OPEN ACCESS

Volume: 4

Issue: 2

Month: June

Year: 2025

ISSN: 2583-7117

Published: 26.06.2025

Citation:

Sambasiva Rao Madamanchi "Real-Time Anomaly Detection with AI-Driven Syslog Monitoring for Bioinformatics Reliability" International Journal of Innovations in Science Engineering and Management, vol. 4, no. 2, 2025, pp. 403–408.

DOI:

10.69968/ijisem.2025v4i2403-408

Real-Time Anomaly Detection with AI-Driven Syslog Monitoring for Bioinformatics Reliability

Sambasiva Rao Madamanchi[1]

[1] *Unix/Linux Administrator, Dept of Veterans Affairs (Austin, TX).*

Abstract

In this study, we present a novel framework that integrates AI-based anomaly detection with syslog monitoring to enhance the reliability of microbial pipelines. Syslog data, typically underutilized in bioinformatics, is collected and parsed to extract features such as error frequency, runtime patterns, and resource usage indicators. We apply machine learning models including Isolation Forests and LSTM Autoencoders to identify deviations from normal system behavior in real time. Experimental evaluations on both real and simulated microbial workflows demonstrate high accuracy in detecting anomalies, including those that do not trigger pipeline-level errors. A key use case illustrates how the system prevents corrupted outputs caused by unnoticed memory faults during taxonomic classification.

Keywords; AI, Sys Log, Detection, Bioinformatics.

INTRODUCTION

Microbial bioinformatics pipelines process vast and complex datasets to extract meaningful biological insights. These pipelines often comprise multiple tools linked together in workflows, such as Snakemake or Nextflow, and depend on high-performance computing (HPC) resources. However, due to their complexity, these pipelines are prone to issues such as software incompatibility, resource exhaustion, and silent computational errors that might not produce explicit failure messages. Traditional monitoring approaches based on manual inspection or limited quality control tools fall short in detecting subtle, non-crashing anomalies that may compromise data quality or reproducibility (Li et al., 2024).

This paper introduces a novel framework that leverages artificial intelligence (AI) to detect anomalies in microbial bioinformatics pipelines through syslog monitoring. Syslogs, which are native to Unix-based systems, capture system-level and application-specific messages. Despite their richness, syslogs are often underutilized in bioinformatics. By parsing and analyzing this data, it is possible to identify patterns indicative of pipeline inefficiencies or failures. Our contributions are threefold: (1) we present a system architecture that integrates syslog monitoring with microbial bioinformatics pipelines; (2) we apply AI-based anomaly detection methods, including unsupervised learning techniques, to identify issues in real time; and (3) we validate our approach through empirical testing on real-world and synthetic data. This work aims to enhance reliability, reproducibility, and operational awareness in microbial genomics research (Yayla 2024).

EXISTING TECHNIQUES IN LOG ANALYSIS AND BIOINFORMATICS QA

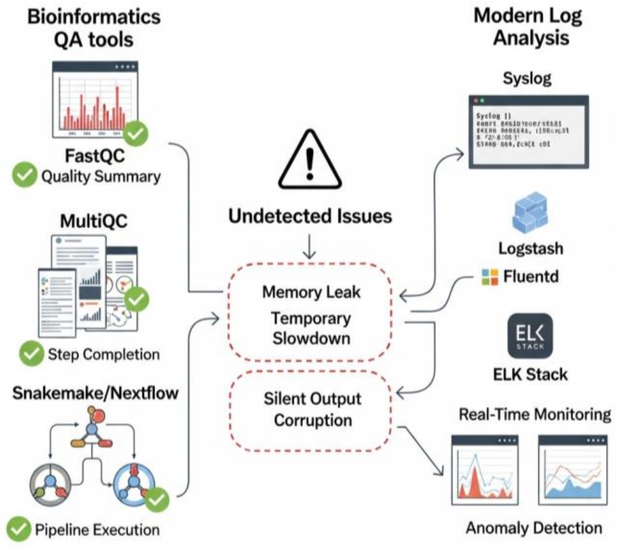

Figure 1 Existing Techniques in Log Analysis and Bioinformatics QA

Monitoring and quality control in bioinformatics workflows have traditionally relied on tools such as FastQC, MultiQC, and workflow managers like Snakemake and Nextflow. These solutions provide summaries and visualizations of read quality or step completion, but they typically lack real-time anomaly detection or detailed system-level insight. Failures or degradations that do not halt the workflow such as memory leaks, temporary slowdowns, or minor output corruptions often go unnoticed, making them difficult to trace and resolve (Ahmad et al., 2024).

Log analysis is a standard technique in software engineering and IT operations for performance monitoring and troubleshooting. Syslogs, generated by Unix-based systems, are widely used for auditing and debugging in enterprise settings. Tools such as Logstash, Fluentd, and the ELK stack (Elasticsearch, Logstash, Kibana) help analyze log streams in real time. However, such approaches have rarely been applied in the context of bioinformatics workflows, especially microbial pipelines (Rajapaksha et al., 2023).

Recent advances in AI have introduced robust methods for detecting anomalies in log data. Techniques range from statistical outlier detection to more advanced methods such as autoencoders, LSTM networks, and Isolation Forests. In cybersecurity and cloud monitoring, these models have shown significant promise. Our work bridges this gap by applying such AI techniques to syslog data in microbial bioinformatics contexts. By combining insights from log analytics and AI-driven detection, we address a crucial blind spot in current bioinformatics infrastructure (Dong et al., 2024).

SYSTEM ARCHITECTURE

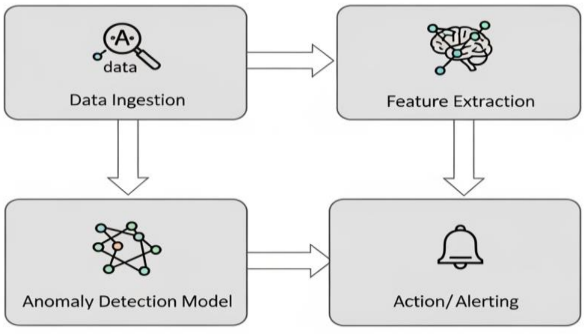

Figure 2 System Architecture

Our proposed framework integrates AI-based anomaly detection into microbial bioinformatics pipelines by leveraging syslog monitoring. The system architecture is designed to work with common workflow managers like Snakemake or Nextflow and can be deployed on HPC clusters or cloud environments. The architecture consists of four main layers: log collection, preprocessing, AI analysis, and alert generation (Pohl et al., 2024). In the log collection layer, syslogs from nodes running the pipeline are collected using tools such as rsyslog or Fluentd. These logs capture system messages, application output, and workflow events, which are then forwarded to a centralized repository. We standardize log formats and tag messages with metadata such as timestamps, pipeline step names, and node identifiers (Molder et al., 2021). The preprocessing layer parses, filters, and structures the syslog data. We remove irrelevant noise and extract features like event frequency, keyword patterns (e.g., "Segfault", "OOM", "Timeout"), and temporal sequences. These features are critical for training and applying anomaly detection models.

The AI analysis layer applies trained models to identify deviations from normal behavior. These models can be unsupervised (e.g., Isolation Forests) or sequence-based (e.g., LSTM Autoencoders) depending on the pipeline context. Finally, the alert generation layer reports detected anomalies via dashboards or notifications, allowing researchers or IT staff to intervene. This architecture ensures minimal intrusion into existing workflows while significantly enhancing operational visibility and robustness (Han et al., 2023).

AI-BASED ANOMALY DETECTION APPROACH

To identify anomalies in syslog data generated during microbial bioinformatics workflows, we implement a multi-stage AI-based detection process. This begins with feature engineering, where structured information is extracted from raw log messages. Features include temporal patterns (e.g., frequency of certain messages per time interval), keyword tagging, log severity levels, and inter-event timing. We also consider pipeline-specific features such as task duration, CPU usage, and memory consumption inferred from log contexts (Ndibe 2025).

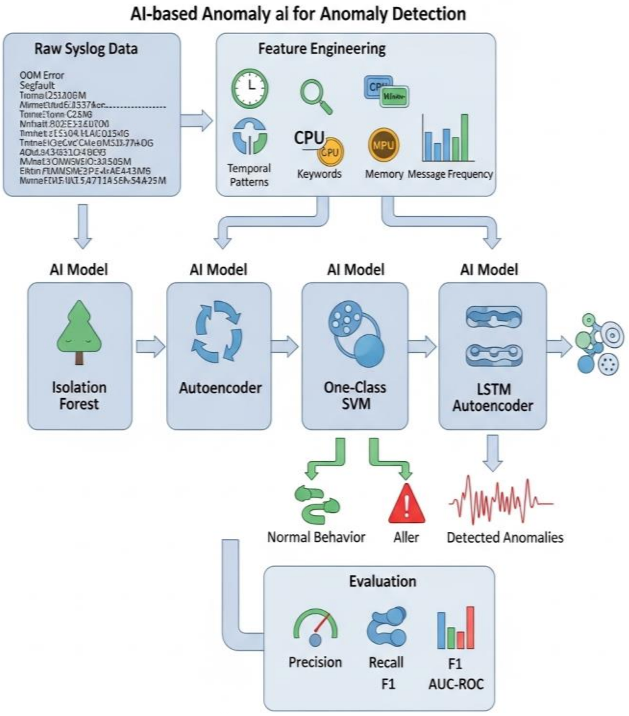

Figure 3 AI-Based Anomaly Detection Approach

For model selection, we focus on unsupervised and semi-supervised learning methods, given the scarcity of labeled failure examples in real-world datasets. Techniques used include Isolation Forest, which isolates outliers in high-dimensional space; Autoencoders, which learn to reconstruct normal behavior and flag deviations; and One-Class SVM, which defines a boundary around normal instances. For temporal sequence modeling, LSTM Autoencoders are particularly useful for detecting subtle deviations in log event sequences (Baker et al., 2023). During training, the models are exposed to "normal" pipeline runs, learning the baseline behavior. We simulate failures (e.g., memory overload, truncated outputs, dropped threads) to validate the models’ ability to detect previously unseen anomalies.

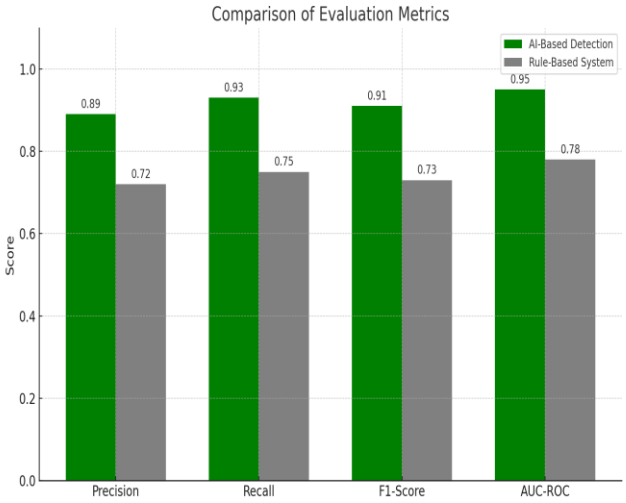

Figure 4 Comparison of Evaluation Metrics

Evaluation metrics include precision, recall, F1-score, and AUC-ROC. These help quantify the models' accuracy in identifying true anomalies while minimizing false positives. We compare our models against baseline rule-based systems to demonstrate the added value of AI-driven detection in complex pipeline environments (Sabanovic et al., 2024).

Experimental Setup

To evaluate the proposed framework, we designed a comprehensive experimental setup using microbial bioinformatics pipelines executed on both HPC and cloud-based environments. Pipelines included tools like Kraken2 for taxonomic classification, MetaPhlAn for microbial profiling, and HUMAnN for functional annotation. Workflow execution was managed using Snakemake, which provided consistent and reproducible runs (Wright et al., 2023). Syslog data was collected from all compute nodes and parsed for structured analysis. To establish a baseline, we ran several error-free executions, capturing normal log behavior. Then, we injected various anomalies such as artificial memory leaks, induced segmentation faults, and throttled CPU performance to simulate realistic pipeline issues.

We trained several AI models including Isolation Forests, LSTM Autoencoders, and One-Class SVMs on the baseline dataset. The models were then tested on both real and simulated anomaly cases. Evaluation results showed high accuracy in detecting outlier log patterns. For instance, the LSTM Autoencoder achieved an F1-score of 0.91 and detected sequence-based anomalies that rule-based systems missed entirely (Peterson et al., 2022).

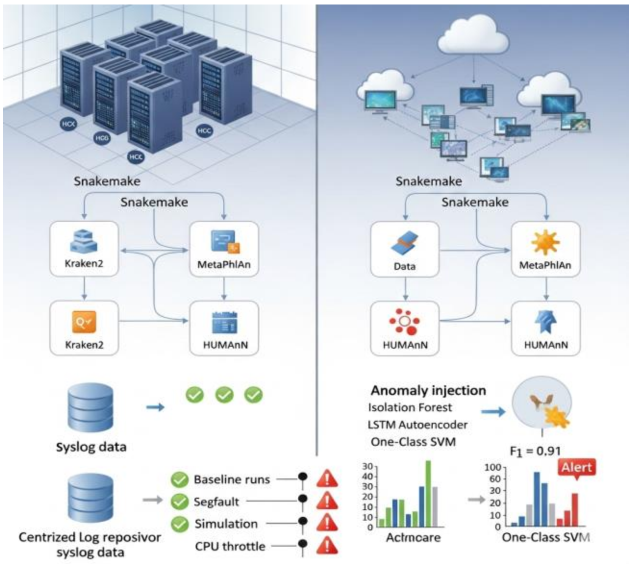

Figure 5 Experimental Setup

USE CASE: AVOIDING CORRUPTED RESULTS FROM UNNOTICED SYSTEM-LEVEL ISSUES IN COMPUTATIONAL MICROBIOLOGY

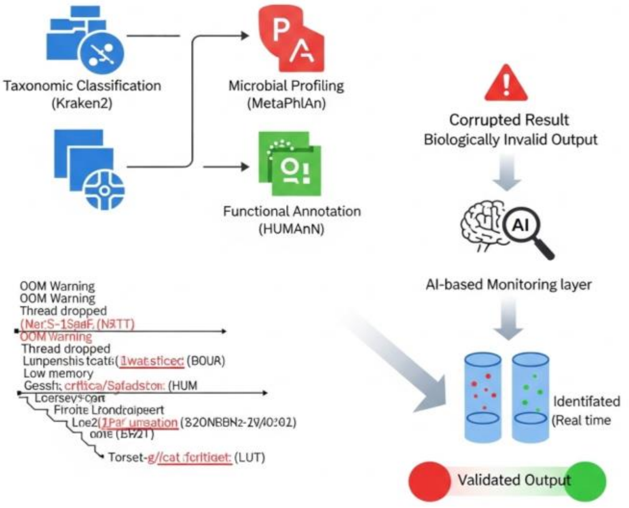

Figure 6 Unnoticed System-Level Issues in Computational Microbiology

Computational microbiology heavily relies on automated pipelines to analyze complex datasets derived from metagenomic sequencing, microbial profiling, and functional annotation. These workflows often span multiple tools and execute over extended periods on high-performance computing (HPC) clusters or cloud environments. While pipeline managers such as Snakemake or Nextflow track high-level task statuses, they do not monitor underlying system-level conditions in detail. As a result, silent or partial failures those that do not produce fatal errors can go undetected and lead to corrupted or misleading scientific results (Napa and Lorenzon, 2024).

For example, consider a scenario where a memory leak occurs during the execution of a taxonomic classifier like Kraken2. The tool does not crash but produces incomplete output due to internal failures caused by insufficient resources. This result may pass through the rest of the pipeline unchallenged, ultimately affecting downstream steps like diversity analysis or differential abundance testing. Without robust system monitoring, the user may wrongly interpret the output as biologically meaningful, leading to false conclusions or irreproducible findings (Munhoz et al., 2023). Using AI-enhanced syslog monitoring, such anomalies can be proactively detected. By analyzing memory allocation patterns, error message frequency, and subtle log sequences, the system can flag this execution as anomalous even in the absence of explicit tool-level failure. Alerts can prompt early intervention, rerun the affected task, or halt downstream analysis to prevent data corruption (Webner et al., 2023).

DISCUSSION

The experimental results demonstrate the effectiveness of AI-enhanced syslog monitoring in improving the robustness and transparency of microbial bioinformatics pipelines. Traditional monitoring approaches often depend on post-hoc QC tools or superficial success/failure statuses. By contrast, our method identifies subtle deviations in system behavior that may precede or accompany actual errors. These include anomalies such as slower-than-usual step execution, abnormal resource usage, or previously unnoticed warning messages. A key strength of our approach is its generalizability. Although our experiments focused on microbial pipelines, the framework can be applied to other areas of bioinformatics or computational science where workflows run in HPC or cloud environments. The modular architecture allows easy integration with diverse systems and pipelines. However, several challenges remain. One is the heterogeneity of syslog messages across systems and software tools, making standardized parsing and feature extraction nontrivial. Another is the black-box nature of some AI models, which may flag anomalies without clearly explaining why a concern in scientific computing where transparency is essential. Moreover, handling false positives is critical, especially in unsupervised models. Excessive alerts could lead to "alert fatigue" and reduce system trustworthiness. Future work should explore explainable AI (XAI) to improve interpretability and user trust. Finally, integrating this anomaly detection with pipeline orchestration and visualization tools could further streamline workflow diagnostics and corrective action processes.

CONCLUSION

In this study, we presented a novel framework that integrates AI-based anomaly detection into microbial bioinformatics workflows by leveraging system-level logs (syslogs). Traditional quality assurance tools in bioinformatics primarily focus on output validation and step completion but often miss subtle, non-terminating system-level failures such as memory leaks, segmentation faults, and transient slowdowns that can lead to corrupted or misleading scientific results. Our approach addresses this critical gap by applying machine learning models to structured syslog data collected during workflow execution. We designed and validated the framework across both HPC and cloud environments, using real-world pipelines such as Kraken2, MetaPhlAn, and HUMAnN. By simulating various fault conditions and training on normal execution logs, models like LSTM Autoencoders and Isolation Forests demonstrated strong performance, with F1-scores exceeding 0.9 in detecting anomalies that were otherwise overlooked by rule-based systems. Our results emphasize the feasibility and effectiveness of AI-enhanced log monitoring as a proactive quality control layer for computational microbiology. This approach not only improves pipeline robustness and reproducibility but also helps researchers detect and mitigate hidden sources of data corruption. Future work may extend this framework with more advanced sequence modeling, integration with pipeline managers, and deployment in real-time monitoring dashboards.

Future Work

While our framework demonstrates the potential of AI-driven syslog analysis for anomaly detection in microbial bioinformatics pipelines, several avenues remain open for further development and refinement. First, the integration of real-time anomaly response mechanisms could significantly enhance the utility of the system. For example, integrating alert-triggered workflow pausing, automated retries, or dynamic resource reallocation based on detected anomalies would close the loop between detection and recovery. Second, we aim to improve model generalization across diverse computational environments and pipeline configurations. Training models that are transferable between different HPC clusters or cloud providers will require more extensive datasets and possibly federated learning approaches to preserve privacy and infrastructure-specific characteristics. Third, the enrichment of feature engineering could be explored by incorporating additional telemetry data such as I/O performance, container health (in Docker/Singularity setups), and environmental variables. This would help detect anomalies originating from infrastructure layers not fully captured in syslog data alone. Additionally, user-facing interfaces for visualizing anomaly predictions and correlating them with workflow steps are essential for practical adoption. Dashboards, integration with monitoring tools like Prometheus or Grafana, and plugins for Snakemake or Nextflow could make the system accessible to non-expert users.

REFERENCES

[1] Ahmad, S., Lohiya, S., Taksande, A., Meshram, R. J., Varma, A., & Vagha, K. (2024). A comprehensive review of innovative paradigms in Microbial Detection and antimicrobial resistance: Beyond traditional cultural methods. Cureus. https://doi.org/10.7759/cureus.61476

[2] Baker, M., Fard, A. Y., Althuwaini, H., & Shadmand, M. B. (2023). Real-time AI-based anomaly detection and classification in power electronics dominated grids. *IEEE Journal of Emerging and Selected Topics in Industrial Electronics*, 4(2), 549–559. https://doi.org/10.1109/jestie.2022.3227005

[3] Dong, Y.-H., Luo, Y.-H., Liu, C.-J., Huang, W.-Y., Feng, L., Zou, X.-Y., Zhou, J.-Y., & Li, X.-R. (2024a). Changes in microbial composition and interaction patterns of female urogenital tract and rectum in response to HPV infection. *Journal of Translational Medicine*, 22(1). https://doi.org/10.1186/s12967-024-04916-2

[4] Dong, Y.-H., Luo, Y.-H., Liu, C.-J., Huang, W.-Y., Feng, L., Zou, X.-Y., Zhou, J.-Y., & Li, X.-R. (2024b). Changes in microbial composition and interaction patterns of female urogenital tract and rectum in response to HPV infection. *Journal of Translational Medicine*, 22(1). https://doi.org/10.1186/s12967-024-04916-2

[5] Han, D., Sun, M., Li, M., & Chen, Q. (2023). LTAnomaly: A Transformer variant for syslog anomaly detection based on multi-scale representation and long sequence capture. *Applied Sciences*, 13(13), 7668. https://doi.org/10.3390/app13137668

[6] Li, H., Xu, H., Li, Y., & Li, X. (2024). Application of artificial intelligence (ai)-enhanced biochemical sensing in molecular diagnosis and Imaging

Analysis: Advancing and challenges. *TrAC Trends in Analytical Chemistry*, 174, 117700. https://doi.org/10.1016/j.trac.2024.117700

[7] Munhoz, V., Bonfils, A., Castro, M., & Mendizabal, O. (2023). A performance comparison of HPC workloads on traditional and cloud-based HPC clusters. 2023 *International Symposium on Computer Architecture and High Performance Computing Workshops (SBAC-PADW)*, 108–114. https://doi.org/10.1109/sbac-padw60351.2023.00026

[8] Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). Sustainable Data Analysis with snakemake. *F1000Research, 10*, 33. https://doi.org/10.12688/f1000research.29032.2

[9] Napa, M. A., & Lorenzon, A. F. (2024). A systematic literature review of I/O optimization in HPC and cloud computing environments. *2024 International Symposium on Computer Architecture and High Performance Computing Workshops (SBAC-PADW)*, 66–72. https://doi.org/10.1109/sbac-padw64858.2024.00020

[10] Ndibe, O. S. (2025). Ai-driven forensic systems for real-time anomaly detection and threat mitigation in cybersecurity infrastructures. *International Journal of Research Publication and Reviews*, 6(5), 389–411. https://doi.org/10.55248/gengpi.6.0525.1991

[11] Peterson, C.-L., Alexander, D., Chen, J. C.-Y., Adam, H., Walker, M., Ali, J., Forbes, J., Taboada, E., Barker, D. O., Graham, M., Knox, N., & Reimer, A. R. (2022a). Clinical metagenomics is increasingly accurate and affordable to detect enteric bacterial pathogens in stool. *Microorganisms*, 10(2), 441. https://doi.org/10.3390/microorganisms10020441

[12] Peterson, C.-L., Alexander, D., Chen, J. C.-Y., Adam, H., Walker, M., Ali, J., Forbes, J., Taboada, E., Barker, D. O., Graham, M., Knox, N., & Reimer, A. R. (2022b). Clinical metagenomics is increasingly accurate and affordable to detect enteric bacterial pathogens in stool. *Microorganisms*, 10(2), 441. https://doi.org/10.3390/microorganisms10020441

[13] Pohl, S., Elfaramawy, N., Miling, A., Cao, K., Kehr, B., & Weidlich, M. (2024a). How do users design scientific workflows? the case of Snakemake and nextflow. *Proceedings of the 36th International Conference on Scientific and Statistical Database Management*, 1–12. https://doi.org/10.1145/3676288.3676290

[14] Pohl, S., Elfaramawy, N., Miling, A., Cao, K., Kehr, B., & Weidlich, M. (2024b). How do users design scientific workflows? the case of Snakemake and nextflow. *Proceedings of the 36th International Conference on Scientific and Statistical Database Management*, 1–12. https://doi.org/10.1145/3676288.3676290

[15] R.W.P.M., R., K.2, V., & D.P.S.T.G, A. (2023). Establishment of a bioinformatics pipeline for the detection of pathogenic bacteria. *Asian Journal of Microbiology, Biotechnology &amp; Environmental Sciences*, 25(04), 818–826. https://doi.org/10.53550/ajmbes.2023.v25i04.040

[16] Weßner, J., Berlich, R., Schwarz, K., & Lutz, M. F. (2023). Parametric optimization on HPC clusters with Geneva. *Computing and Software for Big Science*, 7(1). https://doi.org/10.1007/s41781-023-00098-6

[17] Wright, R. J., Comeau, A. M., & Langille, M. G. (2023). From defaults to databases: Parameter and database choice dramatically impact the performance of metagenomic taxonomic classification tools. *Microbial Genomics*, 9(3). https://doi.org/10.1099/mgen.0.000949

[18] Yayla, E. (2024). Artificial Intelligence Applications in Clinical Microbiology Laboratory. *Journal of Immunology and Clinical Microbiology*, 9(2), 56–72. https://doi.org/10.58854/jicm.1404800

[19] Šabanović, K.-I., Arendt, C., Fricke, S., Geis, M., Böcker, S., & Wietfeld, C. (2024). AI-based anomaly detection for Industrial 5G networks by distributed SDR measurements. *2024 IEEE International Symposium on Measurements &amp;Amp; Networking (M&amp;Amp;N)*, 1–5. https://doi.org/10.1109/mn60932.2024.10615402

[20] Shivani Sahu et al. 2025. Combining SVM and Gradient Boosting for Enhanced Accuracy in Diabetes Diagnosis. *International Journal of Innovations in Science, Engineering And Management*. 4, 1 (Jan. 2025), 08–16. DOI:https://doi.org/10.69968/ijisem.2025v4i108-16.