International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

# Predictive Modelling for Network Threat Detection using Artificial Intelligence Techniques

**Shaik Mehaboob[1], R. Divya Sree[2], Dr. G. Soniya Priyatharsini[3], Dr. M. Sujitha[4], Dr. M. Nisha[5]**

[1, 2]Students, [3, 4, 5]Professor
Department of Computer Science and Engineering,
Dr. M.G.R Educational and Research Institute of Technology, Maduravoyal, Chennai-95, Tamil Nadu, India
**Corresponding Author: Shaik Mehaboob**

**Abstract**
**The integration of artificial intelligence (AI) techniques has transformed network security by enabling predictive modeling for proactive threat detection. This research proposes a novel approach to enhancing network security through advanced AI-driven predictive analytics. By analyzing vast volumes of network traffic data, AI algorithms can identify patterns indicative of cyber threats, including malware, intrusions, and anomalous activities. The predictive models developed in this study can anticipate potential network vulnerabilities and detect emerging threats before they escalate into security breaches. This proactive approach strengthens network defenses, reduces the risk of cyberattacks, and safeguards critical data. By combining AI and predictive modeling, this research aims to establish a more resilient and adaptive network security framework in an increasingly interconnected digital landscape.**

**Keywords: Predictive modelling, Network security, Artificial intelligence, Threat detection, Cybersecurity, Machine learning, Anomaly detection, Predictive analytics, Network traffic analysis, Cyber threats**

## I.INTRODUCTION

Data science is an interdisciplinary field that applies scientific methods, processes, algorithms, and systems to extract valuable insights from structured and unstructured data. It plays a crucial role across various domains by transforming raw data into actionable knowledge.

The term **"data science"** was first introduced by **Peter Naur** in 1974 as an alternative name for computer science. However, it wasn't until 1996 that the **International Federation of Classification Societies (IFCS)** formally recognized data science as a distinct discipline. Despite this, the definition remained fluid.

In 2008, **D.J. Patil** and **Jeff Hammerbacher**, the pioneering data and analytics leaders at LinkedIn and

Facebook, popularized the term "data science." Over the past decade, it has evolved into one of the most in-demand professions in the tech industry.

Data science combines **domain expertise, programming skills, and mathematical and statistical knowledge** to derive meaningful insights from data. It is a blend of mathematics, business acumen, machine learning techniques, and computational tools, all of which help uncover hidden patterns in raw data. These insights are essential for making informed business decisions.

### Data Scientist:

A **data scientist** identifies critical questions, locates relevant data sources, and applies analytical techniques to derive insights. They possess a combination of **business acumen, analytical skills, data mining expertise, and data visualization abilities**. Organizations rely on data scientists to handle vast amounts of unstructured data, ensuring it is cleaned, processed, and analyzed effectively to drive strategic decisions.

### Artificial Intelligence (AI):

Artificial Intelligence (AI) is the simulation of human intelligence in machines, enabling them to learn, reason, and solve problems. AI systems analyze data, recognize patterns, and make predictions, mimicking cognitive functions like perception and decision-making

### 1.1 Scope of the Project

The scope of this project encompasses the design, development, and implementation of an AI-based network threat detection system. This system will focus on identifying a variety of network attacks such as DDoS, phishing, malware, and insider threats. The project will include the following key

**Objective Of The Project :**1.2 Data Collection and Organization: We gathered an expansive set of network traffic data containing standard operations as well as hazardous exploits. The data underwent sorting and cleaning to guarantee it was structured suitably for advancing AI models. Feature Determination: Relevant features were recognized and extracted from the network traffic data that could adequately separate standard behavior from attacks.

Model Crafting: An assortment of machine learning and profound learning models, for example Random Forest, Support Vector Machine, and Convolutional Neural Networks were created and prepared to precisely order various sorts of digital threats with expanding precision. The convoluted models were intended to find concealed examples and designs in the information and enhance attack location and anticipation.

## 2. LITERATURE SURVEY

A literature review is a body of text that examines and summarizes key findings from existing research and methodological approaches on a particular topic. It relies on secondary sources and discusses published information within a specific subject area, sometimes focusing on a particular time period. The primary goal of a literature review is to provide an overview of current knowledge on the topic, serving as a foundation for future research. It often precedes a research proposal and may range from a simple

summary of sources to a more detailed analysis.
A literature review typically follows an organized structure, incorporating both summary and synthesis. A summary provides a concise recap of essential information from sources, while a synthesis reorganizes and integrates that information, offering new interpretations or tracing the intellectual development of the field, including major debates. Depending on its purpose, the literature review may also evaluate sources and guide readers toward the most relevant and significant works in the field.

**2.1 Title: Malware Classification and Composition Analysis - A Survey of Recent Development** Author: Adel Abusitta, Miles Q. Li and Benjamin C. M. Fung McGill University, Montreal Year malware patterns and evolution. Additionally, new composition analysis techniques provide deeper insights into malware functionality and behavior, aiding in identifying attackers' objectives.

This survey reviews and compares key findings in malware classification and composition analysis, discusses evasion techniques and feature extraction methods, and evaluates research based on algorithms and features used. It also highlights strengths, limitations, challenges, and future research directions in malware analysis.

**2.2 Title : Malware Classification with Improved Convolutional Neural Network Model** Author- Sumit S. Lad, Amol C. Adamuthe year 2020 Malware poses a serious threat in cyberspace, stealing personal data and harming systems. Security experts continuously develop new detection strategies, with machine learning playing a key role in malware classification. However, existing solutions often require high computing resources and struggle with large datasets.

This paper introduces an improved **Convolutional Neural Network (CNN) model** with preprocessing and augmentation techniques for classifying malware grayscale images. The study uses the **Malimg dataset**, consisting of **9,339 images** from **25 malware families**. The proposed **CNN and Hybrid CNN+SVM** model efficiently extracts features, reducing resource consumption and processing time.

The proposed CNN model achieves **98.03% accuracy**, outperforming existing models such as **VGG16 (96.96%)**, **ResNet50 (97.11%)**, **InceptionV3 (97.22%)**, and **Xception (97.56%)**. Additionally, integrating **SVM** instead of the **Softmax activation function** further improves classification performance, achieving an accuracy of **99.59%**. The **fine- tuned CNN model**, with a **256-neuron fully connected (FC) layer**, generates optimized feature vectors for SVM classification. The **Linear SVC kernel** converts the binary SVM classifier into a multi-class SVM using the **one- against-one** method, ensuring highly accurate malware detection with faster execution times than existing CNNmodel

This paper introduces an improved **Convolutional Neural Network (CNN) model** with preprocessing and augmentation techniques for classifying malware grayscale images. The study uses the **Malimg dataset**, consisting of **9,339 images** from **25 malware families**. The proposed **CNN and Hybrid CNN+SVM** model efficiently extracts features, reducing resource consumption and processing time.

**2.3 Title: Camouflaged Adversarial Malware Example Generation Based on Conv-GANs**

**Against Black-Box Detectors** Author: **Fangtian Zhong, Xiuzhen Cheng** Year **2023**Deep learning has revolutionized various fields by enabling computers to learn from experience and understand complex patterns. Leveraging its power, this paper introduces **MalFox**, a **Convolutional Generative Adversarial Network (Conv-GAN)-based framework** designed to generate adversarial malware examples that can bypass **black-box malware detectors**.

Inspired by the ongoing battle between malware authors and detection systems, MalFox adopts a **confrontational approach**, generating **perturbation paths** using three techniques: **Obfusmal, Stealmal, and Hollowmal**. To evaluate its effectiveness, a large dataset of malware and benignware programs was analyzed, measuring **accuracy, detection rate, and evasive rate**.

Results show that **MalFox achieves 99.01% accuracy**, outperforming **12 well-known machine learning models**. Additionally, it significantly reduces the detection rate by **45.1% on average** and enhances the evasive rate by **up to 56.0%**, demonstrating its effectiveness in bypassing malware detection systems.

2.4 **Title: Modeling and Analyzing Malware Propagation Over Wireless Networks Based on Hypergraphs Project**. Author Jiaxing Chen, Shiwen Sun, Chengyi x i a , Dinghua Shi, Guanrong Chen **Year** 2023 With the increasing use of wireless networks, malware threats to cyberspace are growing. This paper presents a **hypergraph-based model** to analyze **malware propagation** in large-scale wireless networks. The model effectively represents **limited-range, internet- independent transmission**, where malware infects a device, spreads to the **wireless router**, and then infects all connected devices.

**Using a** heterogeneous mean-field approach**, the study identifies the** malware outbreak threshold**. Simulations show that** malware spread depends on network size and malware characteristics**, and isolating internet connections** does not fully prevent outbreaks**. Real-world data confirms that** heterogeneous **device distributions** make wireless networks more vulnerable to malware. Additionally, malware spreads **faster in wireless networks** than on traditional internet connections, especially when the initial number of infected devices is high. **. Real-world data confirms that** heterogeneous **device distributions** make wireless networks more vulnerable to malware. Additionally, malware spreads **faster in wireless networks** than on traditional internet connections, especially when the initial number of infected devices is high.

## III. PROPOSED METHODOLOGY

3.1 **EXISITING SYSTEM:** More than 150 cellular networks worldwide have adopted LTE-M (LTE-Machine Type Communication) and NB-IoT (Narrow Band Internet of Things) technologies to support large-scale IoT applications like smart metering and environmental monitoring. These IoT services run on the same cellular network infrastructure as regular mobile devices, such as smartphones. However, improper integration of IoT services can introduce new security vulnerabilities.

In this study, we identify security risks in cellular IoT from both system and service integration perspectives. Our analysis reveals multiple weaknesses, including flaws in cellular standards, network operation errors, and IoT device implementation issues. These vulnerabilities could allow attackers to

remotely access IP addresses and phone numbers of IoT devices, disrupt power-saving features, and launch various attacks like data/text spamming, battery draining, and forced device shutdowns.

We tested certified cellular IoT devices across five major carriers in the U.S. and Taiwan to validate these vulnerabilities. Our results show that attackers can increase an IoT data bill by up to $226 with just 120 MB of spam traffic, inflate text messaging costs at a rate of $5 per second, and interfere with a device's power-saving mode, potentially causing service disruptions. In some cases, IoT devices may even become completely unusable.

To address these security threats, we propose, develop, and test effective solutions to improve the security of cellular Bitcoin IoT networks.

**3.2 PROPOSED SYSTEM:** The proposed system for predicting network threats uses artificial intelligence (AI) to strengthen cybersecurity. By combining advanced machine learning with in-depth network data analysis, it can detect and stop potential threats before they reach critical systems. The system continuously monitors network traffic and uses anomaly detection to spot unusual activities that may indicate cyber threats like malware, hacking attempts, or data breaches.

Additionally, deep learning models trained on past cyber incidents help the system learn and improve over time, making it more effective at identifying new and evolving threats. AI-powered predictive modeling also speeds up response times, enabling security teams to act quickly and minimize potential damage. By proactively protecting networks, this system plays a crucial role in defending against cyber threats, ensuring strong security and uninterrupted operations for organizations across different industries.

### 3.3 ADVANTAGES OF PROPOSED SYSTEM:

1. We use structured data for network traffic attack classification using advance machine learning method.
2. We build a framework based application for deployment purpose.
3. Accuracy was improved.
4. We classify more than 5 attack.
5. We compared more than a two algorithms to getting better accuracy level.

### 4.1 SYSREM ARCHITECTURE:

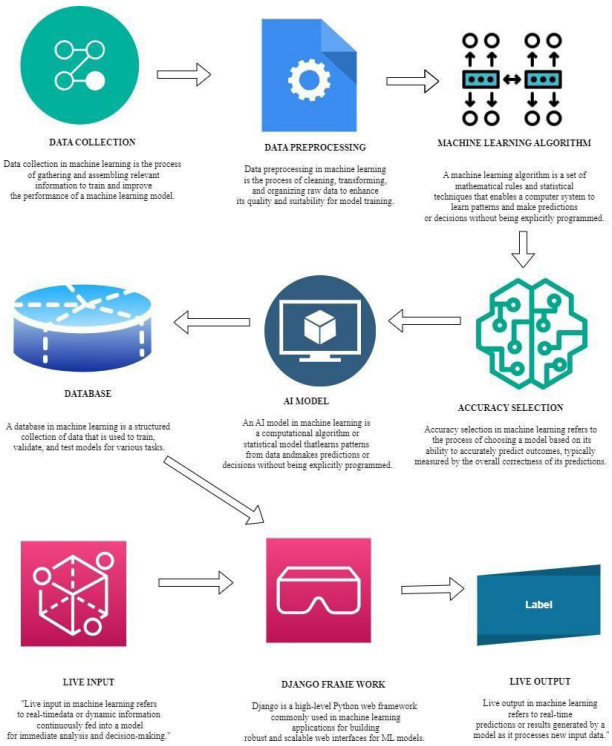


**Figure: 4.1**

**Data visualization:** Data visualization is a crucial aptitude in mathematics and machine learning. While statistics regularly emphasize quantitative descriptions and approximations, data visualization offers effective instruments for cultivating a qualitative apprehension of the knowledge. It plays a primary role during the exploratory phase, helping to uncover designs, identify anomalies, spot outliers, and determine issues such as corrupted facts.With some area expertise, visualizations can be strikingly compelling in highlighting significant relationships through intuitive plots and charts. . Other times, longer or more complex sentences are needed to fully explore ideas, alongside shorter ones to emphasize key points. Outliers in the data can also represent rare but important occurrences worthy of enhanced investigation and representation.

.

### 4.2 WORK FLOW DIAGRAM:

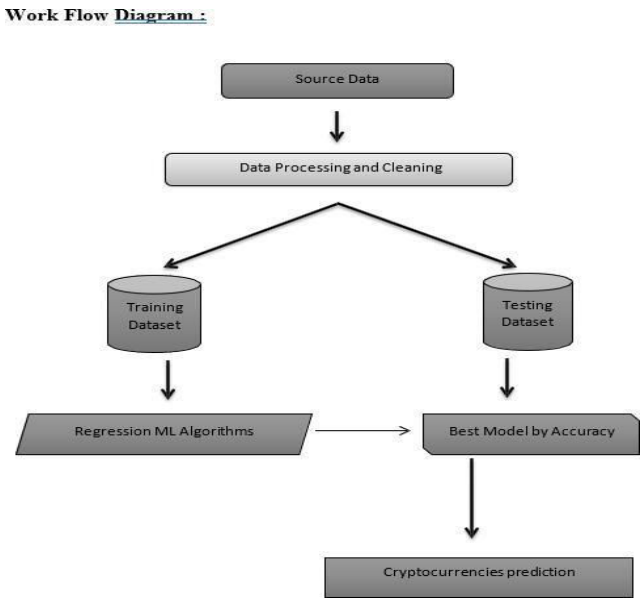


**Figure: 4.2**

**Data Collection: 4.3:** The dataset collected for predicting the given information was partitioned into a Training set comprising 70% of the data and a Test set with the remaining 30%. Models utilizing various algorithms, including Random Forest, logistic regression, decision trees, and support vector classification, were fitted on the Training set. Each model's accuracy was then evaluated on the Test set. The top-performing model, which achieved the highest test accuracy, was selected to generate predictions for new observations.

4.3 **Data Preprocessing:** The raw data collected contained missing values and potential outliers that could introduce inconsistencies and degrade predictive performance. To optimize algorithm efficiency and results quality, data preprocessing steps were undertaken. Variables were also examined and normalized when necessary to facilitate model training and enhance pattern detection across the full parameter space.

### 4.3 Data Validation/ Cleaning/Preparing Process

Importing the necessary packages and loading the given dataset allowed for analysis to begin. Examination of variable identification included assessing data shape and type as well as probing for missing or duplicate values. A validation dataset, data held back from model training, could provide an estimation of model skill during parameter tuning. Such datasets let users best leverage validation and test data when evaluating models. Before analysis, data cleaning commenced through renaming, dropping unneeded columns, and addressing issues found.

## 5. CONCLUSION:

```
# Check our datasets information (ex: data_type, NULL_count, Columns, and so on)
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2991 entries, 0 to 2990
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   High    2991 non-null   float64
 1   Low     2991 non-null   float64
 2   Open    2991 non-null   float64
 3   Close   2991 non-null   float64
 4   Volume  2991 non-null   float64
dtypes: float64(5)
memory usage: 117.0 KB
```

**Figure: 4.5**

## 6. MODULE DIAGRAM :

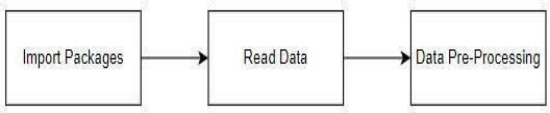


**Figure: 5**

## 7. ALGORITHM:

In machine learning and statistics, classification is a type of supervised learning where a computer model analyzes labeled input information and utilizes that understanding to categorize new observations. The dataset employed for coaching can contain binary categorization (e.g., figuring out whether an email is spam or not, or deciding if an individual is male or female) or multi-class categorization (including over two types).

Common examples of categorization responsibilities incorporate speech recognition, handwriting recognition, biometric identification, and file categorization. The algorithm learns by taking in labeled facts—details that involves each input features and the proper output labels. By identifying designs in the coaching facts, the algorithm understands to link specific designs with explicit labels, enabling it to foresee the proper label for new, unseen facts.

**Random Forest:**

Random Forest is an ensemble learning algorithm that can be used for both classification and regression tasks. It builds multiple decision trees and merges them together to get a more accurate and stable prediction.

**ADC:** AdaBoost, short for Adaptive Boosting, is an ensemble learning algorithm that is used to boost the performance of weak learners (individual models that perform slightly better than random chance) to create a strong classifier. It was introduced by Yoav Freund and Robert Schapire in 1996

In conclusion, our predictive modeling approach for network threat detection leverages advanced artificial intelligence techniques to enhance security measures. The integration of AI allows for real-time analysis and swift identification of potential threats, significantly reducing response time. This system not only improves accuracy in threat detection but also adapts continuously to evolving cyber threats. Ultimately, it offers a robust solution for maintaining network integrity and safeguarding sensitive information.

## 8. REFERENCES:

[1] (2020). Cellular IoT Market. [Online]. Available: https://www. marketdataforecast.com/market-reports/cellulariot-market

[2] Clp.29: LTE-M Deployment Guide to Basic Feature Set Requirements, GSMA, London, U.K., 2019.

[3] Clp.28: Nb-IoT Deployment Guide to Basic Feature Set Requirements, GSMA, London, U.K., 2019.

[4] Ericsson. (2016). Cellular IoT Alphabet Soup. [Online]. Available: https://www.ericsson.com/en/blog/2016/2/cellular-iot- alphabet-soup

[5] TS 24.301: Technical Specification Group Core Network and Terminals; Non-Access-Stratum (NAS) Protocol for Evolved Packet System ( EPS); Stage 2, 3GPP, Sophia Antipolis, France, 2020.

[6] C. Peng, C.-Y. Li, H. Wang, G.-H. Tu, and S. Lu, "Real threats to your data bills: Security loopholes and defenses in mobile data charging," in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., Nov. 2014, pp. 1–12.

[7] Y. Go, J. Won, D. F. Kune, E. Jeong, Y. Kim, and K. Park, "Gaining control of cellular traffic accounting by spurious TCP retransmission," in Proc. Netw. Distrib. Syst. Secure. Symp., 2014, pp. 1–15.

[8] C.-Y. Li et al., "Insecurity of voice solution VoLTE in LTE mobile networks," in Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur., Oct. 2015, pp. 1–12.

[9] H. Kim et al., "Breaking and fixing VoLTE: Exploiting hidden data channels and mis-implementations," in Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur., Oct. 2015, pp. 1–12.

[10] T. Xie, C.-Y. Li, J. Tang, and G.-H. Tu, "How voice service threatens cellular-connected IoT devices in the operational 4G LTE networks," in Proc. IEEE Int. Conf. Commun. (ICC), May 2018, pp. 1 – 6.

[11] Y. Li, K.-H. Kim, C. Vlachou, and J. Xie, "Bridging the data charging gap in the cellular edge," in Proc. ACM SIGCOMM, 2019, pp. 15–28.

[12] T. Xie, G.-H. Tu, C.-Y. Li, and C. Peng, "How can IoT services pose new security threats in operational cellular networks?" IEEE Trans. Mobile Comput., vol. 20, no. 8, pp. 2592–2606, Aug. 2021.

[13] Fcm.01: Volte Service Description and Implementation Guidelines V1.1, GSMA, London, U.K., 2014.

[14] (2021). Can I Get Unlimited Data? [Online]. Available: https://www.xfinity.com/support/articles/exp-unlimited- data

[15] (2018). Mobile IoT in the 5G Future—Nb-IoT and LTE-M in the Context of 5G. [Online]. Available: https://www.gsma.com/iot/wpcontent/uploads/2018/05/ GSMA-5GMobile-IoT.pdf

[16] Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control ( RRC ) Protocol Specification, document TS 36.331, 3GPP,

2020.

[17] Evolved Universal Terrestrial Radio Access (E- UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description; Stage 2, document TS 36.300, 2016.

[18] Ir.92: IMS Profile for Voice and SMS. Version 13.0, GSMA, London, U.K., 2019.

[19] D. Rupprecht, K. Kohls, T. Holz, and C. Pöpper, "Breaking LTE on layer two," in Proc. IEEE Symp. Secur. Privacy (SP), May 2019, pp. 1121–1136.

[20] D. Rupprecht, K. Kohls, T. Holz, and C. Poepper, "IMP4GT: IMPersonation attacks in 4G networks," in Proc. Netw. Distrib. Syst. Secur. Symp., 2020, pp. 893– 907.

[21] M. Stute, A. Heinrich, J. Lorenz, and M. Hollick, "Disrupting Continuity of Apple's wireless ecosystem security: New tracking, DoS, and MitM attacks on iOS and macOS through Bluetooth low energy, AWDL, and WIFI," in Proc. USENIX Security, 2021, pp. 1–19.

[22] (Oct. 26, 2016). Understanding Physical Internet Infrastructure Vulnerabilities. [Online]. Available: https://cip.gmu.edu/2016/10/26/ understanding- physicalinternet-infrastructure-vulnerabilities/

[23] H. Yang, S. Bae, and M. Son, "Hiding in plain signal: Physical signal overshadowing attack on LTE," in Proc. USENIX Security, 2023, pp. 1–12.

[24] (2023). WIO LTE Cat M1/Nb-IoTTracker.

[Online]. Available: https://wiki.seeedstudio.com/Wio_LTE_Cat_M1_NB-IoT_Tracker/

[25] (Nov. 22, 2016). Pycom Fipy Testbed. [Online]. Available: https://pycom. io/product/fipy/

[26] (Oct. 23, 2019). Mangoh Yellow Testbed. [Online].

Available: https://mangoh.io/mangoh-yellow

[27] (2023). Sixfab CIOT Hat. [Online]. Available: https://sixfab.com/ product/arduino-lte-m-nb-iot-egprs- cellularshield/

[28] (2021). Arduino MKR Nb 1500 Testbed. [Online]. Available: https://store.arduino.cc/usa/arduino-mkrnb-1500

[29] (2023). Telit Charlie Evaluation Kit. [Online]. Available: https://www.telit.com/support-tools/developmentevaluation-kits/charlieevaluation-kit-for- cellular-lpwa/

[30] (2023). Waveshare CIOT Kit. [Online]. Available: https://www. waveshare.com/wiki/SIM7080G_CatM/NB- IoT_HAT [31] (2020). 2020 IoT Developer Survey Key Findings. [Online]. Available: https://iot.eclipse.org/community/resources/iotsurveys/assets/iotdeveloper-survey-2020.pdf

[32] J. Postel, Transmission Control Protocol, document RFC 793, Sep. 1981.

[33] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Large-scale measurement and characterization of cellular machine-to-machine traffic," IEEE/ACM Trans. Netw., vol. 21, no. 6, pp. 1960–1973, Dec. 2013.

[34] I Cisco Systems. (2023). Dynamic ARP Inspection. [Online]. Available: https://www.cisco.com/c/en/us/td/docs/switches/lan/catalys t4500/12 - 2 /25ew/configuration/guide/conf/dynarp.html

[35] J. Arkko, J. Kempf, and B. Zill. (2005). Secure Neighbor Discovery (Send). [Online]. Available: https://tools.ietf.org/html/rfc3971

[36] TS 23.401: General Packet Radio Service (GPRS) Enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Access, 3GPP, Sophia Antipolis, France, 2020.

[37] TS 32.240: Technical Specification Group Services and System Aspects; Telecommunication Management; Charging Management; Charging Architecture and Principles, 3GPP, Sophia Antipolis, France, 2020.

[38] Technical Specification Group Services and System Aspects; Policy and Charging Control Architecture, document TS 23.203, 2020.

[39] Technical Specification Group Services and System Aspects; Telecommunication Management; Charging Management; IP Multimedia Subsystem (IMS) Charging, document TS 32.260, 2020.

[40] Charging Management; Charging Data Record (CDR) Parameter Description, document TS 32.298, 2020.