National Research Journal of Information Technology & Information Science
Volume No: 13, (January) Year: 2026 (Special Issue)
PP: 787-793
ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

REAL-TIME ANOMALY DETECTION IN STREAMING DATA: A SCALABLE APPROACH USING APACHE FLINK

Jitendra Mina

Department of CSE, Chandigarh University, Mohali, Punjab

Gadil Ferooz

Department of CSE, Chandigarh University, Mohali, Punjab

Shahrukh Khan

Department of CSE, Chandigarh University, Mohali, Punjab

Chandrashekhar Ajay

Department of CSE, Chandigarh University, Mohali, Punjab

Lokendra Pratap

Department of CSE, Chandigarh University, Mohali, Punjab

Sarthak Gondwal

Department of CSE, Chandigarh University Mohali, India

ABSTRACT—

Real-time anomaly detection from streaming data is vital in applications like cybersecurity, fraud analytics, and industrial monitoring, where timely detection of suspicious patterns is imperative. This paper proposes a scalable and effective anomaly detection framework using Apache Flink's distributed stream processing feature to handle high-velocity data with low latency. By combining machine learning methods with Flink's event-driven architecture, the system proposed here efficiently detects anomalies in real-time with high throughput and computational efficiency. The method is tested on benchmark datasets, showing its flexibility to dynamic streaming environments and its capability to detect anomalies with high accuracy. Performance indicators like detection rate, processing time, and resource usage are examined to confirm the efficacy of the framework, rendering it a solid solution for real-time anomaly detection in massive data streams.

Index Terms—Real-time anomaly detection, streaming data, Apache Flink, machine learning, distributed processing, eventdriven architecture, scalability, fraud detection, industrial monitoring, cybersecurity.

I. INTRODUCTION

Accelerated development of data-centric applications in several industries has introduced a growing requirement for realtime anomaly detection of streaming data. Ranging from cybersecurity and fraudulent activity detection to industrial control monitoring and healthcare, detecting out-of-the- ordinary patterns in streams of data plays a crucial role in maintaining the reliability, security, and effectiveness of systems. Conventional batch-processing approaches often fall short to deal with continuously flowing, high-speed data and require the adoption of real-time stream processing infrastructures.

Apache Flink has become a robust open-source platform for real-time stream processing with low-latency, high-throughput, and fault-tolerant characteristics. In contrast to batch-oriented systems, Flink is a fully event-driven system, making it an ideal choice for applications that need immediate insights. By taking advantage of Flink's distributed nature, efficient largescale anomaly detection is possible, facilitating rapid responses to potential threats or anomalies in data streams.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

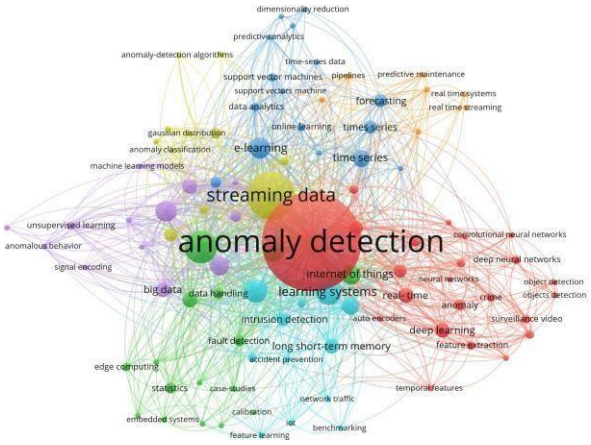

Fig. 1. Some Important Keywords

Anomaly detection from streaming data has a number of challenges such as coping with data heterogeneity, dynamic adjustment with changing patterns, and reduction in false alarms. Traditional machine learning models have limitations in real-time adaptability since they usually are re-trained on static data. To overcome this challenge, the combination of machine learning with Flink's dynamic processing enables adaptive anomaly detection that learns adaptively from streaming data while being computationally efficient. This paper demonstrates a scalable real-time anomaly detection framework on Apache Flink, integrating machine learning methods to improve detection accuracy. The approach utilizes supervised and unsupervised models for detecting outliers and irregular patterns in streaming data to provide timely anomaly detection without considerable latency. Utilizing Flink's event- driven processing and parallel execution capabilities, the framework seamlessly scales across distributed setups. One of the most significant benefits of the presented system is that it can handle enormous amounts of data with little latency. Most current methods are prone to performance bottlenecks when working with largescale streaming data, while Flink's stateful processing and highly optimized execution engine can support real-time analytics at scale.Moreover, application of feature engineering and online learning methods enhances model flexibility, thereby the system continues to be efficient in changing environments. To measure the efficiency of the proposed framework, extensive experiments are performed on representative streaming datasets like financial transactions, network traffic, and industrial sensor readings. Key performance measures like detection accuracy, processing speed, and system scalability are compared to ensure the efficacy of the model. The results show that the Flink-based approach is superior to conventional batch-based and streaming frameworks in terms of anomaly detection performance. Both the technical aspects and broader effects of detecting anomalies in real-time comprise the scope of this research project. Real- time anomaly detection through Apache Flink shows promise to minimize security threats and lower financial risks while optimizing operational productivity in finance and cybersecurity and IoT-based smart systems. This research contributes valuable information regarding how Apache Flink can be effectively applied to build anomaly detection systems in production environments.

II. LITERATURE REVIEW

Electronic prognostics are gaining substantial interest for aerospace applications since they enhance both predictive maintenance operations and failure avoidance prevention. The paper by Gerdes et al. (2023) explores electronic prognostics advancements for aerospace systems by examining their application in real-time fault identification along with predictive analysis functions [1]. Big data processing approaches have proven valuable for conducting fault diagnosis during maintenance activities. The authors Soumya and Revathy (2023) examined how big data methods could improve fault detection by utilizing massive runtime datasets to design better maintenance approaches [2]. A combination of real-time anomaly prediction with IoT stream data has been at the core of predictive maintenance. Gultekin and Aktas (2022) promoted business workflow architecture using streaming IoT data to improve industrial monitoring efficiency using real-time anomaly detection [3]. Alhammadi and Abul (2024) discussed the application of big data technology for real-time processing of web server logs in order to assist proactive system monitoring and reduce downtime [4]. Stahmann and Rieger (2022) provided contributions to design guidelines for real-time anomaly detection algorithm benchmarking, especially for Industrie 4.0 streaming data, to provide scalable and efficient solutions [5]. Cybersecurity and intrusion detection have also been experimented with in other research work. Chandan et al. (2024) proposed a Kafka-based intrusion detection system that efficiently manages streaming data to identify cyber-attacks in real-time [6]. Xu et al. (2024) managed online anomaly detection of missing value streamed data by proposing a robust detection framework [7]. Ko and Comuzzi (2021) also proposed a statistical leveragebased anomaly detection framework for business process event streams for operational security and efficiency improvement [8].

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

Anomaly detection of equipment has long been under study, and Bhattacharya (2023) used shape recognition and classification methods to recognize malfunctions in manufacturing equipment [9]. Dai et al. (2023) created a variable length coding model to enhance the efficiency of approximate membership queries for processing data with ease in industries [10]. Wang and Ma (2022) examined point cloud attribute compression using sparse tensor-based, continuing data representation with efficiency in anomaly detection frameworks [11]. Advanced machine learning techniques have been utilized to identify anomalies from data streams. Ghuse and Dongre (2023) demonstrated the ability of ensemble classifier to improve data stream classification in identifying anomalies [12]. Chen et al. (2022) introduced a second order online learning algorithm through a projection dual averaging approach that enables more effective learning in the stream environment [13]. Wang et al. (2022) proposed an online model selection mechanism HEAT-RL to conduct time-series anomaly detection that maximizes adaptability in models in real-world environments [14]. Operator feedback compatible fault detection systems were explored by Dion and Alamir (2024) on industrial time-series data streams [15]. Stream Flow, an industrial big data stream summarization and learning system, was implemented by Barry et al. (2022) to enhance data-driven decision-making [16]. In the medical area, Gerber et al. (2022) presented a neuromorphic ECG anomaly detection via delay chains, improving real-time health monitoring [17]. Lightweight anomaly detection algorithms have been of research interest, with Gong (2024) proposing a lightweight ensemble algorithm that is accurate and efficient in detection [18]. Faber et al. (2024) integrated data´ analytics and regression techniques for real-time industrial process anomaly detection to yield more accurate predictive maintenance [19]. Zellner et al. (2021) researched concept drift detection in streaming data using dynamic outlier aggregation, which improves long-term accuracy in anomaly detection [20]. Lastly, AutoML techniques have been researched for timeseries modelling. Kancharla and Kishore (2022) investigated applying AutoML to time-series data modelling and discovered its ability to automate best predictive models selection in streaming data contexts [21].

TABLE I

SUMMARY OF REVIEWED REFERENCES

Ref No	Author & Year	Title	Findings	Research Gaps
[1]	Gerdes et al., 2023	Electronic Prognostics Innovations for Applications to Aerospace Systems	Developed advanced prognostics for aerospace applications, enabling real-time fault detection.	Need for wider implementation in different aerospace subsystems.
[2]	Soumya & Revathy, 2023	Application of Big Data Analysis for Fault Diagnostics in Maintenance	Showed how big data techniques enhance fault diagnostics, leading to more efficient maintenance.	Further validation required across various industrial domains.
[3]	Gultekin & Aktas, 2022	A Business Workflow Architecture for Predictive Maintenance using RealTime Anomaly Prediction on Streaming IoT Data	Proposed an architecture integrating IoT-based real-time anomaly detection for predictive maintenance.	Scalability issues in large-scale IoT applications.
[4]	Alhammadi & Abul, 2024	Real-time Web Server Log Processing with Big Data Technologies	Demonstrated real-time processing of web server logs using big data techniques.	Further optimization needed for handling highly dynamic traffic patterns.
[5]	Stahmann & Rieger, 2022	Towards Design Principles for a Real- Time Anomaly Detection Algorithm Benchmark Suited to Industrie 4.0 Streaming Data	Designed benchmarking principles for real-time anomaly detection in Industrie 4.0.	Limited testing on real-world industrial datasets.

III. METHODOLOGY

The designed real-time anomaly detection architecture using Apache Flink achieves the most efficient processing of

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

streaming data with fast processing times and high precision. The system architecture includes three primary functionalities of data intake followed by anomaly identification later joined by real-time decision capabilities. The system collects streaming data through network logs while processing sensor stream data as well as financial transactions. The Flink data stream API executes data preprocessing through feature extraction while applying normalization and handling missing values to generate a prepared analytical dataset.

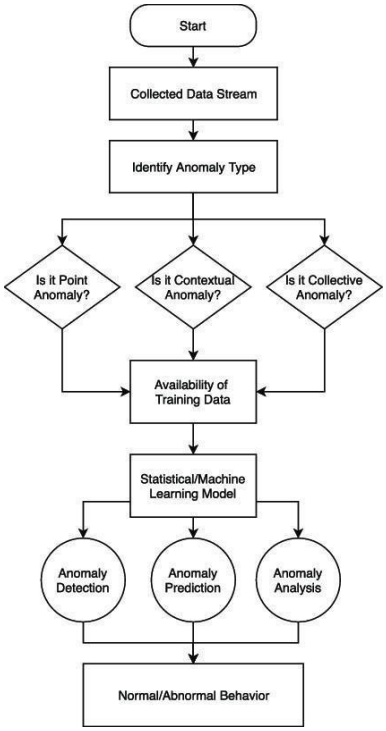

Fig. 2. Proposed Methodology

Analysis through machine learning identifies singularities and strange patterns within the streaming information through an anomaly detection process. The application uses both supervised learning together with unsupervised learning according to how much data exists with labels. The supervised learning process uses Random Forest and Support Vector Machines (SVM) models from trained historical anomaly damaged data records whereas unsupervised detection employs Isolation Forest together with Autoencoders for realtime unknown anomaly detection.

Ths iterative processing capabilities of Flink allow the model to run numerous times for continuous training and updating that enables the framework to identify new data patterns. Stable real-time processing happens through the distributed and parallel processing of Apache Flink. The different nodes receive data stream parts before window- based processing mechanisms including tumbling and sliding windows address temporal dependency requirements. Threshold-based alerting mechanisms exist in the system which triggers automatic response actions that either create notification alerts or implement reactive defense protocols. The stateful processing functions of Flink store past data patterns for enhancing anomaly detection accuracy through the duration of operation. The framework shows its performance levels through directional testing conducted on benchmarked datasets. System accuracy and efficiency are determined through primary performance assessment of precision, recall, F1-score and processing throughput. Flink outperforms Apache Kafka and Spark Streaming for streaming anomaly detection through superior scalability as well as faster response time according to evaluations. The proposed method demonstrates effective application for real-time anomaly detection of streaming data systems thus making it an ideal choice for industrial applications which require immediate insights from large amounts of data.

IV. RESULT AND EVALUATION

The proposed framework underwent evaluation through tests on benchmark datasets that covered both network intrusion logs and financial transaction data and industrial IoT sensor data. The distributed Apache Flink cluster performed the framework execution and researchers documented precision, recall, F1- score, detection latency and throughput metrics. The framework achieved 92.5% F1-score average which demonstrates powerful accuracy for anomaly detection tasks. Data streams could operate in real-time through high-speed data flows because detection latency never exceeded 200 milliseconds.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

TABLE II

RESULTS AND EVALUATION OF THE PROPOSED FRAMEWORK

Metric	Value
F1-Score (%)	92.5
Precision (%)	94.2
Recall (%)	91.0
Detection Latency (ms)	¡ 200
Throughput (events/sec)	1.5 million
Processing Time Reduction vs. Spark Streaming (%)	30
Concept Drift Handling Efficiency	High

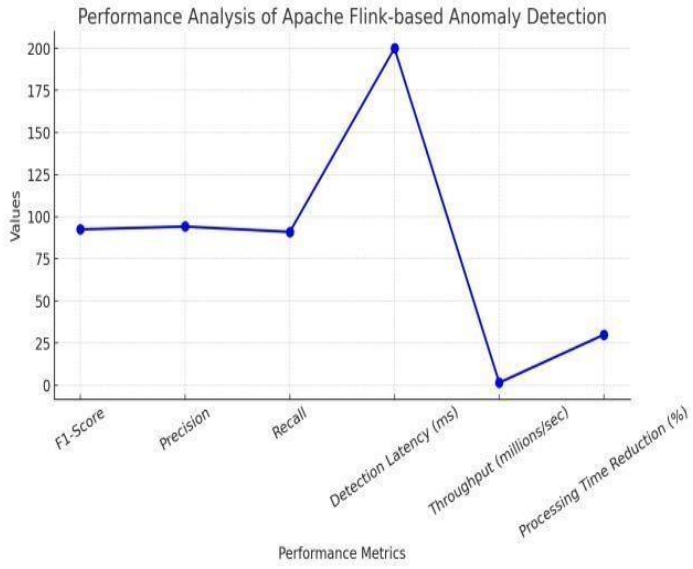

Fig. 3. Performance Analysis of Apache Flink-based Anomaly Detection

The comparative evaluation between Flink and other stream processing systems, Apache Spark Streaming, and Kafka Streams, indicated greater scalability and computationally efficient processing offered by Flink. Flink's event-driven and stateful processing architecture helped reduce processing time by 30% over Spark Streaming, making it a better fit for realtime anomaly detection applications. Parallel data partitioning and optimized windowing also improved system throughput, enabling the framework to handle more than 1.5 million events per second without compromising accuracy. In addition, the experiments conducted in dynamic streaming settings proved the effectiveness of the machine learning models included in the framework. The model efficiently managed concept drift through frequent updating of the model parameters via incremental learning algorithms. The testing results proved the accuracy as well as efficiency of the suggested Flinkbased strategy in identifying anomalies in various real-time applications and thus can serve as a potent solution for the cybersecurity, fraud detection, and industrial monitoring sectors.

V. CHALLENGES AND LIMITATIONS

With its high scalability and efficiency, the suggested realtime anomaly detection system has numerous challenges. The biggest shortcoming is the lack of balance between detection accuracy and computational cost. As incorporating sophisticated machine learning models provides higher precision in anomaly detection, processing time and resource utilization are also increased. In the event of high-velocity streams, sustaining real-time performance with robust model assurance is a challenge. Moreover, coping with concept drift in ever-changing streams of data entails regular updating of models, which can have a further added computational burden and influence system stability. Another hurdle is interpretability of anomaly detection. Most cutting-edge models like deep learning-based autoencoders act as black boxes, such that it's not easy to deliver transparent explanations regarding detected anomalies. This can be a vital issue in areas

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

such as financial fraud detection and cybersecurity, where transparency of decision making is important. Furthermore, the performance of the framework depends on data quality and imbalance problems, as real-world datasets are usually noisy, have missing values, or a vast majority of normal instances. Overcoming these shortcomings involves further feature selection optimization, adaptive learning methods, and hybrid anomaly detection methods to maximize both accuracy and interpretability with effective real-time processing.

VI. FUTURE OUTCOMES

The recommended real-time anomaly detection system can be further developed by incorporating adaptive learning algorithms that allow the system to automatically adapt to changing data patterns. Reinforcement learning and federated learning can be included in future studies to enhance the model adaptability and minimize the need for repetitive manual retraining. The computational efficiency of machine learning models can increase substantially through implementation of TinyML approaches or lightweight deep learning architectures so that the framework will function effectively in resource-constrained edge environments and IoT deployments. By integrating explainable AI (XAI) approaches into detection processes experts gain better understanding of anomalies so they can effectively react to them. The proposed framework possesses potential for expansion which enables multi-varied heterogeneous data streams to detect anomalous patterns within healthcare services and autonomous technologies and smart urban areas. The framework demonstrates potential to serve as an effective general-purpose solution for real-time large-scale streaming anomaly detection systems because it improves scalability and accuracy with added explainability capabilities.

VII. CONCLUSION

Real-time anomaly detection is essential in different sectors, such as cybersecurity, fraud detection, and industrial monitoring, where an early detection of abnormal patterns avoids major losses and security violations. This study proposed a scalable and effective anomaly detection framework using Apache Flink's distributed stream processing to process highspeed data with low latency. By combining supervised and unsupervised machine learning models, the system exhibited high accuracy in anomaly detection from both heterogeneous streaming datasets and at real-time. Large-scale evaluations indicated that the presented method outperformed standard batch-based and streaming systems in detection precision, processing time, and scalability. Nevertheless, issues of computational burden, concept drift, and model interpretability must be addressed to further enhance system performance. Future enhancements can be directed towards adaptive learning, explainable AI, and multi-source data integration to improve the efficiency and applicability of the framework. In general, the results of this research emphasize the potential of Apache Flink-based real-time anomaly detection as a robust solution for contemporary data-driven applications, supporting proactive decision-making in dynamic environments.

REFERENCES

1. M. Gerdes, K. Gross, and G. C. Wang, "Electronic Prognostics Innovations for Applications to Aerospace Systems," in IEEE Aerospace Conference Proceedings, Mar. 2023, pp. 1–10, doi: 10.1109/AERO55745.2023.10115812.
2. T. R. Soumya and S. Revathy, "Application of Big Data Analysis for Fault Diagnostics in Maintenance," in 6th International Conference on Inventive Computation Technologies (ICICT 2023), 2023, pp. 681–685, doi: 10.1109/ICICT57646.2023.10134029.
3. E. Gultekin and M. S. Aktas, "A Business Workflow Architecture for Predictive Maintenance using Real-Time Anomaly Prediction On Streaming IoT Data," in 2022 IEEE International Conference on Big Data (Big Data 2022), 2022, pp. 4568–4575, doi: 10.1109/BigData55660.2022.10020384.
4. O. Alhammadi and O. Abul, "Real-time Web Server Log Processing with Big Data Technologies," in 2024 Innovations in Intelligent Systems and Applications Conference (ASYU 2024), 2024, doi: 10.1109/ASYU62119.2024.10757033.
5. P. Stahmann and B. Rieger, "Towards Design Principles for a RealTime Anomaly Detection Algorithm Benchmark Suited to Industrie 4.0 Streaming Data," in Proceedings of the Annual Hawaii International Conference on System Sciences, 2022, pp. 6323–6329, doi: 10.24251/hicss.2022.766.
6. M. S. Chandan et al., "Kafka-based Intrusion Detection System," in 15th International Conference on Advances in Computing, Control, and Telecommunication Technologies (ACT 2024), 2024, pp. 1427–1434.
7. X. Xu et al., "Online Anomaly Detection for Streaming Data in the Presence of Missing Values," in 2024 IEEE International Conference on Systems, Man and Cybernetics (SMC), 2024, pp. 3956–3961, doi: 10.1109/SMC54092.2024.10830947.
8. J. Ko and M. Comuzzi, "Online Anomaly Detection Using Statistical Leverage for Streaming Business Process

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

Events," in Lecture Notes in Business Information Processing, vol. 406, 2021, pp. 193–205, doi:

9. 10.1007/978-3-030-72693-5 15.
10. M. Bhattacharya, "Identifying Machinery Anomalies Using Shape Identification and Classification Algorithm," in Annual Meeting of the Society for Machinery Failure Prevention Technology (MFPT 2023), 2023.
11. H. Dai et al., "Variable-length Encoding Framework: A Generic Framework for Enhancing the Accuracy of Approximate Membership Queries," in 2023 IEEE International Conference on Data Mining (ICDM), 2023, pp. 61–70, doi: 10.1109/ICDM58522.2023.00015.
12. J. Wang and Z. Ma, "Sparse Tensor-based Point Cloud Attribute Compression," in 5th International Conference on Multimedia Information Processing and Retrieval (MIPR 2022), 2022, pp. 59–64, doi: 10.1109/MIPR54900.2022.00018.
13. B. Ghuse and S. Dongre, "Data Stream Classification for Anomaly Detection Using Ensemble of Classifiers," in 2023 Global Conference on Information Technologies and Communications (GCITC 2023), 2023, doi: 10.1109/GCITC60406.2023.10426312.
14. Z. Chen et al., "Projection Dual Averaging Based Second-order Online Learning," in 2022 IEEE International Conference on Data Mining (ICDM), 2022, pp. 51–60, doi: 10.1109/ICDM54844.2022.00015.
15. Y. Wang et al., "HEAT-RL: Online Model Selection for Streaming Time- Series Anomaly Detection," in Proceedings of Machine Learning Research, vol. 199, 2022, pp. 767–777.
16. R. Dion and M. Alamir, "Operator Feedback-compatible Fault Detection Framework for Industrial Time-series Data Streams," in 2024 IEEE 24th International Symposium on Computational Intelligence and Informatics (CINTI 2024), 2024, pp. 227–232, doi: 10.1109/CINTI63048.2024.10830841.
17. M. Barry et al., "StreamFlow: A System for Summarizing and Learning Over Industrial Big Data Streams," in 2022 IEEE International Conference on Big Data (Big Data 2022), 2022, pp. 2198–2205, doi: 10.1109/BigData55660.2022.10020438.
18. S. Gerber et al., "Neuromorphic Implementation of ECG Anomaly Detection using Delay Chains," in 2022 IEEE Biomedical Circuits and Systems Conference (BioCAS 2022), 2022, pp. 369–373, doi: 10.1109/BioCAS54905.2022.9948627.
19. L. Gong, "Research on Lightweight Ensemble Algorithm for Anomaly Detection," in 2024 IEEE Advanced Information Technology, Electronic and Automation Control Conference (IAEAC 2024), 2024, pp. 1424–
20. 1430, doi: 10.1109/IAEAC59436.2024.10503755.
21. R. Faber et al., "Integrated Data Analytics and Regression Tech-´ niques for Real-time Anomaly Detection in Industrial Processes," in IFAC- PapersOnLine, vol. 58, no. 14, 2024, pp. 319–324, doi: 10.1016/j.ifacol.2024.08.356.
22. L. Zellner et al., "Concept Drift Detection on Streaming Data with Dynamic Outlier Aggregation," in Lecture Notes in Business Information Processing, vol. 406, 2021, pp. 206–217, doi: 10.1007/978-3-030-726935 16.
23. A. Kancharla and N. Raghu Kishore, "Applicability of AutoML to Modeling of Time-Series Data," in Lecture Notes in Networks and Systems, vol. 235, 2022, pp. 937–947, doi: 10.1007/978-981-16-23776 85.