

**CAPSTONE PROJECT- MARCH-2024**

**COURSE CODE:** CSA4717

**COURSE NAME:** DEEP LEARNING FOR FUZZY SYSTEMS

# 

# PROJECT TITLE

ANOMALY DETECTION

Guided by:

**Dr . N. POONGAVANAM,**

Associate Professor,

Department of Bigdata and Network Security.

Submitted by:

192124137- S. Harshitha Sai Vasuki

192224224- P. Ammal Harshini

**ABSTRACT:**

Anomaly detection, a crucial aspect across diverse domains like cybersecurity, manufacturing, and finance, is increasingly relying on deep learning models for robust and scalable solutions. This abstract presents an overview of deep learning-based anomaly detection, focusing on its applications and challenges. Deep learning techniques, particularly neural networks, have shown remarkable performance in detecting anomalies or outliers in various types of data, including network traffic, manufacturing processes, and financial transactions. These applications encompass identifying fraudulent activities in financial transactions, detecting equipment failures in manufacturing processes, and monitoring for cybersecurity threats in network traffic. The versatility of deep learning models allows for the development of highly accurate anomaly detection systems capable of handling complex and dynamic datasets. By leveraging techniques such as autoencoders, recurrent neural networks (RNNs), and convolutional neural networks (CNNs), these models can effectively learn the underlying patterns of normal behavior and detect deviations indicative of anomalies. However, despite the promising results, deep learning-based anomaly detection encounters several challenges. One significant challenge is the requirement for large amounts of label data, which can be scarce and expensive to obtain, particularly for rare anomalies. Moreover, the interpretability of deep learning models poses challenges in understanding and explaining the detected anomalies, especially in critical applications such as healthcare and finance. Addressing these challenges necessitates continued research efforts toward developing techniques for efficient training with limited label data, enhancing model interpretability, and adapting to evolving data distributions in real-time scenarios. In conclusion, deep learning-based anomaly detection holds immense potential for addressing critical issues in various domains. By overcoming existing challenges and advancing the state-of-the-art, these techniques can contribute significantly to enhancing security, reliability, and efficiency across diverse applications.

**INTRODUCTION:**

Anomalies, often referred to as outliers, abnormalities, rare events, or deviants, are data points or patterns in data that do not conform to a notion of normal behavior. Anomaly detection, then, is the task of finding those patterns in data that do not adhere to expected norms, given previous observations. The capability to recognize or detect anomalous behavior can provide highly useful insights across industries. Hence, anomaly detection has found diverse applications in a variety of domains, including IT analytics, network intrusion analytics, medical diagnostics, financial fraud protection, manufacturing quality control, marketing, and social media analytics. Online large-scale Internet traffic analysis and threat alerts are also urgent to enterprises. With the development of computer techniques, many technologies like big data processing technologies and machine learning methods have been widely used in large-scale networks for cyber situational awareness. The detection process is mainly to continuously monitor various data packets and perform feature analysis on the detected suspicious data packets. Most anomaly detection techniques first establish a data model for normal Internet traffic and the anomalies are those that do not fit perfectly with the model. The statistical method takes adjacent identical parts of data in the time series for comparison. The clustering method assumes that the data far away from the normal sample is abnormal (based on the distance), outlier detection of multivariate Gaussian distribution is a classic example of abnormal point detection. For Internet traffic anomaly detection, detection time delay and accuracy are both significant. Deep learning has achieved excellent results in computer vision, speech recognition, and natural language processing. It has also brought new opportunities for the development of network traffic classification and anomaly detection. In the era of big data, Deep learning has powerful feature extraction and abstraction capabilities, which can integrate multi-source information and process Heterogeneous data,

* We use a big data processing framework for network traffic collection and preprocessing. We use Kafka and Spark Streaming for message delivery and data preprocessing.
* We use LSTM to model time-based network data streams for anomaly detection. We propose a novel method for anomaly detection.

With the advent of Industry, the anomaly detection performance of products in industrial production is the key to ensuring product quality and achieving intelligent manufacturing. Nowadays, detection technology based on machine vision has become more and more mature. With the strengths of high accuracy, low cost, and non-destructive, it has been widely studied and applied in the field of industrial detection instead of manual detection.The traditional machine learning detection method needs feature extraction and classifier. Firstly, the industrial production product images are preprocessed, and then the corresponding features are extracted thereafter to perform classification detection with classifiers. Moreover, in the high-speed and complex industrial production scenario, the traditional machine vision detection technology has the disadvantages of complex model parameter settings, poor robustness, and poor adaptability, which leads to the failure to meet the requirement of real-time speed and high precision in industrial production. Moreover, the performance of detection models based on deep learning is superior to that based on traditional machine learning. It brings new ideas to product anomaly detection under high-speed and complex industrial production scenarios. However, the balanced data set has to collect several anomaly instances. This is especially difficult in scenarios where anomalies occur very rarely like industrial production. Moreover, manual labeling of anomalies is extremely time-consuming and laborious. So, there are a large number of unbalanced product image data sets in industrial production scenarios. For the balanced image data set of industrial production products which provides both sufficient normal and anomaly samples during model training, we propose a supervised anomaly detection model based on YOLOv3. This model uses a new basic network Darknet-53 which simplifies the complexity of network training and improves the detection performance compared with other networks. It better guarantees the model with high precision and real-time performance. For an unbalanced data set, we consider the unsupervised learning model GAN[6] (Generative Adversarial Networks) for image generation This model only requires normal samples for training and detects anomalies by monitoring the anomaly score which is obtained by calculating the difference between the generated image and the test image. The performance is verified in a real industrial production scenario regarding both speed and accuracy.

**STATEMENT OF THE PROBLEM:**

Anomaly detection plays a pivotal role in various domains such as cybersecurity, finance, healthcare, and industrial operations by identifying deviations from expected behavior within complex datasets. Despite significant advancements in machine learning and data analytics, anomaly detection systems encounter several challenges that impede their efficacy and reliability in real-world applications.

**Need for the Study:**

Anomaly detection systems serve as critical components across numerous industries and applications, offering the capability to identify irregularities and deviations from expected patterns within complex datasets. The significance of advancing anomaly detection methodologies is underscored by several pressing factors:

1. Rising Data Complexity: In today's digital age, organizations are inundated with vast volumes of heterogeneous data generated from diverse sources such as sensors, logs, transactions, and network traffic. The increasing complexity and diversity of these datasets pose significant challenges for traditional anomaly detection methods, necessitating the development of more sophisticated and adaptable detection systems.

2. Escalating Threat Landscape: With the proliferation of cyber threats, financial fraud, and operational anomalies, the need for robust anomaly detection systems has become paramount. Cyberattacks, in particular, continue to evolve in sophistication and frequency, highlighting the urgency of enhancing detection capabilities to safeguard critical infrastructure, sensitive information, and organizational assets.

3. Economic Implications: Anomalies, whether they stem from fraudulent activities, system malfunctions, or operational inefficiencies, can have profound economic repercussions for businesses and society at large. Detecting and mitigating anomalies in a timely manner can mitigate financial losses, enhance operational efficiency, and preserve trust and reputation in organizations and institutions.

4. Regulatory Compliance: Various regulatory frameworks mandate stringent requirements for anomaly detection and risk management in sectors such as finance, healthcare, and cybersecurity. Compliance with regulations such as GDPR, HIPAA, PCI-DSS, and SOX necessitates the implementation of robust anomaly detection systems to ensure data integrity, privacy protection, and regulatory compliance.

5. Research Gaps and Opportunities: Despite significant progress in anomaly detection research, several challenges persist, including scalability, interpretability, adaptability to dynamic environments, and domain-specific considerations. Addressing these gaps presents fertile ground for innovation and contributes to the advancement of anomaly detection methodologies, thereby enabling more effective and reliable detection systems.

6. Societal Impact: Beyond organizational contexts, anomaly detection systems have broader societal implications, ranging from early detection of disease outbreaks and environmental anomalies to fraud prevention and public safety. By advancing anomaly detection research, we can harness the potential of data-driven technologies to address pressing societal challenges and improve the well-being of individuals and communities.

7. Continued Evolution of Technology: The rapid evolution of technology, including advancements in machine learning, deep learning, and edge computing, offers unprecedented opportunities to enhance anomaly detection capabilities. Leveraging emerging technologies and methodologies can drive innovation in anomaly detection and unlock new possibilities for detecting anomalies in diverse data streams and environments.

In light of these compelling factors, there exists a clear imperative to conduct research aimed at advancing anomaly detection methodologies, addressing existing challenges, and harnessing the full potential of data-driven approaches to anomaly detection. By undertaking this study, we aim to contribute to the broader research agenda in anomaly detection, address pressing societal and economic needs, and pave the way for more effective and reliable anomaly detection systems in the future.

**LITERATURE REVIEW:**

This literature review technique has already been used and disseminated in several papers published in international journals ([Henriques, Sobreiro, & Kimura, 2018](https://www.sciencedirect.com/science/article/pii/S2666954422000187" \l "bb0350), p. 145; [Lee et al., 2021](https://www.sciencedirect.com/science/article/pii/S2666954422000187" \l "bb0485), p. 68415; [Mariano, Sobreiro, & Do Nascimento Rebelatto, 2015](https://www.sciencedirect.com/science/article/pii/S2666954422000187" \l "bb0550), pp. 34–35; [Masudin & Fernanda, 2019](https://www.sciencedirect.com/science/article/pii/S2666954422000187" \l "bb0560), p. 2; [Nazário, Silva, Sobreiro, & Kimura, 2017](https://www.sciencedirect.com/science/article/pii/S2666954422000187" \l "bb0580), pp. 116–117; [Salim, Rahman, & Wahab, 2019](https://www.sciencedirect.com/science/article/pii/S2666954422000187" \l "bb0685), p. 1449). As scientific studies seek to describe, explain, predict, or evaluate phenomena ([Abreu, Kimura, & Sobreiro, 2019](https://www.sciencedirect.com/science/article/pii/S2666954422000187" \l "bb0005), p. 197), this review method makes tracking the evolution and main contributions within the field of study possible ([Cruz, Kimura, & Sobreiro, 2019](https://www.sciencedirect.com/science/article/pii/S2666954422000187" \l "bb0240), p.80) by mapping and identifying the main characteristics and topics of articles ([Pinheiro et al., 2019](https://www.sciencedirect.com/science/article/pii/S2666954422000187" \l "bb0630), p. 844) and by grouping them for deeper exploration of the analysis theme ([Ferreira, Sobreiro, Kimura, & Barboza, 2016](https://www.sciencedirect.com/science/article/pii/S2666954422000187" \l "bb0280), p. 114), creating a complete view of the existing knowledge ([Henriques, Sobreiro, Kimura, & Mariano, 2020](https://www.sciencedirect.com/science/article/pii/S2666954422000187" \l "bb0355), p. 6). This is done through a search for and selection of articles using a combination of keywords that allows for increasing the scope and reach of the research ([Jabbour et al., 2017](https://www.sciencedirect.com/science/article/pii/S2666954422000187" \l "bb0410), p. 291) and by elaborating a classification framework that proposes guidelines for categorization of the selected articles ([Cabral & Dhar, 2019a](https://www.sciencedirect.com/science/article/pii/S2666954422000187" \l "bb0165), p. 26).

**PROPOSED SYSTEM FOR ANOMALY DETECTION:**

**1. Introduction:**

Provide a brief overview of the proposed system for anomaly detection.

Highlight the significance of developing advanced anomaly detection methodologies to address emerging challenges and meet the evolving needs of various industries.

**2. System Architecture:**

Present a detailed architecture diagram illustrating the components and data flow of the proposed anomaly detection system.

Describe the role of each component, including data preprocessing, feature extraction, anomaly detection algorithms, and result visualization modules.

**3. Data Preprocessing:**

Discuss the preprocessing techniques employed to clean, normalize, and transform raw data into a suitable format for anomaly detection.

Address methods for handling missing values, outliers, and noise in the data.

**4. Feature Extraction:**

Explain the feature extraction methods used to capture relevant patterns and characteristics from the preprocessed data.

Consider techniques such as dimensionality reduction, time-series decomposition, or domain-specific feature engineering.

**5. Anomaly Detection Algorithms:**

Present the anomaly detection algorithms or techniques proposed for implementation in the system.

Justify the selection of specific algorithms based on their suitability for addressing the identified challenges and requirements.

Include traditional statistical methods, machine learning algorithms, and potentially novel approaches if applicable.

**6. Model Training and Evaluation:**

Describe the process of training the anomaly detection models using labeled or unlabeled data.

Discuss the evaluation metrics and methodologies used to assess the performance of the proposed system, including accuracy, precision, recall, and computational efficiency.

Address strategies for cross-validation, hyperparameter tuning, and model selection.

**7. Adaptability and Resilience:**

Explain how the proposed system adapts to concept drift and changes in the underlying data distribution over time.

Discuss mechanisms for model retraining, online learning, or ensemble techniques to maintain detection efficacy in dynamic environments.

**8. Interpretability and Explainability:**

Emphasize the importance of interpretability in anomaly detection and describe methods for enhancing the interpretability of the proposed system.

Consider techniques such as model explanation, feature importance analysis, or rule extraction to provide insights into detected anomalies.

**9. Scalability and Efficiency:**

Address strategies for ensuring the scalability and efficiency of the proposed system, particularly in handling large-scale or streaming data.

Discuss optimization techniques, parallel processing, or distributed computing frameworks to enhance system performance.

**10. Integration and Deployment:**

- Outline considerations for integrating the proposed anomaly detection system into existing infrastructure or applications.

- Discuss deployment options, including on-premises deployment, cloud-based solutions, or edge computing architectures.

**11. Case Studies and Experiments:**

- Present case studies or experimental results demonstrating the efficacy and performance of the proposed anomaly detection system.

- Provide comparative analyses with existing methods or benchmarks to showcase the advantages and limitations of the proposed approach.

**12. Discussion and Future Directions:**

- Reflect on the findings from the proposed system's implementation and evaluation.

- Discuss potential areas for further research, enhancements, or extensions to the proposed system, considering emerging trends and technological advancements.

**PROGRAM:**

import numpy as np

import tensorflow as tf

from tensorflow.keras.layers import Input, Dense

from tensorflow.keras.models import Model

# Generate some dummy data (replace this with your actual dataset)

# Assuming data is normalized

# X\_train: Training data without anomalies

# X\_test: Test data with anomalies

X\_train = np.random.normal(loc=0, scale=1, size=(1000, 10))

X\_test\_normal = np.random.normal(loc=0, scale=1, size=(200, 10))

X\_test\_anomaly = np.random.normal(loc=5, scale=2, size=(50, 10))

# Autoencoder architecture

input\_layer = Input(shape=(10,))

encoded = Dense(5, activation='relu')(input\_layer)

decoded = Dense(10, activation='sigmoid')(encoded)

autoencoder = Model(input\_layer, decoded)

encoder = Model(input\_layer, encoded)

autoencoder.compile(optimizer='adam', loss='mse')

# Train the autoencoder

autoencoder.fit(X\_train, X\_train, epochs=50, batch\_size=32, shuffle=True, validation\_data=(X\_test\_normal, X\_test\_normal))

# Detect anomalies

predictions\_normal = autoencoder.predict(X\_test\_normal)

mse\_normal = np.mean(np.power(X\_test\_normal - predictions\_normal, 2), axis=1)

threshold = np.mean(mse\_normal) + 2 \* np.std(mse\_normal)

predictions\_anomaly = autoencoder.predict(X\_test\_anomaly)

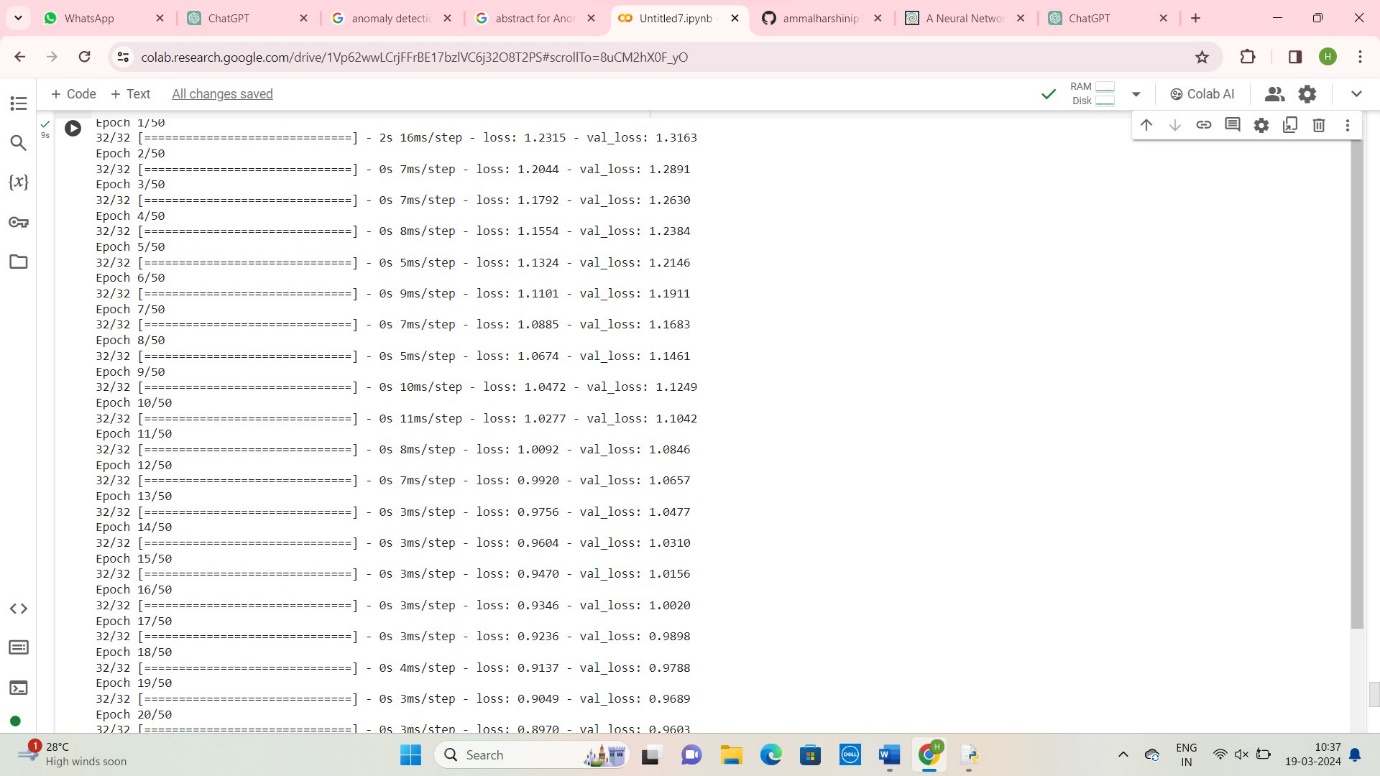
mse\_anomaly = np.mean(np.power(X\_test\_anomaly - predictions\_anomaly, 2), axis=1)

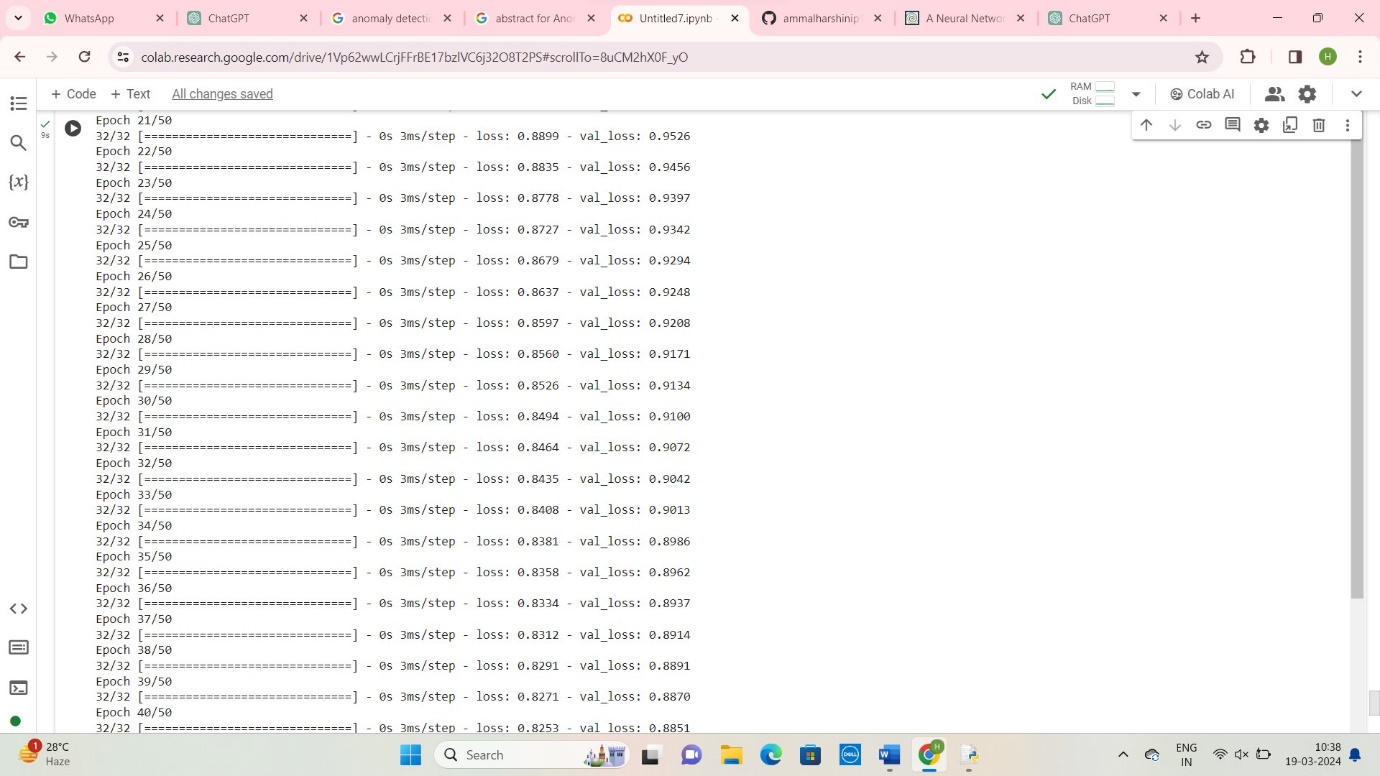
# Identify anomalies based on threshold

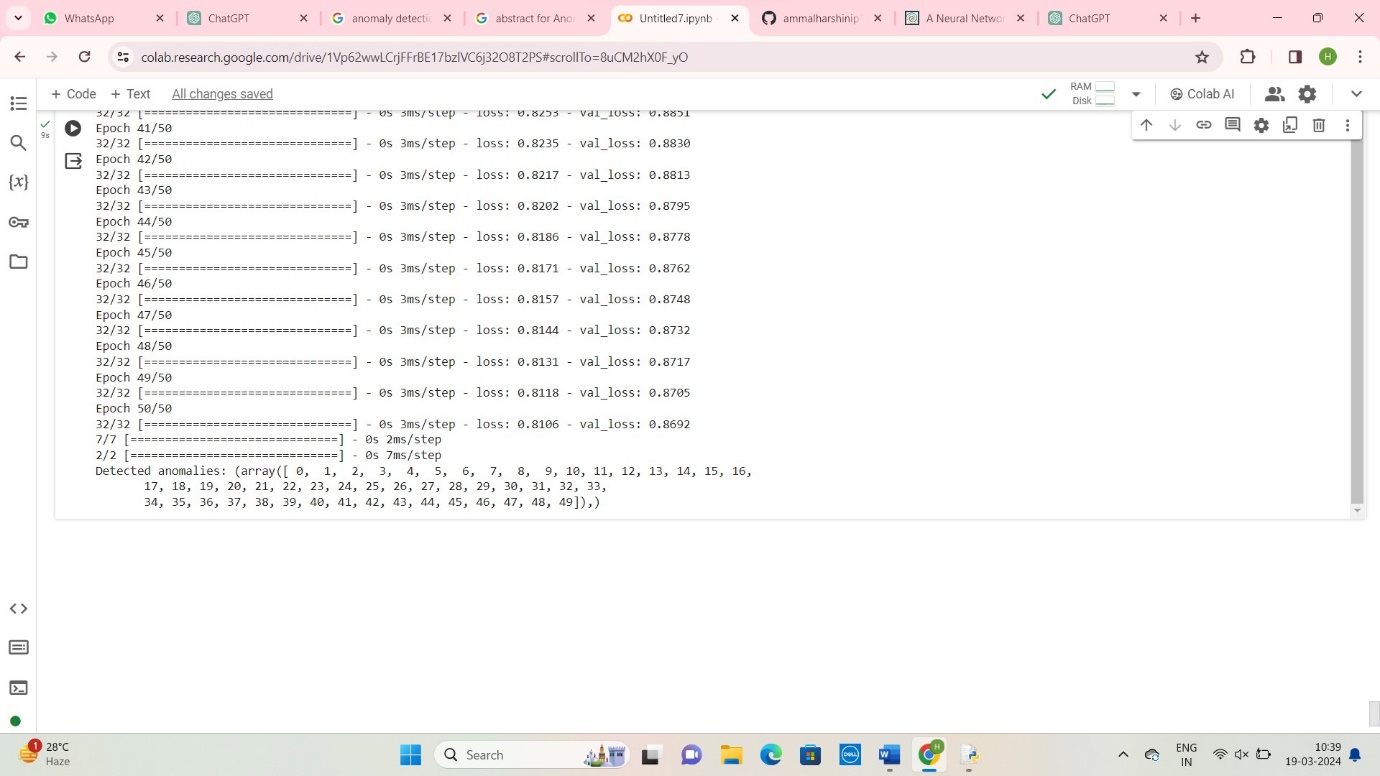
anomalies = mse\_anomaly > threshold

print("Detected anomalies:", np.where(anomalies))

**OUTPUT:**







**CONCLUSION:**

This survey paper attempts to provides a systematic review on machine and deep learning techniques for anomaly detection. We build our systematic review around two objectives which we then attempt to achieve in the remaining sections. Further, we devote a complete section on accuracy metrics used in the literature In addition, we provide a list of application areas where anomaly detection is employed. We put a special focus on latest techniques of anomaly detection that are driven by the most advanced applications of machine and deep learning. Due to the fast-evolving nature of the filed, we include only the most recent papers

We also mention several datasets that were utilized in experiments of relevant research publications, with the majority of the experiments using real-world datasets as training or testing datasets for their models. Our review reveals that many avenues are still in the infancy phase and require significant research. Moreover, many datasets are becoming obsolete and are being replaced with newer and more relevant real-world datasets and hence are more valuable. We believe that this review could be a valuable starting point for researchers and AI community to get up-to-date and relevant information on anomaly detection using machine learning techniques.

**REFERENCE:**

[1]Frank E Grubbs. 1969. Procedures for detecting outlying observations in samples. Technometrics 11, 1 (1969), 1–21.

[2]S. Agrawal and J. Agrawal, ``Survey on anomaly detection using data mining techniques,'' Procedia Comput. Sci., vol. 60, no. 1, pp.

708713, 2015, doi: 10.1016/j.procs.2015.08.220.

[3]P. Gogoi, D. K. Bhattacharyya, B. Borah, and J. K. Kalita, ``A survey of outlier detection methods in network anomaly identification,''

Computer. J., vol. 54, no. 4, pp. 570588, Apr. 2011, doi: 10.1093/comjnl/bxr026.

[4]V. Chandola, A. Banerjee, and V. Kumar, ``Anomaly detection: A survey,'' ACM Computer. Surev., vol. 41, no. 3, pp. 7197, 2009,

doi:10.1145/1541880.1541882.

[5] V. Chandola, A. Banerjee, and V. Kumar, ``Anomaly detection for discrete sequences: A survey,'' IEEE Trans. Knowl. Data Eng., vol. 24,

no. 5, pp. 823839, Nov. 2012.

[6]R. A. A. Habeeb, F. Nasaruddin, A. Gani, I. A. T. Hashem, E. Ahmed, and M. Imran, ``Real-time big data processing for anomaly detection: A survey,'' Int. J. Inf. Manage., vol. 45, pp. 289307, Apr. 2019, doi: 10.1016/j.ijinfomgt.2018.08.006.

[7] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, ``Network anomaly detection: Methods, systems and tools,'' IEEE Community. Surveys Tuts., vol. 16, no. 1, pp. 303336, 1st Quart., 2013

[8] U. Fiore, F. Palmieri, A.Castiglione, and A. De Santis, ``Network anomaly detection with the restricted Boltzmann machine,'' Neuro

computing, vol. 122, pp. 1323, Dec. 2013, doi: 10.1016/j.neucom.2012.11.05.

[9] Pecht, Michael G., and Myeongsu Kang. "Machine learning: Anomaly detection." (2019): 131-162

[10] Cortes, C. and Vapnik, V.N. (1995). Support-vector networks. Machine Learning 20 (3): 273–297.

[11] Sutrisno, E., Fan, Q., Das, D., and Pecht, M. (2012). Anomaly detection for insulated gate bipolar transistor under power cycling using principal component analysis and k-nearest neighbour algorithm. Journal of the Washington Academy of Sciences 98 (1):1–8.