

**ENHANCING CYBERSECURITY WITH INTELLIGENT AI AND MACHINE  
LEARNING USING AUTOML TECHNIQUES**

**Dr.R.Madhubala<sup>1</sup>, Dr.Akhila K.R<sup>2</sup>, P.S.G. Aruna Sri<sup>3</sup>, B Madhav Rao<sup>4</sup>,  
Dr.A.Deepa<sup>5</sup>**

<sup>1</sup>Lecturer, Software Engineering and Data Technologies Unit, Department of Computing and Information Sciences, University of Technology and Applied Sciences, Shinas, Oman

<sup>2</sup>Lecturer, Cybersecurity and Networks Unit, Department of Computing and Information Sciences, University of Technology and Applied Sciences, Shinas, Oman

<sup>3</sup>Professor, Department of Internet of Things, Koneru Lakshmaiah Education Foundation, Vaddeswaram

<sup>4</sup>Professor, Department of CSE, Sir C R Reddy College of Engineering, Eluru

<sup>5</sup>Associate Professor, Department of Computer Science and Engineering, School of Computing, Sathyabama Institute of Science and Technology, Chennai

E-mail: r.madhubala@utas.edu.om<sup>1</sup>, akhila.rajeswary@utas.edu.om<sup>2</sup>,  
arunasri\_2012@kluniversity.in<sup>3</sup>, madhavraob@gmail.com<sup>4</sup>, nraj.deepa@gmail.com<sup>5</sup>

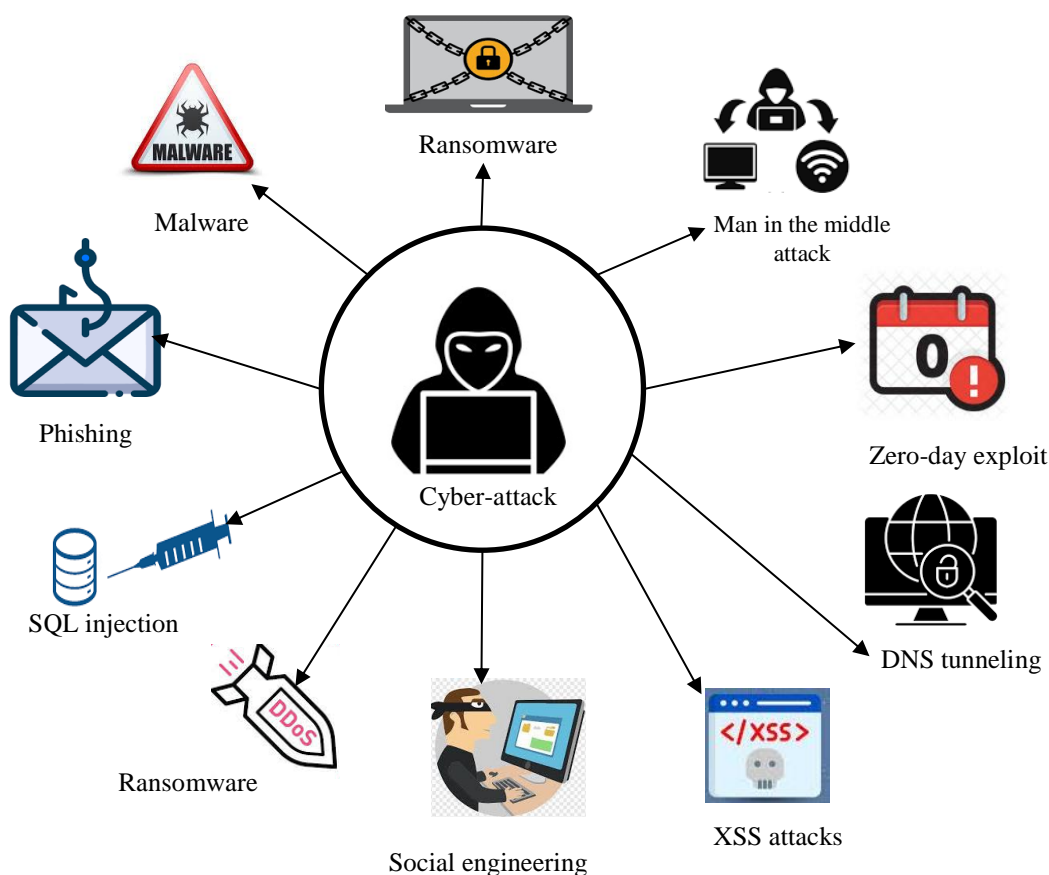
**Abstract**

Automated Machine Learning, also known as AutoML, is a collection of methods and procedures designed to make machine learning understandable to non-experts. For specific datasets, autoML can recommend the best models or show optimal improvement of an existing model. The goal of the new discipline of automated machine learning, or autoML, is to automate the process of creating machine learning models. When machine learning is used, autoML was developed to boost efficiency and production by automating as much of the repetitive, unproductive work as feasible. Research has long been done on technologies that can efficiently create high-quality models with the least amount of model creators' involvement in the process, from selecting and fine-tuning algorithms to preparing data. The data processing requirements for AutoML techniques are compiled and explained in detail in this semantic review investigation. Through the internet, the latter can safely make the local Jenkins address available to the public. As a result, the suggested pipeline is a hybrid software design that incorporates elements from the machine learning operations (MLOps) theme. Additionally, it looks into security intelligence modeling as a way to provide useful information for enhancing organizational resilience. This paper also aims to give readers a comprehensive grasp of how artificial intelligence is changing cybersecurity and to suggest future directions for this ever-evolving field's research and development.

**Keywords:** autoML, hyperparameter, automation, AutoML pipeline, Machine Learning, cybersecurity.

## 1. Introduction

The digital age in which we live has advantages and disadvantages like anything else. The primary disadvantage is the risk to security [1]. More and more of our private data is moving online, which increases the likelihood of disastrous security breaches.



**Figure 1.** A several frequent threats or attacks in the domain of cybersecurity

As Figure 1 illustrates, malware [2], phishing, ransomware, denial-of-service (DoS), zero-day attacks, etc. are frequent.

Automating model optimization procedures, such as feature engineering, algorithm selection, and parameter tweaking, is the focus of the quickly emerging field of automated machine learning (AutoML) [3]. The overall requirement for human input and modeling skill is decreased by autoML systems, which generate high-performance models using a variety of complex algorithms and preprocessing techniques. Similarly, AutoML is reducing the knowledge and technical barriers that prevent machine learning from being more accessible across a range of industries, including biomedicine. By enabling scientists and clinicians to train and use potent models, AutoML has the potential to significantly speed up clinical machine learning applications, which are becoming more and more popular in clinical research due to the growing breadth, depth, and accessibility of clinical health data. However, for ML models to be clinically useful, they must be methodologically and inferentially reproducible, interpretable for biological causes, and verified by practitioners.

This is how the remainder of the paper is organized. The relevance of machine learning in current cybersecurity research and applications is explained in Section 2, which also serves to motivate and define the scope of our study. Section 3 delves deeply into cybersecurity data, while Section 4 thoroughly examines several machine learning algorithms. We discuss key research directions and the potential applications of machine learning in cybersecurity in the future. Section 5 brings the research to a close.

## **2. Related Works**

In this research, we apply AutoML to the problem of malware identification. Two AutoML approaches, Microsoft NNI and AutoGluon-Tabular, are empirically evaluated here. The Light Gradient Boosted Machine (LightGBM) model we train is trained on two malware data sets. A publicly accessible labeled benchmarking data set for training machine learning models for malware detection based on static analysis of Windows portable executable (PE) files is the first data set, called EMBER [4]. This data set, which has 2 million samples, is sufficiently large, open, and generic to cover a number of intriguing application cases. SecureAge, the second analysis, includes recently gathered malware and innocuous PE files that we purchased from a nearby business.

Cyber security systems driven by AI and ML will be crucial in addressing the ongoing rise in threat complexity and quantity, the changing nature of threats, and the demand for quick and highly automated threat responses [5]. Defense systems driven by AI/ML, for instance, may instantly detect suspicious patterns and abnormalities in massive data sets. Large-scale cyberattacks can be avoided by automatically updating current software based on advanced real-time analysis using AI/ML. AI approaches are being used by major email providers to stop unwanted images, identify phishing, spyware, and forged payments. Models based on artificial neural networks (ANNs) are being used by other providers to identify and categorize malware and phishing emails. Furthermore, AI/ML is superior to static signatures, which are utilized in traditional anti-virus systems, for finding malware and anti-virus defense.

Several cutting-edge machine learning models have previously been developed to detect intrusions in network profile data in order to protect user data. Furthermore, whereas standard machine learning algorithms are good at categorizing small and low dimensional data, they are impractical for resolving problems involving huge dimensionality data [6]. Furthermore, current methods only accurately depict classes with substantial dispersion because to issues with data imbalance, which prevents them from offering great performance. Additionally, ExploreKit and AutoLearn are free source. Although there are minor differences in how each approach manages the feature transformations required for analysis and the feature selection stage, none of them consistently outperforms the others.

Since the need for computers is always growing, cyber security is a major issue in the computer industry. Additionally, it is well known that millions of zero-day attacks are appearing in the IoT space. It is challenging for cutting-edge mechanisms like machine learning to identify cyberattacks. However, the success of DL addresses the problems that cyber security faces [7]. Deep learning can now be used practically because to advancements

in CPU and neural network algorithms. Due of its powerful extraction capabilities, DL is being used in cyberattacks that are successful in minor mutation and unique attacks. The primary method for identifying attacks and differentiating them from innocuous traffic is the deep learning architecture's self-taught and compressing capabilities.

These tools' development has made it possible for interdisciplinary study and the use of deep learning and machine learning methodologies, which has produced encouraging outcomes and showed how machine learning models may be used to address a range of issues. But it has also become clear that creating a machine learning model is a difficult process that calls for technical know-how, experience, and intuition to adjust the model's hyper-parameters [8]. Researchers are motivated to investigate the potential for developing a method to automate the creation of machine learning models due to the significant dependence of machine learning development on human specialists. Researchers and machine learning professionals first made these attempts to concentrate on AutoML projects, and firms that offer Auto-ML frameworks as part of their business models then followed suit.

Many various kinds of businesses have successfully implemented AI-driven solutions to enhance their security protocols, proving that AI has practical applications in cybersecurity. These case studies show how AI may improve threat detection and response, effectively address specific problems, and ultimately strengthen an organization's cybersecurity posture as a whole. Several significant case studies that show how AI has been successfully integrated into cybersecurity will be looked at in this section [9]. Leading cybersecurity company Dark Trace protects businesses from advanced cyberattacks with AI-driven solutions. Dark Trace uses a cutting-edge method known as "self-learning AI" to create a digital immune system that learns the normal behavior of every individual and device on a network inside an organization.

ML has completely changed how organizations identify and react to cyber threats. While traditional threat detection methods often rely on static rule-based systems that can be effective against known threats, they are unable to address the ever-evolving tactics of cybercriminals. ML offers a dynamic and adaptive solution because of its capacity to analyze large datasets and identify patterns. One of the main uses of ML in cybersecurity is anomaly detection, where organizations can establish a baseline of "normal" behavior for their systems, networks, or users by training ML algorithms on historical data [10]. An alert is sent out in the event that this baseline is broken, suggesting possible malicious behavior. In this field, methods like neural networks, clustering, and classification are frequently employed.

### **3. Methods and Materials**

#### **3.1. Academic framework**

The theoretical framework of this study incorporates several fundamental ideas from automation, enterprise systems integration, and machine learning. By combining these ideas, we aim to create a systematic method for comprehending how AutoML may enhance enterprise AI pipelines. The framework provides insights into the strategic and technical

aspects of deploying AutoML in enterprise settings by incorporating theories about system optimization [11], automated machine learning, and organizational change management.

○ **Automation Theory**

The trade-off between efficiency and control: more automated processes require less human oversight, which could lead to increased efficiency but less oversight. Understanding how AutoML impacts enterprise AI pipelines is crucial. Routine task automation may improve speed and scalability, but it must be closely watched to maintain control over decision-making and model quality.

○ **Machine Learning Pipeline Theory**

To cut down on the time and effort needed to create efficient AI models, autoML seeks to provide an optimal and automated approach (as much as possible) in each of these processes. You could think of it as the steps of a pipeline or a series of related processes that need to be coordinated to function well. Every machine learning pipeline depends on the continuous cycle, and understanding and responding to their interplay is essential for achieving the best results. Automating individual processes is simply one aspect of the difficulty in the context of AutoML; another is making sure that the automated components function as a cohesive unit to produce high-quality models that satisfy organizational requirements.

○ **Systems Theory**

Knowledge about pieces alone is insufficient if we do not understand how each component interacts to make the whole, as systems theory suggests. This theory emphasizes the necessity of coupling automated machine learning tools to data infrastructure, business processes, and legacy enterprise systems in the context of AutoML. The successful integration of AutoML in business AI pipelines requires careful management of this interaction to guarantee that the automation process meets organizational goals [12], responds to evolving business requirements, and delivers timely input for decision-making.

○ **Resource-Based View (RBV) of the Firm**

The Resource-Based View (RBV) provides a theoretical framework for evaluating the strategic benefits of AutoML use in businesses. RBV focuses on how businesses can use their resources and competencies to obtain a competitive edge. As a cutting-edge technological tool, autoML can give businesses a special opportunity to improve their AI capabilities without depending on costly and in-demand data science expertise. AutoML can help businesses access and implement AI solutions more rapidly, make better decisions, and extract value from data by lowering the obstacles to AI adoption. In this regard, AutoML is viewed as a strategic tool that enables businesses to promote innovation, optimize their AI processes, and maybe obtain a competitive advantage in a market that is driven by data.

○ **Diffusion of Innovation Theory**

The Diffusion of Innovation (DOI) hypothesis is essential to comprehending the adoption of AutoML technology in businesses. The DOI theory describes how innovations proliferate within a community, taking into account variables including relative advantage, trialability,

observability, complexity, and compatibility with current habits. While compliance with current corporate systems continues to be a significant concern, AutoML's relative benefit is its capacity to streamline the creation of machine learning models. The rate of adoption may be impacted by the perceived difficulty of implementing AutoML technologies, particularly in companies with little experience with AI. The DOI framework sheds light on the elements that either facilitate or impede the broad adoption of AutoML in businesses by highlighting the roles of corporate culture, leadership, and outside forces in the diffusion process.

- **Decision Theory**

Decision theory is essential to comprehend the connection between business results and AutoML-generated models, especially when it comes to decision support systems (DSS). The process of rational decision-making in intricate, unpredictable situations is the main emphasis of decision theory. AutoML solutions enable businesses to more effectively use data-driven insights by automating the model selection and optimization process, which lessens subjectivity and cognitive burden in decision-making. The core of AutoML models' value proposition in enterprise AI pipelines is their capacity to improve business performance and decision-making. Decision theory states that to maximize business success, model predictions must be in line with strategic goals. This is an essential idea for businesses using AutoML in practical applications.

### **3.2. AutoML Challenges in Cybersecurity**

Machine learning could be used to model cybersecurity threats as data-driven issues. We face several difficulties when using machine learning to address these well-known security risks; these difficulties are fueled by the particulars of cybersecurity issues [13]. Because of this, using ML in cybersecurity is seen to be more difficult than in other fields. Numerous research studies in the literature provide clear explanations of these difficulties. Based on our review of the literature and our assessment of the chosen AutoML tools, we exclusively address the most prevalent AutoML issues that are not addressed in the body of current research.

#### **3.2.1. Feature Engineering and Extraction**

In cybersecurity applications, raw data is extremely varied and typically necessitates specialized and time-consuming feature engineering and extraction processes. The characteristics are not conventionally categorized, nominal, or numerical. Rich intelligence is typically encoded by the features. For example, port numbers cannot be considered numerical data even though they are numerical values. HTTP is no more advanced than SMTP or SSH; binary file or protocol names are not only strings or categorical properties. The semantics of the features are unlikely to be captured by standard or best practice methods for feature engineering like one-hot encoding.

Consequently, automated or semi-automatic feature engineering techniques must be supported by AutoML tools. We can determine which characteristics are usually designed for a particular cybersecurity application. These are the characteristics that security and machine learning specialists have suggested and employed in the literature. The feature might be included in a feature store or collection of design patterns. AutoML tools might then be used

to automate and implement these feature engineering design patterns. While preparing some of the chosen datasets for this study, we investigated this strategy.

### **3.2.2. Reliability and Dependability**

ML models in cybersecurity have particular difficulties that make them more sensitive to dependability and credibility than models in other fields. It is the result of various intrinsic features present in cybersecurity applications and their environments of operation. An hostile environment, data quality, and model drift are important variables affecting the reliability and consistency of ML models in cybersecurity.

### **3.2.3. Model Drift**

The majority of machine learning models used in cybersecurity applications function in a very dynamic and ever-evolving context [14]. Consequently, the ML model's predicted accuracy tends to deteriorate with time and ultimately drop significantly. Model drift or decay is the term used to describe this process. The primary causes of the drift or decay of ML models are concept and data drift. While data drift occurs when the target class remains constant but the characteristics or features that are used to predict it vary, concept drift occurs when the target we wish to predict changes over time. In cybersecurity applications, both drift of concepts and data drift are common.

For deployed models, autoML pipelines need to include ongoing monitoring and verification systems. The problem of model drift has become a more pressing subject of recent AutoML research.

### **3.2.4. Adversarial and Hostile Settings**

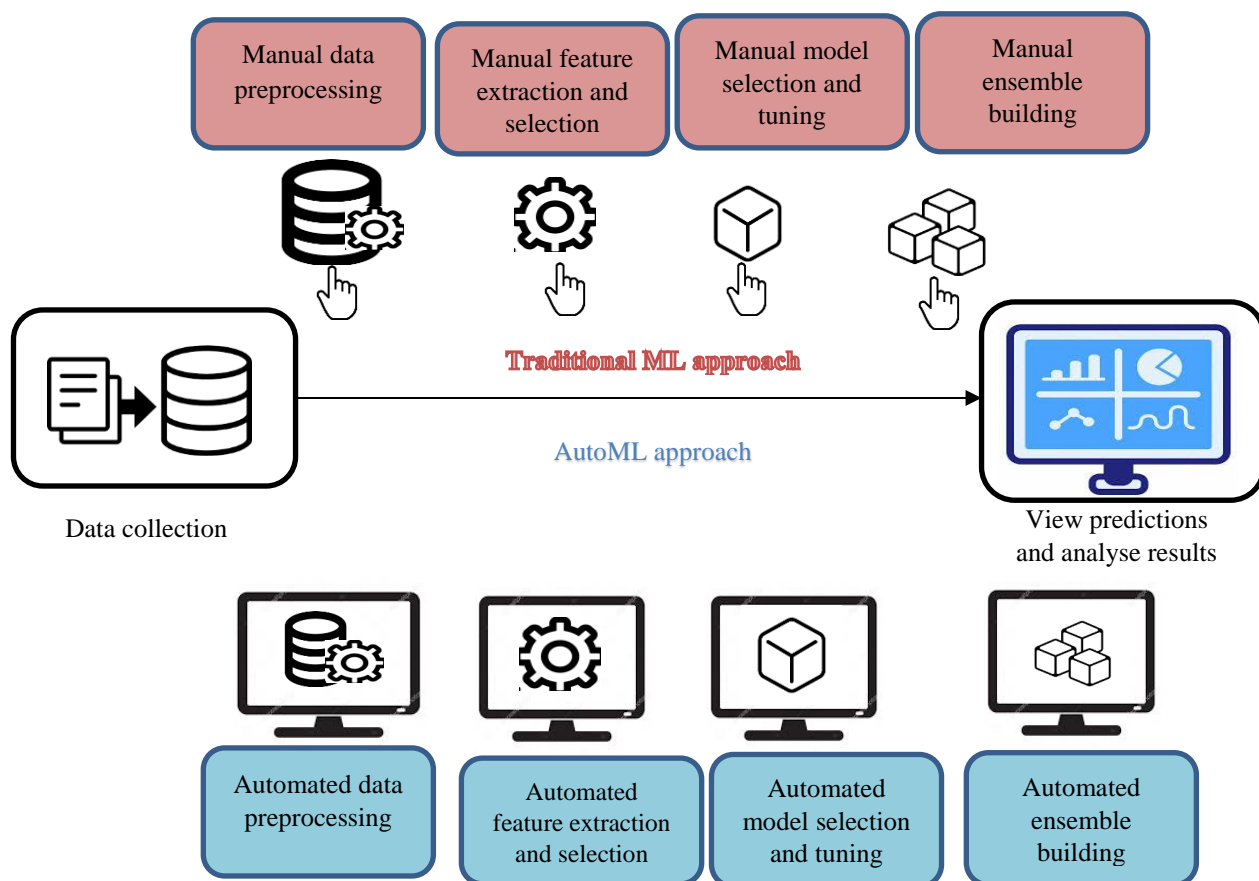
ML models are not made to operate in adverse environments or with adversaries. On the other hand, adversary attacks and hostile operating environments are common in cybersecurity applications. Malicious actors and attackers typically create adversarial samples to feed into machine learning models. The adversarial cases purposefully lead to an error in the model. More complex adversarial instances might fool the model into producing a desired output. Compared to manually created ML models by ML professionals, an adversary's chances of successfully reverse engineering AutoML-generated ML models are higher. The ease with which an adversary could use reconnaissance and probing techniques to discover the structure and characteristics of the AutoML model is a major security concern associated with AutoML.

The AutoML model can be replicated by the adversary through active or passive probing. This enables the attacker to thoroughly examine the ML model and find any potential weaknesses. Thus, handcrafted deep-learning models are more safe against adversarial attacks than models built using AutoML, whereas AutoML models are more susceptible to adversarial attacks.

The most common problems affecting data quality, particularly in cybersecurity applications, are labeling errors, observer bias, exclusion bias, and selection bias. As a result, cybersecurity datasets often have unbalanced, insufficient, and noisy training data [15]. The difficulties with

data quality and its effects on machine learning applications in cybersecurity are highlighted by several research in the literature. Despite the community's and cybersecurity researchers' ongoing attempts to provide new, pertinent, high-quality datasets, obtaining a representative training dataset remains extremely difficult. Model drift causes these fresh datasets to quickly become incomplete and irrelevant.

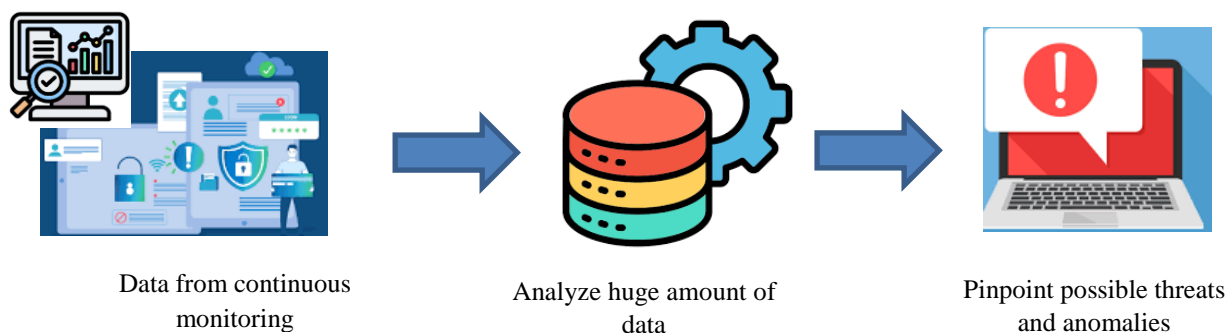
Apart from traditional techniques for managing unbalanced data, several recent studies suggested automated techniques to identify or fix problems with data quality in cybersecurity data sources. The framework uses a multifaceted method to assess the data's quality. The analyst must determine how to address the data quality problems that the framework identifies. Other recent studies suggested employing other techniques, like generative models and augmentation, to automatically generate high-quality datasets. We believe that AutoML tools should enable data quality assurance and assessment, even though we do not anticipate them to directly support data generation and collection. This is especially crucial for AutoML frameworks and applications that make use of meta-learning techniques.



**Figure 2.** A comparison between AutoML and standard machine learning techniques

Automated machine learning techniques have become strong substitutes for traditional machine learning by automating the ML workflow, from data preparation to model validation,

as illustrated in Figure 2. This removes the need for manual feature selection, improves performance, and adjusts to changing data properties.



**Figure 3.** The Role of AI in Security

**4. Implementation and Experimental Results**

The outcomes of the simulations run using the suggested AutoML pipeline are shown in this section. Table 1 displays the properties of the ten data sets that were extracted from the Kaggle ML database and used in the experiments using the Kaggle API.

**Table 1.** Possessions of used datasets

No	Dataset	Number of instances	Number of inputs
1	Diabetes	764	7
2	Surgical	14,621	22
3	Anemia	1410	5
4	Heart Attack	3588	12
5	Room Occupancy	2663	5
6	Blood Transfusion	742	3
7	Bank Note Authentication	1369	3
8	Ionosphere	342	32
9	Brain Tumor	33	7465
10	Phishing Website	1351	8

With the exception of the "Surgical" data set, which has a noticeably higher number of instances, nine out of ten datasets meet the aforementioned condition, as indicated in Table 1. Once more, the goal was to determine whether the quantity of input features had a substantial effect on the amount of time the pipeline needed to run the simulations. The fact that all data sets pertain to binary classification problems is now stressed.

**Table 2.** The hyper-parameter search space

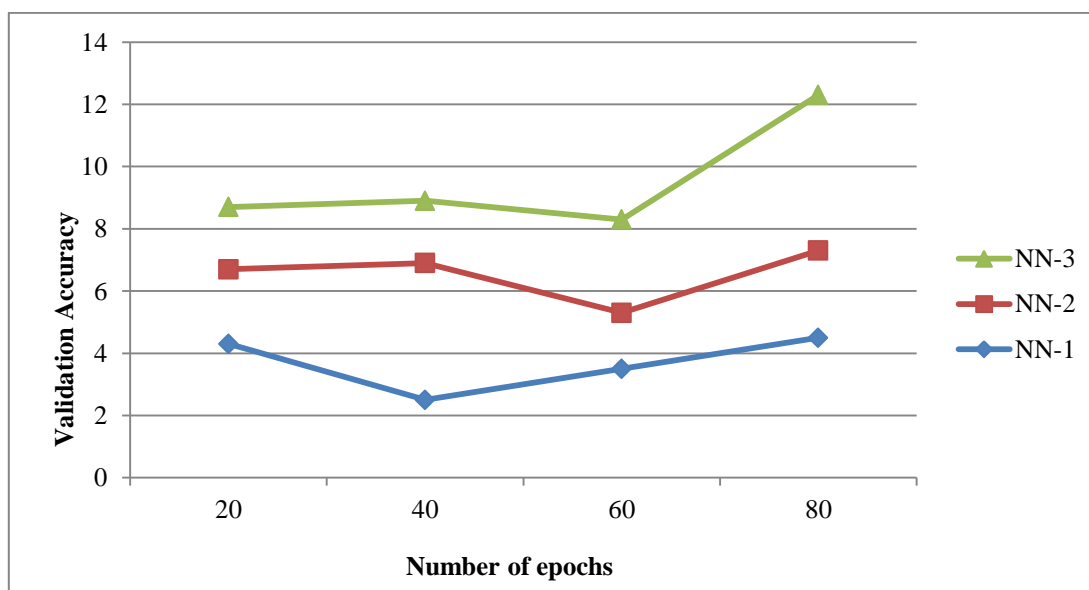
Hyper parameter	Domain of Values
Learning rate	{0.0001, 0.001, 0.01}
Number of layers	{3, 4, 5}
Number of neurons	with step 5
Types of activation function	“ReLU”, “Tanh”, “2nd order Hermite polynomial (H2)”

Keep in mind that according to Table 2, the domains of three of the four types of hyperparameters are discrete set values. Because it combines real and discrete domains, the hyperparameter search space is a hybrid space. Particularly noteworthy is the final parameter in Table 2.

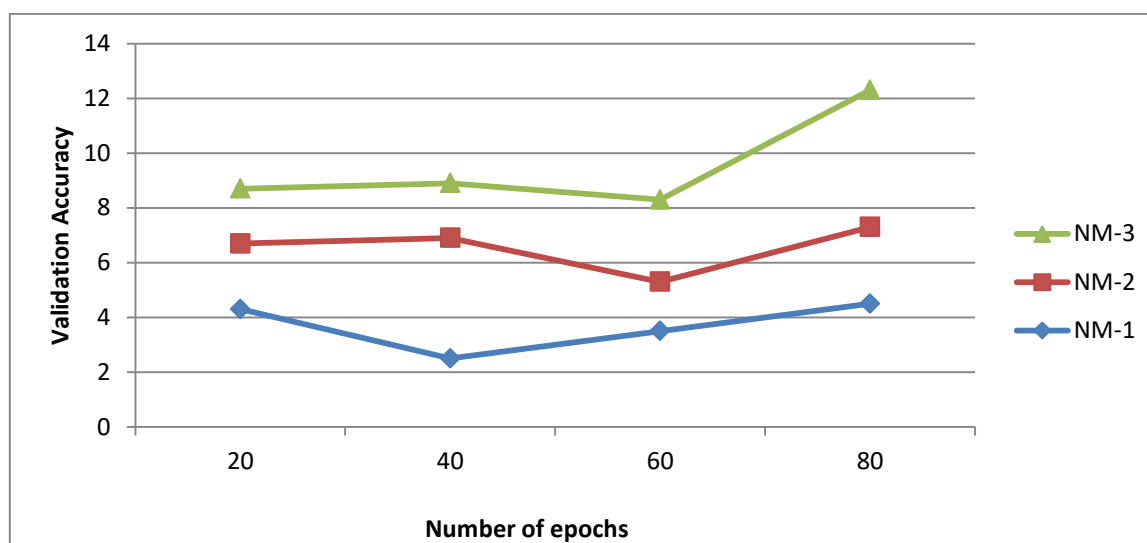
**Table 3.** Values allocated to the Keras Bayesian optimization tuner's inputs

Argument	Value
max_trials	100
num_initial_points	2
alpha	0.001
beta	2.6
seed	random.seed

Additionally, random seeding was employed to improve the optimizer's stochastic behavior across many runs. The values for the Keras Bayesian optimization tuner's parametric arguments, which were specified in Figure 4.



**Figure 4.** The NN\_1, NN\_2, and NN\_3 neural network models are the top three



**Figure 5.** (Second Experiment) The accuracy of classifications (i.e., validation) is assessed on the testing data in relation to the number of training epochs that the optimal neural network produced

As with the previous experiment, the learning rates that are obtained in the majority of cases are equal to 0.01; both models have the fewest number of layers (i.e., three layers) in the Banknote Authentication dataset; In light of this, the behaviors of the models shown in Figure 5.

## 5. Conclusion

The development of machine learning models is being revolutionized by the discipline of autoML. Businesses and individuals can more effectively leverage the potential of data and machine learning by using AutoML to automate the complex processes involved.

An AutoML pipeline was created in this paper to manage the creation and assessment of ML models. Utilizing the pipeline to carry out architectural and hyperparameter optimization was the primary responsibility. Through the use of Git and Jenkins technologies, the pipeline's implementation produced a quick model building process that was also easily replicable for other datasets. At the same time, it allowed our team members to collaborate synchronously and effectively.

One important experimental outcome was the measurement of the time difference between the manual execution of different steps in the overall process and their corresponding automated implementation under the scope of the proposed pipeline. This time difference is equal to the time saved by utilizing the recommended AutoML solution. The results illustrated the importance of this discrepancy, which is expected to increase dramatically when more complex machine learning models—such as deep learning models—and larger data sets are used.

**References**

- [1] Chou, A., Torres-Espin, A., Kyritsis, N., Huie, J. R., Khattry, S., Funk, J., ... & TRACK-SCI Investigators. (2022). Expert-augmented automated machine learning optimizes hemodynamic predictors of spinal cord injury outcome. *PloS one*, 17(4), e0265254.
- [2] Saad, S., Shi, K., Mamun, M., & Elmiligi, H. An Empirical Study on the Effectiveness of Automated Machine Learning Tools for Cybersecurity. Available at SSRN 5096180.
- [3] Xu, H., Sun, Z., Cao, Y., & Bilal, H. (2023). A data-driven approach for intrusion and anomaly detection using automated machine learning for the Internet of Things. *Soft Computing*, 27(19), 14469-14481.
- [4] Kundu, P. P., Anatharaman, L., & Truong-Huu, T. (2021, April). An empirical evaluation of automated machine learning techniques for malware detection. In *Proceedings of the 2021 ACM Workshop on Security and Privacy Analytics* (pp. 75-81).
- [5] Glavan, A. F., & Croitoru, V. (2023, June). Autoencoders and AutoML for intrusion detection. In *2023 15th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)* (pp. 1-4). IEEE.
- [6] Priyanka, E. B., Thangavel, S., Mohanasundaram, R., & Subramaniam, S. (2024). Artificial intelligence approaches in healthcare informatics toward advanced computation and analysis. *The Open Biomedical Engineering Journal*, 18(1).
- [7] Geetha, R., & Thilagam, T. (2021). A review on the effectiveness of machine learning and deep learning algorithms for cyber security. *Archives of Computational Methods in Engineering*, 28(4), 2861-2879.
- [8] Purwanto, R., Pal, A., Blair, A., & Jha, S. (2021). Man versus Machine: AutoML and Human Experts' Role in Phishing Detection. *arXiv preprint arXiv:2108.12193*.
- [9] Khan, M. I., Arif, A., & Khan, A. R. A. (2024). The Most Recent Advances and Uses of AI in Cybersecurity. *BULLET: Jurnal Multidisiplin Ilmu*, 3(4), 566-578.
- [10] Karamchand, G. K. (2023). Automating Cybersecurity with Machine Learning and Predictive Analytics. *Journal of Computational Innovation*, 3(1).
- [11] Verma, D. (2024). Enhancing Cybersecurity Through Adaptive Anomaly Detection Using Modern AI Techniques.
- [12] Saad, S., Shi, K., Mamun, M., & Elmiligi, H. An Empirical Study on the Effectiveness of Automated Machine Learning Tools for Cybersecurity. Available at SSRN 5096180.
- [13] Roshanaei, M., Khan, M. R., & Sylvester, N. N. (2024). Enhancing cybersecurity through AI and ML: Strategies, challenges, and future directions. *Journal of Information Security*, 15(3), 320-339.
- [14] Thapaliya, S., & Bokani, A. (2024). Leveraging artificial intelligence for enhanced cybersecurity: Insights and innovations. *Sadgamaya*, 1(1), 46-52.
- [15] Sezgin, A., & Boyacı, A. (2023). AID4I: An Intrusion Detection Framework for Industrial Internet of Things Using Automated Machine Learning. *Computers, Materials & Continua*, 76(2).