IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

An Organized Analysis Of Cloud Computing Environment Anomaly Detection

[1]Sobhitha S P,[2]Ms.Aparna A
[1]MCA Scholar,[2]Assistant Professor
[1] Department of MCA,
[1] Nehru College of Engineering and Research Centre, Pampady,India

Abstract: The technique for anomaly detection in cloud-based networks is fundamental. Its usage is more specific to detecting intrusion and hardware failures, among other applications. The literature review has drawn 215 papers from the past decade focused on machine learning, deep learning, and statistical methods for anomaly detection in a cloud environment. Techniques of machine learning include three types: supervised, unsupervised, and semi-supervised learning for pattern identification; deep learning techniques include CNNs and RNNs, which are more accurate and efficient. Statistical methods are mathematical and probabilistic in nature and detect anomalies from the standard behavior. This review is based on important contributions, new research trends, and potential future directions for the improvement of the effectiveness and reliability of anomaly detection in cloud computing.

Index Terms **- Anomaly detection, Intrusion detection, Cloud computing, Machine learning.**

I. Introduction

Anomaly detection is crucial for the cloud computing environment for issues such as cyber intrusions, malware attacks, hardware malfunctions, and degradation in performance. Following an increased adoption of cloud services, early and accurate detection becomes increasingly important over time. Most significant research has started nearly a decade ago on techniques for anomaly detection in cloud computing, and there is no complete review available in the literature. This review fills the gap by examining 215 publications of the past ten years on machine learning, deep learning, and statistical approaches. Techniques of machine learning include supervised, unsupervised, and semi-supervised learning that identify anomalies by data patterns and relations. Deep learning methods like CNNs and RNNs provide better accuracy and efficiency. Anomalies in behavior are determined using statistical methods, based on mathematical and probabilistic methods. This review of key publications, identification of research trends, and insights into future directions may enhance the security and reliability of cloud-based networks.

II. LITERATURE SURVEY

Alarqan and Zaaba (2020), "Detection Mechanisms of DDoS Attack in Cloud Computing Environment: A Survey" Explores
methods to detect and eliminate DDoS attacks through statistical anomaly-based techniques toward the real-time detection and mitigation. These include a judgment of deviations from regular network traffic patterns as signatures of potential DDoS attacks. Early detection is necessary to neutralize attacks before considerable disruption is done, which preserves cloud systems' integrity, availability, and performance.

Faisal Shahzad et al. (2022), "Cloud-based Multiclass Anomaly Detection and Categorization Using Ensemble Learning" An ensemble machine learning-based cloud anomaly detection using CNN-LSTM proposed, which had high accuracy regarding anomaly detection, and CAD can be divided into two main blocks: an ensemble machine learning-based model for a binary classification anomaly and a CNN-LSTM to categorize a multiclass anomaly. The study tests CAD on a challenging UNSW dataset and reveals that CAD excels over other models by achieving a maximum accuracy of 97.06% in binary anomaly detection and 99.91% in multiclass anomaly detection.

Dehraj and Sharma (2020), "A Review on Architecture and Models for Autonomic Software Systems" provides an insightful survey on the issue of autonomic decision-making across various applications such as intrusion detection, cloud-based data security, and the Internet of Things. They emphasized the significance of the "autonomic computing" which develops system efficiency and reliability through self-configuration, self-healing, self-optimization, and self-protection. These capabilities are crucial in developing autonomously manageable complex systems, without constant human oversight.

M. P. G. K. Jayaweera et al. (2023), "Detect Anomalies in Cloud Platforms by Using Network Data: A Review" This paper gives a comprehensive review of various techniques for detecting anomalies in cloud platforms using network data. The study highlights the efficacy of machine learning methods in anomaly detection and aims to identify the best-suited algorithm for analyzing cloud network data. It systematically reviews scholarly articles published between 2017 and 2023, discussing different approaches and techniques for anomaly detection in cloud environments.

Ramachandra et al., 2020, "Literature Survey on Log-Based Anomaly Detection Framework in Cloud", Discusses applying log data to anomalous detection mechanisms in cloud and points out this use for faulting system glitches or security attack issues. Specifically, the paper reports on intrusion-based anomaly detection techniques as well as security anomalies. It concludes that although log data is vital in the detection of some anomalies, no thorough survey covers the whole range of anomaly detection methods in cloud infrastructures.

III. RESEARCH METHODOLOGY

3.1 Research Questions

This survey explores the state-of-the-art in cloud computing environments for anomaly detection through research questions into three different scientific databases—SpringerLink, Web of Science (WoS), and the open-access archive ArXiv. These databases embrace the different dimensions of the research being pursued. The research questions put forth are as follows:
What anomalies can be detected in a cloud environment?
What are the specific objectives of anomaly detection in a cloud environment?
How has this field of study evolved with time?
The text mining techniques are used to get a thematic clustering of the output papers in order to complement the manual query pipeline.

3.2 Identification and Selection

3.2.1 Search Terms

The literature review is to gather and evaluate the works that apply anomaly detection techniques to identify the trends in cloud-based infrastructure. Two search terms are applied :
Cloud Computing and Anomaly Detection
Cloud Monitoring and Anomaly Detection

3.2.2 Inclusion Criteria

Focus on Unique Approaches: Exclude full book chapters or survey studies to focus on unique, well-defined approaches and their evaluation.
Publisher Inclusion: WoS contains publications from a number of publishers, including Springer.
Duplicate Elimination: Eliminate duplicates based on titles or DOI.

Manual Screening: Screen the papers generated manually and exclude those that do not meet the evaluation criteria.

A total of 215 relevant articles were obtained, which are discussed and evaluated according to the specific issues addressed and the methods used to address them.

3.3. Synthesis and Analysis

This statistical analysis of identified references suggests that the search returned virtually the same percentage of references (~45%) from both SpringerLink and Web of Science (WoS), which excluded SpringerLink. In contrast, the proportion of identified publications in ArXiv was much smaller (9.9%), which might have resulted from its relatively smaller size. Figure 5 illustrates a more in-depth analysis of the results from WoS, grouped by areas of application, and Figure 4 illustrates the chronological distribution of the publications. Analysis also indicates that the oldest relevant publication is from 2011, marking the beginning of the observed dataset. Publication numbers over time have maintained a trend of upward change, positive. The increase has been steady up to 2019. It then continued upward through May of 2020. According to this pattern, a new high in the number of publications is expected in 2020. Continuous growth in relevant research can be attributed to several key factors. The most prominent are the recent eagerness and proliferation of cloud-based infrastructures, providing scalable and flexible R&D platforms. Moreover, the progress in state-of-the-art machine learning technologies has opened an avenue for exploring new, innovative solutions and, subsequently, creating new possibilities across various applications. The proliferation of use cases and methodologies represents growing interest and activity in this field, further explaining consistent scholarly contributions over time.

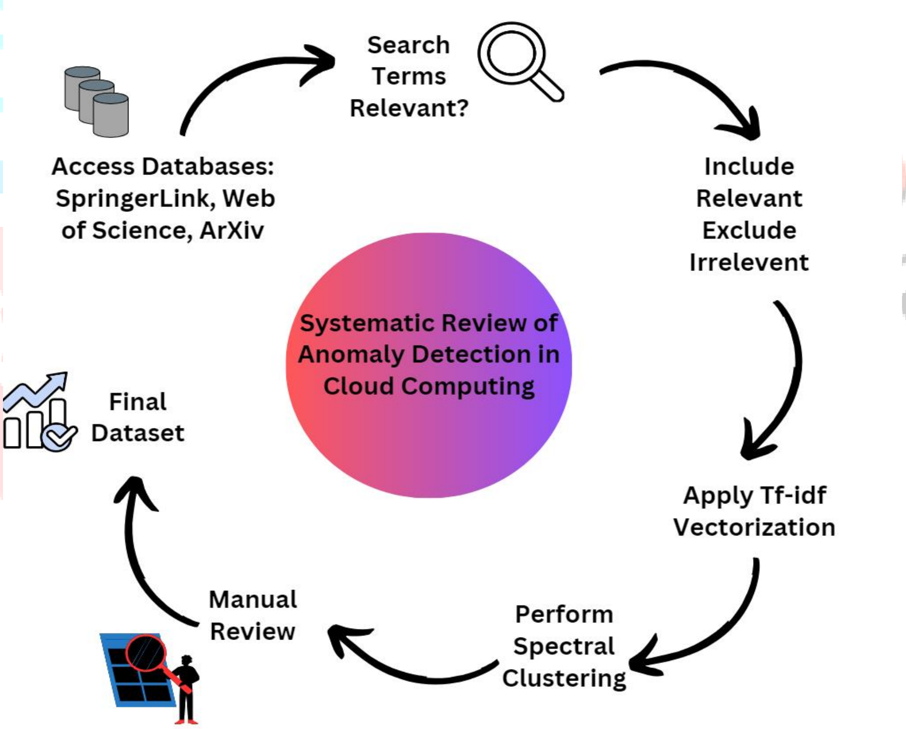
Search Terms Relevant?
Include Relevant Exclude Irrelevent
Apply Tf-idf Vectorization
Perform Spectral Clustering
Manual Review
Final Dataset
Access Databases: SpringerLink, Web of Science, ArXiv
Systematic Review of Anomaly Detection in Cloud Computing

Figure 1 : Steps of Methodology

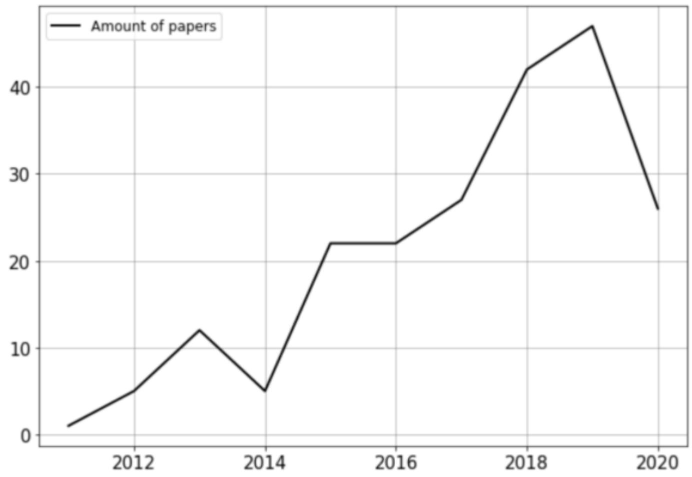

Figure 2: The total number of reviewed Papers dispersed among dataset

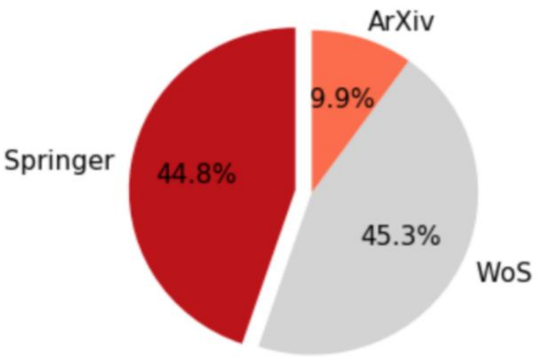

Figure 3 : Total number of publications published after all reviews. Keep in mind that the 2020 results are only available through May.

IV. RESULTS AND DISCUSSION

Three major categories of anomaly detection techniques were identified in the literature review: statistical methods (23%), deep learning (19.7%), and traditional machine learning (28.3%). These domains sometimes overlap because many approaches are often combined or contrasted with each other. The dominant model that detects anomalies is emphasized when multiple approaches are used. Also, the methods developed with one architecture cannot be easily applied to others; that happens in 27.9% of articles.

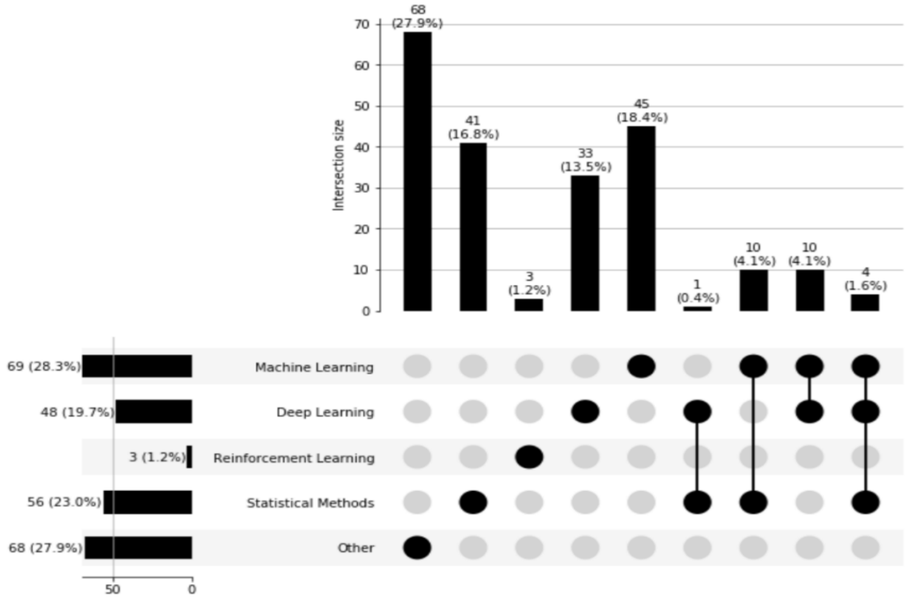

Figure 4 : The distribution of the identified methodological domains (left) across the 215 examined papers is displayed in an upset plot (horizontal bars). The quantity of articles primarily employing techniques from a single field, denoted by the black dot, is displayed in the first five vertical bars at the top. The total number of publications addressing multiple approaches from various domains is indicated by the final four vertical bars.

4.1 Classical Machine Learning

Classical machine learning includes techniques based on the assumption of data learning and the formation of rules instead of using predefined rules. They have been mainly classified into two categories: unsupervised and supervised learning. Here, models are learned without being tagged and only try to find a pattern and correlation within the data. A good example is clustering algorithms, which combine similar data points into groups and outliers as those data points that do not fit into any formed cluster. Clustering can either be soft, where data points can belong to more than one cluster, or hard, where every data point falls into only one cluster. Supervised learning, in contrast, focuses on training models using labeled data, making it a prediction-oriented field. It comprises of some examples, including SVM-the algorithm working to find the best boundary which would separate some classes. Techniques are k-NN, or classification based upon the nearest neighbors, and tree-based approaches which indeed decide using the structure of the decision trees and forests. Applied neural network methods and ensemble approaches, which comprise boosting, bagging, etc., come in supervised learning. These methods have widely been applied in various fields such as image and speech recognition, natural language processing, and financial forecasting. In a nutshell, there are different strategies of supervised and unsupervised learning which are used to find patterns and predict outcomes in traditional machine learning. It shows how diverse and effective it is in solving complex problems. Classical machine learning methods are very robust and versatile with a wide application in tasks. They are helpful in providing foundational knowledge to advance the understanding of more complex techniques of machine learning. These versatile techniques can be adapted to different application domains and industries, making them incredibly versatile. The basis of classical machine learning methods is through data-driven insight, which identifies hidden patterns in datasets and unseen relationships. This becomes crucial for prediction models and hence, in deriving a data-informed decision. A classical machine learning model is extremely effective in most practical applications where it is useful for solving realistic problems. Due to this advantage, with evolving technologies and time, they are not getting out of importance. Instead, these will act as simple as well as rugged tools in conducting analysis on large amounts of data and making proper predictions. Ensuing developments in this field ensure these classical machine learning methods remain continuously relevant and very useful in contemporary artificial intelligence disciplines. Combining supervised and unsupervised techniques would form a fundamental approach to doing data analysis on classical machine learning. This implies that complex relationships and future prospects can be revealed with high accuracies. Classical machine learning techniques are basically the backbone to the advancement of artificial intelligence and the development of intelligent systems. They are pivotal in shaping the future of technology and innovation.

4.2 Deep Learning

Deep learning involves using the neural network in its multi-layer form called "hidden layers". The networks extract feature information from the raw data and develop a succession of representations on top of each other. Unsupervised learning within deep learning often incorporates an autoencoder. An autoencoder involves an encoder that modifies the input data and a decoder that reconstructs the original form. Reconstruction errors can sometimes reflect anomalies while Recurrent Neural Networks are specially adapted for sequential data like time-series. They look at the entire sequence and guess the next member in the chain as well as anomaly detection capabilities. Long Short-Term Memory (LSTM) networks specialize in overcoming limitations of RNNs, notably the vanishing gradient problem. Self-Organizing Maps look for common aspects in data that allow for simplification of information to a small low-dimensional grid for increased anomaly detection performance. Deep learning supervised techniques include MLP, which behaves similarly to logistic regression classifiers, and CNNs, where convolutional layers reduce the feature space before output classification, typically on grid-like inputs such as images or time series. Extreme Learning Machines is effective for regression or classification applications because the hidden neurons are randomly assigned and not updated in training. Deep learning uses multilayered networks to detect anomalies in data and complex patterns. Techniques are specialized to specific problems and types of data, but these generally indicate application in areas like image recognition, speech recognition, natural language processing, and financial forecasting. For instance, in the above sets of results, deep learning succeeded as shown. Therefore, there is a lot deep learning can offer on complex problems. Deep learning models can obtain outstanding precision and performance using very large datasets and big computer powers. Advances in hardware, especially GPUs and TPUs, have pushed deep learning development. It is with the possibility of training large neural networks that advances in hardware provided. Huge datasets also played a major role in the success of deep learning models.

4.3 Statistical Methods

Statistical methods include a wide scope such as regression analysis, probabilistic graphical models, and time series analysis methods, which are based on probability. Such methods form very strong frameworks for the identification and interpretation of anomalies in complex data. Time series analysis is considerably an applied technique when the process is being tracked or monitored in terms of metrics over time. It accounts for the underlying structure that comes with the collection of data over time, which may include autocorrelation, trends, and seasonal patterns. As with LSTMs-based methods, time series analysis, involves a comparison of forecasted values and actual occurrences in which anomalies can be identified as less frequent patterns previously. Another important class of statistical techniques is the Bayesian ones, all stemming from Bayes' rule. Bayesian classifiers classify the most probable class for a given example with respect to its feature vector. Bayesian networks are probabilistic graphical models that are also referred to as belief networks or Bayes nets. Anomaly detection can be achieved by learning a joint probability distribution and finding anomalous states within it. Regression analysis has been an accepted statistical method in learning methods estimating how a given set of independent variables depends upon another, different set of real, dependent values. In other words, it lets one know of the changes caused in the dependent variable due to alteration in independent variables. It would therefore play a very critical role in this anomaly detection due to its usage in predicting modern values of yet-to-happen events. With these predicted and observed deviations, regression analysis identifies anomalies that characterize data points with patterns of inconsistency. These methods provide strong frameworks for anomaly identification and interpretation in complex data. Time series analysis is a highly applied technique, especially when the process is being tracked or monitored in terms of metrics over time. Bayesian methods are another important class of statistical techniques, all based on Bayes' rule. Regression analysis is a well-known statistical method in learning methods that helps in estimating how a set of independent variables depends on another set of actual, dependent values. Essentially, it provides insights into changes in the dependent variable resulting from changes in independent variables. It would therefore be instrumental for anomaly detection since it also helps to predict up-to-date values of future events. With these predicted and observed deviations, regression analysis identifies anomalies that characterize data points with patterns of inconsistency.

V. CHALLENGES AND FUTUREWORK

- Simulating cloud environments for reinforcement learning.
- Detecting real vs. fake data in GANs.
- Evaluating attention mechanisms for sequential cloud data.
- Graph-based modeling of topological information.
- Overcoming the labeling bottleneck in active learning.
- Preventing targeted errors in neural networks in adversarial learning.
- Interpretability in Explainable AI (XAI).

Future work must address these challenges to enhance anomaly detection in cloud computing.

VI. CONCLUSION

Altogether, the overview of 215 articles on the anomaly detection techniques in cloud computing reveals the comparative strengths and weaknesses among machine learning and deep learning or statistical approaches in intrusion detection performance monitoring and failure detection. Machine learning, especially techniques like supervised learning and unsupervised learning can effectively identify a pattern in the data. CNNs and RNNs are examples of deep learning models that have shown to have the highest accuracy and efficiency. In contrast, robust frameworks are presented in statistical methods in anomaly detection within cloud systems. The comprehensive review overall provides the key contributions, identifies gaps in research, and gives guidance for further advancements to be made in terms of security, performance, and reliability in cloud-based networks.

VII. REFERENCES

[1] AC Ramachandra, Subhrajit Bhattacharya, et al. 2020. Literature Survey on LogBased Anomaly Detection Framework in Cloud. In Computational Intelligence in Pattern Recognition. Springer, 143–153.

[2] S Sandosh, V Govindasamy, and G Akila. 2020. Enhanced intrusion detection system via agent clustering and classification based on outlier detection. Peerto-Peer Networking and Applications (2020), 1–8.

[3] Siva Rama Krishna Tummalapalli and ASN Chakravarthy. 2020. Intrusion detection system for cloud forensics using bayesian fuzzy clustering and optimization based SVNN. Evolutionary Intelligence (2020), 1–11.

[4] "Detection Mechanisms of DDoS Attack in Cloud Computing Environment: A Survey" by Alarqan and Zaaba (2020)

[5] "Cloud-based Multiclass Anomaly Detection and Categorization Using Ensemble Learning" by Faisal Shahzad et al. (2022)

[6] "A Review on Architecture and Models for Autonomic Software Systems" by Dehraj and Sharma (2020)

[7] "Detect Anomalies in Cloud Platforms by Using Network Data: A Review" by M. P. G. K. Jayaweera, W. M. C. J. T. Kithulwatta, and R. M. K. T. Rathnayaka (2023)

[8] "Literature Survey on Log-Based Anomaly Detection Framework in Cloud" by Ramachandra et al. (2020).

[9] Sahil Garg, Kuljeet Kaur, Shalini Batra, Gagangeet Singh Aujla, Graham Morgan, Neeraj Kumar, Albert Y Zomaya, and Rajiv Ranjan. 2020. En-ABC: An ensemble artificial bee colony based anomaly detection scheme for cloud environment. J. Parallel and Distrib. Comput. 135 (2020), 219–233.

[10] Nurudeen Mahmud Ibrahim and Anazida Zainal. 2020. A Distributed Intrusion Detection Scheme for Cloud Computing. International Journal of Distributed Systems and Technologies (IJDST) 11, 1 (2020), 68–82.

[11] Mohammad Islam and Andriy Miranskyy. 2020. Anomaly Detection in Cloud Components.

[12] Aws Naser Jaber and Shafiq Ul Rehman. 2020. FCM–SVM based intrusion detection system for cloud computing environment. Cluster Computing (2020), 1–11.

[13] Souhila Benmakrelouf, Cédric St-Onge, Nadjia Kara, Hanine Tout, Claes Edstrom, and Yves Lemieux. 2020. Abnormal behavior detection using resource level to service level metrics mapping in virtualized systems. Future Generation Computer Systems 102 (2020), 680–700.

[14] Lelio Campanile, Mauro Iacono, Fabio Martinelli, Fiammetta Marulli, Michele Mastroianni, Francesco Mercaldo, and Antonella Santone. 2020. Towards the Use of Generative Adversarial Neural Networks to Attack Online Resources. In Workshops of the International Conference on Advanced Information Networking and Applications. Springer, 890–901.

[15] Marta Catillo, Massimiliano Rak, and Umberto Villano. 2020. 2L-ZED-IDS: A Two-Level Anomaly Detector for Multiple Attack Classes. In Workshops of the International Conference on Advanced Information Networking and Applications. Springer, 687–696.

[16] Hongyang Chen, Pengfei Chen, and Guangba Yu. 2020. A Framework of Virtual War Room and Matrix Sketch-Based Streaming Anomaly Detection for Microservice Systems. IEEE Access 8 (2020), 43413–43426.