Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2021, 8(8):98-104

Research Article ISSN: 2394 - 658X

Real Time Data Anomaly Detection in OTT Streaming Data Services

Arjun Mantri

Independent Researcher Seattle, USA
*mantri.arjun@gmail.com

ABSTRACT

The rapid growth of Over-the-Top (OTT) streaming services such as Netflix, Hulu, and Amazon Prime has significantly transformed media consumption patterns, generating vast amounts of data that need to be processed and analyzed in real-time. Ensuring seamless streaming quality and reliable user activity tracking is critical for maintaining user satisfaction and operational efficiency. This paper provides a comprehensive review of machine learning models used for real-time anomaly detection in OTT streaming services. It explores various techniques, including clustering-based models (K-Means, DBSCAN), classification models (Support Vector Machines, Random Forests), and neural networks (Autoencoders, Long Short-Term Memory networks). The integration of these models with Apache Spark Streaming is discussed to form a robust pipeline for real-time anomaly detection. Challenges such as data quality, model interpretability, scalability, and real-time processing are addressed, along with future directions for research. The review underscores the potential of machine learning in enhancing the resilience and reliability of OTT streaming services by effectively identifying and mitigating anomalies in user activity and streaming quality.

Key words: Real-time anomaly detection, OTT streaming services, Machine learning, Spark Streaming, User activity monitoring

---

INTRODUCTION

The proliferation of Over-the-Top (OTT) streaming services has revolutionized the media consumption landscape. Platforms like Netflix, Hulu, Amazon Prime, and Disney+ have transformed how users access and enjoy content, making it possible to stream movies, TV shows, and other video content on-demand from virtually any device connected to the internet [1,2]. This shift has led to an exponential increase in data generation, necessitating advanced techniques for managing and analyzing streaming data in real-time. One critical aspect of this data management is the detection of anomalies, which can indicate various issues ranging from network problems to unusual user behaviour [3].

The Rise of OTT Streaming Services

The advent of OTT streaming services can be traced back to the early 2000s, with Netflix being one of the pioneers in this domain. Initially a DVD rental service, Netflix transitioned to streaming in 2007, setting the stage for the OTT revolution. OTT platforms bypass traditional broadcast and cable television, delivering content directly to consumers over the internet. This direct-to-consumer model has several advantages, including the ability to offer a vast library of content, personalized recommendations, and the convenience of on-demand viewing [3,4]. The global OTT market has grown rapidly, driven by advancements in broadband technology, increased internet penetration, and the proliferation of smart devices. According to Statista, the global revenue for OTT video streaming was projected to reach $171.77 billion by 2020, with a user penetration rate of 10.7% (5). This growth has led to an increased demand for robust data processing and analytics solutions to ensure the seamless delivery of content and optimal user experience [5,6].

Importance of Anomaly Detection in OTT Services

In the context of OTT streaming services, anomalies refer to irregularities or deviations from the expected patterns in data. These anomalies can arise from various sources, including network disruptions, server failures, unauthorized access, and unexpected user behaviour. Detecting these anomalies in real-time is crucial for several reasons:

- **Quality of Service (QoS):** Anomalies can negatively impact the quality of service, leading to issues such as buffering, poor video quality, and playback interruptions. Identifying and addressing these issues promptly is essential to maintain a high level of user satisfaction (7).
- **Security:** Unauthorized access and fraudulent activities can be detected through anomaly detection. By monitoring user activity and identifying unusual patterns, OTT platforms can enhance security measures and protect user data (8).
- **Operational Efficiency:** Anomaly detection helps in identifying and diagnosing operational issues quickly, thereby reducing downtime and improving the overall efficiency of the streaming service (9).

Machine Learning for Anomaly Detection

Machine learning (ML) has emerged as a powerful tool for anomaly detection in streaming data. ML models can learn from historical data to identify patterns and detect deviations in real-time. Several ML techniques are commonly used for anomaly detection:

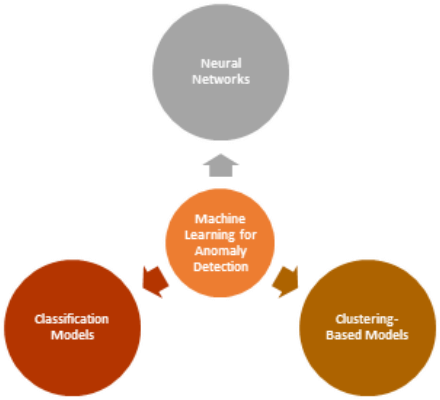

Figure 1: Machine Learning for Anomaly Detection

- **Clustering-Based Models:** Techniques such as K-means and DBSCAN group data points into clusters based on similarity. Data points that do not fit well into any cluster are flagged as anomalies (9).
- **Classification Models:** Supervised learning models, including Support Vector Machines (SVM) and Random Forests, are trained on labelled datasets to classify data points as normal or anomalous (10).
- **Neural Networks:** Deep learning models, particularly Autoencoders and Long Short-Term Memory (LSTM) networks, are effective in capturing complex temporal patterns in streaming data, making them suitable for anomaly detection (11).

Apache Spark Streaming

Apache Spark Streaming is an extension of the core Spark API that enables scalable and fault-tolerant stream processing of live data streams. Spark Streaming divides incoming data into batches and processes each batch using Spark's computational model (11,12). This framework is particularly well-suited for real-time anomaly detection in OTT streaming services due to its ability to handle large-scale data processing with low latency.

Integrating ML and Spark Streaming for Real-Time Anomaly Detection

The integration of machine learning models with Spark Streaming forms a powerful pipeline for real-time anomaly detection in OTT streaming services. The typical workflow involves several steps:

- **Data Ingestion:** Tools like Apache Kafka are used to ingest streaming data from various sources, such as user activity logs and streaming quality metrics (12).
- **Stream Processing:** Spark Streaming processes the ingested data in real-time, applying necessary transformations and aggregations.

- **Model Application:** Pre-trained machine learning models are applied to the processed data to identify anomalies. This can be achieved by leveraging Spark's MLlib library for distributed machine learning (13).
- **Alerting and Mitigation:** Detected anomalies are flagged, and alerts are generated for further investigation. Automated mitigation strategies can be implemented to address common issues, such as adjusting bitrate for improved streaming quality or blocking suspicious user activities (13,14).

MACHINE LEARNING MODELS FOR ANOMALY DETECTION IN OTT STREAMING SERVICES

The use of machine learning (ML) models for anomaly detection in Over-the-Top (OTT) streaming services is essential for maintaining the quality of service, ensuring security, and optimizing operational efficiency. This section delves into the various machine learning models employed for real-time anomaly detection, explaining their methodologies, strengths, and applications within the context of OTT streaming.

A. Clustering-Based Models

Clustering-based models group data points into clusters based on their similarities, making it easier to identify outliers or anomalies. These models are unsupervised, meaning they do not require labelled training data, which is particularly advantageous when dealing with vast amounts of streaming data where anomalies are infrequent and not pre-labelled (15).

B. K-Means Clustering

K-Means clustering is one of the simplest and most widely used clustering techniques. It partitions the data into K clusters, where each data point belongs to the cluster with the nearest mean. The algorithm iterates to minimize the variance within each cluster. This method's primary advantage lies in its scalability and computational efficiency, making it suitable for large datasets typical of OTT streaming services. Additionally, its simplicity makes it easy to implement and understand (16). In OTT applications, K-Means can be used for user behaviour analysis by grouping user activities to detect unusual patterns or behaviours. It can also cluster streaming quality metrics to flag outliers indicating potential issues. However, the choice of the number of clusters (K) can be challenging without domain knowledge, and the algorithm's sensitivity to initialization can lead to different solutions based on initial cluster centres.

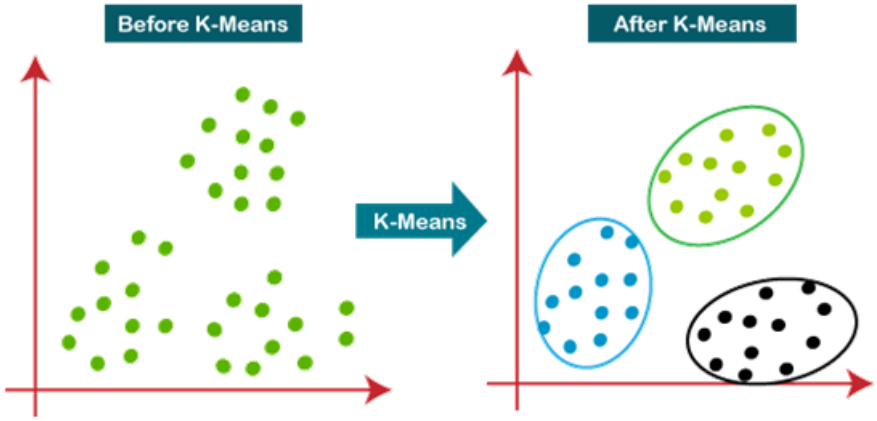

Figure 2: K-Means clustering

C. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN groups data points that are closely packed together, marking points that lie alone in low-density regions as outliers. One of its primary advantages is its ability to detect clusters of varying shapes and sizes, unlike K-Means. Additionally, DBSCAN effectively identifies and handles noise, making it suitable for anomaly detection in dynamic environments. In OTT streaming, it can be applied to network anomaly detection, identifying unusual network traffic patterns that could indicate security threats, and content delivery issues, detecting irregularities in performance metrics. Despite its advantages, DBSCAN's performance is sensitive to parameter selection, particularly epsilon and minPoints, which must be carefully chosen to achieve optimal results (17,18).

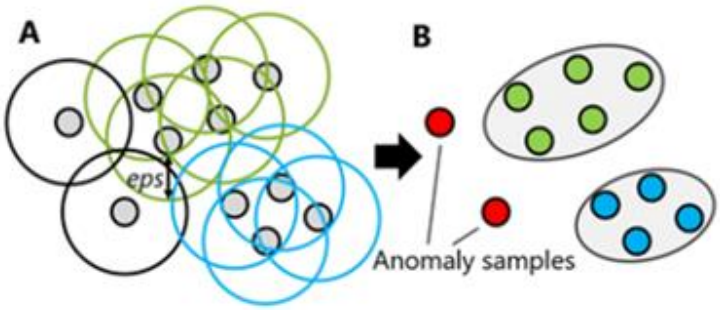

Figure 3: DBSCAN groups

CLASSIFICATION MODELS

Classification models involve supervised learning techniques where the model is trained on labelled data to classify new observations as normal or anomalous. These models are powerful when historical labelled data is available.

A. Support Vector Machines (SVM)

Support Vector Machines are effective for high-dimensional spaces and work well for both linear and non-linear classification. They provide high classification accuracy, especially in high-dimensional spaces, and the use of kernel functions allows SVMs to handle non-linear relationships. In OTT streaming, SVMs can be used for user authentication by detecting fraudulent access attempts through classification of login patterns, and for content usage monitoring by identifying unusual streaming sessions. However, SVMs can be computationally intensive, especially with large datasets, and require a significant amount of labeled training data, which can be challenging to obtain (19,20).

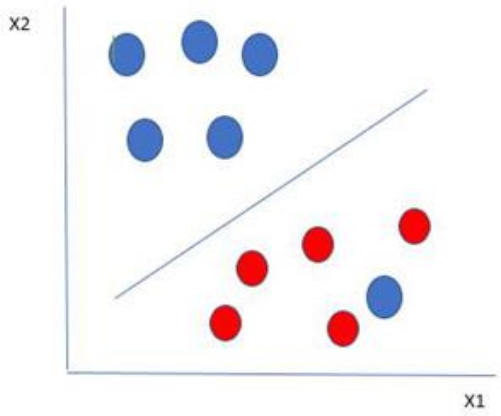

Figure 4: Support Vector Machines (SVM)

B. Random Forests

Random Forests are an ensemble learning method that builds multiple decision trees and merges them to get a more accurate and stable prediction. They are known for their robustness and resistance to overfitting, providing reliable predictions across various applications. Furthermore, Random Forests offer insights into feature importance, helping to understand the underlying data patterns. In OTT services, they can be applied to real-time streaming quality monitoring, classifying sessions to detect anomalies in quality metrics, and user activity classification, detecting unusual behaviours based on activity logs. The complexity of the model can increase with a large number of trees, making it computationally intensive, and while more interpretable than some models, the ensemble nature can still pose interpretability challenges (21, 22).

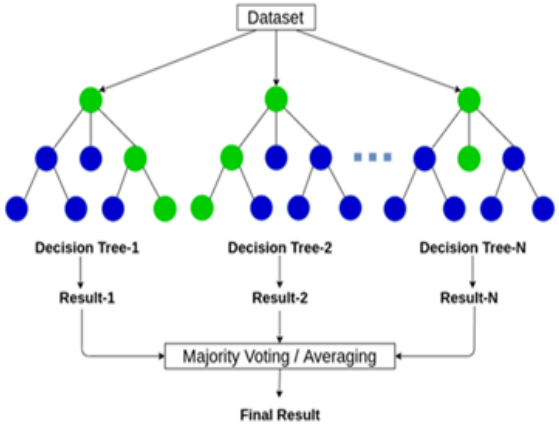

Figure 5: Random Forests

C. Neural Networks

Neural networks, particularly deep learning models, have shown great promise in capturing complex patterns in data, making them suitable for anomaly detection in OTT streaming services (23,24).

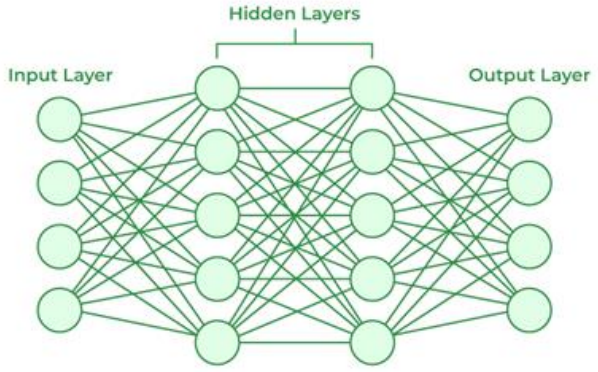

Figure 6: Neural networks

D. Autoencoders

Autoencoders are neural networks designed to learn a compressed representation of data. They consist of an encoder that compresses the input into a lower-dimensional space and a decoder that reconstructs the input from this compressed representation. Anomalies are detected based on the reconstruction error, with high errors indicating deviations from normal patterns. Autoencoders can be trained without labelled data, making them suitable for unsupervised learning. They are capable of capturing intricate patterns, making them effective for detecting anomalies in streaming quality metrics and user behaviour. However, training autoencoders can be computationally intensive and requires large datasets. Additionally, performance heavily depends on hyperparameter tuning, which can be challenging (23,24)

E. Long short-term memory (LSTM) networks

LSTM networks are a type of recurrent neural network (RNN) designed to capture long-term dependencies in sequential data. They are particularly effective for time-series anomaly detection, handling temporal data and capturing dependencies over long sequences. This capability makes LSTMs well-suited for handling streaming data in OTT services, where they can be used for temporal anomaly detection in streaming quality and user session analysis. Despite their effectiveness, training LSTMs requires significant computational resources and time, and they are prone to overfitting, especially with limited data (24).

HYBRID APPROACHES

Combining different machine learning models can enhance the robustness and accuracy of anomaly detection systems. Hybrid approaches leverage the strengths of multiple models to improve performance.

Example: Clustering with Neural Networks

A hybrid approach might involve using clustering algorithms to identify potential anomalies and then applying neural networks to further refine the detection. For instance, clustering can initially flag data points as potential outliers, and an autoencoder can be used to analyze these points in detail. This method combines the scalability and simplicity of clustering with the pattern recognition capabilities of neural networks, providing enhanced accuracy and robustness. However, implementing and maintaining hybrid systems can be complex and resource-intensive, and ensuring seamless integration and coordination between different models can be challenging (23,24).

CHALLENGES AND CONSIDERATIONS

While machine learning models offer significant advantages for anomaly detection, several challenges must be addressed. The effectiveness of anomaly detection models depends heavily on the quality of the input data. Poor data quality can lead to inaccurate detections, which may result in missed anomalies or false positives. Moreover, some machine learning models, particularly deep learning models, can be challenging to interpret (22,23). Understanding why a model flagged a particular data point as an anomaly is crucial for trust and further action. As OTT services continue to grow, the models and systems must scale efficiently to handle increasing data volumes without compromising performance. Ensuring that models can process data in real-time with low latency is critical for timely anomaly detection and mitigation. Additionally, anomaly detection models should adapt to evolving data patterns. Implementing mechanisms for continuous learning and model updating is essential to maintain effectiveness.

FUTURE DIRECTIONS

Future research and development in machine learning for anomaly detection in OTT streaming services should focus on enhanced scalability, developing more scalable algorithms and architectures to handle the growing data volumes and user base of OTT services. Improving the interpretability of machine learning models to provide clear explanations for detected anomalies, enhancing trust, and facilitating corrective actions is also crucial. Exploring more sophisticated hybrid models that combine the strengths of different approaches to improve robustness and accuracy is another area of interest. Leveraging edge computing to perform anomaly detection closer to the data source can reduce latency and improve real-time capabilities. Lastly, enhancing models to be robust against adversarial attacks ensures reliable anomaly detection in the presence of malicious attempts to evade detection.

CONCLUSION

Machine learning models play a pivotal role in real-time anomaly detection for OTT streaming services, helping to ensure high-quality user experiences, enhance security, and optimize operational efficiency. From clustering-based models like K-Means and DBSCAN to classification models such as SVMs and Random Forests, and advanced neural networks like Autoencoders and LSTMs, each technique offers unique strengths and applications. While challenges exist, ongoing advancements and future research promise to further enhance the effectiveness of these models, paving the way for more resilient and reliable OTT streaming services.

REFERENCES

[1]. Ahmad, S., Lavin, A., Purdy, S., & Agha, Z. (2017). Unsupervised real-time anomaly detection for streaming data. Neurocomputing, 262, 134-147. doi: 10.1016/j.neucom.2017.04.070

[2]. Talagala, P. D. (2019). Anomaly Detection in Streaming Time Series Data.

[3]. Ahmad, S., & Purdy, S. (2016). Real-Time Anomaly Detection for Streaming Analytics. ArXiv, abs/1607.02480.

[4]. Munir, M., Siddiqui, S. A., Dengel, A., & Ahmed, S. (2019). DeepAnT: A Deep Learning Approach for Unsupervised Anomaly Detection in Time Series. IEEE Access, 7, 1991-2005. doi: 10.1109/ACCESS.2018.2886457

[5]. Hill, D., & Minsker, B. (2010). Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. Environ. Model. Softw., 25, 1014-1022. doi: 10.1016/j.envsoft.2009.08.010

[6]. Tan, S. C., Ting, K., & Liu, F. (2011). Fast Anomaly Detection for Streaming Data. IJCAI11-254, 1511-1516. doi: 10.5591/978-1-57735-516-8/IJCAI11-254

[7]. Huang, H., & Kasiviswanathan, S. (2015). Streaming Anomaly Detection Using Randomized Matrix Sketching. Proc. VLDB Endow., 9, 192-203. doi: 10.14778/2850583.2850593

[8]. Chen, Z., Yu, X., Ling, Y., Song, B., Quan, W., & Hu, X. (2018). Correlated Anomaly Detection from Large Streaming Data. 2018 IEEE International Conference on Big Data (Big Data), 982-992. doi: 10.1109/BigData.2018.8622004

[9]. Rana, A. I., Estrada, G., Solé, M., & Muntés, V. (2016). Anomaly Detection Guidelines for Data Streams in Big Data. 2016 3rd International Conference on Soft Computing & Machine Intelligence (ISCMI), 94-98. doi: 10.1109/ISCMI.2016.24

[10]. Zhu, S., Yuchi, H., & Xie, Y. (2019). Adversarial Anomaly Detection for Marked Spatio-Temporal Streaming Data. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 8921-8925. doi: 10.1109/ICASSP40776.2020.9053837

[11]. Liu, N., & Rebentrost, P. (2017). Quantum machine learning for quantum anomaly detection. Physical Review A, 97.

[12]. Wahyono, T., & Heryadi, Y. (2019). Machine Learning Applications for Anomaly Detection. In Computational Intelligence in the Internet of Things.

[13]. Murphree, J. (2016). Machine learning anomaly detection in large systems. 2016 IEEE AUTOTESTCON, 1-9.

[14]. Zhang, J., Gardner, R., & Vukotic, I. (2019). Anomaly detection in wide area network meshes using two machine learning algorithms. Future Gener. Comput. Syst., 93, 418-426.

[15]. Shon, T., & Moon, J. (2007). A hybrid machine learning approach to network anomaly detection. Inf. Sci., 177, 3799-3821.

[16]. S. Wermter and R. Sun, (eds.) Hybrid Neural Systems. Springer-Verlag, Heidelberg. 2000. http://www.cogsci.rpi.edu/~rsun/book4-ann.html Archived 2009-09-24 at the Wayback Machine

[17]. Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M. (eds.). A density-based algorithm for discovering clusters in large spatial databases with noise (PDF). Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231. CiteSeerX 10.1.1.121.9220. ISBN 1-57735-004-9.

[18]. "Microsoft Academic Search: Papers". Archived from the original on April 21, 2010. Retrieved 2010-04-18. Most cited data mining articles according to Microsoft academic search; DBSCAN is on rank 24.

[19]. Vapnik, Vladimir N. (1997). Gerstner, Wulfram; Germond, Alain; Hasler, Martin; Nicoud, Jean-Daniel (eds.). "The Support Vector method". Artificial Neural Networks — ICANN'97. Berlin, Heidelberg: Springer: 261–271. doi:10.1007/BFb0020166. ISBN 978-3-540-69620-9.

[20]. Ben-Hur, Asa; Horn, David; Siegelmann, Hava; Vapnik, Vladimir N. ""Support vector clustering" (2001);". Journal of Machine Learning Research. 2: 125–137.

[21]. Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). The Elements of Statistical Learning (2nd ed.). Springer. ISBN 0-387-95284-5.

[22]. Kleinberg E (1990). "Stochastic Discrimination" (PDF). Annals of Mathematics and Artificial Intelligence. 1 (1–4): 207–239. CiteSeerX 10.1.1.25.6750. doi:10.1007/BF01531079. S2CID 206795835. Archived from the original (PDF) on 2018-01-18.

[23]. Vincent, Pascal; Larochelle, Hugo (2010). "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion". Journal of Machine Learning Research. 11: 3371–3408.

[24]. Welling, Max; Kingma, Diederik P. (2019). "An Introduction to Variational Autoencoders". Foundations and Trends in Machine Learning. 12 (4): 307–392. arXiv:1906.02691