A Hybrid Approach of ConvLSTM-DT and GPT-4 for Real-time Anomaly Detection Decision Support in Edge-Cloud Environments

Radityo Fajar Pamungkas, Ida Bagus Krishna Yoga Utama, Khairi Hindriyandhito, Yeong Min Jang*

Department of Electronics Engineering, Kookmin University, Seoul, South Korea

Abstract

Anomaly detection is crucial in various domains for early identification of abnormal behavior. This research introduces an innovative approach that combines prediction-based detectors with Large Language Models (LLMs) for anomaly detection, focusing on indoor air quality data from multiple sensors. The hybrid approach integrates Convolutional Long Short-term Memory with non-parametric dynamic thresholding (ConvLSTM-DT) for prediction-based anomaly detection and fine-tuned GPT-4 for generating human-understandable explanations. Each sensor parameter has its specific model for accurate predictions. Furthermore, Dynamic thresholding and continuous learning adapts to the dynamic environment, update model and setting non-parametric confidence intervals for anomaly detection in rapidly changing scenarios. The system deploys anomaly detection on the edge for reduced latency and fast detection, while LLM processing occurs on the cloud for resource optimization. The results demonstrate accurate anomaly detection and well-explained reasoning for real-time decision-making, offering a novel approach for comprehensive anomaly detection solutions in various applications.

Keywords: Anomaly detection, Large Language Models, ConvLSTM-DT, Human-understandable explanations, Dynamic thresholding

1. Introduction

The increasing complexity of modern IoT systems and data-driven applications highlights the critical importance of anomaly detection. While traditional methods have their merits, their heavy reliance on expert analysis poses scalability challenges and is susceptible to human bias. This becomes even more pronounced when anomalies are multi-dimensional, defying easy identification by established patterns. If left unaddressed, these complexities can result in numerous false alarms, undermining the effectiveness of detection systems.

In the realm of Indoor Air Quality (IAQ) error management, pinpointing the reasons behind an inaccurate reading is paramount. When an IAQ monitoring system signals poor air quality indoors, users seek insight into the key attributes or factors influencing this determination. Conducting a sensitivity analysis becomes essential for identifying potential adjustments, such as tweaking sensor placement or considering environmental factors like ventilation. However, certain IAQ-influencing factors, such as external pollutants or building design, may lie beyond control. Hence, a robust error management system is crucial for elucidating the decision-making process, offering insights into the rationale behind specific air quality assessments, and proposing actionable steps for improvement. This is particularly pertinent for ensuring the precision of IAQ data interpretation and guiding effective corrective measures.

To address these challenges, this research introduces a hybrid approach that integrates prediction-based detector algorithms with the capabilities of Large Language Models (LLMs). Focusing on anomalies in air data such as temperature, humidity, and PM levels is crucial for minimizing potential health risks. By combining Convolutional Long Short-Term Memory with

*Corresponding author

Email addresses: radityofajar@gmail.com (Radityo Fajar Pamungkas), idabaguskrishnayogautama@gmail.com (Ida Bagus Krishna Yoga Utama), khairi@kookmin.ac.kr (Khairi Hindriyandhito), yjang@kookmin.ac.kr (Yeong Min Jang)

non-parametric dynamic thresholding for prediction-based anomaly detection and utilizing fine-tuned GPT-4 for LLM to generate explanations, the model enhances interpretability, facilitates quick decision-making, and ensures adaptability. This innovative method seeks to offer a comprehensive solution for robust anomaly detection in complex IoT systems, particularly in contexts like IAQ monitoring, where accuracy and actionable insights are paramount.

2. Literature Review

Many researchers now use deep learning architecture for anomaly detection in time series due to its powerful data dependency modeling. There are two commonly used deep learning techniques: prediction-based approaches, and reconstruction-based approaches. Several prediction-based models have been developed, including the LSTM-AD that has excellent memory capabilities, and single-handedly performs anomaly detection without requiring any prior knowledge of temporal length. DeepLSTM, developed by Chauhan et al., also employs a LSTM recurrent network, offering benefits such as direct application to raw time series data, eliminating preprocessing steps. Hundmand et al.'s LSTM-NDT has superior predictive performance by integrating dynamic thresholding and anomaly scoring. Lastly, TCN-AE and GANF are used for prediction-based anomaly detection and spatial feature correlations respectively.

Reconstruction-based approaches, on the other hand, such as Autoencoders (AE), LSTM-AE, and LSTM Variational Autoencoders (LSTM-VAE), have also been acknowledged. Unlike prediction-based anomaly detection, these approaches involve recreating the original input through a latent low-dimension representation, ideal for capturing collective anomalies.

However, anomaly detection's integration with LLM in Internet of Things (IoT) and air quality monitoring is yet to be observed. Existing methods either rely heavily on prediction or reconstruction-based anomaly detection without providing comprehensive explanations. To fill this gap, the study introduces a hybrid approach combining prediction-based anomaly detection with LLMs within IoT systems' air quality monitoring context. The contributions of our proposed method are as follows:

1. The hybrid approach integrating ConvLSTM-DT and Fine-tuned GPT-4 that effectively generates anomaly reasoning and possible solutions.
2. The outlined method deploys the model in edge-cloud environments, accounting for seamless real-world edge-computing integrations.
3. To assess the anomaly detection performance, a comparative evaluation uses machine learning anomaly detection algorithms, including Isolation Forest (iForest), One-Class Support Vector Machine (OC-SVM), Histogram-Based Outlier Score (HBOS), and Local Outlier Factor (LOF), as benchmarks for the hybrid approach.

3. Proposed Method

3.1. Data Collection and Data Preprocessing

Jelasin data collection, data cleaning or missing data, feature selection, normalization

3.2. ConvLSTM-PBNN

Bang Krish Mohon bantuannya

3.3. Dynamic Thresholding

To effectively monitor thousands of telemetry channels influenced by changing environmental conditions and command sequences, there's a critical need for a fast and unsupervised approach to detect anomalies in predicted values. While common methods like distance-based techniques are available, they often come with high computational costs, involving comparisons between each data point and a set of k neighbors. Another frequently employed method relies on the assumption of Gaussian distributions for past smoothed errors, which enables swift comparisons with concise representations of previous errors. However, this approach can face challenges when parametric assumptions are violated.

To address this issue, non-parametric dynamic thresholds proposed by Hundmand et al. play a crucial role. These non-parametric dynamic thresholds can be divided into three key processes: error and smoothing, threshold calculation, and anomaly scoring.Therefore, in this architecture we implemented this method as the anomaly detector.

3.4. GPT-4 Integration

GPT-4 is a useful large-scale model, but its general-purpose design may not fully cater to specific tasks, such as time-series anomaly detection. To enhance its specificity, GPT-4 can be fine-tuned with a custom dataset, improving accuracy and reasoning. This also allows for rapid deployment for tasks requiring immediate responses. Integrating Language Model (LLM) with

anomaly detection involves examining data from each sensor using the ConvLSTM-DT algorithm. Detected anomalies trigger a simplified explanation, which is formatted into a prompt, sent to the cloud, and processed by the fine-tuned GPT-4. This produces understandable feedback and possible solutions for anomalies, providing operators with critical insights. Anomaly detection is executed at the edge for quick responses, while more computationally intensive LLM tasks are run in the cloud. This division optimizes efficiency and system performance.

4. Result and Discussion

The main document with the manuscript text, figures, and tables should be prepared in an MS World or LaTeX format in English. The manuscript should be written in 10-point font with single line spacing on A4 sized (21.0 x 29.7 cm) paper with 1.3 cm margins on the top, bottom, right, and left. The standard order of sections in the manuscript is title page, abstract, introduction, system model and methods, results, discussion, references. Figures and tables with legends should be embedded in the proper place in the text. The number of all manuscript pages should be specified, starting with the title page as page 1. A single file is permitted for initial submission, but figures and tables can be uploaded separately.

5. Conclusion

In conclusion, this research introduces a novel hybrid approach that combines prediction-based anomaly detection using Convolutional Long Short-term Memory with non-parametric dynamic thresholding (ConvLSTM-DT) and the fine-tuned GPT-4 model for Large Language Model (LLM)-based explanation generation. This hybrid approach significantly improves the interpretability and adaptability of anomaly detection models. By integrating the strengths of both prediction-based algorithms and LLMs, the model not only detects anomalies but also provides human-understandable explanations and potential solutions, making it a powerful tool for real-world applications where quick decision-making is crucial.

Furthermore, the implementation of this approach on an edge-cloud architecture enhances its performance by reducing latency and ensuring fast detection. The use of custom fine-tuning with GPT-4 tailored to the specific anomaly detection task further amplifies its accuracy. This study emphasizes the importance of domain-specific fine-tuning in achieving precise and meaningful explanations. As the field of anomaly detection continues to evolve, the presented hybrid approach, along with the use of advanced LLMs, promises to play a pivotal role in improving anomaly detection across various domains, ultimately leading to safer and more efficient operations in complex systems.

For the future work, there is significant potential for enhancing anomaly detection by incorporating commonsense reasoning trees. This addition could further bolster the accuracy and depth of anomaly reasoning and provide a more comprehensive understanding of complex situations. Moreover, the integration of multimodal sensor fusion, which combines time-series data with visual and auditory inputs, could lead to more holistic anomaly detection solutions capable of capturing anomalies that span various data modalities. Additionally, exploring the integration of reliable control systems to facilitate intelligent decision-making services in real-time scenarios offers an exciting avenue for improving anomaly detection systems, ultimately advancing their capabilities and impact across diverse domains.

Acknowledgments

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IIITP-2023-2018- 0-01396) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00590, Industrial small cell system supporting 5G multi-band).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

5.1. Title page

The Title page should include a full title, running title (no more than 40 characters in length) of the article and authors' information.

Title. should be as concise as possible but informative enough to facilitate information retrieval.

Table 1
An example of a table.

Column heading	Column A	Column B
And an entry	1	2
And another entry	3	4
And another entry	5	6
And another entry	7	8

Author names and affiliations. Please clearly indicate the given name(s) and family name(s) of each author and check that all names are accurately spelled. Present the authors' affiliation addresses (where the actual work was done) below the names. Indicate all affiliations with a lower-case superscript letter immediately after the author's name and in front of the appropriate address. Provide the full postal address of each affiliation, including the country name and, if available, the e-mail address of each author.

Corresponding author. Clearly indicate who will handle correspondence at all stages of refereeing and publication, also post-publication. This responsibility includes answering any future queries. Ensure that the e-mail address is given and that contact details are kept up to date by the corresponding author.

Present/permanent address. If an author has moved since the work described in the article was done, or was visiting at the time, a 'Present address' (or 'Permanent address') may be indicated as a footnote to that author's name. The address at which the author actually did the work must be retained as the main affiliation address. Superscript Arabic numerals are used for such footnotes.

5.2. Text

The text is recommended to be arranged in this order, if possible: Introduction, System Model and Methods, Result, Discussion, and Conclusions.

Introduction. the purpose and the background should be written simply and lucidly.

System Model and Methods. the methodology should be written precisely so that others may use some or all of the methods in another study or judge the scientific merit of your work.

Result. a detailed description of the study results should be objectively presented, in an orderly and logical sequence using both text and illustrative materials (Tables and Figures).

Figure 1. Example of a figure.

Discussion and Conclusions. author's interpretation of the results, author's opinion. Conclusion should be written simply.

5.3. Text section heading

Divide your article into clearly defined and numbered sections. Subsections should be numbered as 1.1 (then 1.1.1, 1.1.2, ...), 1.2, etc. (the abstract is not included in the section numbering). Use this numbering also for internal cross-referencing. Any subsection may be given a brief heading. Each heading should appear on its own separate line.

5.4. References in text

References should be obviously related to documents. Indicate references by number(s) in square brackets in line with the text. The numbered references are listed in the order in which they appear in the text. The actual authors can be referred to, but the reference number(s) must always be given. Example: '..... as demonstrated [3,6]. Barnaby and Jones [8] obtained a different result ....'

5.5. Text equations

The Equations should be punctuated and aligned to bring out their structure and numbered on the right as

$$y = Hx + n. \tag{1}$$

Mathematical operation signs indicating continuity of the expression should be placed at the left of the second and succeeding lines. Use $\times$ rather than a centered dot,

except for scalar products of vectors. The solidus (/) should be used instead of built-up fractions in running text. Furthermore, the Notation must be legible, clear, compact, and consistent with standard usage. All unusual symbols whose identity may not be obvious must be identified when they first appear, and whenever confusion might arise. Superscripts are normally set directly over subscripts; authors should note where readability or the meaning requires a special order. In the text, numbers should be Arabic numerals, except when beginning a sentence. Numbers greater than 999 should have commas, e.g., 13,970. The 24-hour system is used to indicate time, e.g., 18:00 hr. If you are using Word for Math, use either the Microsoft Equation Editor or the MathType add-on (http://www.mathtype.com) for equations in your paper. "Float over text" should not be selected.

5.6. Units and abbreviations

Units of measure should be presented according to the International System (SI) of Units. If other units are mentioned, please give their equivalent in SI.

Abbreviations must be used as an aid to the reader, rather than as a convenience of the author, and therefore their use should be limited. Acronyms and abbreviations should be defined at the first time when they are used in text.

5.7. Tables

Each Table should be numbered with Arabic numerals in the order of their appearance in the text. Tables should have a concise and informative title with the table content between horizontal lines. The structure should be clear, with simple column headings giving all units. A table should not exceed one page when printed. Use lower case letters in superscripts a, b, c ... for special remarks. An example of a table is given in Table 1.

5.8. Figures

Figures are numbered consecutively in the sequence mentioned in the text and must have a caption written in one paragraph style. The caption should contain an explanation of all abbreviations and symbols used, and indicate the size value of lines or bars unless shown directly on the figure. An example of a figure is given in Fig. 1. The Figure number should be placed at the lower-left corner of each figure, and the numbering order must be from left to right, and from upper to lower. Citations of figures in the text or parentheses are abbreviated, e.g., Fig. 1, Figs. 1 and 2, Figs. 1-3, (Fig. 1), (Figs. 1 and 2), (Figs. 1-3). When the text refers to both figures and tables, they should be mentioned in parentheses, e.g., (Table 1; Fig. 2) and (Tables 1-3; Figs. 4-6).

5.9. Footnotes

Footnotes should be used sparingly. Number them consecutively throughout the article. Many word processors can build footnotes into the text, and this feature may be used. Otherwise, please indicate the position of footnotes in the text and list the footnotes themselves separately at the end of the article. Do not include footnotes in the Reference list.

Appendix

Appendix section should be placed before References section. If there is more than one appendix, they should be identified as A, B, etc. Formulae and equations in appendices should be given with separate numbering: Eq. (A.1), Eq. (A.2), etc.; in a subsequent appendix, Eq. (B.1) and so on. The numbering of tables and figures in the appendix is similar: Table A.1; Fig. A.1, etc.

References

[1] H. Kwon, K. Kim, C. Lee, The unified UE baseband modem hardware platform architecture for 3GPP specification, J. Commun. Net., 13 (1) (2011) 70-76.

[2] T. Roos, P. Myllymaki, H. Tirri, A statistical modeling approach to location estimation, IEEE Trans. Mobile Comput. 1 (1) (2002) 59-69.

[3] H. Liu, G. Li, OFDM-Based Broadband Wireless Networks: Design and Optimization. Hoboken, NJ: Wiley-Interscience, 2005.

[4] T. L. Marzetta, How much training is required for multiuser MIMO?, in: 2006 Asilomar Conf. Signal. Syst. Comput., Pacific Grove, 2006, pp. 359–363.

[5] J. Arrillaga, B. Giessner, Limitation of short-circuit levels by means of HVDC links, in: 1990 IEEE Summer Power Meeting, Los Angeles, 1990, pp. 1-8.

[6] J. O. Williams, Narrow-band analyzer, Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, 1993.

[7] J. H. Davis, J. R. Cogdell, Calibration program for the 16-foot antenna, Elect. Eng. Res. Lab., Univ. Texas, Austin, Tech. Memo. NGL-006-69-3, Nov. 15, 1987.

[8] RealVNC Ltd. Remote control software [Online]. Available: http:// www.realvnc.com