Email: editor@ijerst.com or editor.ijerst@gmail.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

INTUTIVE HYBRID CNN MODELWITH CATBOOST FOR ANAMOLY AND INTRUSION DETECTION

Parepalli Nageswara Rao[1], Dr. K.Radhika[2]

[1]Research Scholar, Department of Computer Science and Engineering, Osmania University, Hyderabad, Telangana 500007

[2]Professor, Department of Information Technology, Chaitanya Bharati Institute of Technology, Osman Sagar Rd, Kokapet, Gandipet, Hyderabad, Telangana 500075

ABSTRACT:

Anomaly detection is of paramount importance in Anomaly detection is crucial for safeguarding critical systems across various domains, yet existing methods often face challenges in accurately identifying anomalies while minimizing false alarms. This research proposes a novel hybrid approach, merging Convolutional Neural Networks (CNN) with an CAT BOOST FILTER using ML Filter technique, to address these limitations across network security, financial fraud detection, and sensor data analysis. By leveraging CNNs to capture spatial patterns and LSTM networks to capture temporal dependencies, the model enhances anomaly detection accuracy for both static and dynamic datasets. Additionally, the Hybrid Filter with Interpolation technique dynamically combines information from CNN and LSTM components, adapting to dataset characteristics for improved versatility and effectiveness. In network traffic data, the hybrid model identifies intrusions and anomalies by analyzing spatial features within network packets and behaviors, while in financial transactions, it detects irregular activities indicative of fraud. For sensor datasets, including IoT sensor readings, the model extracts spatial and temporal features to identify anomalies caused by malfunctions or unusual conditions. Through extensive experiments, the hybrid CNN-CAT BOOST FILTER using ML Filter approach consistently outperforms traditional methods and standalone CNN models, offering a comprehensive solution for ensuring the security and reliability of networked systems, financial transactions, and sensor-based applications. To further enhance anomaly detection and address research gaps, the integration of a CAT boost filter design with machine learning algorithms, such as the Intuitive Hybrid CNN, holds promise in improving detection accuracy and reducing false alarms, thereby advancing anomaly detection capabilities in critical domains.

INTRODUCTION:

In the realm of network data processing, ensuring the integrity and security of datasets is paramount, with anomaly detection serving as a critical task. Anomalies within network data, whether indicative of errors, fraud, or potential threats, can significantly compromise data quality and decision-making processes. Traditional statistical methods like mean-variance analysis and clustering, although effective to some extent, often struggle to cope with the high-dimensional and non-linear nature of network data, leading to elevated false-positive rates and diminished detection accuracy.

Machine learning-based approaches, such as one-class SVMs and isolation forests, have emerged as popular choices for anomaly detection in network data. While these methods exhibit improved performance over traditional techniques, they may still fall short when confronted with complex and evolving anomalies within large-scale network datasets. Moreover, deep learning techniques, including autoencoders and recurrent neural networks (RNNs), have shown promise in enhancing anomaly detection accuracy. However, these approaches often focus solely on either spatial or temporal features, limiting their ability to comprehensively address the multifaceted nature of anomalies in network data.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

In response to the limitations of existing methods, our research proposes a holistic solution for anomaly detection in network data. By synergistically combining Convolutional Neural Networks (CNNs) for spatial feature extraction, Long Short-Term Memory (LSTM) networks for temporal dependency modeling, and a novel approach utilizing CatBoost filters with machine learning algorithms, we aim to provide a comprehensive framework capable of effectively detecting anomalies in network data, adapting to their dynamic nature, and reducing false positives.

The integration of CNNs and LSTM networks in our approach allows for the capture of both spatial and temporal features within network data. CNNs excel at detecting spatial patterns, making them well-suited for identifying anomalies in multidimensional and unstructured network data. Conversely, LSTM networks are proficient at modeling sequences, enabling the detection of anomalies that evolve over time. By combining these two components, our approach harnesses a synergistic effect, enhancing anomaly detection performance.

The unique aspect of our approach lies in the utilization of CatBoost filters with machine learning algorithms, which dynamically adapt to the unique characteristics of each network dataset. This innovative addition optimizes the anomaly detection process by aligning the model's behavior with the intrinsic properties of the data. By leveraging the capabilities of CatBoost filters, our approach not only enhances accuracy but also reduces false positives, rendering it a versatile and effective tool for anomaly detection in network data.

Problem Statement:

Problem Statement: The escalating complexity and diversity of network-related datasets pose challenges for anomaly detection techniques. Existing methods often struggle to effectively handle the spatial and temporal aspects of network data, leading to compromised accuracy and elevated false-positive rates. Addressing this gap requires innovative solutions capable of leveraging advanced machine learning techniques to enhance anomaly detection accuracy and efficiency.

Objectives: This study aims to achieve the following objectives:

1. Develop an advanced anomaly detection model by integrating CNN and LSTM architectures with CatBoost filter design.
2. Evaluate the performance of the proposed model on extensive real-world network-related datasets, focusing on achieving an accuracy of at least 98%.
3. Compare the effectiveness of the proposed model against traditional anomaly detection techniques, assessing its ability to reduce false positives and improve anomaly detection accuracy in network-related datasets.
4. Explore the adaptability and potential applications of the hybrid CNN-LSTM-CatBoost anomaly detection model across various network-related domains, including network security and anomaly detection in network traffic data.

Overview:

The subsequent chapters of this research work will delve into each component of the proposed solution. Section 2 provides an in-depth review of related work, highlighting the limitations of existing methods. Section 3 describes the methodology and architecture of the proposed model. Section 4 presents the experimental setup and results, showcasing the superiority of the model. Section 5 discusses the implications and potential applications of the hybrid anomaly detection model, and Section 6 concludes the research, summarizing the contributions and future directions.

LITERATURE SURVEY:

In the realm of computing, machines have traditionally followed predetermined instructions set by humans or users, a process known as machine learning. While

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

humans rely on memory to retain and acquire knowledge, machines primarily use statistical analysis to process data stored in databases. For instance, music applications utilize machine learning algorithms to recommend songs based on users' listening history. This integration of machine learning algorithms, specifically Support Vector Machines (SVM) and Learning Management Systems, is also being explored in higher education, where it holds the potential to revolutionize teaching methodologies and student learning experiences.

Machine Learning (ML) and network data analytics represent rapidly evolving fields, driven by the increasing availability of huge datasets related to network patterns. These datasets pose challenges for traditional machine learning techniques due to the demands of real-time data processing. Consequently, ML is undergoing transformation to adapt to the complexities of analyzing extensive network-related data. Recent advancements in ML for processing large-scale network datasets are being reviewed, emphasizing the need for innovative strategies to address network data challenges. This review identifies deep neural networks, support vector machines, and decision trees as commonly used techniques in network data analytics, highlighting various applications and challenges in the field.

Network data analytics plays a crucial role in enabling precise real-time analysis of large datasets. Studies propose comprehensive frameworks leveraging ML to analyze cognitive patterns among students and lecturers in the context of Decision Support Systems (DSS) at educational institutions like UIN Jakarta. By employing machine learning algorithms such as K-Means clustering, these frameworks aim to enhance academic environments by providing insights into learning patterns and optimizing educational strategies.

The integration of machine learning with large network datasets presents both challenges and opportunities. Efforts are being made to explore the utilization of ML for analyzing extensive network data, providing insights into its qualitative and quantitative characteristics. Research trends and issues in ML for network data are being identified, emphasizing the need for efficient platforms and methodologies to handle the complexities of network analytics.

Furthermore, data streams, particularly those from network sensors, are gaining attention for their potential to provide real-time insights. Initiatives like the Real-time Machine Learning Competition on Network Data Streams focus on forecasting network activity in real-time. Leveraging fast incremental learners, such competitions aim to address critical issues in data stream mining related to network patterns.

In conclusion, machine learning offers immense potential for analyzing extensive network datasets, presenting challenges and opportunities for research and innovation in network data analytics. The ongoing exploration and development of ML techniques in this domain are poised to drive advancements in various fields, from education to healthcare, by unlocking valuable insights from network data.

PROPOSED MODEL:

In our comprehensive anomaly detection system, we have developed a versatile and highly effective hybrid approach that leverages Convolutional Neural Networks (CNNs) in combination with the innovative CAT BOOST FILTER using ML Filter technique for the purpose of detecting anomalies within three diverse and critical domains: network traffic data, financial transactions, and sensor datasets. For network traffic data, the foundation of our approach lies in the utilization of CNNs to capture spatial features and intricate patterns within network packets and behaviours. These CNNs are adept at recognizing deviations in network traffic that may signify intrusions or anomalies. They provide the system with the ability to discern even the most subtle deviations from normal network behaviour, thus enhancing network security. In the realm of financial transactions, our hybrid model relies on CNNs to scrutinize transaction patterns, enabling the system to flag unusual financial activities that could potentially be indicative of fraudulent behavior. This integration of CNNs facilitates the detection of anomalies in a

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

highly dimensional and complex dataset, essential for maintaining the integrity of financial systems.

In sensor datasets, particularly those originating from IoT environments, our approach leverages CNNs to extract both spatial and temporal features from the data. This enables the identification of anomalies caused by sensor malfunctions or unforeseen environmental changes. Additionally, we incorporate the CAT BOOST FILTER using ML Filter technique into the model, ensuring that the system can adapt dynamically to changing sensor data patterns, ultimately safeguarding the reliability of critical monitoring and control systems. The key innovation in our approach lies in the seamless integration of deep learning capabilities offered by CNNs with the adaptive interpolation techniques of the CAT BOOST FILTER using ML Filter. This combination empowers our model to not only excel in anomaly detection accuracy but also to adapt readily to evolving data patterns. Through extensive experimentation and evaluation on real-world datasets, our hybrid CNN-CAT BOOST FILTER using ML Filter approach consistently demonstrates superior performance when compared to conventional methods and standalone CNN models, signifying its potential as a robust and invaluable tool for ensuring the security and reliability of networked systems, financial transactions, and sensor-based applications in today's data-driven world.

Concept:

This comprehensive exploration delves into the practical implementation of our Hybrid CNN-CAT BOOST FILTER using ML Filter approach for anomaly detection in network traffic data, financial transactions, and sensor datasets. We elucidate the underlying concept, design approach, design steps, and provide detailed descriptions, including the key formulations for the Hybrid Filter component. At the core of our implementation is the fusion of deep learning, facilitated by Convolutional Neural Networks (CNNs), with the dynamic interpolation technique known as the CAT BOOST FILTER using ML Filter. These fusion forms a versatile approach capable of capturing spatial and temporal features, adapting to evolving data patterns, and effectively identifying anomalies. It harnesses the strengths of CNNs for spatial feature extraction and LSTM-like functionality for temporal dependency modeling.

Block Diagram:

Description:

Our block diagram for anomaly detection in network traffic data, financial transactions, and sensor datasets using the Hybrid CNN-CAT BOOST FILTER using ML approach illustrates the holistic architecture of our system. This diagram encapsulates the key components, their interactions, and the flow of data within our approach.

Data Input

At the outset of our block diagram, we depict the data input stage, where diverse datasets from network traffic, financial transactions, and sensor readings are ingested into the system. These datasets serve as the foundation for our anomaly detection framework, each with its unique characteristics and data structures.

Preprocessing and Feature Extraction

Following data input, the next step involves preprocessing and feature extraction. For network traffic data, preprocessing includes packet parsing and protocol identification, while financial transactions may require data standardization and feature engineering. In the case of sensor datasets, the preprocessing stage involves data cleaning and alignment. After preprocessing, spatial and temporal features are extracted using Convolutional Neural Networks (CNNs) for spatial patterns and Long Short-Term Memory (LSTM) networks for temporal dependencies. These extracted features serve as the foundation for anomaly detection.

Hybrid CNN-CAT BOOST FILTER using ML Components:

The heart of our approach, the Hybrid CNN-CAT BOOST FILTER using ML Filter component, is depicted prominently in the block diagram. This component combines the spatial and temporal features extracted from the CNN and LSTM networks. The

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

Hybrid Filter dynamically calculates adaptive weights based on the significance of each component's output for each data instance. These weights guide the weighted combination of CNN and LSTM features, creating a feature vector that seamlessly integrates both spatial and temporal information. This integration is essential for achieving robust and adaptable anomaly detection.

Anomaly Detection

The feature vector generated by the Hybrid CNN-CAT BOOST FILTER using ML Filter is then used for anomaly detection. A threshold-based mechanism is applied to classify instances as anomalies or normal data points. This step serves as the final decision-making stage, where anomalies are flagged for further investigation or action.

Network Traffic Data Path

In our block diagram, we highlight separate paths for each data domain. In the case of network traffic data, the path emphasizes the interaction between preprocessing, feature extraction, and the Hybrid CNN-CAT BOOST FILTER using ML Filter component. This path demonstrates how our approach adapts to the intricacies of network data, effectively capturing spatial and temporal anomalies.

Financial Transactions Data Path

For financial transactions, the dedicated path showcases the preprocessing and feature extraction specific to this domain. It emphasizes the model's ability to recognize fraud patterns within financial transactions by integrating spatial and temporal information through the Hybrid Filter.

Sensor Datasets Data Path

The sensor datasets path illustrates how preprocessing and feature extraction cater to sensor data characteristics. It emphasizes the model's capacity to adapt to evolving sensor patterns, identifying anomalies caused by sensor malfunctions or environmental changes through the Hybrid Filter.

Our block diagram for anomaly detection encapsulates the robustness and versatility of our Hybrid CNN-CAT BOOST FILTER using ML Approach. By combining spatial and temporal features and dynamically adapting to changing data patterns, our system offers a comprehensive solution for enhancing security, reliability, and data-driven decision-making in the domains of network traffic, financial transactions, and sensor datasets. This block diagram represents a visual representation of the power and adaptability of our approach in the era of big data, providing actionable insights and valuable tools for critical applications.

Design Approach:

Our design approach emphasizes adaptability and real-time interpolation. The model commences with a CNN-based feature extractor to capture spatial patterns in the data. Subsequently, these spatial features are concatenated with temporal features generated by the LSTM component. The innovative aspect of our approach lies in the CAT BOOST FILTER using ML Filter, which dynamically combines CNN and LSTM information using adaptive weights, ensuring real-time adaptation to data dynamics.

Design Steps:

The Hybrid Filter serves as the linchpin of our approach, offering adaptability and real-time interpolation capabilities. It operates as follows:

1. **Weight Calculation**: For each data instance, the CAT BOOST FILTER using ML computes weights based on the CNN and LSTM output values. These weights are determined dynamically, ensuring the relative importance of each component varies according to the data's characteristics.

2. **Weighted Combination**: Subsequently, the CAT BOOST FILTER using ML Filter combines the CNN and LSTM outputs through a weighted sum. This results in a single feature representation for each data instance that effectively integrates spatial and temporal information.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

3. **Anomaly Detection**: The combined feature is then employed for anomaly detection. A threshold-based approach is employed, classifying instances as anomalies if the feature value exceeds a predefined threshold; otherwise, they are deemed normal.

Dataset Evaluation Description

- *Network Traffic Data*

During the implementation study, we applied our approach to a large-scale network traffic dataset, comprising millions of network packets. Our model exhibited outstanding performance with an anomaly detection accuracy of 98.3%. This success can be attributed to the model's ability to capture intricate spatial and temporal patterns in network behaviour and its real-time adaptation capabilities using the CAT BOOST FILTER using ML Filter.

Formulations For Hybrid Filter:

The above block diagram in figure 1 implicates the different features that are governed and estimated with adaptive filter weights as proposed below:

1. $F(x) = 1 - e^{-2*\pi\delta\mu/2\sigma^2}$ Which improvises the randomness for each set of $\delta\mu\sigma$ estimated with the expectation probability for each feature model considered.

2. Using the above equation we calculate the filter weights as the

   a. $T(n) = \sum_{i=1}^{N} x_i W_i * F(x_i) / \sum_{j=k}^{M-k} y_i W_{j-k} y(x_{i-j+k})$ (1)

   b. Here k is list of factors for which PDF is one for all the elements considered.

   c. W is weights estimated with predicted data y and input random variable operated data F(x).

With use of these weights, we initiate an optimization feature for the entire dataset using Hybrid Filter (Gaussian and CAT BOOST FILTER using ML).

EXPERIMENTATAL SETUP:

In our extensive experimental study, we rigorously evaluated the performance of the Hybrid CNN-CAT BOOST FILTER using ML Filter approach for anomaly detection across the domains of network traffic data, financial transactions, and sensor datasets. We present the key details of our experiments, including model specifications, input data sizes, and the remarkable accuracy observed in all cases, consistently hovering around an impressive 98%.

Model Specifications and Training

For our hybrid model, we employed a deep convolutional neural network (CNN) architecture with multiple convolutional and pooling layers, optimized for spatial feature extraction, followed by a Long Short-Term Memory (LSTM) layer for capturing temporal dependencies. The model's adaptive component, the CAT BOOST FILTER using ML Filter, was dynamically integrated to facilitate real-time interpolation of evolving data patterns during both training and testing phases.

In the case of network traffic data, our model was trained and tested on a sizable dataset comprising millions of network packets. The dimensions of the input data for this domain included spatial features extracted from network traffic metrics and temporal sequences of packet behaviour, resulting in a high-dimensional input size. Similarly, for financial transactions, we utilized a substantial dataset containing millions of transactions, while for sensor datasets, the input comprised sensor readings from numerous devices over an extended period. The varying input data sizes and complexity across these domains underscore the adaptability and scalability of our approach.

Metrics:

Throughout our experimental study, a consistent and noteworthy result emerged—anomaly detection accuracy exceeding 98% across all domains. In the

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

network traffic data domain, our approach showcased an accuracy of 98.3%, effectively identifying network intrusions and abnormal behaviour patterns with exceptional precision. In financial transactions, our model achieved a detection accuracy of 98.1%, successfully flagging fraudulent transactions while maintaining a low false positive rate. Sensor datasets displayed a similar trend, with an anomaly detection accuracy of 98.2%, proving the model's efficiency in identifying anomalies resulting from sensor malfunctions or environmental variations.

These remarkable results not only highlight the effectiveness of our Hybrid CNN-CAT BOOST FILTER using ML Filter approach but also underscore its adaptability and versatility across diverse and complex data domains. The consistently high accuracy demonstrates its potential to enhance security, reliability, and decision-making processes in critical applications. Furthermore, the model's ability to maintain such performance across different data sizes and types reinforces its practical applicability and scalability. This experimental study substantiates the viability of our approach as a powerful anomaly detection tool in the era of big data, providing robust solutions for network security, financial integrity, and sensor-driven applications.

RESULTS AND DISCUSSION:

In this section, we present a comprehensive discussion of the results achieved through the implementation of our Hybrid CNN-CAT BOOST FILTER using ML Filter approach for anomaly detection in network traffic data, financial transactions, and sensor datasets. We will focus on both the training and testing aspects to provide a holistic view of the approach's performance and its adaptability to various data sources.

Training and Testing Process

The success of our Hybrid CNN-CAT BOOST FILTER using ML Filter approach can be attributed to its robust training process, which incorporates both supervised and unsupervised learning. During the training phase, the model was exposed to labelled anomalies to grasp the underlying patterns across the diverse datasets. CNNs effectively captured spatial and temporal features, while the CAT BOOST FILTER using ML Filter dynamically adapted to changing data characteristics during training.

Network Traffic Data

In the network traffic data domain, our model's training and testing performance were outstanding. We trained the model on a large dataset consisting of both normal and anomalous network behaviours. During testing, the model demonstrated an impressive detection accuracy of 98.3%, indicating its proficiency in identifying malicious intrusions and abnormal network patterns. False alarms were notably reduced, with a false positive rate falling by 42%. This exemplifies the model's adaptability to evolving network behaviours and its ability to maintain a low false alarm rate even as the network environment changes.

Financial Transactions

For financial transactions, the training process involved exposure to a myriad of legitimate and fraudulent transactions, ensuring the model learned to distinguish between them. In testing, our approach excelled with a detection accuracy of 99.9%. Fraudulent activities were consistently identified, emphasizing the model's efficacy in safeguarding financial systems. Moreover, false alarms decreased by approximately 38%, a testament to the model's adaptability to emerging fraud tactics and its ability to reduce unnecessary alerts.

Sensor Datasets

In the sensor dataset domain, our approach with IOT dataset from UCI website, underwent training on these featured datasets encompassing normal sensor readings as well as anomalous instances resulting from sensor malfunctions or environmental fluctuations. Testing revealed an impressive anomaly detection accuracy exceeding 98.2%. False alarms were also significantly reduced, dropping by approximately 32%. This underscores the model's capability to adapt to shifting sensor data patterns, ensuring that it reliably identifies anomalies while minimizing false alarms in various sensor-driven applications.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

Table 1: Representing the Overall Comparison with KDD ANAMOLY for Existing and Proposed CNN hybrid Algorithms

DATASET	ALGORITHM	SENSIVITY	SECIFICITY	F1-SCORE	RECALL	PRECISION	AUC	ROC	ACCURACY
KDD	LR	90.45	85.24	88.63	84.23	87.25	0.868	0.894	89.3
KDD	SVM	89.62	90.15	85.23	84.63	87.84	0.8561	0.8834	88.36
KDD	RFC	91.17	90.75	88.41	87.63	89.74	89.96	0.894	90.86
KDD	ENSEMBLE SVM	90.91	91.75	90.75	92.75	89.75	0.914	0.904	91.85
KDD	RFC+SVM	90.32	90.41	89.56	88.74	91.23	0.912	0.9025	91.05
KDD	CNN HYBRID	99.85	97.58	97.56	98.56	98.89	0.986	0.981	97.3
KDD	CATB+CNN-HYBRID	98.56	97.24	98.75	97.48	98.96	0.99	0.986	98.5
KDD	LSTM	99.38	97.86	98.63	99.56	97.45	0.98	0.979	98.1
KDD	UNET	95.16	95.72	96.52	97.19	95.28	0.963	0.972	96.85
KDD	TRASFER LEARNING RESNET	93.75	95.63						94.86

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

In our study on the KDD dataset, we have employed a diverse set of anomaly detection algorithms, including traditional machine learning techniques and advanced deep learning models, to comprehensively assess their performance. Here, we provide a detailed comparison of our proposed CNN Hybrid and IF CNN algorithms with other algorithms based on sensitivity, specificity, F1-score, recall, precision, AUC, ROC, and accuracy metrics.

Performance of Traditional Algorithms:

Our analysis begins with traditional machine learning algorithms, such as Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest Classifier (RFC). These algorithms exhibit respectable performance with sensitivity, specificity, and accuracy scores ranging from 85% to 91%. However, they seem to struggle with achieving high F1-scores and AUC values, indicating room for improvement in precision and model discrimination.

Ensemble Models and RFC+SVM:

To enhance anomaly detection, we explored ensemble methods, such as Ensemble SVM and RFC+SVM. These models demonstrate improved sensitivity, specificity, F1-scores, and AUC values, with sensitivity reaching up to 92.75%. Ensemble models showcase the benefits of combining multiple base classifiers for better overall performance.

Deep Learning Approaches:

Transitioning to deep learning, our proposed CNN Hybrid and IF CNN algorithms deliver outstanding results. CNN Hybrid achieves exceptional sensitivity and specificity values of 99.85% and 97.58%, respectively, surpassing other models in these aspects. It also excels in F1-score, recall, precision, AUC, ROC, and accuracy, with F1-score reaching 97.56% and accuracy at 97.3%. IF CNN performs impressively as well, with a sensitivity of 98.56%, specificity of 97.24%, and AUC of 0.99.

LSTM and UNET (SOA):

We also evaluated the performance of Long Short-Term Memory (LSTM) and UNET, which are recurrent and convolutional neural networks, respectively. These models exhibit strong anomaly detection capabilities, with LSTM achieving a sensitivity of 99.38% and UNET showing competitive performance. LSTM's recall of 99.56% and UNET's F1-score of 96.52% highlight their effectiveness in capturing anomalies.

Transfer Learning:

Lastly, we explored Transfer Learning with ResNet, showing a sensitivity of 93.75% and specificity of 95.63%. While not matching the performance of our proposed CNN Hybrid and IF CNN, it demonstrates the potential of leveraging pre-trained models for anomaly detection.

In conclusion, our comprehensive evaluation of various algorithms on the KDD dataset highlights the superiority of deep learning-based approaches, particularly our CNN Hybrid and IF CNN models. These models outperform traditional algorithms in terms of sensitivity, specificity, F1-score, AUC, and accuracy, indicating their robustness in identifying anomalies within network traffic data. These findings emphasize the importance of leveraging deep learning techniques for effective intrusion detection and network security applications.

CONCLUSIONS:

Our study on "Anomaly Detection in Network Traffic Data, Financial Transactions, and Sensor Datasets Using a Hybrid CNN-CAT BOOST FILTER using ML Filter Approach" presents compelling evidence of the effectiveness and versatility of our proposed method, particularly when applied to the Network KDD dataset. The introduction of our novel Hybrid CNN-CAT BOOST FILTER using ML Filter Approach aims to address the challenges of anomaly detection in diverse datasets, emphasizing the importance of robust and accurate anomaly detection capabilities for maintaining the integrity and security of critical domains. Our experimental results highlight exceptional performance, with the Network KDD dataset consistently achieving an accuracy rate of 98%, affirming the efficacy of our hybrid approach in detecting anomalies within network traffic data. Moreover, the utilization of AUC and ROC metrics further substantiates the reliability and robustness of our model's predictive capabilities, showcasing its ability to effectively discriminate between normal and anomalous network traffic patterns.

ISSN 2319-5991 www.ijerst.com
Vol. 17, Issue 2, 2024

An important aspect of our approach is its versatility, as demonstrated by its successful application across three distinct domains: network traffic data, financial transactions, and sensor datasets. The consistent high performance achieved across these diverse datasets emphasizes the adaptability and generalizability of our hybrid model. In conclusion, our research underscores the tremendous potential of the Hybrid CNN-CAT BOOST FILTER using ML Filter Approach for anomaly detection, particularly in the context of the Network KDD dataset. The consistent accuracy rate of 98% and high AUC and ROC values attained for network traffic data validate the effectiveness and versatility of our approach, contributing to enhancing security and data integrity in critical domains.

REFERENCES:

1. A. S. Genale, B. B. Sundaram, A. Pandey, V. Janga, D. Aweke and P. Karthika, "Big Data Analysis for knowledge based on Machine Learning using Classification Algorithm," 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2022, pp. 108-113, doi: 10.1109/ICSCDS53736.2022.9760898.

2. M. Kirola, M. Memoria, M. Shuaib, K. Joshi, S. Alam and F. Alshanketi, "A Referenced Framework on New Challenges and Cutting-Edge Research Trends for Big-Data Processing Using Machine Learning Approaches," 2023 International Conference on Smart Computing and Application (ICSCA), Hail, Saudi Arabia, 2023, pp. 1-5, doi: 10.1109/ICSCA57840.2023.10087686.

3. I. K. Nti, J. A. Quarcoo, J. Aning and G. K. Fosu, "A mini-review of machine learning in big data analytics: Applications, challenges, and prospects," in Big Data Mining and Analytics, vol. 5, no. 2, pp. 81-97, June 2022, doi: 10.26599/BDMA.2021.9020028.

4. Nurhayati, Busman and V. Amrizal, "Big Data Analysis Using Hadoop Framework and Machine Learning as Decision Support System (DSS) (Case Study: Knowledge of Islam Mindset)," 2018 6th International Conference on Cyber and IT Service Management (CITSM), Parapat, Indonesia, 2018, pp. 1-6, doi: 10.1109/CITSM.2018.8674354.

5. K. Abe, "Data Mining and Machine Learning Applications for Educational Big Data in the University," 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Fukuoka, Japan, 2019, pp. 350-355, doi: 10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00071.

6. M. Assefi, E. Behravesh, G. Liu and A. P. Tafti, "Big data machine learning using apache spark MLlib," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 2017, pp. 3492-3498, doi: 10.1109/BigData.2017.8258338.

7. S. Kour and S. Z. Dey Babu, "A Comparative Analysis of Big Data Technologies using Machine Learning Techniques," 2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART), MORADABAD, India, 2021, pp. 546-550, doi: 10.1109/SMART52563.2021.9676291.

8. A. C. Onal, O. Berat Sezer, M. Ozbayoglu and E. Dogdu, "Weather data analysis and sensor fault detection using an extended IoT framework with semantics, big data, and machine learning," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 2017, pp. 2037-2046, doi: 10.1109/BigData.2017.8258150.

9. D. Boulegane et al., "Real-Time Machine Learning Competition on Data Streams at the IEEE Big Data 2019," 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 3493-3497, doi: 10.1109/BigData47090.2019.9006357.

10. S. Mittal and O. P. Sangwan, "Big Data Analytics using Machine Learning Techniques," 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2019, pp. 203-207, doi: 10.1109/CONFLUENCE.2019.8776614.

11. D. Tian, "Simulation of Distributed Big Data Intelligent Fusion Algorithm Based on Machine Learning," 2022 International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS), Bristol, United Kingdom, 2022, pp. 421-424, doi: 10.1109/AIARS57204.2022.00101.

12. S. Madan, P. Kumar, S. Rawat and T. Choudhury, "Analysis of Weather Prediction using Machine Learning & Big Data," 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE), Paris, France, 2018, pp. 259-264, doi: 10.1109/ICACCE.2018.8441679.

13. K. R. Swetha, N. M, A. M. P and M. Y. M, "Prediction of Pneumonia Using Big Data, Deep Learning and Machine Learning Techniques," 2021 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatre, India, 2021, pp. 1697-1700, doi: 10.1109/ICCES51350.2021.9489188.

14. A. Sheshasaayee and J. V. N. Lakshmi, "An insight into tree based machine learning techniques for big data analytics using Apache Spark," 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kerala, India, 2017, pp. 1740-1743, doi: 10.1109/ICICICT1.2017.8342833.

15. G. Siwach, A. Haridas and D. Bunch, "Inferencing Big Data with Artificial Intelligence & Machine Learning Models in Metaverse," 2022 International Conference on Smart Applications, Communications and Networking (SmartNets), Palapye, Botswana, 2022, pp. 01-06, doi: 10.1109/SmartNets55823.2022.9994013.

16. Z. Xiang, C. Jinghua and W. Tao, "Review of Machine Learning Algorithms for Health-care Management Medical Big Data Systems," 2020 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2020, pp. 651-654, doi: 10.1109/ICICT48043.2020.9112458.

17. M. Klymash, O. Hordiichuk-Bublivska, M. Kyryk, L. Fabri and H. Kopets, "Big Data Analysis in IIoT Systems Using the Federated Machine Learning Method," 2022 IEEE 16th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), Lviv-Slavske, Ukraine, 2022, pp. 248-252, doi: 10.1109/TCSET55632.2022.9766908.

18. J. -z. Liu, "Research on Network Big Data Security Integration Algorithm Based on Machine Learning," 2021 International Conference of Social Computing and Digital Economy (ICSCDE), Chongqing, China, 2021, pp. 264-267, doi: 10.1109/ICSCDE54196.2021.00067.

19. S. R. Swarna, S. Boyapati, P. Dixit and R. Agrawal, "Diabetes prediction by using Big Data Tool and Machine Learning Approaches," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 2020, pp. 750-755, doi: 10.1109/ICISS49785.2020.9315866.

20. B. Raviteja, K. A. Pandya, F. Khan, Z. Tufail Khan, R. Prajwal and A. Kekatpure, "Smart Supply Chain Management using Big Data Analysis and Machine Learning," 2022 International Conference on Edge Computing and Applications (ICECAA), Tamilnadu, India, 2022, pp. 190-193, doi: 10.1109/ICECAA55415.2022.9936359.