

**SCHOOL OF COMPUTER SCIENCE ENGINEERING AND INFORMATION SYSTEMS**

**Winter Semester – 2023-24**

**M.Tech (SE)**

**SWE1904 - Capstone Project**

**1st Review**

|  |  |
| --- | --- |
| **Register Number** | 19MIS0184 |
| **Student Name** | VIVEK R |
| **Project Code**  **(Course Code)** | SWE1904 |
| **Project Domain (Capstone Project)** | DATA SCIENCE |
| **Project Title (Capstone Project)** | Comprehensive approach of Static and Dynamic Data Analytics through Automated Machine Learning (AutoML) |
| **Guide Name** | Dr. CHADRASEGAR T |
| **Guide Signature** |  |

**Comprehensive approach of Static and Dynamic Data Analytics through Automated Machine Learning (AutoML)**

**ABSTRACT:**

The Intrusion Detection Evaluation Dataset (CICIDS2017) consists of normal network traffic and simulated abnormal data, caused by deliberate attacks on a test network. The success of Intrusion Detection Systems and Intrusion Prevention Systems, which are vital for detecting and mitigating computer network attacks, hinges on having current and relevant training data. This often figures out whether a network remains secure or becomes compromised. With the proliferation of sensors and smart devices, data generation speed in Internet of Things (IoT) systems has surged. These systems regularly process, transform, and analyse large volumes of data to enable various IoT services and functionalities. Machine Learning (ML) approaches have proven their effectiveness in IoT data analytics. However, applying ML models to such tasks is still challenging, particularly in terms of model selection, design/tuning, and updating. This has created a significant demand for experienced data scientists. Additionally, the dynamic nature of IoT data may lead to concept drift issues, negatively affecting model performance. Alleviate these issues, Automated Machine Learning (AutoML) has appeared as a field aimed at automatically selecting, constructing, tuning, and updating ML models to optimize performance on specific tasks. This project reviews existing methods for model selection, tuning, and updating in AutoML. The aim is to find and summarize optimal solutions for applying AutoML Techniques to the Intrusion detection evaluation dataset (CIC-IDS2017) and IoT Network Intrusion Dataset.

Keywords:   
Intrusion Detection system, Network traffic, IoT data analytics, Machine Learning, AutoML

**PROBLEM DEFINITION:**

The CICIDS2017 intrusion detection dataset is a valuable public benchmark for developing network security systems using machine learning. However, it lacks rigorous profiling and optimization for building robust models. Issues such as class imbalance, redundant records, and missing guidelines limit its effectiveness. Analysing massive and dynamic IoT data using machine learning also has its challenges. The large volume and speed of heterogeneous IoT data streams make manual model management difficult. Changes in data or concept drift degrades model performance over time. There is a need for Automated Machine Learning (AutoML) solutions that can reduce human effort in developing analytic models. For intrusion detection, this involves evaluating datasets thoroughly, correcting deficiencies, and establishing best practices. For IoT data, this involves automating steps such as data preprocessing, feature engineering, algorithm selection, hyperparameter optimization, and drift-adaptive model updating. By improving benchmark datasets and demonstrating end-to-end AutoML capabilities, we can rapidly develop more robust intrusion detection systems and optimized IoT data analytics. This will better equip us to handle evolving security threats and streaming heterogeneous data.

**LITERATURE SURVEY:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **Title of the Paper** | **Merits** | **Demerits** |
| **1.** | IoT data analytics in dynamic environments: From an automated machine learning perspective | * Comprehensive review of AutoML for IoT analytics * Analyzes range of methods for distributed data, concept drift, edge deployment * Structured taxonomy of existing techniques | * Rapidly evolving field means some recent advances not covered * Does not include experimental comparisons |
| **2.** | AutoML-ID: automated machine learning model for intrusion detection using wireless sensor network | * Focus on important problem of IoT intrusion detection * Reviews specific AutoML techniques applied in this domain * Discusses open challenges and future research directions | * Scope limited to intrusion detection * Does not include quantitative evaluations * Light on technical details of AutoML implementations |
| **3.** | Methods for network intrusion detection: Evaluating rule-based methods and machine learning models on the CIC-IDS2017 dataset | * Leverages the commonly used CIC-IDS2017 benchmark dataset reflecting modern network attacks. * Examines a diverse set of models including neural networks and ensemble methods in addition to more basic classifiers. * Analysis of performance on different attack categories offers useful insights for selection of appropriate techniques. | * Only certain standard classification algorithms are evaluated rather than more recent advanced methods. * The study is limited to binary classification of network flows; multi-class detection is not examined. * Lacks analysis of the efficiency and computational requirements of the different approaches. |
| **4.** | AMLBID: An auto-explained Automated Machine Learning tool for Big Industrial Data | * Innovative AutoML tool for Big Industrial Data. * Meta-learning-based recommendation system for efficient ML algorithm selection. * Interactive visualization module enhances user understanding. | * Limited empirical validation on real-world datasets. * Dependency on meta-features may limit effectiveness. * Scalability concerns for handling large-scale datasets. |
| **5.** | AMC: AutoML for Model Compression and Acceleration on Mobile Devices | * Innovation in leveraging reinforcement learning for automated model compression. * Support for both accuracy-guaranteed and resource-constrained compression. * Significant improvements in inference speed demonstrated on mobile devices. | * Lack of discussion on potential limitations or failure modes of the proposed approach. * Limited comparison with other state-of-the-art techniques in model compression and acceleration. |
| **6.** | AutoML: A Survey of the State-of-the-Art | * Provides an extensive exploration of Neural Architecture Search (NAS), a key area in AutoML, with performance comparisons on CIFAR-10 and ImageNet datasets. * Encompasses a wide array of AutoML topics, including automated data augmentation, hyperparameter optimization, and evolutionary algorithms. | * Lacks coverage of recent AutoML advancements, necessitating an update to incorporate newer methods * NAS receives disproportionate attention. * Lacks critical analysis and comparison. |
| **7.** | Dynamic Hyperparameter Allocation under Time Constraints for Automated Machine Learning | * Paper introduces a novel diversification strategy for HPO, emphasizing dynamic hyperparameter space allocation for a sampler based on the remaining time budget. * Proposed solution claims better performance in terms of both time and resource awareness compared to previous resource-aware solutions | * Effectiveness of the proposed strategy may depend on the specific characteristics of the datasets and models used in the benchmarks. The generalizability of the approach to diverse scenarios should be considered. |
| **8.** | AutoML for Multi-Label Classification: Overview and Empirical Evaluation | * The paper conducts an extensive experimental study evaluating multiple optimization methods on a suite of MLC problems, providing a thorough analysis of each approach's performance * By extending existing AutoML approaches to tackle multi-label classification problems, the paper contributes valuable insights into the applicability and effectiveness of these techniques in complex learning scenarios. | * primarily focuses on theoretical comparisons and the performance of optimization methods in controlled experiments. Real-world application scenarios or case studies could provide more practical insights into the usability of these approaches. * While the paper outlines potential research directions to improve AutoML for MLC, it could further elaborate on practical implementations, challenges |
| **9.** | Online AutoML: an adaptive AutoML framework for online learning | * OAML is the first system to propose a flexible and practical AutoML system for adaptive online learning pipelines, combining algorithm selection, hyperparameter optimization, and preprocessing. * OAML leverages drift detection to trigger pipeline redesign, adapting to changing data distributions * OAML offers different adaptation strategies (ensemble, model store) and supports various online learning algorithms and preprocessors | * The current implementation focuses on classification tasks and might require adjustments for other supervised learning problems * The paper acknowledges the lack of research on hyperparameter optimization in online settings and uses manually defined search intervals. Exploring more advanced techniques could further improve performance. |
| **10.** | AutoML-GPT: Automatic Machine Learning with GPT | * Proposed AutoML-GPT framework offers a systematic approach to automating the training pipeline for various AI tasks by leveraging the language capabilities of GPT and dynamically generating task-oriented prompts * This can potentially reduce the manual effort required in model selection and hyperparameter tuning | * Lack of external validation on real-world tasks * Scalability and resource requirements |

**OBJECTIVES:**

* Developing automated solutions for data preprocessing, feature engineering, algorithm selection, and hyperparameter optimization using AutoML techniques.
* Addressing class imbalance, redundant records, and missing guidelines in the CICIDS2017 dataset to improve its effectiveness as a benchmark for intrusion detection systems.
* Investigating techniques for handling concept drift in IoT data analytics and develop automated model updating procedures to maintain model performance over time.
* Implementing an end-to-end AutoML pipeline for both static intrusion detection and dynamic IoT data analytics tasks.
* Comparing the performance of AutoML models with traditional machine learning approaches to assess the benefits of automation in model development.
* Demonstrating the feasibility and effectiveness of AutoML in improving the efficiency and accuracy of data analytics tasks in both static and dynamic environments.

**SCOPE OF THE PROJECT:**

The project aims to enhance intrusion detection systems and IoT data analytics through Automated Machine Learning (AutoML). It involves evaluating the CICIDS2017 intrusion detection dataset, automating data preprocessing, feature engineering, algorithm selection, and hyperparameter optimization. The project will also address concept drift in IoT data and implement an end-to-end AutoML pipeline for both static and dynamic data. Performance comparison with traditional approaches will be conducted, and recommendations for deploying AutoML in real-world applications will be provided. Ultimately, the project seeks to demonstrate the feasibility and effectiveness of AutoML in improving data analytics tasks in diverse environments

1. AutoML for Algorithm Selection & Hyperparameter Optimization: Investigating automated methods to select ML algorithms and optimize hyperparameters for intrusion detection tasks.
2. Handling Concept Drift in IoT Analytics: Developing automated procedures to detect and adapt to changes in data distribution over time in IoT analytics.
3. End-to-End AutoML Pipeline: Implementing a comprehensive AutoML pipeline covering data preprocessing to model deployment for both static intrusion detection and dynamic IoT analytics.

A screenshot of a computer

Description automatically generated**ARCHITECTURE:**

Static dataset – Intrusion Detection Evaluation Dataset (CICIDS2017)

A diagram of a company

Description automatically generated with medium confidence

Dynamic dataset – IoT Data Analytics in Dynamic Environments

A diagram of a cloud computing system

Description automatically generated

**REFERENCES:**

1. Yang, L., & Shami, A. (2022). IoT data analytics in dynamic environments: From an automated machine learning perspective. Engineering Applications of Artificial Intelligence, 116, 105366.
2. Singh, A., Amutha, J., Nagar, J., Sharma, S., & Lee, C. C. (2022). AutoML-ID: Automated machine learning model for intrusion detection using wireless sensor network. Scientific Reports, 12(1), 9074.
3. Lindstedt, H. (2022). Methods for network intrusion detection : Evaluating rule-based methods and machine learning models on the CIC-IDS2017 dataset (Dissertation). Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-479347
4. Garouani, M., Ahmad, A., Bouneffa, M., & Hamlich, M. (2022). AMLBID: an auto-explained automated machine learning tool for big industrial data. SoftwareX, 17, 100919.
5. He, Y., Lin, J., Liu, Z., Wang, H., Li, L. J., & Han, S. (2018). Amc: Automl for model compression and acceleration on mobile devices. In Proceedings of the European conference on computer vision (ECCV) (pp. 784-800).
6. He, X., Zhao, K., & Chu, X. (2021). AutoML: A survey of the state-of-the-art. Knowledge-Based Systems, 212, 106622.
7. Lee, J., Ahn, S., Kim, H., & Lee, J. R. (2022). Dynamic Hyperparameter Allocation under Time Constraints for Automated Machine Learning. Intelligent Automation & Soft Computing, 31(1).
8. Wever, M., Tornede, A., Mohr, F., & Hüllermeier, E. (2021). AutoML for multi-label classification: Overview and empirical evaluation. IEEE transactions on pattern analysis and machine intelligence, 43(9), 3037-3054.
9. Celik, B., Singh, P., & Vanschoren, J. (2023). Online automl: An adaptive automl framework for online learning. Machine Learning, 112(6), 1897-1921.
10. Zhang, S., Gong, C., Wu, L., Liu, X., & Zhou, M. (2023). AutoML-GPT: Automatic Machine Learning with GPT. arXiv preprint arXiv:2305.02499.