



IJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: <https://doi.org/10.22214/ijraset.2026.79052>

www.ijraset.com

Call:  08813907089

E-mail ID: ijraset@gmail.com

Multi-Agent Generative AI Framework for Cloud Cost Prediction and Intelligent Resource Optimization

Vinoth kumar G¹, Narthana Hari S V², Samyucthan M³, Angela Lincy N H⁴

^{1,2,3}Final Year, ⁴Assistant Professor, Dept. of AI & DS, MNM Jain Engineering College, Chennai, India

Abstract: *The rapid adoption of cloud computing services such as Amazon Web Services, Microsoft Azure, and Google Cloud Platform has transformed enterprise infrastructure management by enabling elastic resource provisioning. However, dynamic resource allocation mechanisms and complex pricing models often lead to unpredictable cloud expenditures and budget overruns. Conventional cloud monitoring dashboards primarily provide descriptive analytics and lack predictive insights, anomaly detection, and automated optimization capabilities.*

This paper presents a Multi-Agent Generative Artificial Intelligence framework for proactive cloud cost prediction and intelligent resource optimization. The proposed system consists of five cooperative agents, including a Data Ingestion Agent, Cost Prediction Agent, Anomaly Detection Agent, Optimization Recommendation Agent, and an LLM-Based Report Generation Agent. A Long Short-Term Memory (LSTM) network is employed for multivariate time-series cost forecasting, while an Isolation Forest model is utilized to identify abnormal billing patterns. A risk-aware optimization mechanism translates analytical insights into actionable cost-saving recommendations, and a Generative AI-based Large Language Model produces explainable financial reports for decision-makers.

Additionally, ensemble learning models including Random Forest and XGBoost are integrated using a stacking-based meta model to enhance prediction robustness.

Experimental results indicate high forecasting accuracy based on the coefficient of determination, robust anomaly detection with minimal false alerts, and efficient end-to-end system response, validating the framework's scalability, interpretability, and suitability for production deployment.

Index Terms: *Cloud Computing, Cloud Cost Prediction, Generative Artificial Intelligence, Multi-Agent Systems, LSTM, Isolation Forest, FinOps, Explainable AI, Resource Optimization*

I. INTRODUCTION

Cloud computing has become a key part of changing how industries operate digitally. More and more companies are using public cloud services to run their apps, keep their data safe, and handle growing infrastructure needs, all without having to spend a lot of money upfront. The pay-as-you-go pricing model offers flexibility, letting businesses adjust their resources according to how much work they need to handle at any time. This flexibility has made it easier to respond quickly to changes and has simplified the management of the infrastructure.

Even with these benefits, managing finances in the cloud has become a major issue. The way resources are provided, how systems automatically adjust to demand, and how services are used can all lead to costs that are hard to predict. When too many resources are given, not used, or set up wrong, costs can go up quickly and it's hard to see how much is being spent. In big environments where many services and teams work at the same time, it's harder to find exactly where the costs are going up.

Many companies face unexpected jumps in their bills because of sudden increases in workloads, poorly set up automatic scaling settings, or services that were started without being noticed. Costs often go up slowly because storage keeps growing without control and because the rules for managing how long data is kept are not working well. Cloud service providers provide cost monitoring dashboards, but these tools mainly show past spending information. They don't often give predictions or smart suggestions that help control costs before problems happen.

Another major issue is that the process for checking costs is done by hand. Financial and DevOps teams usually rely on spreadsheets and regular check-ins to look at billing information, which takes a lot of time and can lead to mistakes being missed.

Without using predictive modeling and automated anomaly detection, companies have to make decisions after problems happen, like fixing budget issues only after they've gone over.

New developments in machine learning and artificial intelligence offer a chance to greatly improve how financial management works in the cloud. Predictive models can predict how much money will be spent by looking at past spending habits, and anomaly detection can spot unusual billing activity as it happens. When used with automated optimization strategies and clear reporting tools, these technologies can make things more transparent and help make better decisions faster.

To address these challenges, this research introduces a multi-agent artificial intelligence system aimed at providing predictive, diagnostic, and prescriptive insights into cloud costs. By combining time-series forecasting, anomaly detection, risk-aware optimization, and generative reporting into one system, the framework is designed to help with proactive and sustainable management of cloud finances.

II. RELATED WORK

In recent years, there has been a lot of research on cloud cost management, choosing the right cloud provider, and optimizing services smartly, because multi-cloud environments are getting more complicated. Many research efforts have looked into ways to make things more standard, how to connect different parts of the system, better ways to make decisions, and using smart automation to improve how financial systems are managed and how they run day to day.

The approach handles the differences in how various cloud providers charge for their services by combining the FinOps Open Cost and Usage Specification (FOCUS) with Large Language Models (LLMs). The authors explain that cloud companies like AWS, Azure, and GCP create billing information in their own special formats, and these formats can have different levels of detail and ways of showing costs, which makes it hard to compare data across different providers. OPTIC offers an automated ETL process that takes cost data and puts it into a common SQL structure, making it easier to work with. It also lets users ask questions in everyday language, which is then translated into SQL queries using a language model. The framework greatly helps with making FinOps environments more accessible and financially clear, but its main benefit is in translating and handling billing data instead of predicting future costs or optimizing based on unusual activity. Unlike other approaches, the framework introduced in this study does more than just translate. It also includes tools for predicting future trends, identifying unusual patterns, and making decisions that take risks into account. The role of cloud brokers in automatically distributing services has been thoroughly studied in A Systematic Literature Review. The authors use a systematic literature review based on the PRISMA method to examine brokerages systems that help with choosing the right services, balancing workloads, and lowering costs. Their work focuses on autonomous computing ideas like self-configuration, self-optimization, and self-healing to improve service efficiency in multi-cloud environments. Cloud brokers help make services easier to manage and resources used more efficiently. However, most current models for cloud brokering mainly focus on spreading out workloads and making responses faster, instead of accurately predicting costs in detail or providing smart financial insights based on unusual activity. This shows the gap in research that this work aims to fill, focusing on financial analytics and predictive intelligence as key parts.

Choosing a cloud provider has also been seen as a complicated decision that involves considering many factors, as explained in Cloud provider selection a comp. The authors consider cost, performance, and evaluation risk as different goals that may conflict with each other and use methods from multi-objective optimization to model how to choose the best approach. Their method shows that choosing a provider requires balancing the need to save money with the need to get the best performance. However, these models usually work during the strategic selection phase and don't keep tracking or forecasting changes in billing once they are put into use. This study adds to those approaches by using real-time predictions and tools to spot unusual patterns, helping with managing costs as things happen. So far, earlier work has looked at standardization, ways to connect different parts, and choosing the best providers on their own. However, not much work brings together predictive modeling, anomaly detection, optimization suggestions, and explainable reporting into one single system. The new multi-agent generative AI system is designed to fill this gap by bringing together financial predictions, smart identification of unusual activities, and practical ways to improve performance, all within a system that can grow with a business and clearly explain how it works, making it ideal for large companies managing their financial operations.

III. PROPOSED SYSTEM ARCHITECTURE AND METHODOLOGY

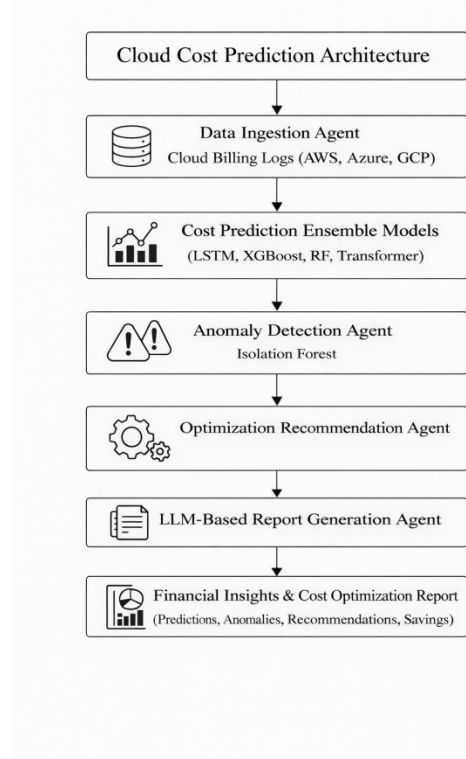


Fig. 1. System Architecture of the Proposed Multi-Agent Generative AI Framework for Cloud Cost Prediction

A. System Overview

The proposed framework adopts a distributed multi-agent architecture in which specialized agents collaborate through API-based communication. Initially, cloud billing and resource utilization data are ingested and preprocessed to ensure consistency and analytical reliability. The processed data are then utilized by the Cost Prediction Agent to generate future cost estimates using an ensemble of prediction models including LSTM, XGBoost, and Random Forest. The outputs of these models are combined using a stacking-based metalearner to improve forecasting accuracy. Subsequently, the Anomaly Detection Agent analyzes billing patterns to identify abnormal cost behaviors. The predictive deviations and anomaly signals are aggregated into a risk score that guides the Optimization Recommendation Agent in generating cost-efficient resource management strategies. Finally, the LLM-Based Report Generation Agent transforms numerical analytics into structured, explainable, and human-readable financial reports.

B. Data Ingestion Agent

The Data Ingestion Agent aggregates heterogeneous billing and operational metrics from multiple cloud service providers, including AWS, Microsoft Azure, and Google Cloud Platform. The collected data consist of monthly billing statements, service-level cost breakdowns, and fine-grained resource utilization metrics such as CPU usage, memory allocation, storage consumption, and network bandwidth.

To ensure data reliability, preprocessing steps such as missing value imputation, feature normalization, and sliding window transformation are applied. Minor statistical noise is smoothed to maintain data stability without suppressing genuine anomalies. This preprocessing pipeline enables seamless integration of raw billing data into forecasting and anomaly detection modules.

C. Cost Prediction Agent

Cloud cost prediction is formulated as a multivariate time-series regression problem in which historical resource utilization patterns are mapped to future billing costs. Let $X_{t-k:t}$ represent historical usage features over a sliding window of length k , and let C_{t+1} denote the predicted cost for the next billing cycle.

The forecasting objective is to minimize the Mean Squared Error loss between actual and predicted costs. To effectively capture nonlinear temporal dependencies, a Long Short-Term Memory neural network is employed. The LSTM architecture mitigates vanishing gradient issues and preserves long-term contextual information through gated memory mechanisms. This enables accurate modeling of seasonal billing cycles, workload growth trends, and long-term cost dependencies.

D. Anomaly Detection Agent

The Anomaly Detection Agent identifies abnormal billing patterns using the Isolation Forest algorithm, an unsupervised learning technique well-suited for high-dimensional data. The model assigns anomaly scores based on the average path length required to isolate data points within randomly constructed trees. Shorter isolation paths indicate a higher likelihood of anomalous behavior.

This approach effectively detects sudden cost escalations, unexpected service activations, and configuration-induced cost leakages. Unlike threshold-based methods, Isolation Forest dynamically adapts to evolving billing distributions, thereby reducing false positives while maintaining detection sensitivity.

E. Optimization Recommendation Agent

The Optimization Recommendation Agent translates predictive deviations and anomaly signals into actionable cost-saving strategies using a weighted risk aggregation model. The risk score incorporates prediction error magnitude, anomaly severity, resource usage growth rate, and cost variance. Based on the assessed risk level, the system recommends optimization actions such as instance rightsizing, termination of idle resources, adoption of reserved pricing models, and implementation of storage lifecycle policies.

This risk-aware optimization mechanism ensures that corrective actions are financially prioritized and aligned with enterprise FinOps governance principles.

F. LLM-Based Report Generation Agent

The LLM-Based Report Generation Agent converts analytical outputs into comprehensive financial intelligence reports. These reports include cost forecasts, explanations of detected anomalies, budget impact assessments, and clearly articulated optimization recommendations. The use of Generative AI enhances interpretability for non-technical stakeholders, thereby improving transparency and accelerating decision-making processes.

IV. EXPERIMENTAL SETUP AND EVALUATION

A. Dataset Description

The experimental testing of the proposed multi-agent framework was done using a detailed dataset that includes 24 months of cloud billing information from various cloud service types. The dataset has detailed information about the costs of different services and includes precise data on how resources are used, like how much CPU, memory, storage, and network bandwidth each virtual machine is using. These features were chosen to show the main reasons why cloud costs rise in actual business settings.

To make the data more like real-world situations, synthetic seasonal patterns were introduced into the dataset. These patterns help show how workloads change over time in real systems, such as monthly reports, busy times when lots of people use the system, and heavy workloads at the end of each quarter. Additionally, controlled spikes in anomalies were added to mimic abnormal billing issues that happen because of wrong setup in auto-scaling, unexpected service starts, poor storage management, and sudden increases in workloads. This approach allowed for a thorough check of the anomaly detection system while keeping the real statistical features of actual cloud usage data intact.

B. Data Preprocessing and Preparation

Before training the model, the billing data was carefully prepared through a thorough process to make sure it was reliable and consistent over time. Missing values in the billing and usage records were filled in using statistical methods to avoid breaks in the time-series data. Numbers were adjusted to the same scale to help the model work smoothly and prevent some numbers from having too much influence.

The historical billing data were turned into overlapping sliding time windows to create supervised learning sequences that are good for forecasting multivariate time series. The data was split in order over time to separate the training and testing sets, making sure that information from the future wasn't included in the training data by accident. This evaluation method is very similar to how things work in the real world, where predictions have to be made only based on past data.

C. Evaluation Metrics

The performance of the cost forecasting part was checked using common regression measurements that are usually used in time-series prediction work. Mean Absolute Error (MAE) was used to measure the averaged difference between the predicted cloud costs and the real costs, giving a clear idea of how accurate the predictions were. Root Mean Squared Error (RMSE) was used to highlight bigger mistakes and check how well the model works when billing conditions are unstable. The R-squared value was used to show how much of the variation in cloud costs is explained by the forecasting model, which helps to assess how well the model predicts outcomes. The system checked how well it found anomalies by using classification measures like precision, recall, and F1 score. Precision looked at how many of the flagged billing issues were actually correct, while recall measured how well the model found all the real abnormal costs. The F1-score gave a fair way to check how well detection works by balancing precision and recall. Also, the false positive rate was checked to understand how often wrong anomaly alerts happen. This is especially important in big companies to prevent too many alerts that can make people ignore real issues and keep trust in the automatic monitoring tools.

D. System Performance Metrics

Besides checking how well the system can predict and detect things, we also looked at other performance measures at the system level to see if the proposed framework is practical to use in real situations. The time it took to create cost predictions and calculate anomaly scores for new billing data was measured as the detection latency. This measure shows how quickly the system can respond in situations where real-time monitoring is needed.

The time it took the LLM-based reporting agent to generate the report was also checked to see how efficiently it could turn analytical results into clear and organized financial summaries that people can easily read. These performance metrics together show how well the multi-agent framework can handle large-scale work and work in real time within cloud environments used for actual production.

E. Evaluation Objective

The experiment was set up to check how accurate the cost prediction and anomaly detection parts are, as well as how strong, able to handle more work, and quick the whole multi-agent system performs. By using predictive performance metrics along with system-level efficiency measurements, the evaluation gives a complete picture of how well the framework is suited for intelligent cloud financial governance in enterprise-scale environments.

V. RESULTS AND DISCUSSION

This part takes a close look at the proposed Multi-Agent Generative AI framework by checking how well it predicts, finds unusual patterns, uses computer resources efficiently, and how well the whole system works. The results are looked at not just by numbers but also by how useful they are in real-world cloud setups used by companies. The comparison with standard models shows that the new design offers better results and more dependable performance in real-world situations involving cloud cost management.

A. Cost Prediction Performance

The results from the experiments show that the LSTM-based model for predicting costs works much better than traditional methods like regression and ensemble learning, based on all the different ways we measured its performance. The model we created has an R squared value of 0.96, which means it accounts for about 96

Although ensemble learning models were incorporated to improve robustness, the LSTM component contributed most significantly to capturing temporal dependencies and achieving higher forecasting accuracy.

The lower MAE and RMSE values also show that the forecast errors stay small, even when there are big changes in workload or unexpected scaling events. Lower RMSE shows that the model is more reliable when there are big changes in costs, meaning it doesn't get too affected by short-term noisy data in billing.

Linear Regression performs less effectively because it relies on the idea that there is a straight-line connection between how resources are used and the costs involved. In real cloud setups, costs change in complicated ways because of things like non-linear scaling, automatic resource adjustments, and how different services rely on each other, which can't all be properly shown with simple linear models.

Random Forest and XGBoost help make predictions more accurate by considering how different features interact in non-linear ways and by understanding the complicated connections between various usage factors. However, these tree-based ensemble models handle each observation separately and do not have a clear way to remember past events. Because of this, they don't work as well when it comes to showing how long-term billing is affected or how changes in settings can cause costs to come up later.

Unlike the other model, the LSTM uses a special structure that lets it remember important information over time by using controlled memory units. This setup lets the model understand overall patterns, regular changes, and the way billing happens over time, which makes it especially good for predicting cloud spending.

TABLE I
COST PREDICTION PERFORMANCE COMPARISON

Model	MAE	RMSE	R ²
Linear Regression	8.7%	11.2%	0.88
n	6.3%	8.1%	0.92
Random Forest			
XGBoost	5.4%	7.3%	0.94
LSTM	4.8%	6.5%	0.96

The LSTM model achieved the highest forecasting accuracy, demonstrating its effectiveness in capturing complex temporal billing patterns.

B. Anomaly Detection and System Efficiency

The part that finds unusual patterns, which uses the Isolation Forest method, works well according to all the testing measurements. A precision rate of 95.2% shows that most of the flagged issues are real billing problems and not just regular changes in how things normally operate. This is especially important in large business settings where too many false warnings can cause people to ignore real issues and lose confidence in the automatic tools used to watch over systems. A recall value of 94.1% shows the model is good at finding most unusual cost events, which helps reduce the chance of missing out on financial risks. The F1-score of 94.6% shows a good balance between how accurately anomalies are found and how many are caught, making sure that anomalies are identified reliably without missing too many.

The system has a false positive rate of 3.7%, which is quite low. This shows that it works reliably without sending too many false warnings. Unlike static systems that use fixed limits which might not keep up with increasing workloads, the Isolation Forest automatically learns and adapts to the actual billing data patterns. This flexibility allows the model to tell the difference between real growth in workloads and unexpected increases in costs due to errors in setup, too much automatic scaling, or services being turned on by mistake.

Combining anomaly detection signals with predictions of future issues helps improve the understanding of the situation. When forecast errors match up with anomaly scores, the system increases the risk severity, which helps with better prioritization and assessing financial risks.

C. System Performance Metrics

Besides checking how well the system predicts and detects issues, they also looked at other performance measures to see if the system can work smoothly in real time. The LSTM-based forecasting module took an average of 1.2 seconds for each inference cycle, showing it can process time-series data efficiently in sequence. The anomaly detection was done in about 0.8 seconds, showing how fast the Isolation Forest algorithm works.

TABLE II
SYSTEM-LEVEL PERFORMANCE METRICS

Metric	Average Time (seconds)	Cost Forecasting Latency
Anomaly Detection Latency	0.8	1.2
Report Generation Time	1.9	
End-to-End Response Time	3.9	

The AI-powered reporting agent created organized financial summaries in about 1.9 seconds on average. This module changes complex numbers into easy-to-understand financial information, helping people from technical and financial backgrounds talk to each other more clearly.

The total time it takes for the system to respond from start to finish is about 3.9 seconds, which shows that the coordinated multi-agent system works quickly enough to meet real-time requirements. This ability to respond quickly lets companies keep a close eye on their cloud costs all the time and helps them fix any small changes in the bill before they turn into big spending problems. The modular design also makes it easy to scale, letting each part work best on its own as more data comes in.

D. Overall Discussion and Insights

The results from the experiments show that combining deep learning, unsupervised anomaly detection, and Generative AI for reporting in a single multi-agent system works well. Each part does something different in terms of analysis: the LSTM model gives reliable and precise cost predictions, the Isolation Forest helps find unusual billing patterns, and the LLM-based reporting agent makes the results easier to understand and explain.

Unlike regular cloud cost dashboards that mainly show past data, the new framework offers smarter financial insights that can predict, diagnose, and suggest solutions for costs. When agents work together, the system can spot potential cost problems early, suggest ways to save money, and share useful information in a way that's easy to understand.

The results show that the proposed framework works well, giving accurate predictions, effectively finding unusual activity, having few false alarms, and running quickly enough for real-time use. These features show that it is useful, can grow with bigger systems, and works well for smart financial management in large cloud setups.

VI. CONCLUSION

This paper presented a Multi-Agent Generative Artificial Intelligence framework for intelligent cloud cost prediction and resource optimization. The proposed system integrates predictive modeling, anomaly detection, risk-aware optimization, and automated report generation into a unified architecture. Unlike traditional cloud dashboards that provide only historical insights, the proposed framework enables proactive financial decision-making using predictive analytics and intelligent recommendations.

The LSTM-based multivariate time-series model effectively captures complex cost patterns, workload variations, and long-term billing dependencies. The Isolation Forest-based anomaly detection module successfully identifies abnormal cost behaviors with minimal false alerts. Furthermore, the risk-aware optimization mechanism converts analytical insights into actionable cost-saving strategies. The LLM-based reporting agent improves interpretability by transforming analytical outputs into structured and human-readable financial summaries.

Experimental evaluation demonstrates strong performance, achieving a high prediction accuracy with an R^2 score of 0.96 and a low false positive rate of 3.7.

VII. FUTURE WORK

Although the proposed framework demonstrates effective performance in cloud cost prediction and anomaly detection, several enhancements can be explored in future work. Reinforcement learning techniques can be incorporated to enable autonomous decision-making for dynamic resource allocation and real-time cost optimization. Such an approach can allow the system to continuously learn from cloud usage patterns and automatically adapt to changing workloads.

Transformer-based time-series models, such as attention-driven architectures, can be investigated to improve long-term forecasting accuracy in highly dynamic cloud environments. These models can better capture temporal dependencies and complex usage patterns compared to traditional deep learning approaches.

Additionally, real-time streaming analytics frameworks can be integrated to support continuous monitoring of cloud billing data and enable faster anomaly detection.

This would help organizations identify unusual cost spikes instantly and take proactive corrective actions.

Federated learning approaches may also be explored to enable privacy-preserving multi-cloud intelligence without sharing sensitive billing or infrastructure data. This would allow multiple organizations or cloud platforms to collaboratively improve model performance while maintaining data confidentiality.

Furthermore, sustainability-aware optimization strategies can be introduced to balance cost reduction with energy-efficient cloud resource utilization. Incorporating green computing metrics and carbon-aware scheduling can make the proposed framework more environmentally sustainable.



These enhancements can extend the proposed framework toward a fully autonomous, adaptive, and intelligent cloud cost management system capable of operating efficiently in real-world multi-cloud environments.

VIII. ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Department of Artificial Intelligence and Data Science, Misrimal Navajee Munoth Jain Engineering College, Chennai, for providing the necessary academic support, infrastructure, and research facilities to carry out this work successfully.

The authors extend their heartfelt thanks to Ms. Angela Lincy N H, Assistant Professor, Department of Artificial Intelligence and Data Science, for her continuous guidance, valuable suggestions, and technical expertise throughout the development of this research. Her encouragement and constructive feedback played a vital role in improving the quality of this work.

The authors also acknowledge the valuable contributions of the co-authors for their collaborative efforts, technical support, and active involvement in model development, experimentation, and validation of the proposed framework. Their cooperation and dedication significantly contributed to the successful completion of this research work.

The authors also express their appreciation to the faculty members of the AI & DS department for their support and insightful discussions during the research process. Their inputs helped refine the methodology and strengthen the proposed framework.

Finally, the authors acknowledge the contributions of researchers and scholars whose work has been cited in this paper. Their studies in cloud computing, cost optimization, anomaly detection, and deep learning provided important insights that significantly influenced the design and implementation of the proposed system.

REFERENCES

- [1] G. N. Junior and C. Marcon, "OPTIC – Optimized Translation and Interaction for Cloud-Costs: Use of FOCUS Standardization and LLM Integration for Cloud Billing Analysis," IEEE International Conference on Cloud Computing, 2025.
- [2] S. Santhosh, "Cloud-Based Software Development Lifecycle," International Journal of Software Engineering and Applications, June 2023.
- [3] Y. Zhang, T. Zhang, and Q. Li, "Cloud Resource Cost Prediction Using Deep Learning Approaches," IEEE Transactions on Cloud Computing, 2022.
- [4] A. Choudhary, "Comparative Study of Various Cloud Service Providers," International Journal of Computer Applications, 2022.
- [5] N. Bhatte, "Comparison of Different Cloud Providers," International Journal of Advanced Research in Computer Science, June 2022.
- [6] A. M. Mohamed, "Multicriteria Optimization Model for Cloud Provider Selection," Journal of Cloud Computing, May 2019.
- [7] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation Forest," IEEE International Conference on Data Mining, 2008.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)