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

In recent years, the rapid increase in data-driven applications and the growing complexity of modern systems have highlighted the critical significance of anomaly detection. Anomalies, often indicative of potentially harmful behavior within systems, can have long-term effects and wide-ranging consequences in various domains, such as industrial operations. Traditional anomaly detection methods, while effective to some extent, often heavily rely on expert analysis, making them resource-intensive and susceptible to human bias. This dependency on expert input can result in scalability, reliability, and consistency challenges within anomaly detection systems[1].

Furthermore, traditional anomaly detection methods based on statistical and machine learning approaches exhibit limitations, particularly when handling time-series data from multiple sensors. Anomalies in this context may be multi-dimensional and not easily characterized by conventional patterns. Consequently, traditional approaches can produce a significant number of false alarms, compromising the effectiveness and trustworthiness of anomaly detection systems[2].

In response to these challenges, this study aims to contribute to the field of anomaly detection by proposing a hybrid approach that harnesses the strength of prediction-based detector algorithms and the capabilities of Large Language Models (LLMs). This work centers on monitoring and detecting anomalies in air quality data collected from various sensors, including temperature, humidity, PM (Particulate Matter), and CO2 levels, all situated within a single indoor room environment. The ability to predict and detect anomalies in this air quality data is of paramount importance, given its profound implications for human health, particularly for

*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)

those occupying the room.

The proposed hybrid approach combines Convolutional Long Short-Term Memory with non-parametric dynamic thresholding (ConvLSTM-DT) for prediction-based anomaly detection and utilizes fine-tuned GPT-4 for LLM-based explainability. This hybrid approach is essential to address the aforementioned limitations and enhance the interpretability and adaptability of anomaly detection models. It not only facilitates anomaly detection across a variety of sensors but also provides human-understandable explanations and potential solutions for addressing anomaly situations. This makes it a powerful tool for real-world implementation where swift decision-making is imperative.

2. Literature Review

Multiple techniques can be employed for anomaly detection on time series and currently, a significant number of researchers are exploring the utilization of various deep learning architectures for anomaly detection. This is due to the inherent power of deep learning methods in modeling data dependencies, particularly in data with complex structures. In deep learning techniques, there are two different approaches that are commonly used for anomaly detection, prediction-based approaches and reconstruction-based approaches.

Malhotra et al [5] introduced LSTM-AD, a model renowned for its long-term memory capabilities. A noteworthy innovation is the combination of hierarchical recurrent processing layers, a groundbreaking approach for detecting anomalies within univariate time series data. A captivating aspect of LSTM-AD is the first deep learning model that has the ability to perform anomaly detection without requiring labeled training data or unsupervised learning. By stacking recurrent hidden layers, this architecture empowers the model to learn higher-order temporal trends without any prior knowledge of their temporal length.

Similarly, in the realm of anomaly detection, Chauhan et al [6] proposed DeepLSTM, a model that leverages a stacked LSTM recurrent network. This approach begins by training on normal data and subsequently employs maximum likelihood estimation to predict the error vector based on multivariate Gaussian distributions. Following this, the model is used to forecast a combination of both abnormal and normal validation data, while concurrently recording the probability density function (PDF) values associated with the error. Notably, this method offers the unique benefit of direct application to raw time series data, bypassing the need for preprocessing steps.

On another front, Hundmand et al. [7] embarked on the implementation of a non-parametric, dynamic, and unsupervised thresholding technique to assess errors for anomaly detection. LSTM-NDT combines multiple techniques, including LSTM, to achieve superior predictive performance. A key feature of this approach is the integration of dynamic thresholding and anomaly scoring, both of which can be automatically adjusted to accommodate the diverse, unstable, and noisy characteristics often encountered in dynamic data.

Thill et al [8] introduced the Temporal Convolution Network coupled with an autoencoder framework (TCN-AE), for prediction-based anomaly detection. In contrast to a vanilla autoencoder, TCN-AE replaces the dense layer architecture with a more potent and flexible CNN architecture, making it adaptable to varying input sizes. This architecture incorporates two temporal convolutional networks (TCNs)[9] for both encoding and decoding processes. Concurrently, Dai et al [10] developed the Graph-Augmented Normalizing Flow (GANF) as an anomaly detection framework, grounded in graph neural networks (GNNs), to harness spatial feature correlations. Normalizing flow functions as a profound generative model in unsupervised learning, enabling the investigation of the inherent data distribution and addressing the issue of limited labels. GANF is cast as a Bayesian model, offering an estimate of the density for each data instance. This estimation aligns with the hypothesis that anomalies are more likely to be found in low-density regions.

In the case of reconstruction-based approaches, models such as Autoencoders (AE), LSTM-AE, and LSTM Variational Autoencoders (LSTM-VAE) have gained recognition. In contrast to prediction-based anomaly detection, which typically relies on identifying anomalies through deviations between actual and predicted values, this approach involves learning a latent low-dimensional representation to reconstruct the original input. Consequently, reconstruction-based approaches are often better suited for capturing contextual and collective anomalies[11]. While there has been extensive research on time series anomaly detection, the author has identified only one study that incorporated anomaly detection through explanation (ADE)[12]. In this context, Gilpin[12] employed reasonableness monitors based on anomaly detection within the context of simulated semi-autonomous vehicles. These simulations were designed to replicate real-world, anomalous driving scenarios. Furthermore, Gilpin utilized argument trees to generate reasoning and explanations for the detected anomalies. Gilpin[12] work represents a pioneering effort that has introduced the emerging field of explanatory anomaly detection, particularly in the realm of system-level explanations.

In the domain of Internet of Things (IoT) and air quality monitoring, the author could not find any study that integrates both anomaly detection and LLMs. Most existing approaches primarily rely on either prediction-based or reconstruction-based anomaly detection methods to identify anomalous behavior, without offering comprehensive explanations or reasoning. This study addresses this gap by introducing a hybrid approach that combines prediction-based anomaly detection with LLMs to provide a robust framework for constructing explanations and reasoning in the context of air quality monitoring within IoT systems.

3. Proposed Method

3.1. Data Collection and Data Preprocessing

3.2. ConvLSTM-PBNN

3.3. Dynamic Thresholding

3.4. GPT-4 Integration

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.

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

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.

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.

Figure 1. Example of a figure.

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).

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 × 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