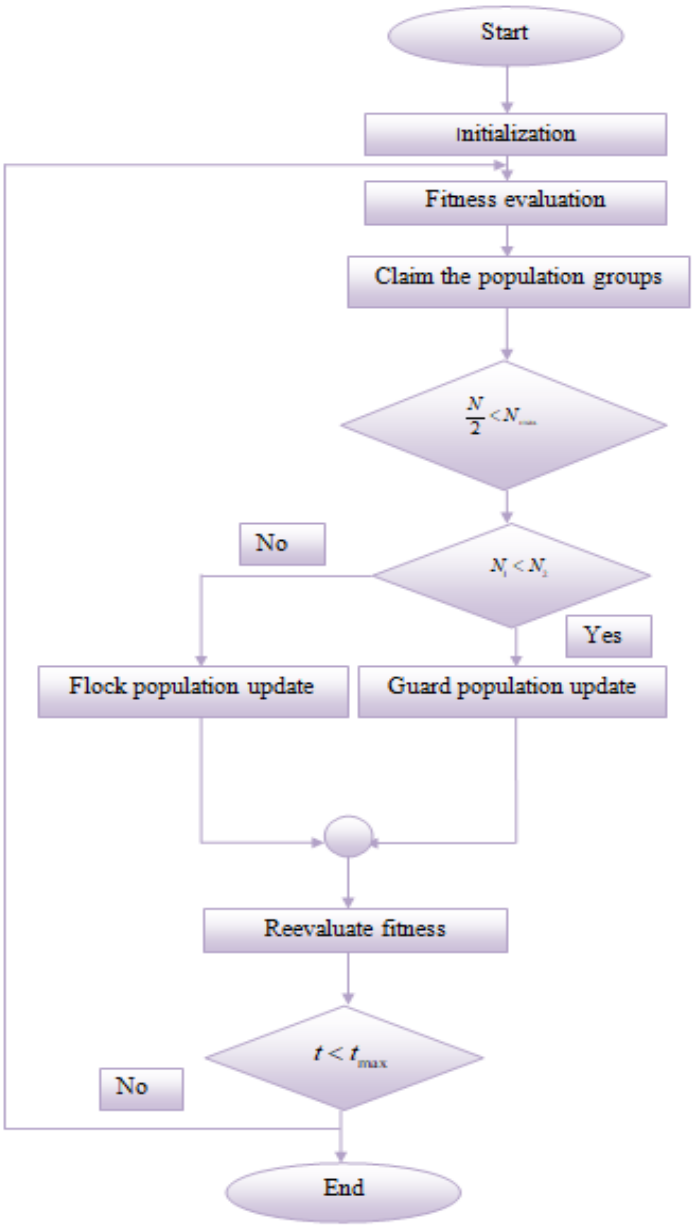


Figure 29: Flowchart for SDHO Optimization

of every individual in the inhabitants to model arbitrarily varying the distance from the solution during the enemy approaching. The position of enemies is modulated as,

$$\begin{cases} X_0^{new} = ag.\cos(\phi_1) + X_0^{actual} \\ X_1^{new} = ag.\sin(\phi_1) + ag.\cos(\phi_2) + X_1^{actual} \\ X_2^{new} = ag.\sin(\phi_1) + ag.\sin(\phi_2) + ag.\cos(\phi_3) + X_2^{actual} \\ .... \\ X_{n-2}^{new} = ag.\sum_{k=1}^{n-2}\sin(\varphi_k) + ag.\cos(\phi_{n-1}) + X_{n-2}^{actual} \\ X_{n-2}^{new} = ag.\sin(\phi_1) + ag.\sin(\phi_2) + ... + ag.\sin(\phi_{n-1}) + X_{n-1}^{actual} \end{cases} \quad (47)$$

The location is updated based on the position of the enemy to attain the optimal solution. This model signifies the characteristics of an enemy after it spots a prey and attempts to approach as near as possible to strike and if fails when exposed to attempts to access the alternate one likewise.

6.4.4 Termination

If $t < t_{max}$, the loop ends, else the process continues to fitness evaluation population split and so on. Figure 29 shows the overall flowchart for the SDHO Optimization algorithm and Algorithm 1 demonstrates the pseudocode of the proposed SDHO-KGNN model.

Table 6: Pseudocode for the proposed SDHO-KGNN model

```
Input: Data obtained from the CDR Graph Dataset
Output: Fraud Call or Non-Fraud Call
Read the Call logs
    Call Pre-Process D_c#Pre-processing step initiated.
      {
      P_Dc = Im putter (D_c)
       return  P_Dc
        }
        Call blacklist table #verifies the Blocklist table
        Execute Feature Extraction(D_c) #If incoming is not available in the Blocklist table
          {
          FE=f_V1, f_V2, ... ... f_Vn
           return FE
            }
                Call GraphNeuralNetwork (FE) #Features as Input- Fraudulent call detection.
                    {
                     GCN Layer (FE, Hidden Size)
                      GCN Layer (Hidden Size, 2)
                      }
                    Call GNN Training (FE, Label) #Generate Trained Model
                    {
                     Model= GraphNeuralNetwork ()
                     Model.Fit (train data, train label)
                     Call Optimization (Model)
                      {
```

```
                        W_g=Model.get weights ()
                         Model=Update hyperparameters (W_g)
                         Return model
                         }
Call Evaluation (model, testing_data) #Model Performance Evaluation
{
 Prediction =model.predict(testing data)
 if prediction==Fraud
    {
    Update blacklist table
     }

        Return
          End- }
```

6.5 Result and discussion

This section presents the research findings of the SDHO-enabled knowledge-enhanced optimal GNN classifier for fraudulent call detection clearly and objectively.

6.5.1 Experimental setup

The experimental setup employed the PYTHON on a Windows 11 operating system equipped with 8GB of RAM. These selections of programming language and the operating system were deliberately chosen to align with the needs of the implementation. The system's 8GB of RAM provided the necessary computational resources for the successful execution of the experiments.

6.5.2 Dataset Description

CDR graph dataset [39]: A CDR (Call Details Record) Graph Dataset is a collection of structured data that represents telecommunications activities, particularly call and message interactions between users. This dataset provides valuable insights into the communication patterns and behaviors of individuals or entities in a network. The dataset comprises information on 1,01,174 customers, encompassing 17 attributes, including an indicator of whether a customer has churned. Among the customers, there are a total of 8,830 churners. Some of the variables in the dataset include state, account length, phone number, international plan, mail plan, number of voice mail messages, total day minutes, and so on. They comprise various data fields including called number, caller number, time, and date of the call. The attributes of the dataset are Gmail message, day charge, eve mins, night charge, Cust Serv Calls, eve calls, Intl charge, Intl Mins, night calls, day mins, eve charge, night mins and Intl calls.

6.5.3 Performance Metrics

The metrics utilized to assess the effectiveness of fraud call detection are accuracy, precision, and recall, which are formulated as,

$$accuracy = \frac{C_{TP}+C_{TN}}{C_{TP}+C_{TN}+C_{FP}+C_{FN}} \quad (48)$$

.

$$precision = \frac{C_{TP}}{C_{TP}+C_{FP}} \quad (49)$$

$$recall = \frac{C_{TP}}{C_{TP}+C_{FN}} \quad (50)$$

where C_{TP} denotes the True positive, C_{TN} indicates the True Negative, C_{FP} is False Positive, and C_{FN} is the False Negative.

6.5.4 Experimental Results

In this section, Figure 30 displays the input for K-GNN, this shows the initial formation of the model process, potentially identifying the key factors influencing its predictions. Figure 4 comprehends the nature of input data in a more complex manner improves transparency and makes it easier to grasp the KCNN's following outputs.

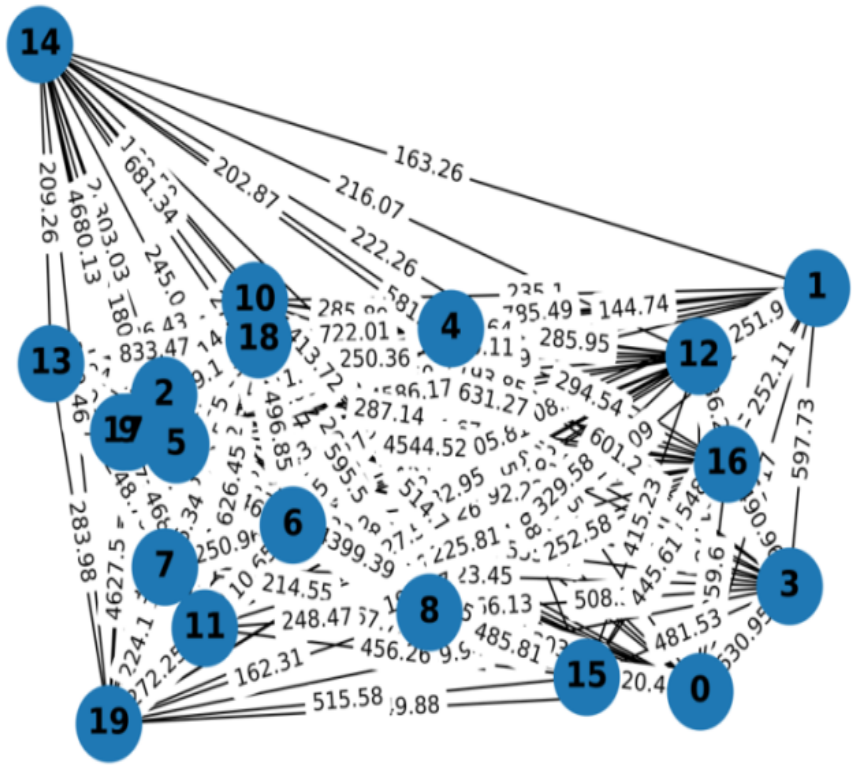
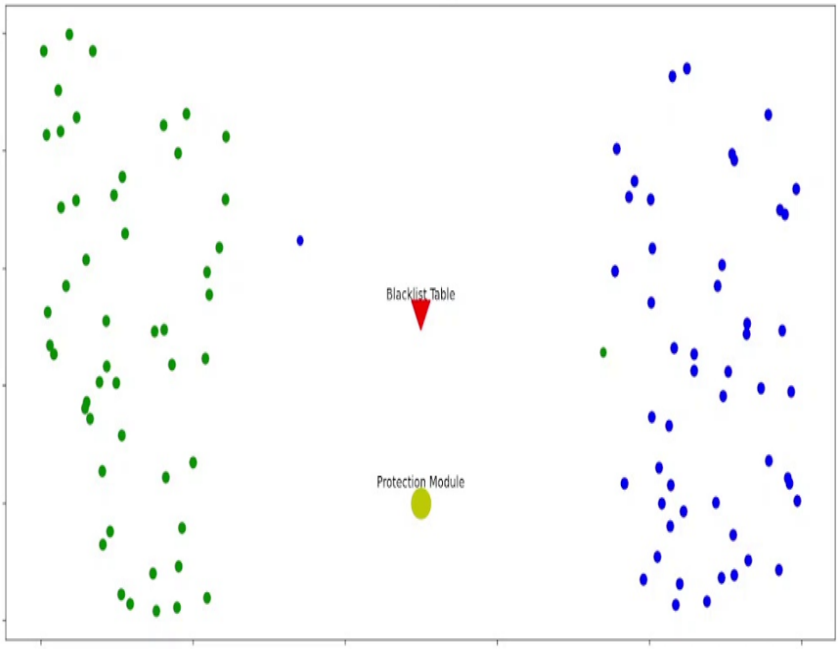
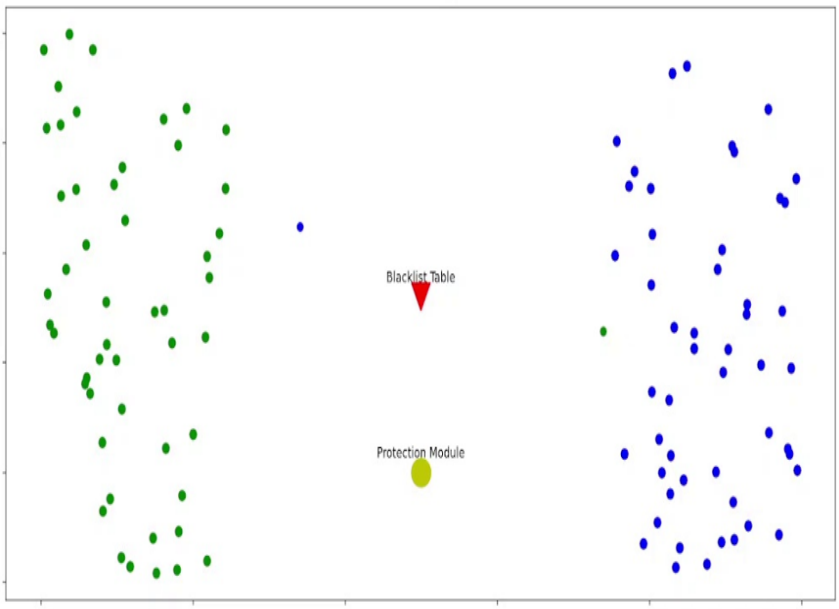


Figure 30: Input graph for k-GNN model

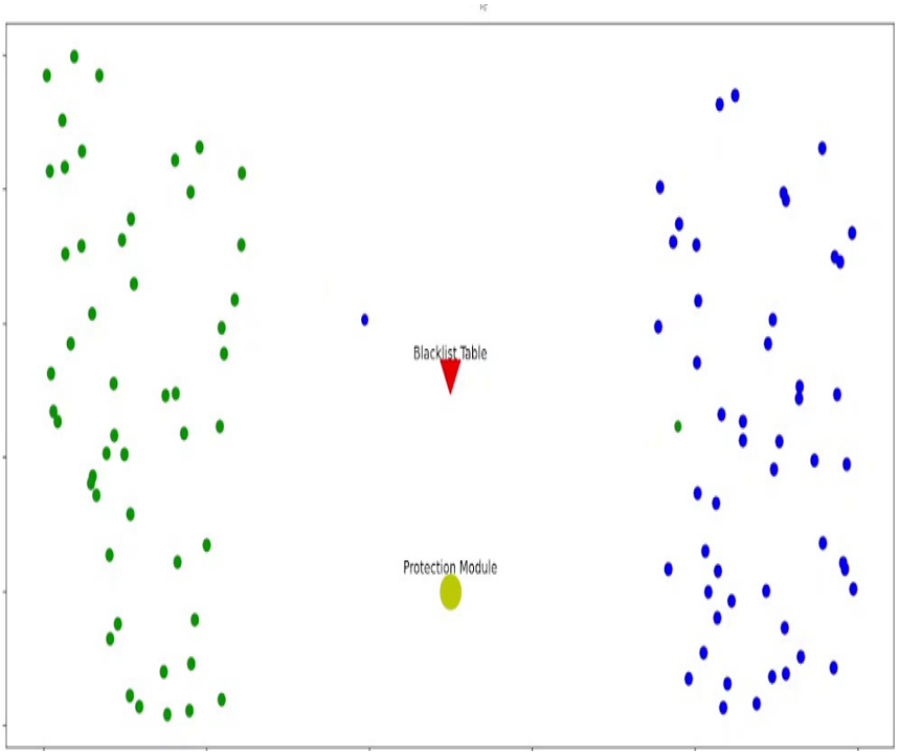
Figure 31 demonstrates the image results obtained from the experiment. Green points represent the sending nodes, while blue points represent the receiving nodes. Initially, calls from the sender are analyzed using the blacklist table. If a call is identified as fraud, it is routed to the protection module and terminated; otherwise, it proceeds to reach its destination. This identification of fraud calls indicates the model's ability to effectively detect fraudulent calls. This visual representation makes it easier to quickly and thoroughly comprehend the effectiveness of the fraud call detection mechanism. The system's ability to successfully route valid calls to their intended recipient while blocking or ending the fraudulent ones using the blacklist table can be seen in Figure 31.



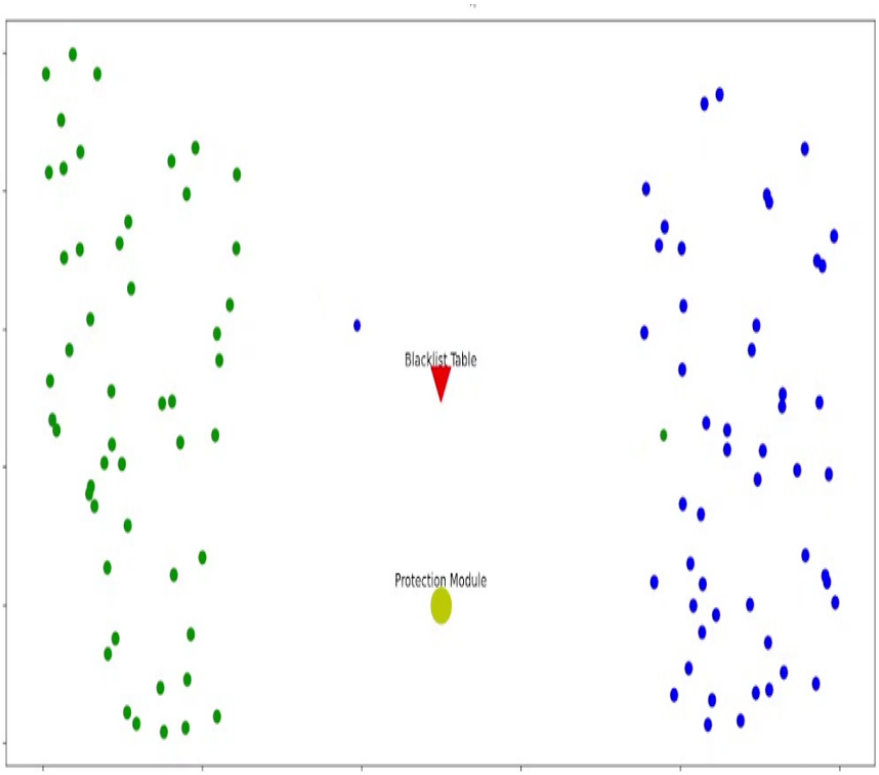
Frame 1



Frame 2



Frame 3



Frame 4

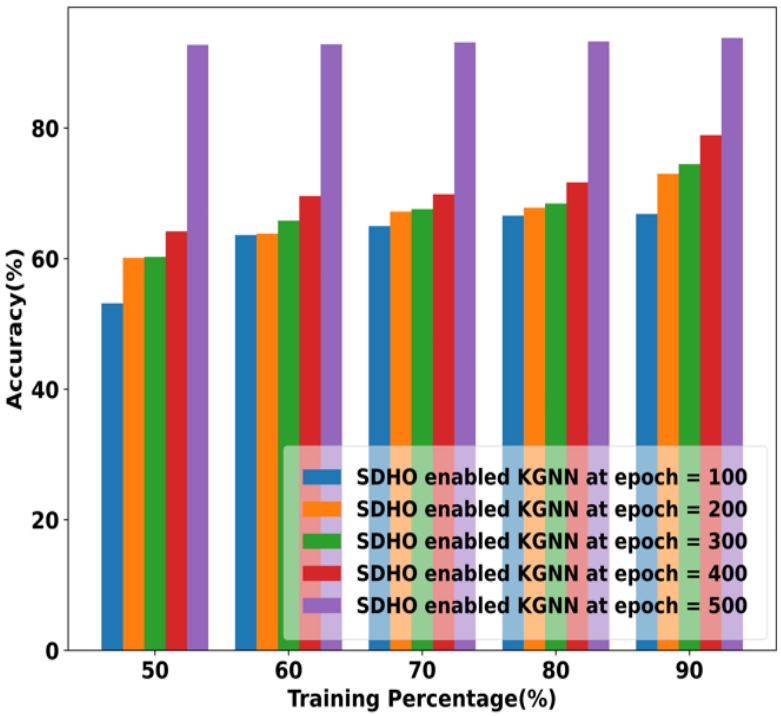
Figure 31: Experimental results of the developed model

6.6 Performance analysis for SDHO enabled KGNN

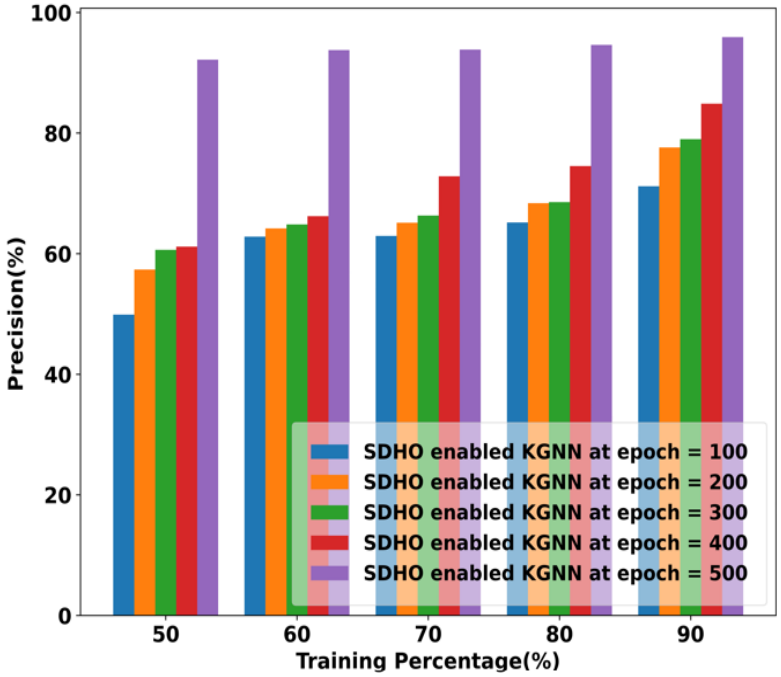
The efficacy of SDHO-enabled KGNN is assessed using a CDR Dataset, employing the TP metric, and employing various k-fold cross-validation schemes with folds labeled 6, 8, and 10. Furthermore, TP values ranging from 50 to 90 are taken into account, along with different epoch sizes of 100 to 500.

6.6.1 Performance analysis with Training Percentage

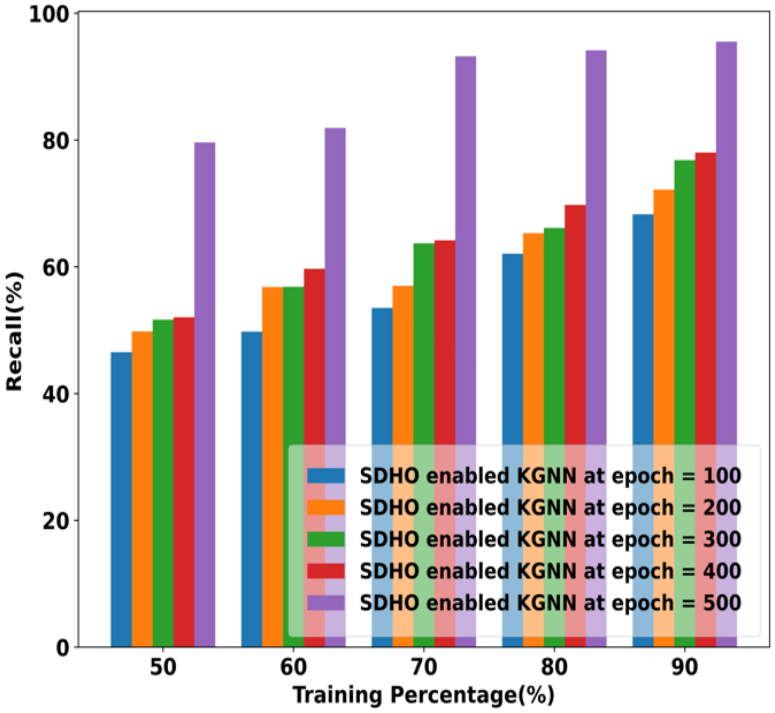
The performance outcomes of SDHO-enabled KGNN in fraud call detection, considering various epoch values, are visualized in Figure 32. In Figure 32(a), the accuracy is measured in terms of TP 90 for epochs 100, 200, 300, 400, and 500 yielding values of 66.83%, 73%, 74.47%, 78.91%, and 93.8%. Similarly, Figure 32 (b) shows precision for TP 90, where the SDHO-enabled KGNN achieves values of 71.17%, 77.61%, 78.98%, 84.88%, and 95.91% for respective epochs. Moreover, in Figure 32 (c), the SDHO-enabled KGNN shows specificity values of 68.28%, 72.17%, 76.82%, 78.02%, and 95.53% for TP 90, corresponding to the different epoch values.



(a) Accuracy



(b) Precision



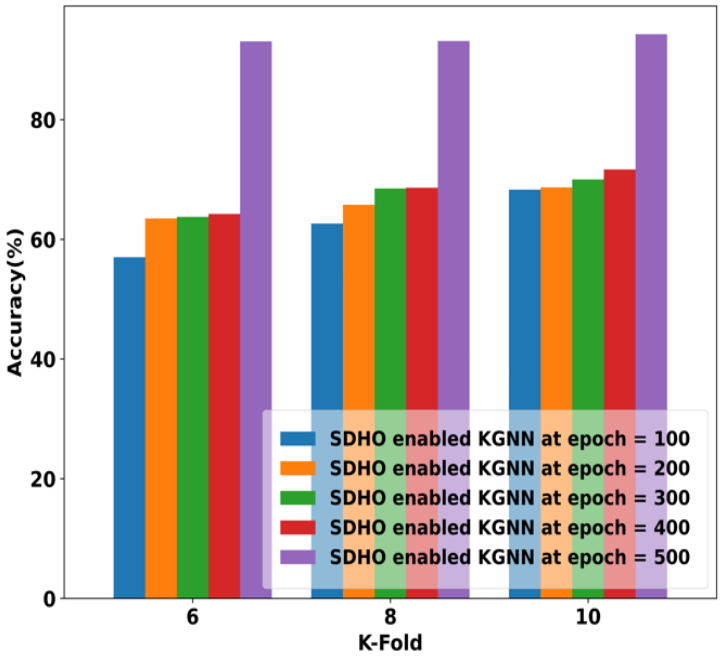
(c) Recall

Figure 32: Performance analysis with TP

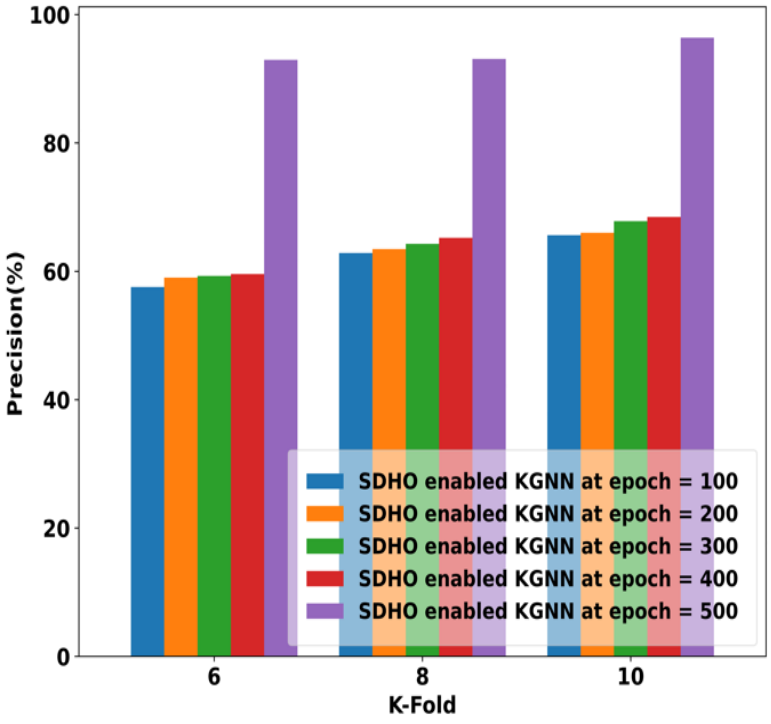
6.6.2 Performance analysis with K-fold

The results depict the SDHO-enabled KGNN's performance in detecting fraud calls based on various epoch values as shown in Figure 33. Figure 33(a), the evaluation of SDHO-KGNN

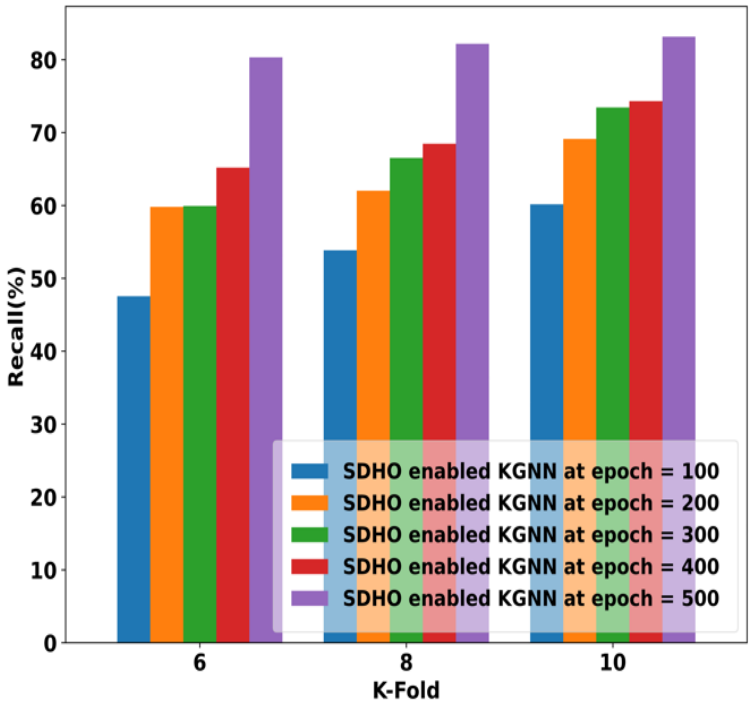
with epoch values 100, 200, 300, 400, and 500 attains an accuracy of 68.27%, 68.67%, 70%, 71.65%, and 94.25% respectively for k fold 10. Figure 33(b) presents the evaluation of SDHO-enabled KGNN with epoch values of 100, 200, 300, 400, and 500, focusing on precision within the context of a 10 k-fold cross-validation. The corresponding sensitivity values achieved are 65.61%, 65.99%, 67.79%, 68.46%, and 96.4%. Moreover, Figure 33(c) demonstrates the SDHO-enabled KGNN's performance, the attained recall values are 60.15%, 69.12%, 73.43%, 74.31%, and 83.15% in terms of k fold 10 for respective epochs.



(a) Accuracy



(b) Precision



(c) Recall

Figure 33: Performance analysis with k-fold

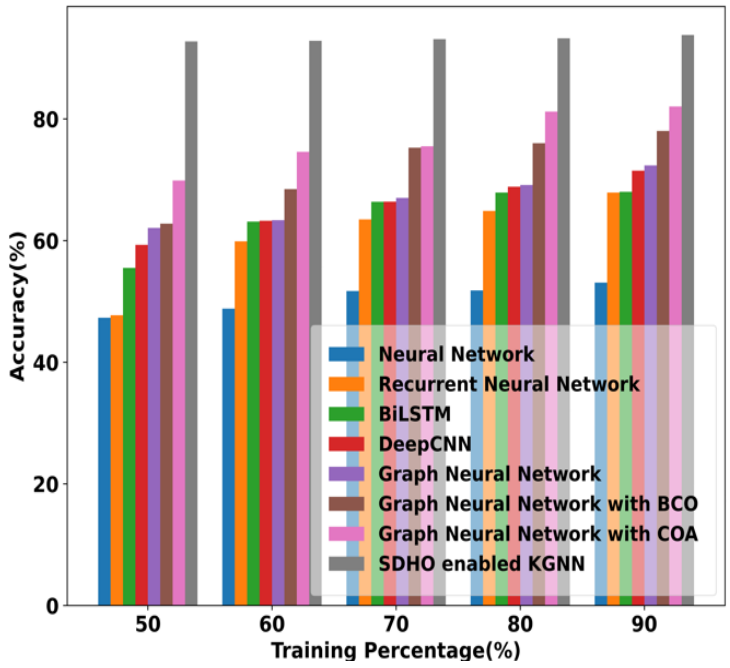
6.7 Comparative Methods

The following techniques are used in the comparison: Neural Network (NN) [48], Recurrent Neural Network (RNN) [17], Bi-LSTM [35], Deep CNN (DCNN) [49], Graph Neural Network (GNN) [50], Graph Neural Network with BCO (GNN-BCO) [51], Graph Neural Network with COA (GNN-COA) [44], SDHO enabled KGNN.

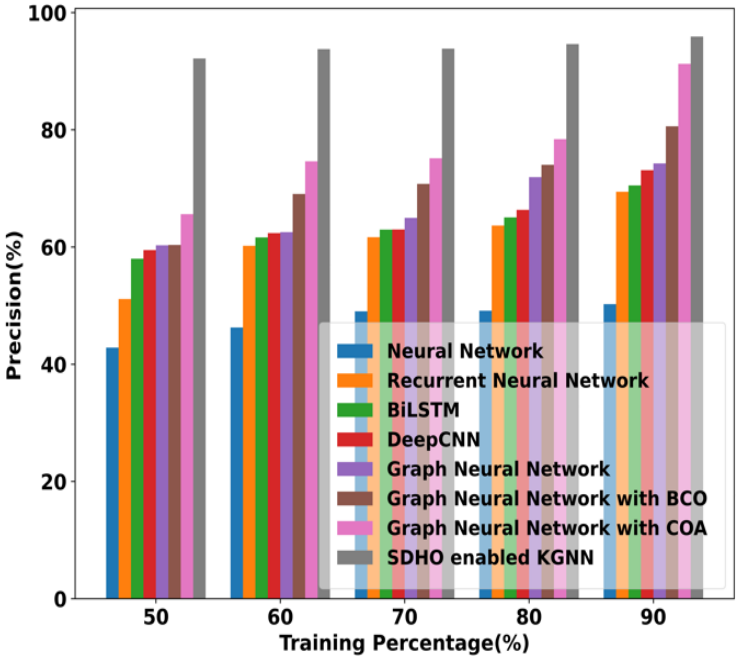
6.7.1 Comparative Evaluation with Training Percentage

Figures 8 visually illustrate a comparison of the SDHO-enabled KGNN accuracy, precision, and recall strategies with those of other methods. As shown in Figure 34(a), the accuracy of SDHO-enabled KGNN at TP 90 is 93.8%, an improvement of 43.41% over the current NN, 27.61% over RNN, 27.49% over BI-LSTM, 23.77% over DCNN, 22.86% over GNN, 16.81% over GNN-BCO and 12.53% over GNN-COA. This improvement is due to the use of SdHO which enables the KGNN to make accurate predictions by tuning its parameters. A similar increase of 47.61% compared to NN, 27.62% compared to RNN, 26.49% compared to BI-LSTM, 23.78% compared to DCNN, 22.57% compared to GNN, 15.97% compared to GNN-BCO, and 4.87% compared to GNN-COA can be shown in Figure 34(b), where the precision of

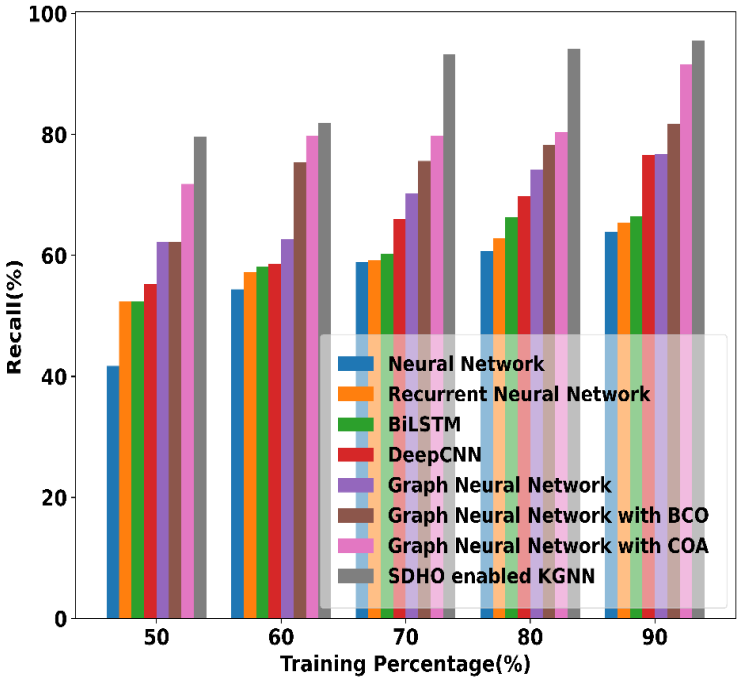
the suggested technique at TP 90 is assessed at 95.91%. This shows that the model reduces the false positives and enhances the model's ability to precisely identify the relevant instances. Additionally, in Figure 34(c), it is evident that the recall achieved by the SDHO-enabled KGNN



(a) Accuracy



(b) Precision



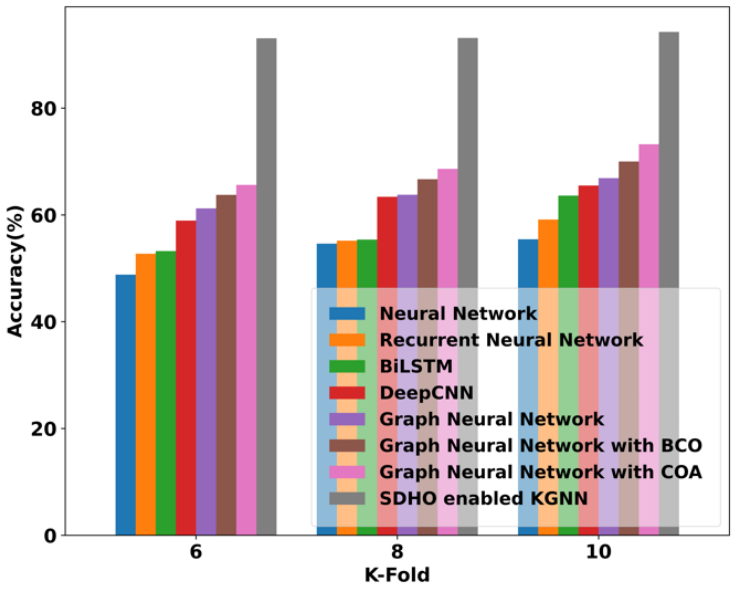
(c) Recall

Figure 34**:** Comparative analysis with TP

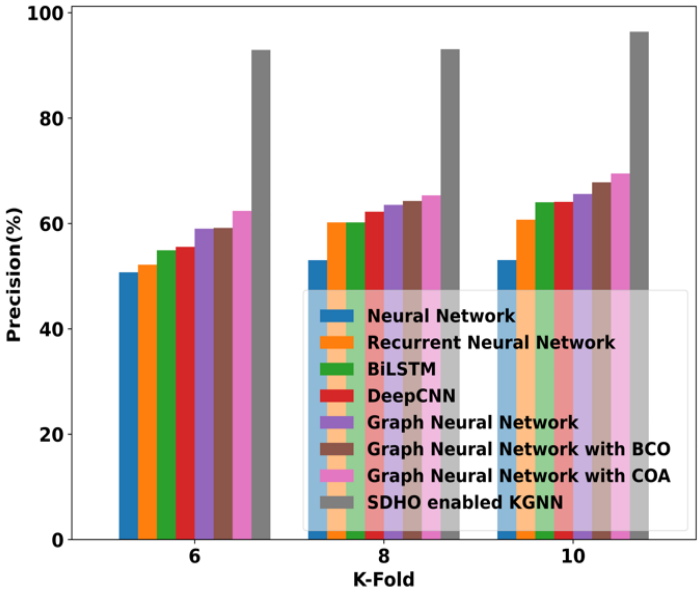
reaches 95.53% at TP 90 showing that the model excels in identifying positive cases. This demonstrates a significant enhancement compared to NN by 33.18%, RNN by 31.56%, BI-LSTM by 37.9%, DCNN by 19.86%, GNN by 19.86%, GNN-BCO by 19.57%, and A7 by 14.39%. The enhancement in these outcomes indicates that the suggested approach surpasses the other conventional techniques, proving its effectiveness in identifying fraudulent calls.

6.7.2 Comparative analysis with K-fold

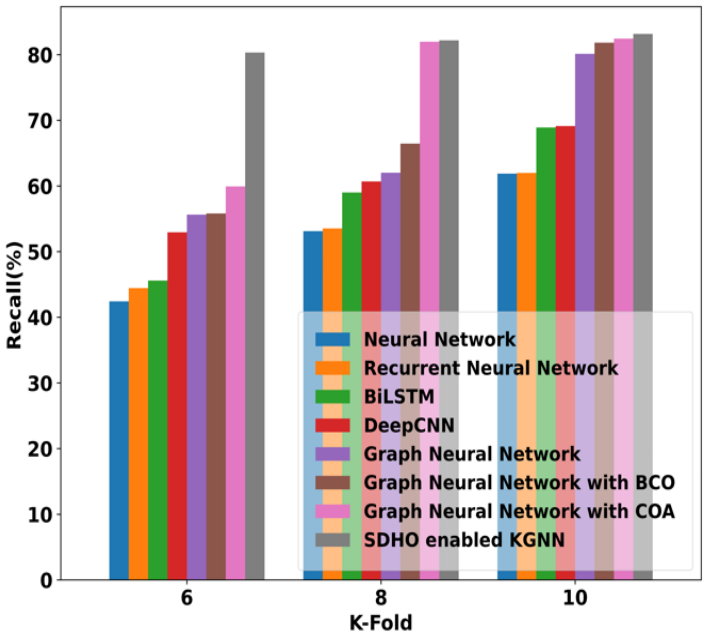
The SDHO-enabled KGNN underwent a comparative analysis, and the results were measured using accuracy, precision, and recall, as illustrated in Figure 9. With a k-fold value of 10, the accuracy achieved by SDHO-enabled KGNN reached 94.25%, showing effective improvement of 41.19% over the existing NN, 37.26% over RNN, 32.51% over BI-LSTM,



(a) Accuracy



(b) Precision



(c) Recall

Figure 35: Comparative Analysis with K-Fold

30.51% over DCNN, 29.03% over GNN, 25.03% over GNN-BCO, and 22.32% over GNN-COA is shown in figure 35(a). SDHO ensures that the model adapts well to diverse subsets of data during cross- validation. This adaptability allows the KNN to make accurate predictions across diverse folds. Similarly, the precision attained by SDHO-enabled KGNN reached 96.4% for k fold 10 showing improvement of 44.96% over the existing NN, 37.02% over RNN, 33.57% over BI-LSTM, 33.50% over DCNN, 31.94% over GNN, 29.67% over GNN-BCO, and 27.91% over GNN-COA is shown in figure 35(b). These improvements show that the model can effectively minimize false positives than other conventional methods. Additionally, for k-fold 10 the recall attained by SDHO-enabled KGNN reached 83.15%, showing improvements of 25.59%, 25.46%, 17.14%, 16.87%, 3.64%, 1.6%, and 0.86% over the existing methods NN, RNN, BI-LSTM, DCNN, GNN, GNN-BCO, and GNN-COA, respectively is shown in figure 35(c). The improvement in these results shows that the developed method outperforms the other conventional methods demonstrating its efficiency in detecting fraudulent calls.

6.7.3 ROC Analysis

The ROC analysis of the proposed SDHO-enabled KGNN model with existing methods is shown in Figure 36. When the False positive rate is 0.6%, the true positive rate of the existing NN, RNN, Bi-LSTM, DCNN, GNN, GNN-BCO, GNN-COA methods, and the proposed SDHO-enabled KGNN model are0.7046%, 0.8427%, 0.8759%, 0.9052%, 0.9501%, 0.9514%, and 0.9518% respectively. Thus, it is clear, that the proposed SDHO-enabled KGNN model achieved

higher accuracy in fraudulent call detection when compared to the existing NN, RNN, Bi-LSTM, DCNN, GNN, GNN-BCO, GNN-COA methods.

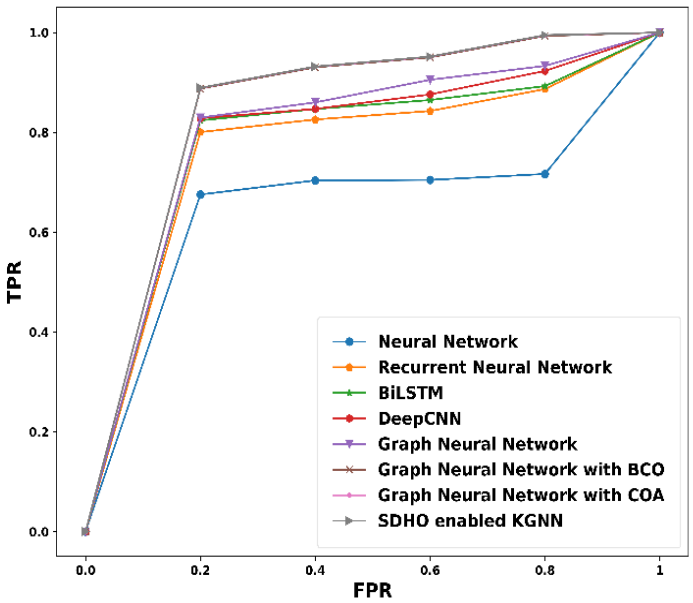


Figure 36: ROC Analysis

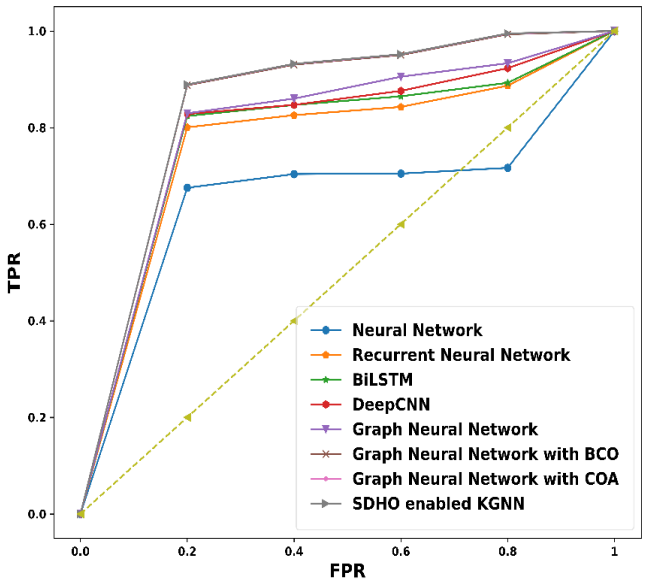


Figure 37: AUC Analysis

6.7.4 AUC Analysis

AUC is utilized to evaluate the performance of the classifier and it is typically calculated using the methods of numerical integration like a trapezoidal rule, which is applied to the results of the ROC analysis. The value of AUC always ranges from 0 to 1 and the variation is clearly shown in Figure 37.

6.7.6 Model Loss and Accuracy Graph Analysis

Figure 38 shows the accuracy analysis of the proposed SDHO-enabled KGNN model with the epoch ranges from 0 to 100. At epoch 0, the accuracy of the proposed method is 0.0964, then it rises to 0.324 for epoch 20, then at epoch 40, the accuracy gradually increases to 0.545, Further, at epoch 60, it reaches 0.691, 0.802 for epoch 80 and lastly the accuracy attains for the epoch 100 is 0.903.

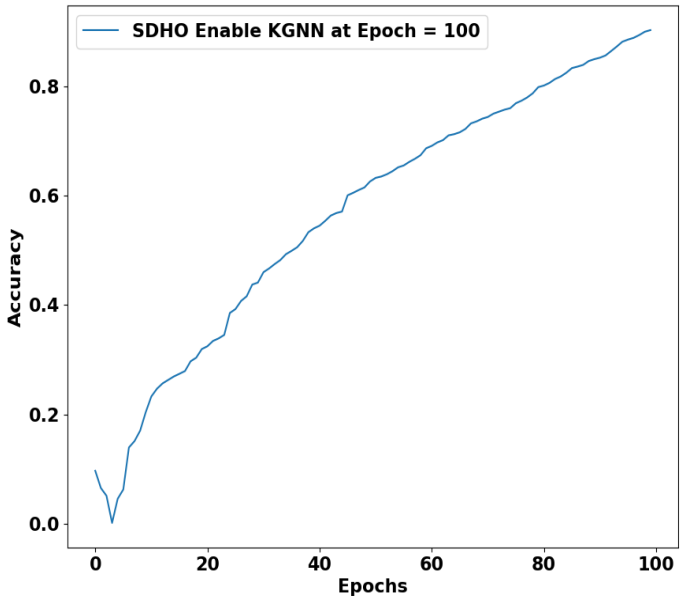


Figure 38: Accuracy Analysis of SDHO enabled KGNN

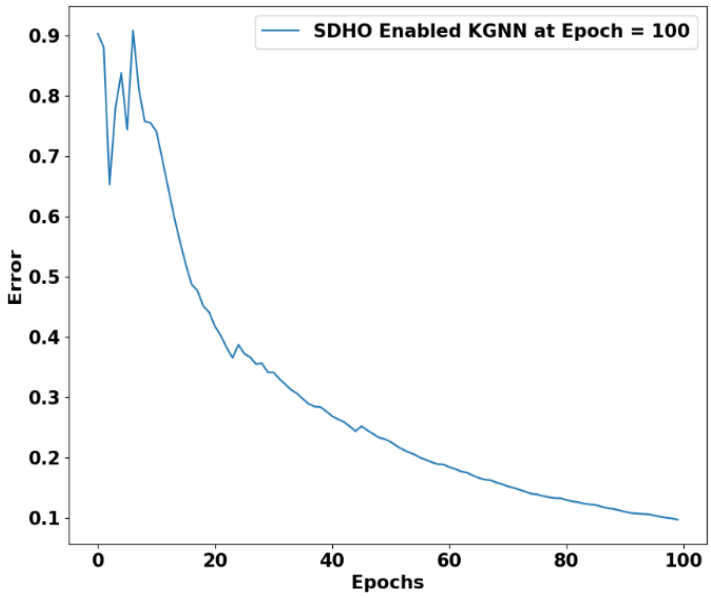


Figure 39: Loss Analysis of SDHO enabled KGNN

Figure 39 shows the loss analysis of the proposed SDHO-enabled KGNN model with epoch ranges from 0 to 100. The loss attained by the proposed SDHO-enabled KGNN model are0.903, 0.417, 0.268, 0.183, 0.129, and 0.096 for epochs 0, 20, 40, 60, 80 and 100 respectively.

6.8 Comparative discussion

In this section, the feasibility of fraudulent call detection through automatic feature learning techniques is analyzed to improve the security of telecommunication. Considering the security issues associated with the 5G and future 6G networks, this section will concentrate on the specific techniques related to fraudulent call detection in the context of 5G/6G networks. From the experimental validation, there are still some challenges that limit the applicability of the competent techniques, and the results are depicted in this section. Neural Network (NN), Recurrent Neural Network (RNN), Bi-LSTM, Deep CNN (DCNN), Graph Neural Network (GNN), Graph Neural Network with BCO (GNN-BCO), and Graph Neural Network with COA (GNN-COA) are the conventional approaches utilized for fraud call detection. These approaches contain various drawbacks; Training deep learning models, such as Deep CNN and GNN, can be computationally intensive and may require access to powerful hardware or cloud resources. RNN, Bi-LSTM, and GNN can be complex and less interpretable. The performance of these models can be notably affected by the quality and accuracy of the data employed for training. Data that is noisy or incomplete can result in less than the optimal outcomes. The SDHO-enabled KGNN model is developed to overcome these limitations, KGNNs offer significant advantages for fraud call detection by leveraging external knowledge to improve accuracy, reduce false positives, adapt to changing fraud patterns, and provide interpretability while handling large-scale and dynamic call networks. Remarkably, the proposed SDHO-enabled KGNN model outperforms the other existing techniques showing its effectiveness in fraud call detection. The incorporation of KGNN helps to learn intricate patterns and relationships, which enhances its ability to detect fraud calls and the use of SDHO increases the model's performance by tuning the classifier and reduces the computational complexity of the model. High accuracy across diverse training percentages and various k-fold ensures that the model performs well consistently demonstrating its ability to adapt and learn diverse amounts of labeled data. The result indicates that the developed method attains a high accuracy of 93.8% in detecting fraudulent calls by adapting to changing fraud patterns. Unlike other conventional methods, the proposed method balances between precision and recall ensuring that the model not only excels in high accuracy but also in identifying fraud calls and minimizes the false positives as well as false negatives. Consequently, the proposed approach learns implicit features that are more resilient to a

fraudulent phone number's constantly changing dynamic and call behavior over time and improves the security of 5G/6G networks. Table 7 depicts the comparative analysis between the SDHO-enabled KGNN and other conventional methods with TP 90 and k-fold 10.

Table 7: Comparative discussion for SDHO-enabled KGNN

Methods/ Metrics	**Training Percentage**			**K-fold**		
	Accuracy (%)	**Precision (%)**	**Recall (%)**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**
NN	53.07	50.24	63.82	55.42	53.05	61.87
RNN	67.9	69.42	65.38	59.12	60.7	61.98
BI-LSTM	68.01	70.5	66.51	63.61	64.03	68.89
Deep CNN	71.5	73.09	76.55	65.49	64.10	69.12
GNN	72.35	74.26	76.82	66.88	65.60	80.12
GNN-BCO	78.03	80.59	81.78	70	67.79	81.81
GNN-COA	82.05	91.24	91.55	73.22	69.48	82.43
SDHO-enabled KGNN	**93.8**	**95.91**	**95.53**	**94.25**	**96.4**	**83.15**

6.9 Summary

In this research, a novel fraudulent call detection approach for 5G/6G networks is proposed with advanced strategies to improve the security and privacy of the user. In the realm of fraudulent call detection, the integration of K-GNN as a classifier represents a cutting-edge and highly effective approach. K-GNNs combine the power of graph-based modeling with external knowledge sources, yielding a range of compelling advantages. By incorporating domain-specific knowledge such as historical fraud patterns and network relationships, K-GNNs significantly enhance accuracy and robustness that excel in distinguishing fraudulent calls from legitimate ones, thus minimizing false positives and reducing operational overhead. Moreover, K-GNNs adapt to evolving fraud tactics by continuously updating their knowledge base, ensuring a proactive stance against emerging threats. Furthermore, the integration of SDHO enabled K-GNNs to excel in few-shot learning and anomaly detection, enabling the identification of new, previously unseen fraudulent patterns. Customizable to specific organizational needs, these models foster industry collaboration, creating a collective defense against fraud. Specifically, the proposed approach facilitates effective fraudulent call detection and is crucial in the 5G/6G era, as it enhances the user privacy and security by terminating the requirement to transfer sensitive information and enhances the network performance in terms of security. The experimental results

demonstrate that the SDHO-enabled K-GNN approach outperforms traditional methods and commonly employed techniques in fraud call detection. Specifically, achieves an impressive overall accuracy, precision, and recall of 93.8%, 95.91%, and 95.53 with a TP rate of 90% respectively. Future work could involve refining knowledge integration techniques, enhancing real-time processing capabilities, exploring federated learning for collaborative fraud detection across organizations, and addressing privacy concerns. Additionally, developing methods to handle adversarial attacks on the knowledge-enhanced GNN model is a crucial area for improvement. The proposed model will be extended with a hybrid deep learning mechanism by focusing on a different number of attributes in the data to improve the performance that serves as the future direction of research.

CHAPTER 7
CONCLUSION AND FUTURE SCOPE AND SOCIAL IMPACT

With the ever-increasing volume of digital data in modern communication, ensuring the privacy and security of customers has become a daunting challenge for data owners. To address this concern, Link Analysis appears to be one of the most viable options as it is able to detect certain fraud calls and hence, it becomes easier for users to gain trust regarding the fraudulent calls. Various link analysis methods have been proposed, aiming to achieve better accuracy, reduced distortion, and enhanced security. These methods are able to note certain patterns through knowledge-based graphs and recognize certain patterns of fraud users.

It was observed that traditional link analysis methods achieved lower accuracy exhibiting lower detection standards and passing certain fraud calls as genuine callers. This thesis has presented several novel link analysis methods such SDHO based KGNNs, PSO-SSO algorithms and hence, they achieve much better results than the previously proposed algorithms.

7.1 Research Contribution 1

Link analysis is a technique of data mining that is especially used to detect useful and interesting patterns. The first challenge in link analysis is to reduce the graphs into manageable portions. In Market-Basket analysis, where intelligent searching of interesting items is done by removing unwanted elements, the link analysis may be helpful to reduce the graphs so that the analysis is manageable. Therefore, graph reduction techniques are needed to be applied to the graphs to obtain meaningful relationships. Identifying the interesting relationships and deciding to reduce the graphs is also an important challenge. Therefore, determining how to apply the link analysis techniques to detect abnormal and suspicious behavior is needed. The other challenge in counter-terrorism analysis is handling the situation with partial information. The research efforts on link analysis have to be conducted for efficient use in counter-terrorism. The availability of good data is another challenge for link analysis in data mining. Hence, this research focusses on various drawbacks and possible solutions for link analysis algorithms, and hence, detection of fraudulent calls become necessary.

7.2 Research Contribution 2

Social network analysis using the proposed Spizella swarm optimization-based Bi-LSTM classifier is performed for the detection of crime rates. Since crimes are rising at an alarming rate, it is difficult to foresee them with any degree of accuracy. Therefore, it is crucial to identify potential crimes now in order to prevent them in the future. Hence the crime rate is detected using the Spizella swarm optimization based Bi-LSTM classifier, where the convergence of the crime rate detection is greatly enhanced. Spizella swarm optimization effectively tuned the parameters and helps in achieving a better output. The proposed classifier could be applicable in determining the behavior of people and helps in reducing the occurrence of crime rates. Compared to the existing methods proposed method gains high accuracy and takes less time for detection. By analyzing the metrics values the Spizella swarm optimization obtained an improvement of 0.5%, 1.16%, and 1.08%, which is more efficient. In the future, the sentimental analysis, and the opinion analysis could also be included for efficient crime rate detection also it is difficult to predict the next crime that is going to take place using Twitter data because the large number of fake information is present in the data is the upcoming future work.

7.3 Research Contribution 3

We are able to show the relationship between crime and policy from the perspective of data science. The project culminated into proof which proves the question:" Does the policy of a state affects another state?". Our contribution lies in proving this linkage between policy and crime using some of the common data mining techniques. The project develops a strong understanding of the need for ETL, model selection, and rule-based inferences for solving any complex and diverse problem. We aimed at providing linkage between states in the USA because, we have observed over the time through different monthly and quarterly magazines, that various state and federal organizations are involved in maintaining peace within the state and neighboring states. Our rule-based deductions sufficiently prove that Twitter tweets play an eminent role in providing useful inferences about the working of the state and its environment. This project is a preliminary approach to policy research but provides an appropriate foundation for further study.

7.4 Research Contribution 4

In this research, a novel fraudulent call detection approach for 5G/6G networks is proposed with advanced strategies to improve the security and privacy of the user. In the realm of

fraudulent call detection, the integration of K-GNN as a classifier represents a cutting-edge and highly effective approach. K-GNNs combine the power of graph-based modeling with external knowledge sources, yielding a range of compelling advantages. By incorporating domain-specific knowledge such as historical fraud patterns and network relationships, K-GNNs significantly enhance accuracy and robustness that excel in distinguishing fraudulent calls from legitimate ones, thus minimizing false positives and reducing operational overhead. Moreover, K-GNNs adapt to evolving fraud tactics by continuously updating their knowledge base, ensuring a proactive stance against emerging threats. Furthermore, the integration of SDHO enabled K-GNNs to excel in few-shot learning and anomaly detection, enabling the identification of new, previously unseen fraudulent patterns. Customizable to specific organizational needs, these models foster industry collaboration, creating a collective defense against fraud. Specifically, the proposed approach facilitates effective fraudulent call detection and is crucial in the 5G/6G era, as it enhances the user privacy and security by terminating the requirement to transfer sensitive information and enhances the network performance in terms of security. The experimental results demonstrate that the SDHO-enabled K-GNN approach outperforms traditional methods and commonly employed techniques in fraud call detection. Specifically, achieves an impressive overall accuracy, precision, and recall of 93.8%, 95.91%, and 95.53 with a TP rate of 90% respectively.

7.5 Social Impact

Link analysis is a way of studying how people, groups, or even online accounts are connected to each other, and it can have a big impact when it comes to social engineering. On the positive side, this technique can be very helpful for society. Security teams and investigators can use link analysis to spot hidden connections between cybercriminals, fraudsters, or fake accounts. For example, by mapping relationships, they might uncover how misinformation spreads through social media or identify the network behind a scam. This makes it possible to stop attacks earlier, protect vulnerable users, and build more trust in digital spaces where people connect, work, and share information every day.

But on the flip side, the same method can also be abused. Social engineers—attackers who exploit human psychology—can use link analysis to learn about a person's social and professional network. By seeing who trusts whom, who interacts frequently, and what kinds of relationships exist, attackers can design highly convincing scams or phishing messages. For instance, they might pretend to be a colleague, a close friend, or even a family member to trick someone into sharing sensitive information. This kind of misuse can damage trust, violate

privacy, and even shape public opinion in harmful ways. In short, link analysis can either protect society by uncovering threats or harm it if used by attackers to manipulate human connections.

7.5 Future Work

Future work could involve refining knowledge integration techniques, enhancing real-time processing capabilities, exploring federated learning for collaborative fraud detection across organizations, and addressing privacy concerns. Additionally, developing methods to handle adversarial attacks on the knowledge-enhanced GNN model is a crucial area for improvement. The proposed model will be extended with a hybrid deep learning mechanism by focusing on a different number of attributes in the data to improve the performance that serves as the future direction of research.

We have used limited classification and clustering techniques in their basic form and didn't include hybrid algorithms. Furthermore, we have restricted ourselves to simple and structural learning without incorporating randomness while making the classifier learn. Lastly, the problem can be formulated in computational intelligence and it can be assumed that much better patterns in terms of rules can be created using Heuristic, Meta-heuristic and Fuzzy methodologies

LIST OF PUBLICATIONS

JOURNAL PUBLICATIONS

1. Pooja Mithoo and Manoj Kumar; Social network analysis for crime rate detection using Spizella swarm optimization based Bi-LSTM classifier, Knowledge Based Systems, Elsevier. Published. (IF-7.2).

2. Pooja Mithoo and Manoj Kumar; SDHO-KGNN: An Effective Knowledge Enhanced Optimal Graph Neural Network Approach for Fraudulent Call Detection. Transactions on Emerging Telecommunications Technologies, Wiley. (IF 2.5)

CONFERENCE PUBLICATIONS

1. Pooja Mithoo and Manoj Kumar; A Role Of Link Analysis in Social Networking: A Survey. 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE). IEEE. December 2022.

2. Pooja Mithoo and Manoj Kumar; Impact of Crime Data Mining on Interstate Policies. 2nd International Conference on Advances in IoT, Security with AI (ICAISA-2025), Springer. April 2023.

MAPPING

Table 8: Objectives Mapping

S. No	Research Objective	Published Papers
RO1	To investigate and analyze all the tools, techniques of link analysis.	▪ Pooja Mithoo and Manoj Kumar, "A Role Of Link Analysis in Social Networking: A Survey", in 2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE), Dec 2022.
RO2	To design an algorithm to generate the associate elements or item sets	▪ Pooja Mithoo and Manoj Kumar, "SDHO-KGNN: An Effective Knowledge Enhanced Optimal Graph Neural Network Approach for Fraudulent Call Detection", in Transactions on Emerging Telecommunications Technologies Willey, April 2025. SCI. DOI: 10.1002/ett.70101.
RO3	To implement the proposed algorithm with the appropriate data set.	▪ Pooja Mithoo and Manoj Kumar, "Social network analysis for crime rate detection using Spizella swarm optimization based Bi-LSTM classifier", in Knowledge Based Systems, Elsevier, June 2023. SCI DOI: https://doi.org/10.1016/j.knosys.2023.110450
RO4	To analyze the performance of the proposed approach using parameters like location, past criminal records, relation with other objects.	▪ (1) Pooja Mithoo and Manoj Kumar, "Social network analysis for crime rate detection using Spizella swarm optimization based Bi-LSTM classifier", in Knowledge Based Systems, Elsevier, June 2023. SCI DOI: https://doi.org/10.1016/j.knosys.2023.110450. ▪ (2) Pooja Mithoo and Manoj Kumar, "Impact of Crime Data Mining on Interstate Policies", in 2nd International Conference on Advances in IoT, Security with AI (ICAISA-2025), April 2025.
RO5	Comparative analysis of proposed approach with existing algorithms on the basics of identified parameters.	▪ Pooja Mithoo and Manoj Kumar, "Impact of Crime Data Mining on Interstate Policies", in 2nd International Conference on Advances in IoT, Security with AI (ICAISA-2025), April 2025.
RO6	To utilize this algorithm for possible inclusion in a real-world problem.	▪ Pooja Mithoo and Manoj Kumar, "SDHO-KGNN: An Effective Knowledge Enhanced Optimal Graph Neural Network Approach for Fraudulent Call Detection", in Transactions on Emerging Telecommunications Technologies, Willey, April 2025. SCI

BIBLIOGRAPHY

[1] M. S. Gerber, "Predicting crime using Twitter and kernel density estimation," *Decision Support Systems*, vol. 61, pp. 115–125, Jun. 2014.

[2] Z. Liang, C. Feng, and L. Chanh-Tien, "Spatiotemporal event forecasting in social media," in *Proc. SIAM Int. Conf. Data Mining (SDM)*, Vancouver, BC, Canada, Apr. 2015, pp. 1–9.

[3] O. O. OIsafiade, "A revised frequent pattern model for crime situation recognition based on floor-ceil quartile function," in *Proc. 3rd Int. Conf. Inf. Technol. Quant. Manage. (ITQM)*, Rio de Janeiro, Brazil, May 2015, pp. 564–570.

[4] H. Loni, O. Uzuner, and C. Kedzie, "Understanding citizens' direct policy suggestions to the federal government: A natural language processing and topic modelling approach," in *Proc. 48th Hawaii Int. Conf. Syst. Sci. (HICSS)*, Kauai, HI, USA, Jan. 2015, pp. 215–224.

[5] C. McCue, *Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis*, 2nd ed. Oxford, U.K.: Butterworth-Heinemann, 2015.

[6] T. Hastie, J. Friedman, and R. Tibshirani, "Additive models, trees and related methods," in *The Elements of Statistical Learning*. New York, NY, USA: Springer, 2016, ch. 9, pp. 295–315.

[7] M. Gupta, "Genetically improved logistic regression with radial basis function for robust software effort prediction," in *Advances in Intelligent Systems and Computing: Machine Intelligence and Signal Processing*, Singapore: Springer, 2015, pp. 315–327.

[8] L. Akoglu, H. Tong, and D. Koutra, "Graph-based anomaly detection and description: A survey," *Data Mining and Knowledge Discovery*, vol. 29, no. 3, pp. 626–688, May 2015.

[9] M. S. Gerber, "Predicting crime using Twitter and kernel density estimation," *Decision Support Systems*, vol. 61, pp. 115–125, Jun. 2014.

[10] D. Olson and G. Lucas, "Link analysis," in *Descriptive Data Mining*, Singapore: Springer, 2019, pp. 107–128, doi: 10.1007/978-981-13-7181-3_7.

[11] M. Winter and J. W. Duncan, "Collaborative learning in networks," *Proc. Natl. Acad. Sci. USA (PNAS)*, vol. 109, no. 3, pp. 764–769, Jan. 2012.

[12] B. Shavers and J. Bryson, *Hiding Behind the Keyboard: Uncovering Covert Communication Methods with Forensic Analysis Techniques*, New York, NY, USA: Elsevier, 2016, pp. 203–221.

[13] R. Nisbet and K. J. Elder, *Handbook of Statistical Analysis and Data Mining Applications*, 2nd ed., Amsterdam, Netherlands: Academic Press, 2018, pp. 783–792.

[14] Z. Wang and J. Liu, “Flamingo search algorithm: A new swarm intelligence optimization algorithm,” *IEEE Access*, vol. 9, pp. 88,564–88,582, Jun. 2021.

[15] G. Carace, J. Gabarro, C. Cioffi, and N. Rubido, “Finding the resistance distance and eigenvector centrality from the network’s eigenvalues,” *Physics A: Statistical Mechanics and its Applications*, vol. 569, Art. no. 125751, Apr. 2021.

[16] H. Liu, “Centrality analysis of online social network big data,” in *Proc. 2018 IEEE 3rd Int. Conf. Big Data Analysis (ICBDA)*, Shanghai, China, Mar. 2018, pp. 256–260.

[17] W. Shafait, S. Aftab, and S. A. Gillani, “Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques,” *IEEE Access*, vol. 9, pp. 70,080–70,094, Jun. 2021.

[18] M. Mittal, L. Garg, J. Singh, and J. D. Velásquez, “Monitoring the impact of economic crisis on crime in India using machine learning,” *Computational Economics*, vol. 53, no. 1, pp. 1–28, Jan. 2019.

[19] A. Saha, M. S. Ullah, M. Srizon, and A. Yousuf, “Twitter data classification by applying and comparing multiple machine learning techniques,” *Int. J. Innovative Research in Computer Science & Technology (IJIRCST)*, vol. 7, no. 2, pp. 2347–5552, Mar.–Apr. 2019.

[20] S. Raut and Vijayalakshmi, “Design and analysis of machine learning algorithms for the reduction of crime rates in India,” *Procedia Computer Science*, vol. 172, pp. 122–127, 2020.

[21] S. Sharma, D. Singh, B. Narang, S. Bansal, M. T. Quasim, and G. R. Sinha, “An empirical analysis of machine learning algorithms for crime prediction using stacked generalization: An ensemble approach,” *IEEE Access*, vol. 9, pp. 67,488–67,500, May 2021.

[22] S. Singh and R. Venkatesh, “Sentimental analysis over Twitter data using clustering based machine learning algorithm,” *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–12, Apr. 2021, doi: 10.1007/s12652-021-03140-2.

[23] A. Saha, M. Zaman, M. S. Ullah, M. Srizon, and A. Yousuf, “Twitter data classification by applying and comparing multiple machine learning techniques,” *Int. J. Innovative Research in Computer Science & Technology (IJIRCST)*, vol. 7, no. 2, pp. 2347–5552, Mar.–Apr. 2019.

[24] K. Ahuja, K. Gajjar, R. Gajjar, and M. Shah, “Application on virtual reality for enhanced education learning, military training and sports,” *Augmented Human Research*, vol. 5, no. 1, pp. 1–9, Jan. 2020.

[25] A. D. Ahmed, “Adaptive sliding mode observer for engine cylinder pressure imbalance under different parameter uncertainties,” *IEEE Access*, vol. 2, pp. 1085–1091, Sep. 2014.

[26] K. Joshi, A. Dodia, P. Patel, and M. Shah, “A comprehensive review on automation in agriculture using artificial intelligence,” *Artificial Intelligence in Agriculture*, vol. 2, pp. 1–12, Dec. 2019.

[27] T. Vu, R. Sharma, R. Kumar, H. Le, B. T. Pham, D. T. Bui, I. Pal, M. Saha, and T. Le, “Crime rate detection using social media of different crime locations and Twitter part-of-speech tagger with Brown clustering,” *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 4, pp. 4287–4299, Apr. 2020.

[28] R. Prakash, J. Byrd, C. Utley, and G. Muresan, “Building knowledge graphs of homicide investigation chronologies,” in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Sorrento, Italy, Nov. 2020, pp. 17–20.

[29] V. Boginski, P. Chen, and S. Sajadmanesh, “Graph embeddings in criminal investigation: Towards combining precision, generalization and transparency,” *World Wide Web*, vol. 25, pp. 2379–2402, Jan. 2022, doi: 10.1007/s11280-021-01001-2.

[30] Y. Chen, “CrimeGraphNet: Link prediction in criminal networks with graph convolutional networks,” *arXiv preprint*arXiv:2311.18543, Nov. 2023.

[31] C. Wang, Z. Li, X. Yan, J. Sun, M. Yang, and C. Shahabi, “HAGEN: Homophily-aware graph convolutional recurrent network for crime forecasting,” *arXiv preprint* arXiv:2109.12846, Sep. 2021.

[32] H. V. Ribeiro, D. D. Leite, A. A. Batista, Á. F. Martins, B. R. da Costa, S. Gualdi, E. K. Lenzi, Q. S. Han, and M. Perc, “Deep learning criminal networks,” *arXiv preprint* arXiv:2304.08457, Apr. 2023.

[33] Y. Chen, “CrimeGNN: Harnessing the power of graph neural networks for community detection in criminal networks,” *arXiv preprint* arXiv:2311.17479, Nov. 2023.

[34] Y. Chen, “CrimeGAT: Leveraging graph attention networks for enhanced predictive policing in criminal networks,” *arXiv preprint* arXiv:2311.18641, Nov. 2023.

[35] J. Brankovic, M. C. Bressan, A. Andreescu, L. von Niederhäusern, E. Arazo, H. Pirim, and K. Atasu, “Graph feature preprocessor: Real-time subgraph-based feature extraction for financial crime detection,” *arXiv preprint*arXiv:2402.08593, Feb. 2024.

[36] E. Kurshan and H. Shen, “Graph computing for financial crime and fraud detection: Trends, challenges and outlook,” *arXiv preprint* arXiv:2103.03227, Mar. 2021.

[37] S. F. Tekin and S. S. Kozat, “Crime prediction with graph neural networks and multivariate normal distributions,” *arXiv preprint* arXiv:2111.14733, Nov. 2021.

[38] B. P. Chamberlain, S. Sanei, E. Rossi, F. Frasca, T. Markovich, N. Hodas, M. M. Bronstein, and M. Herbster, “Graph neural networks for link prediction with subgraph sketching,” *arXiv preprint* arXiv:2209.15486, Sep. 2022.

[39] K. Mann, Y. Mirza, Y. Kaza, R. R. Sahay, and P. Kumaraguru, “Exploring graph neural networks for Indian legal judgment prediction,” *arXiv preprint* arXiv:2310.12800, Oct. 2023.

[40] M. Arif, N. F. Mahmood, and G. A. A. Firmansyah, “Prediction of crime rate in Banjarmasin City using RNN-GRU model,” *Int. J. Intelligent Systems and Applications in Engineering*, vol. 10, no. 3, pp. 1–9, Sep. 2022.

[41] S. P. C. W. Sandagiri, B. T. G. S. Kumara, and K. Banujan, “ANN based crime detection and prediction using Twitter posts and weather data,” in *Proc. 2020 Int. Conf. Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, Sakheer, Bahrain, Oct. 2020, pp. 1–5.

[42] M. Badr and M. Abdel-Aziz, “Crime prediction using a hybrid sentiment analysis approach based on the bidirectional encoder representations from transformers,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 25, no. 2, pp. 1131–1139, Feb. 2022.

[43] I. Idriss, M. Badr, M. Abdel-Aziz, O. Mohamed, and H. E. Farouk, “Toward a deep learning-based intrusion detection system for IoT against botnet attacks,” *IAES Int. J. Artif. Intell. (IJ-AI)*, vol. 10, no. 1, pp. 110–120, Mar. 2021.

[44] S. Hossain, A. Ahmed, I. Khan, M. M. Hossain, and I. S. Hossain, “Crime prediction using spatio-temporal data,” in *Proc. Int. Conf. Computing Science, Communication and Security (COMS2)*, Singapore, Feb. 2020, pp. 277–289.

[45] G. I. Saad, M. M. Salem, and A. E. Hassanien, “A novel melanoma prediction model for imbalanced data using optimized SqueezeNet by bald eagle search optimization,” *Computers in Biology and Medicine*, vol. 136, Art. no. 104712, May 2021.

[46] M. Badr and M. Abdel-Aziz, “Augmented binary multi-labeled CNN for practical facial attribute classification,” *Indones. J. Electr. Eng. Comput. Sci. (IJEECS)*, vol. 23, no. 2, pp. 973–979, Feb. 2021.

[47] C. Sharmila, S. Shrestha, and Y. Li, “Intelligent crime anomaly detection in smart cities using deep learning,” in *Proc. IEEE 4th Int. Conf. Collaboration Internet Comput. (CIC)*, Philadelphia, PA, USA, Oct. 2018, pp. 399–404.

[48] V. N. Umadevi and K. Priyadharshini, “Crime intention detection system using deep learning,” in *Proc. 2018 Int. Conf. Circuits and Systems in Digital Enterprise Technology (ICCSDET)*, Kottayam, India, Dec. 2018, pp. 1–6.

[49] R. C. Krishna and C. Qian, “DeepRan: Attention-based Bi-LSTM and CRF for ransomware early detection and classification,” *Information Systems Frontiers*, vol. 23, no. 2, pp. 299–315, Apr. 2021.

[50] G. Budka, Z. M. Wilimowska, and P. Chlebus, “Time series analysis for crime forecasting,” in *Proc. 2018 26th Int. Conf. Systems Engineering (ICSEng)*, Sydney, Australia, Dec. 2018, pp. 1–10.

[51] N. Shah, N. Bhagat, and M. Shah, “Crime forecasting: A machine learning and computer vision approach to crime prediction and prevention,” *Visual Computing for Industry, Biomedicine, and Art*, vol. 4, no. 1, pp. 1–14, Mar. 2021.

[52] E. A. Kirillova, R. A. Kurbanov, N. V. Svechnikova, T. E. Zul’fugarzade, and S. S. Zenin, “Problems of fighting crimes on the internet,” *Journal of Advanced Research in Law and Economics*, vol. 8, no. 3, pp. 849–856, 2017.

[53] H.-W. Kang and H.-B. Kang, “Prediction of crime occurrence from multimodal data using deep learning,” *PLoS ONE*, vol. 12, no. 4, Art. no. e0176244, Apr. 2017.

[54] Z. M. Wilimowska, S. Jajuga, E. Sroka, Z. Szyjewski, R. Polański, P. Malczewski, and G. Budka, “Data-driven models in machine learning for crime prediction,” in *Proc. 2018 26th Int. Conf. Systems Engineering (ICSEng)*, Sydney, Australia, Dec. 2018, pp. 1–8.

[55] Y. Zhang, X. Li, F. Bai, J. Chen, C. Zhang, and P. Li, “Particle swarm optimization with adaptive learning strategy,” *Knowledge-Based Systems*, vol. 196, Art. no. 105789, Feb. 2020.

[56] D. Shah, R. Desai, A. Shah, P. Shah, and M. Shah, “A comprehensive analysis regarding several breakthroughs based on computer intelligence targeting various syndromes,” *Augmented Human Research*, vol. 5, no. 1, pp. 1–12, Jan. 2020.

[57] J. Xue and B. Shen, “A novel swarm intelligence optimization approach: Sparrow search algorithm,” *Systems Science & Control Engineering*, vol. 8, no. 1, pp. 22–34, Jan. 2020.

[58] H. Patel, D. Patel, D. Makwana, and M. Shah, “Transforming petroleum downstream sector through big data: A holistic review,” *Journal of Petroleum Exploration and Production Technology*, vol. 10, no. 6, pp. 2601–2610, Dec. 2020.