Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 8 Nomor 1, Maret 2024, pp. 263-270
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

Comparative Analysis of Machine Learning Models for Enhanced Chemical Detection in Sensor Array Data

Gregorius Airlangga
Universitas Katolik Indonesia Atma Jaya, Jakarta, Indonesia
e-mail: gregorius.airlangga@atmajaya.ac.id

Abstract

The objective of this study was to compare the efficacy of various machine learning models for classifying chemical substances using sensor array data from a wind tunnel facility. Six widely recognized machine learning algorithms were assessed: Random Forest, Gradient Boosting, Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN). The dataset, consisting of 288 sensor array features, was preprocessed and utilized to evaluate the models based on accuracy, precision, recall, and F1 score through a 5-fold cross-validation method. The results indicated that ensemble methods, particularly Random Forest and Gradient Boosting, outperformed other models, achieving an accuracy and F1 score of over 99%. KNN also demonstrated high efficacy with similar performance metrics. In contrast, Logistic Regression showed modest results in comparison. The study's outcomes suggest that ensemble machine learning models are highly suitable for chemical detection tasks, potentially contributing to advancements in environmental monitoring and public safety. The findings also highlight the importance of quality data preprocessing in achieving optimal model performance. Future research directions include exploring hybrid models, deep learning techniques, and assessing model robustness against environmental variabilities. This research underscores the transformative potential of machine learning in chemical detection and paves the way for developing more sophisticated and reliable detection systems.

Keywords*: Chemical Detection, Machine Learning, Sensor Arrays, Ensemble Methods, Cross-Validation.*

Abstrak

Tujuan dari penelitian ini adalah untuk membandingkan efikasi berbagai model pembelajaran mesin dalam mengklasifikasikan zat kimia menggunakan data array sensor dari fasilitas terowongan angin. Enam algoritma pembelajaran mesin yang diakui luas dinilai: Random Forest, Gradient Boosting, Regresi Logistik, Mesin Vektor Pendukung (SVM), Pohon Keputusan, dan Tetangga Terdekat K (KNN). Dataset yang terdiri dari 288 fitur array sensor diproses dan digunakan untuk mengevaluasi model berdasarkan akurasi, presisi, recall, dan skor F1 melalui metode validasi silang 5-kali lipat. Hasil penelitian menunjukkan bahwa metode ensemble, khususnya Random Forest dan Gradient Boosting, mengungguli model lainnya, mencapai akurasi dan skor F1 di atas 99%. KNN juga menunjukkan efikasi tinggi dengan metrik kinerja yang serupa. Sebaliknya, Regresi Logistik menunjukkan hasil yang sederhana dibandingkan. Hasil penelitian menyarankan bahwa model pembelajaran mesin ensemble sangat cocok untuk tugas deteksi kimia, berpotensi berkontribusi pada kemajuan dalam pemantauan lingkungan dan keselamatan publik. Temuan ini juga menyoroti pentingnya pra-pemrosesan data berkualitas dalam mencapai kinerja model optimal. Arah penelitian masa depan termasuk mengeksplorasi model hibrida, teknik pembelajaran mendalam, dan menilai ketahanan model terhadap variabilitas lingkungan. Penelitian ini menekankan potensi transformatif pembelajaran mesin dalam deteksi kimia dan membuka jalan untuk mengembangkan sistem deteksi yang lebih canggih dan dapat diandalkan.

Kata kunci*: Deteksi Kimia, Pembelajaran Mesin, Array Sensor, Metode Ensemble, Validasi Silang.*

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 8 Nomor 1, Maret 2024, pp. 263-270
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

1. INTRODUCTION

Chemical detection plays a pivotal role in a variety of critical applications ranging from environmental monitoring to public safety and industrial process control [1]–[3]. With the advancement of sensor technologies, the deployment of sensor arrays capable of detecting chemical substances in diverse environments has significantly increased [4]–[6]. These arrays generate vast amounts of data, presenting both opportunities and challenges in chemical discrimination and analysis [7]–[9]. Historically, the analysis of chemical sensor data relied heavily on domain expertise and linear statistical models [10]–[12]. However, the complexity and variability of the data, coupled with the need for high accuracy and real-time processing, have pushed the boundaries beyond the capabilities of traditional approaches. In recent years, machine learning algorithms have emerged as powerful tools for handling complex pattern recognition tasks, leading to significant advancements in the field of chemical detection [13]–[15].

A comprehensive survey of the literature reveals a broad array of techniques applied to chemical sensor data analysis, including classical machine learning models like Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Trees, Random Forests, and advanced ensemble and deep learning approaches [16]–[18]. These studies highlight the potential of machine learning in improving detection accuracy, sensitivity, and specificity across various chemical sensing applications [19]. Despite these advancements, the rapidly evolving nature of chemical threats, coupled with the diverse and dynamic environments where detection systems are deployed, underscores a pressing urgency for more robust, adaptable, and efficient analysis methods [20]. This urgency is further amplified by the critical role that timely and accurate chemical detection plays in safeguarding public health and safety [21].

The state of the art in chemical detection is characterized by a continuous push towards higher accuracy, lower false alarm rates, and the ability to operate in real-time under varying environmental conditions [22]. Machine learning models, particularly those incorporating ensemble and deep learning techniques, represent the forefront of research in achieving these goals [23]. However, the performance of these models is heavily contingent upon the quality and representativeness of the dataset used for training and validation [24]. The present study is motivated by the observation that while there is a substantial body of research on chemical detection using machine learning, there is a noticeable gap in the literature regarding the systematic comparison of different machine learning models on standardized datasets, especially in open-sampling settings [25]–[27]. This gap is critical because it limits our understanding of the relative strengths and weaknesses of various approaches under consistent experimental conditions.

This research aims to bridge this gap by conducting a comprehensive evaluation of several machine learning models, including Random Forest, Gradient Boosting, Logistic Regression, SVM, Decision Tree, and KNN, on a preprocessed dataset derived from sensor arrays deployed in a wind tunnel facility. The dataset captures responses to ten high-priority chemical gaseous substances, presenting a ten-class gas discrimination problem. Our goal is to identify the most effective

models for chemical detection in terms of accuracy, precision, recall, and F1 score, under the conditions represented by the dataset.

In contributing to the body of knowledge, this research provides several key insights. First, it offers a systematic comparison of machine learning models applied to chemical detection, highlighting the best-performing approaches in this specific context. Second, it explores the implications of dataset characteristics, such as feature diversity and class balance, on model performance. Lastly, it discusses the practical considerations for deploying these models in real-world chemical detection systems, including computational efficiency and adaptability to new chemical threats. The remainder of this journal article is organized as follows. Section 2 delves into the methodology, detailing the dataset, preprocessing steps, model selection, and evaluation criteria. Section 3 presents the results of the model comparison, followed by a discussion in Section 4 that interprets the findings in the context of existing research and practical applications. Section 5 outlines the limitations of the current study and suggests directions for future research. Finally, Section 6 concludes the paper by summarizing the key contributions and their implications for the field of chemical detection.

2. RESEARH METHODS

In the pursuit of advancing chemical detection methodologies through machine learning, our research delineates a comprehensive approach encompassing data collection, preprocessing, model selection, evaluation criteria, and experimental procedures. This section explicates the methodological framework employed in our study, ensuring reproducibility and clarity in the exploration of machine learning models applied to chemical detection.

2.1. Dataset Description

The foundation of our research is a meticulously preprocessed dataset derived from the original recordings gathered at the BioCircuits Institute, University of California, San Diego [28]. This dataset, obtained from a sophisticated chemical detection platform situated within a wind tunnel facility, encapsulates the nuanced responses of sensor arrays to a spectrum of ten high-priority chemical gaseous substances. The diversity and complexity inherent in these responses engender a challenging ten-class gas discrimination problem. Encompassing 288 sensor array features labeled from A1 to I8, the dataset provides a rich tapestry of information. It includes two subsets: one with 17921 entries spanning 11 chemicals and another more focused subset comprising 5098 entries associated with 3 chemicals. The data compilation, stretching over 16 months from December 2010 to April 2012, encapsulates a wide array of conditions and chemical exposures, thereby enhancing the robustness and relevance of our analysis.

2.2. Preprocessing Techniques

To prepare the dataset for the analytical rigors of machine learning, we embarked on a preprocessing journey characterized by meticulous data refinement techniques. This phase was pivotal in transforming the raw data into a

structured format amenable to machine learning algorithms. By retaining all 288 sensor array features, we preserved the integrity and fullness of the sensor responses. The categorical chemical names underwent a transformation into numerical values through label encoding, thus facilitating their interpretation by the subsequent models. Normalization of feature values ensured a uniform scale, eliminating any undue influence of disproportionately scaled data. Additionally, a thorough inspection for missing values was conducted, with any identified gaps being filled by the median value of the respective feature, thereby maintaining the continuity and completeness of our dataset.

2.3. Selection of Models

Our exploratory journey through the landscape of machine learning models is guided by the selection of six distinguished models, each representing a unique approach to classification. This eclectic mix includes ensemble methods like Random Forest and Gradient Boosting, known for their robustness and precision, alongside traditional stalwarts such as Logistic Regression and Support Vector Machine (SVM). The simplicity and interpretability of Decision Trees are juxtaposed with the flexibility and locality of K-Nearest Neighbors (KNN), creating a comprehensive palette of models to evaluate. This selection reflects our commitment to exploring a broad spectrum of machine learning methodologies to identify the most efficacious model for chemical detection.

2.4. Evaluation Metrics and Experimental Setup

The crucible of our evaluation process is a meticulously designed experimental setup leveraging k-Fold Cross-Validation. Opting for a 5-fold configuration, this technique ensures an equitable distribution of data across training and testing phases, thereby enhancing the reliability and generalizability of our findings. Each model's performance was scrutinized through a prism of metrics—Accuracy, Precision, Recall, and F1 Score—each providing a unique lens through which the efficacy of the models could be assessed. Accuracy offered a holistic view of model performance, while Precision and Recall illuminated the models' strengths in identifying true positives. The F1 Score, with its harmonious balance between Precision and Recall, served as a critical measure of model robustness.

Our experimental procedures were underpinned by a commitment to fairness and rigor, ensuring that each model was evaluated under identical conditions. The utilization of Python, along with its rich ecosystem of libraries such as Pandas, NumPy, and Scikit-Learn, facilitated a seamless execution of data preprocessing, model training, and performance evaluation. This methodological rigor extends an invitation to peers and practitioners alike, enabling them to replicate our study and contribute to the ongoing dialogue in the realm of chemical detection through machine learning.

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 8 Nomor 1, Maret 2024, pp. 263-270
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

3. RESULT AND DISCUSSION

The evaluation of the selected machine learning models on the chemical detection dataset revealed high performance across all metrics, underscoring the effectiveness of these models in discriminating among different chemical gaseous substances. As shown in the accompanying table, the Random Forest classifier achieved the highest scores in Accuracy, Precision, Recall, and F1 Score, each metric yielding a value of 99.66%. Close behind was the Gradient Boosting model, with a consistent performance across the metrics, nearly matching the Random Forest with scores just above 99.45%. K-Nearest Neighbors (KNN) also performed remarkably well, with scores above 99.23% in all metrics. The Support Vector Machine (SVM) and Decision Tree models displayed robustness with Accuracy, Precision, Recall, and F1 Scores all above 97%. The Logistic Regression model, while still performing admirably, lagged slightly behind the other models with scores approximately ranging from 93.97% to 94.57%.

Tabel 1. Models Comparison

Methods	Accuracy	Precision	Recall	F1 Score
Random Forest	99.66	99.66	99.66	99.66
Gradient Boosting	99.45	99.46	99.45	99.45
Logistic Regression	93.97	94.57	93.97	94.06
SVM	98	98.11	98	98.01
Decision Tree	97.19	97.20	97.19	97.19
K-Nearest Neighbors (KNN)	99.23	99.24	99.23	99.23

The exemplary performance of the Random Forest model can be attributed to its ensemble nature, where a multitude of decision trees work in concert to improve the overall prediction accuracy. This ensemble approach effectively mitigates the risk of overfitting, which is often a concern with individual decision trees. The Gradient Boosting model, leveraging the power of building one tree at a time and learning from the mistakes of previous trees, exhibits a similar level of accuracy. This suggests that ensemble methods are particularly well-suited to the task of chemical detection, possibly due to their ability to capture complex patterns and relationships within the sensor data. The K-Nearest Neighbors algorithm demonstrated high scores that were competitive with the ensemble methods. Its success is likely due to the dataset's well-defined feature space, where chemicals produce distinct and recognizable patterns that KNN can effectively use to classify observations based on their proximity to known instances.

While the Support Vector Machine and Decision Tree models yielded slightly lower performance metrics compared to the ensemble methods and KNN, their results were still impressive, affirming their viability as potential candidates for chemical detection tasks. The SVM, with its disciplined approach to finding the optimal separating hyperplane, has shown that it can handle the high-dimensional data typical of chemical sensor arrays effectively. The Logistic Regression model, typically robust in binary classification problems, appeared less adept in handling this multi-class chemical discrimination task, as indicated by its relatively lower scores. This could be due to the linear nature of Logistic Regression, which might

struggle with the complex and nonlinear relationships in the multi-class setting of the chemical detection data.

It is noteworthy that the high performance across all models could also reflect the quality of the dataset, which was carefully preprocessed and represents a broad range of conditions and exposures. This suggests that the success of machine learning models in chemical detection tasks is highly contingent on the availability of comprehensive and well-curated datasets. The implications of these findings are substantial for the field of chemical detection. The ability of these models to accurately classify chemical substances based on sensor array data could be utilized in developing real-time detection systems, potentially offering significant benefits in environmental monitoring, industrial safety, and homeland security. Moreover, the high performance of ensemble models could be leveraged to develop more advanced systems capable of adapting to new and emerging chemical threats. In light of these results, future research might explore the integration of these models into sensor hardware, the impact of feature engineering to further improve model performance, and the use of deep learning approaches that could offer additional gains in detection accuracy. Furthermore, studies could investigate model performance in on-field conditions where sensor drift and more variable environmental factors come into play.

4. CONCLUSION

This research has provided a comprehensive analysis of various machine learning models for the task of chemical detection using sensor array data. Through a methodical evaluation, our study highlighted the strengths of ensemble models, particularly Random Forest and Gradient Boosting, which demonstrated superior performance in terms of accuracy, precision, recall, and F1 score. The K-Nearest Neighbors (KNN) algorithm also showcased a remarkably high level of performance, asserting its place as a viable and efficient alternative for real-time applications. The findings of this study underscore the critical role that data quality plays in the performance of machine learning models. The extensive preprocessing and careful curation of the dataset were integral to achieving the high levels of model accuracy observed in the research. These results affirm the potential of machine learning approaches in enhancing chemical detection systems, which can be vital for applications in environmental monitoring, industrial safety, and homeland security. The success of the ensemble methods, in particular, suggests that leveraging the collective decision-making of multiple models or algorithms can provide a more nuanced understanding of complex sensor data, leading to more accurate and reliable detection systems. The lower performance of the Logistic Regression model, conversely, indicates that simpler models, while still useful, may be less suitable for complex multi-class discrimination tasks such as those presented by chemical sensor arrays. Our study contributes to the existing body of knowledge by providing a direct comparison of several machine learning models on a standardized dataset in the context of chemical detection. The insights gained from this comparison can guide the selection of models in the development of practical chemical detection solutions. Future work may extend beyond the

scope of this study to explore hybrid models that combine the strengths of the best-performing algorithms, the application of deep learning techniques that could further improve accuracy, and the investigation of model robustness in variable field conditions. Additionally, further research is encouraged to address the challenges of sensor drift and the detection of low-concentration chemical substances in complex backgrounds.

DAFTAR PUSTAKA

[1] J. Chapman *et al.*, "Combining chemometrics and sensors: Toward new applications in monitoring and environmental analysis," *Chem. Rev.*, vol. 120, no. 13, pp. 6048–6069, 2020.

[2] D. Tyagi *et al.*, "Recent advances in two-dimensional-material-based sensing technology toward health and environmental monitoring applications," *Nanoscale*, vol. 12, no. 6, pp. 3535–3559, 2020.

[3] K. C. To, S. Ben-Jaber, and I. P. Parkin, "Recent developments in the field of explosive trace detection," *ACS Nano*, vol. 14, no. 9, pp. 10804–10833, 2020.

[4] M. A. Al Mamun and M. R. Yuce, "Recent progress in nanomaterial enabled chemical sensors for wearable environmental monitoring applications," *Adv. Funct. Mater.*, vol. 30, no. 51, p. 2005703, 2020.

[5] F. Wen, T. He, H. Liu, H.-Y. Chen, T. Zhang, and C. Lee, "Advances in chemical sensing technology for enabling the next-generation self-sustainable integrated wearable system in the IoT era," *Nano Energy*, vol. 78, p. 105155, 2020.

[6] I. Yaroshenko *et al.*, "Real-time water quality monitoring with chemical sensors," *Sensors*, vol. 20, no. 12, p. 3432, 2020.

[7] N. Bhalla, Y. Pan, Z. Yang, and A. F. Payam, "Opportunities and challenges for biosensors and nanoscale analytical tools for pandemics: COVID-19," *ACS Nano*, vol. 14, no. 7, pp. 7783–7807, 2020.

[8] H.-P. Wang *et al.*, "Recent advances of chemometric calibration methods in modern spectroscopy: Algorithms, strategy, and related issues," *TrAC Trends Anal. Chem.*, vol. 153, p. 116648, 2022.

[9] A. Venketeswaran *et al.*, "Recent advances in machine learning for fiber optic sensor applications," *Adv. Intell. Syst.*, vol. 4, no. 1, p. 2100067, 2022.

[10] W. Liu and T. Yairi, "A unifying view of multivariate state space models for soft sensors in industrial processes," *IEEE Access*, 2023.

[11] S. Rangineni, D. Marupaka, and A. K. Bhardwaj, "An examination of machine learning in the process of data integration," *Int. J. Comput. Trends Technol.*, vol. 71, no. 6, pp. 79–85, 2023.

[12] W. Huang, T. Li, J. Liu, P. Xie, S. Du, and F. Teng, "An overview of air quality analysis by big data techniques: Monitoring, forecasting, and traceability," *Inf. Fusion*, vol. 75, pp. 28–40, 2021.

[13] S. Zhong *et al.*, "Machine learning: new ideas and tools in environmental science and engineering," *Environ. Sci. \& Technol.*, vol. 55, no. 19, pp. 12741–12754, 2021.

[14] P. Carracedo-Reboredo *et al.*, "A review on machine learning approaches and trends in drug discovery," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 4538–4558, 2021.

[15] N. Ha, K. Xu, G. Ren, A. Mitchell, and J. Z. Ou, "Machine learning-enabled smart sensor systems," *Adv. Intell. Syst.*, vol. 2, no. 9, p. 2000063, 2020.

[16] E. A. Bamidele *et al.*, "Discovery and prediction capabilities in metal-based nanomaterials: An overview of the application of machine learning techniques and

some recent advances," *Adv. Eng. Informatics*, vol. 52, p. 101593, 2022.
[17] M. Jaleel, A. Amira, and H. Malekmohamadi, "Classification of Gas Sensor Data Using Multiclass SVM," in *Science and Information Conference*, 2023, pp. 1333–1344.
[18] K. Dhibi, M. Mansouri, K. Bouzrara, H. Nounou, and M. Nounou, "An enhanced ensemble learning-based fault detection and diagnosis for grid-connected PV systems," *IEEE Access*, vol. 9, pp. 155622–155633, 2021.
[19] U. Yaqoob and M. I. Younis, "Chemical gas sensors: Recent developments, challenges, and the potential of machine learning—A review," *Sensors*, vol. 21, no. 8, p. 2877, 2021.
[20] A.-A. Bouramdane, "Natural hazards in electricity grids: from landscape dynamics to optimal mitigation and adaptation approaches," *Emerg. Manag. Sci. Technol.*, no. emst-0024-0003, pp. 1–20, 2024.
[21] Y. Li *et al.*, "Recent advances and prospects of persistent luminescent materials in public health applications," *Chem. Eng. J.*, p. 150424, 2024.
[22] D. Khorsandi *et al.*, "Mxene-based nano (bio) sensors for the detection of biomarkers: A move towards intelligent sensors," *Microchem. J.*, p. 109874, 2023.
[23] Y. Cao, T. A. Geddes, J. Y. H. Yang, and P. Yang, "Ensemble deep learning in bioinformatics," *Nat. Mach. Intell.*, vol. 2, no. 9, pp. 500–508, 2020.
[24] M. Mazni, A. R. Husain, M. I. Shapiai, I. S. Ibrahim, D. W. Anggara, and R. Zulkifli, "An investigation into real-time surface crack classification and measurement for structural health monitoring using transfer learning convolutional neural networks and Otsu method," *Alexandria Eng. J.*, vol. 92, pp. 310–320, 2024.
[25] D. P. Elpa, G. R. D. Prabhu, S.-P. Wu, K. S. Tay, and P. L. Urban, "Automation of mass spectrometric detection of analytes and related workflows: A review," *Talanta*, vol. 208, p. 120304, 2020.
[26] T. Liu *et al.*, "Review on algorithm design in electronic noses: Challenges, status, and trends," *Intell. Comput.*, vol. 2, p. 12, 2023.
[27] J. A. Covington, S. Marco, K. C. Persaud, S. S. Schiffman, and H. T. Nagle, "Artificial Olfaction in the 21 st Century," *IEEE Sens. J.*, vol. 21, no. 11, pp. 12969–12990, 2021.
[28] B. R. Mil, "Gas Classification Dataset." 2020.