Automation and Machine Learning (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2023.040307
ISSN 2516-5003 Vol. 4 Num. 3

LSTM-Based Anomaly Detection in Manufacturing Environmental Monitoring Data

Junjing Qiao[1,a,*], Enxian Zhou[1,b]

[1]*Shanghai Zhenhua Heavy Industries Co., Ltd., Shanghai, China*
[a]*qiaojunjing@zpmc.com*, [b]*zhouenxian@zpmc.com*
[*]*Corresponding author*

Keywords: Anomaly detection, Environmental monitoring, LSTM, Manufacturing emission

Abstract: With the proliferation of environmental monitoring data, using machine learning techniques for anomaly detection in environmental time series data has become an active research direction. This study employs Long Short-Term Memory (LSTM) neural network models to detect anomalies in manufacturing emission data. The research first preprocesses the data by handling missing values and conducting stationarity tests. The data will be divided into training and testing sets, with the model trained on normal data and tested for anomalies. Experiments show LSTM outperforms classic methods like Isolation Forest, Matrix Profile, and AutoEncoder in handling enclosed pipeline emission data. This study is primarily due to LSTM's ability to capture long-term dependencies in time series data. Establishing this model facilitates improved environmental protection and safety management, enables automated monitoring and warning, reduces manual intervention, and lowers enterprise environmental compliance risks. This study provides an effective anomaly detection model for monitoring manufacturing emissions, serving as a reliable reference for introducing machine learning into environmental monitoring domains.

1. Introduction

In recent years, the proliferation of sophisticated data collection mechanisms in environmental monitoring has triggered an unprecedented surge in time series data volume. This surge presents a treasure trove of opportunities for uncovering insights into pollution sources, trends, and environmental quality and introduces inherent complexities. Real-world monitoring data, rife with missing values, outliers, and noise, poses formidable challenges to accurate analysis and interpretation.

Addressing these challenges has spurred the exploration of machine learning techniques as robust tools for automated anomaly detection within environmental time series data. This study ventures into this realm, aiming to leverage these techniques to revolutionize the identification of anomalies, thereby fostering a more comprehensive understanding of environmental dynamics.

The focal point of this study lies in environmental anomaly detection, with a specific emphasis on emissions emanating from manufacturing industry outlets. The primary goal is to harness the potential of machine learning, particularly Long Short-Term Memory (LSTM) neural networks, to detect anomalies within enclosed gas emission data. A nuanced understanding of pollutant

concentrations, trends, and cyclical fluctuations is paramount in devising effective anomaly detection models.

2. Literature Review

Environmental monitoring generates extensive time series data, offering insights into pollution sources and trends. However, this real-world data often includes anomalies like missing values, outliers, and noise, which can skew analysis if not appropriately managed. Machine learning methods have emerged as promising tools for automatically detecting environmental time series data anomalies.

Several studies have applied machine learning models like LSTM neural networks to detect anomalies in pollutant concentration data. Housh and Ostfeld[1] developed an integrated logit model using dynamic thresholds and Bayesian sequential probability to detect contamination events in water distribution system data. Their model outperformed previous statistical methods by capturing time dependencies and combining evidence from multiple water quality indicators. Mukherjee et al.[2] compared various classification algorithms like logistic regression and random forest for detecting anomalies in Internet-of-Things sensor data. Lu et al.[3] proposed a sliding window approach to extract time series features and identify outlier subsequences in VOC sensor data, followed by time series decomposition and clustering to pinpoint anomalous values.

Zhong et al.[4] delivered an extensive overview of machine learning in environmental science and engineering, focusing on four main applications: prediction, feature importance identification, anomaly detection (e.g., using DBSCAN for water network contamination and LSTM for pipe burst prediction), and the best practices for implementing machine learning in this field.

Other studies have focused on handling missing values commonly occurring in monitoring data. Du et al.[5] analyzed how different missing data imputation techniques like multiple imputation impact the accuracy of LSTM models in predicting air pollutant concentrations.

The literature demonstrates that machine learning paired with proper data preprocessing can enhance anomaly detection and time series forecasting for environmental monitoring data. Data preprocessing includes identifying and handling outliers, missing values, and latency. Advanced deep learning models like LSTM networks promise to capture complex temporal behaviours. Future opportunities include integrating spatial correlations, investigating anomaly causes, and improving model interpretability.

3. Methodology

Split into training and testing sets, the unsupervised anomaly detection uses only normal samples (0-6000) for training and identifies anomalies in the testing set (6001-40000). Evaluation ensures alignment with current contexts, specifically investigating Volatile Organic Compounds (VOCs) in paint shop exhaust pipes. Gas chromatography (GC) with a flame ionization detector (FID) is employed to detect organic compounds and volatile hydrocarbons efficiently.

3.1. Data Set

This study accessed data from the workshop's environmental monitoring points via an enterprise Internet of Things (IoT) platform. This IoT platform enables retrieving historical and real-time data through API interfaces. The Non-methane volatile organic compounds (NMVOCs) (mg/m3) data from Workshop 3 in June 2023 was selected as the study dataset for this research. The data was collected at a frequency of one sample per minute, resulting in 43,168 samples, each comprising seven attributes. Among these samples, 64 data points contain missing values, while 323 data points

exhibit anomalies. For a comprehensive distribution of different data and a detailed description of a specific feature, please refer to Table 1.

Table 1: Emissions data feature characteristics summary.

Feature	Mean	Std	Min	25%	50%	75%	Max
Smoke Velocity (m/s)	13.339	1.09	0	12.86	13.37	13.88	17.54
Smoke Pressure (Pa)	-0.062	0.081	-2	-0.08	-0.06	-0.04	0.05
Smoke Temperature (℃)	29.386	2.226	0	28	29.2	30.5	37.1
Waste Gas Flow (m3/h)	107991.3	8869.9	0	104048	108247	112450	142064
Smoke Humidity (%)	1.864	0.33	0	1.671	1.837	2.062	3.444
NMVOCs (mg/m3)	31.279	11.419	0	22.975	28.969	37.643	108

3.2. Data Preprocessing

Machine learning relies on numerical input features, usually integers or floating-point numbers. Preprocessing data is crucial for improving machine learning model performance and resilience. We used a direct deletion method to handle missing values, removing 64 from our dataset.

Ensuring time series data is stationary is vital. We conducted a Dickey-Fuller (ADF) test, determining stationarity, a key assumption for statistical models and time series analysis. The test provides ADF statistics, p-value, and critical values (see Figure 1). Our analysis confirms the dataset's stationarity without any apparent trend, affirming its suitability for this research.

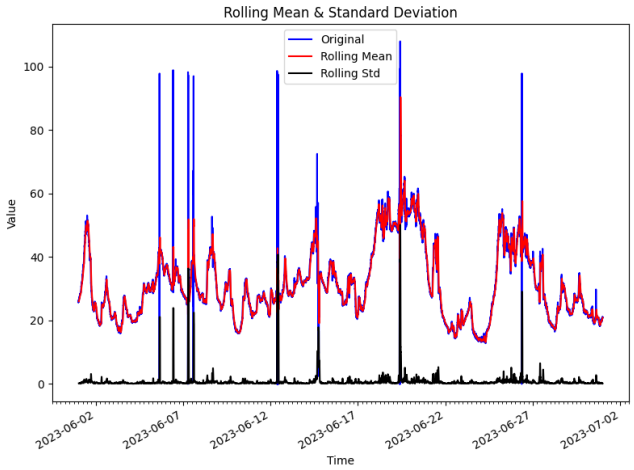

Figure 1: Stationarity Test Results.

Next, we partitioned the dataset into training and testing sets. Specifically, data within the range of 0 to 6000 were designated as the training set, while data ranging from 6001 to 40000 were allocated to the testing set.

4. Evaluation

4.1. LSTM Method Results

In this study, the effectiveness of the Long Short-Term Memory (LSTM) neural network for anomaly detection in emissions data from enclosed pipes is attributed to its robust modelling capability of long-term dependencies within time-series data. Precisely, the gate-controlled memory units within the LSTM model can capture and retain crucial information from historical time-series data. This capability allows a better understanding of the dynamic variations in emission data, such

as trends and seasonal cyclical fluctuations. Consequently, the LSTM model effectively compares current data points with historical baselines, aiding in more accurate anomaly identification and reducing false alarms.

The experimental results depicting the LSTM model's ability to detect anomalies in time-series data are illustrated in Figure 2. The data time interval is one minute, and the red dots are abnormal data. As can be seen from the figure, the LSTM model can identify outliers in time series data.

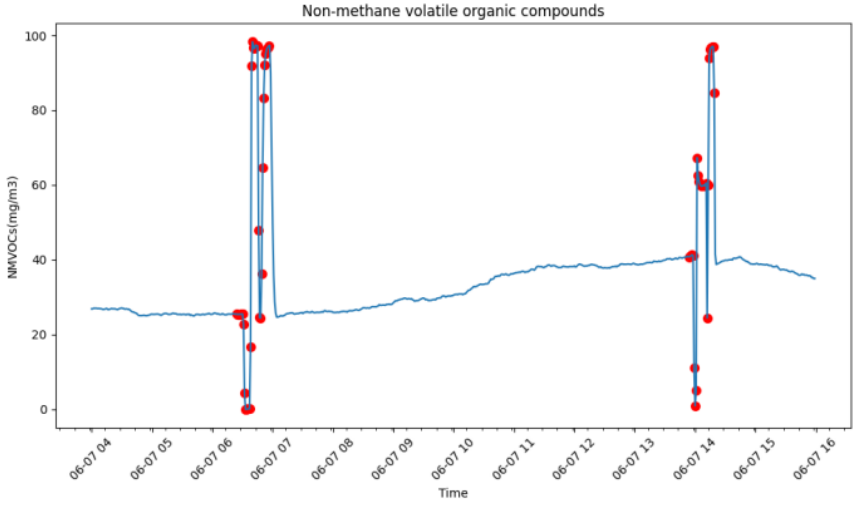

Figure 2: MNVOCs Data from June 7th, 2023, between 04:00 and 16:00.

The LSTM model adjusts parameters using backpropagation, accommodating emission data dynamics by recognizing nonlinear relationships typical in pipe emissions. Unlike linear models, LSTM copes better with these complexities due to its nonlinear functions and multi-layer structure, fitting emission data more effectively and handling sudden anomalies well.

The LSTM model utilized in this study demonstrates strong expressive capabilities through extensive training and optimization. It adapts well to the complex dynamic distribution of emission data, showcasing significantly improved anomaly detection performance compared to traditional methods.

4.2. Compare with Other Methods

Comprehensively assessing the performances of different methods in anomaly detection tasks related to emissions data from enclosed pipes, this study conducted comparative experiments involving several classic anomaly detection algorithms, including isolation forest, matrix profile, and auto-encoder. These methods represent detection strategies based on tree structures, similarity analysis, unsupervised learning, and local density estimation. By comparing the strengths and weaknesses of these methods, this research aims to identify the optimal anomaly detection approach for this specific application scenario.

Compared to other methods, the experimental results indicate that models based on LSTM exhibit the best overall performance when handling emissions data from enclosed pipe outlets. Figure 3 depicts that the LSTM model's ROC curve distinctly outperforms other methods. This superiority primarily stems from its unique advantage in modelling the long-term dependencies within time-series data. Specifically, emissions data from pipe outlets exhibit significant time-related features, such as periodic patterns and trend changes. LSTM effectively learns these temporal patterns, thereby enhancing detection accuracy and robustness. Moreover, there are periodic fluctuations in emissions, which LSTM can better model, thereby reducing false positives.

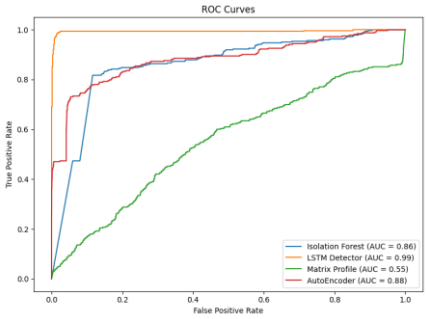

Figure 3: Comparison of ROC of various anomaly detection model

Furthermore, LSTM demonstrates strong generalization capabilities, allowing it to adapt to emissions data from different pipes and facilities with simple deployment. In contrast, other methods may require parameter adjustments for new environments. This research demonstrates that LSTM is the optimal technological route for this anomaly detection task. Its time-series modelling and generalization capabilities are superior to other classic detection algorithms.

5. Conclusions

The primary objective of this study was to establish a machine learning-based anomaly detection model for monitoring emissions from manufacturing industry outlets. Employing Long Short-Term Memory (LSTM) neural networks as the core model, this research made significant strides in detecting anomalies within closed-channel gas emission data.

The main findings indicate that the LSTM-based anomaly detection model performs well in handling emission data. Firstly, the LSTM model captures long-term dependencies within temporal data, enhancing its understanding of the dynamic characteristics of gas emissions, including trends and cyclical fluctuations. Secondly, its ability to handle nonlinear relationships in emission data improves anomaly detection accuracy. In contrast, conventional methods like Isolation Forests exhibit lesser capabilities in handling temporal data and generalization than LSTM.

Hence, this study introduces a potent anomaly detection model for monitoring manufacturing emissions with extensive environmental conservation and safety applications. Future work should emphasize gathering diverse gas emission data to bolster anomaly detection models, enhancing their adaptability across scenarios.

Acknowledgements

This work was supported by Shanghai Science and Technology Committee Rising-Star Program (No.22YF1448700).

References

[1] Housh, M., & Ostfeld, A. (2015). An integrated logit model for contamination event detection in water distribution systems. Water Research, 75, 210-223.

[2] Mukherjee, I., Sahu, N. K., & Sahana, S. K. (2021). Simulation and Modeling for Anomaly Detection in IoT Network Using Machine Learning. International Journal of Wireless Information Networks, 1-13.

[3] Lu, Q., Wang, L., & Huang, G. (2022). Abnormal detection and recovery of pollutant data considering time series characteristics. Journal of Safety and Environment.

[4] Zhong, S., Zhang, K., Bagheri, M., Burken, J. G., Gu, A., Li, B.,& Zhang, H. (2021). Machine Learning: New Ideas and Tools in Environmental Science and Engineering. Environmental Science & Technology.

[5] Du, Y., Zhang, Y., Yuan, Z., Guan, P., & Peng, Y. (2021). Accuracy analysis of air pollution prediction for LSTM network based on data preprocessing. Computer and Digital Engineering, 49(7), 1400-1425.