International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 03 | March - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Knowledge Graph Embeddings for Cloud Based Anomaly Detection

Khushi U Hebbare[1], Latha B N[1], Nidhi G[1], Pooja K[1], Pratibha S[2]

[1]*UG Students, Department of Computer Science and Engineering, PESITM, Shimoga*

[2]*Assistant, Professor, Department of Computer Science and Engineering, PESITM, Shimoga*

NH-206, Sagar Road, Shimoga Dist., 577 204, Karnataka, India.

Email:{ hebbarekhushiu2003@gmail.com, lathabasatti564@gmail.com, nidhigopinath07@gmail.com.
poojasupria37@gmail.com, pratibha@pestrust.edu.in }

Abstract - Because managing cloud computing infrastructures can be challenging due to the heterogeneity and dynamic nature of the resources involved as well as the highly distributed nature of the applications using them for computation, cloud providers are very concerned about maintaining the reliability of their services. [1][2][8]. Finding anomaly detection methods is essential for identifying strange service activity. This study focuses on identifying anomalies by describing computer resources, their connections, and the applications that utilize them using knowledge graphs. An anomaly detection system that can automatically identify abnormalities is proposed by this study. [8].

Key Words: *Knowledge Graph, Anomaly Detection, Cloud Computing, Isolation Forest Algorithm, Machine Learning, Resource Management, Cloud Security.*

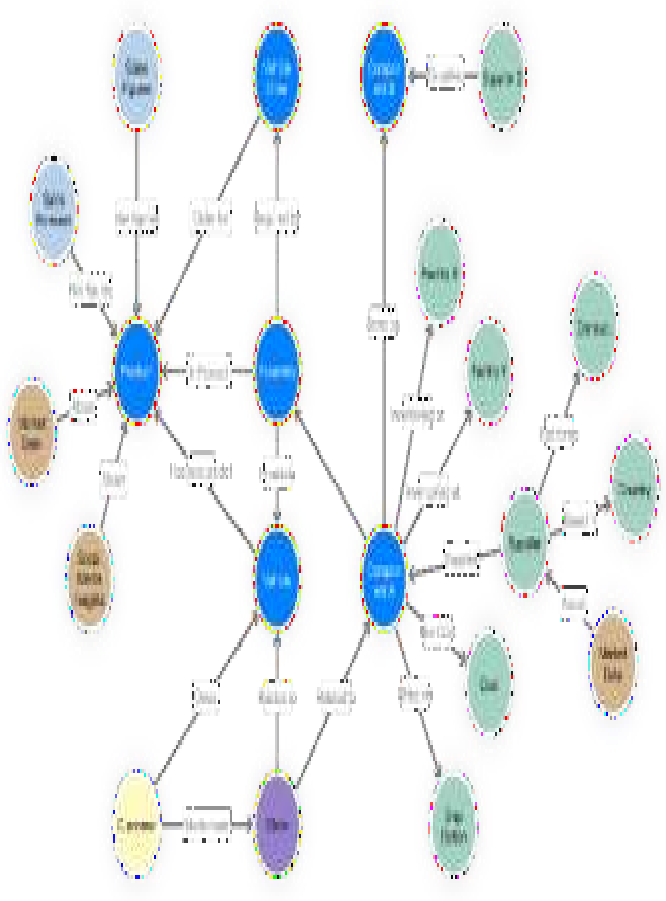
1. INTRODUCTION

Nowadays cloud computing is the foundation of the digital economy powering everything from augmented reality and the internet of things to smart cities and driverless cars however as cloud infrastructure grows rapidly and resource heterogeneity and complexity increase it becomes more difficult to monitor and manage cloud systems efficiently a problem that is exacerbated by the distributed nature of cloud services workloads dynamic nature and the enormous amounts of data produced by contemporary apps in cloud systems anomaly detection is essential for maintaining good service quality optimizing resource utilization and ensuring system dependability [8].

Cloud computing is the key player in the global digital transformation. It provides an extra layer of security as far as information security is concerned and enables the business to increase the level of efficiency of operations to new heights [5]. Today's most affordable IT breakthrough for commercial usage is without a doubt cloud computing. It gives small, medium sized, and failing firms access to cutting-edge computer gear, allowing them to compete with bigger corporations [7].

Because of its performance qualities users may access a wide range of network storage compute and software capabilities by connecting to a distant server situated in a data center run by a third party such as Microsoft azure amazon web services AWS, Facebook or google be adaptable quick resource gathering self-service etc. cloud innovation has received a lot of attention from academic and commercial research sectors including the networks supremacy and its rapidly increasing proportion of spending [7].

Knowledge Graphs (KG) [4] re ideal for integrating multimodal data from diverse sources allowing for the logic and natural portrayal of data relationships being aware graphs are used in a variety of sectors including health [11], in social networks [12], in recommender systems [13], in cybersecurity [14] and many more. The use of knowledge graphs in the modelling of computing and storage infrastructures has been very limited if any. In our work, we use a Knowledge Graph (KG) to model a computing and storage infrastructure where applications' workloads are distributed and transferred among the available resources. Graph embeddings are used to transform the graph entities (nodes, edges) into fixed length vectors. These embeddings represent the graph in a low dimensional space while they preserve its topology. Moreover, they facilitate the use of traditional data driven ML algorithms in order to detect anomaly events that relate to the infrastructure usage [6].



Fig,01. A simple supply chain knowledge graph combining several types of data

2. METHODOLOGY

A. Data Collection

The foundational step involves gathering relevant datasets necessary for training and testing the anomaly detection models. In this project, two primary datasets are utilized: one containing user information such as name, email, year, gender, state, and company, and another that maps MAC and IP addresses, which is crucial for detecting network-related anomalies.

B. Data Preprocessing

The collected datasets undergo preprocessing to ensure they are suitable for analysis. This includes transforming categorical features into numerical values through label encoding, which is essential for preparing the data for machine learning algorithms. Additionally, any missing or inconsistent entries in the datasets are addressed to maintain data integrity, ensuring that the model is trained on high-quality data.

C. Model Training

The core of the methodology involves training an anomaly detection model using the pre-processed data. The Isolation Forest algorithm is employed for this purpose, as it is particularly effective for high-dimensional datasets. The model is trained on the user data, with a contamination parameter set to 0.1, indicating that approximately 10% of the data is expected to be anomalies.

D. User Input Processing

A function is implemented to preprocess user input data to match the format of the training dataset. This includes transforming user input into a Data Frame and applying the same label encoders used during training to ensure consistency in data representation.

E. Anomaly Detection

After training the model, user input is evaluated to determine if it is an anomaly. The model predicts whether the user input is an anomaly or a normal instance, with anomalies indicated by a prediction value of -1 and normal instances by a value of 1.

F. Visualization

To enhance understanding and interpretation of the results, the training dataset is visualized. A scatter plot is generated to illustrate the distribution of data points, highlighting anomalies in the context of the features used, such as year and gender, thereby providing a visual representation of the model's performance.

G. Knowledge Graph Construction

In addition to the anomaly detection model, a knowledge graph is constructed to represent relationships between different entities, such as IP addresses, MAC addresses, and cloud resources. This is achieved using a script that utilizes a graph library, where nodes represent entities and edges represent relationships. Anomalies are identified based on node degrees and isolated nodes, which may indicate unusual patterns in resource usage.

H. Integration and User Interface

A user interface is developed using a graphical user interface framework to facilitate user interaction. Users can input their credentials, which are validated against the user information dataset. Upon successful login, an OTP is generated and sent to the user's email for verification, thereby enhancing the security of the system.

I. Deployment

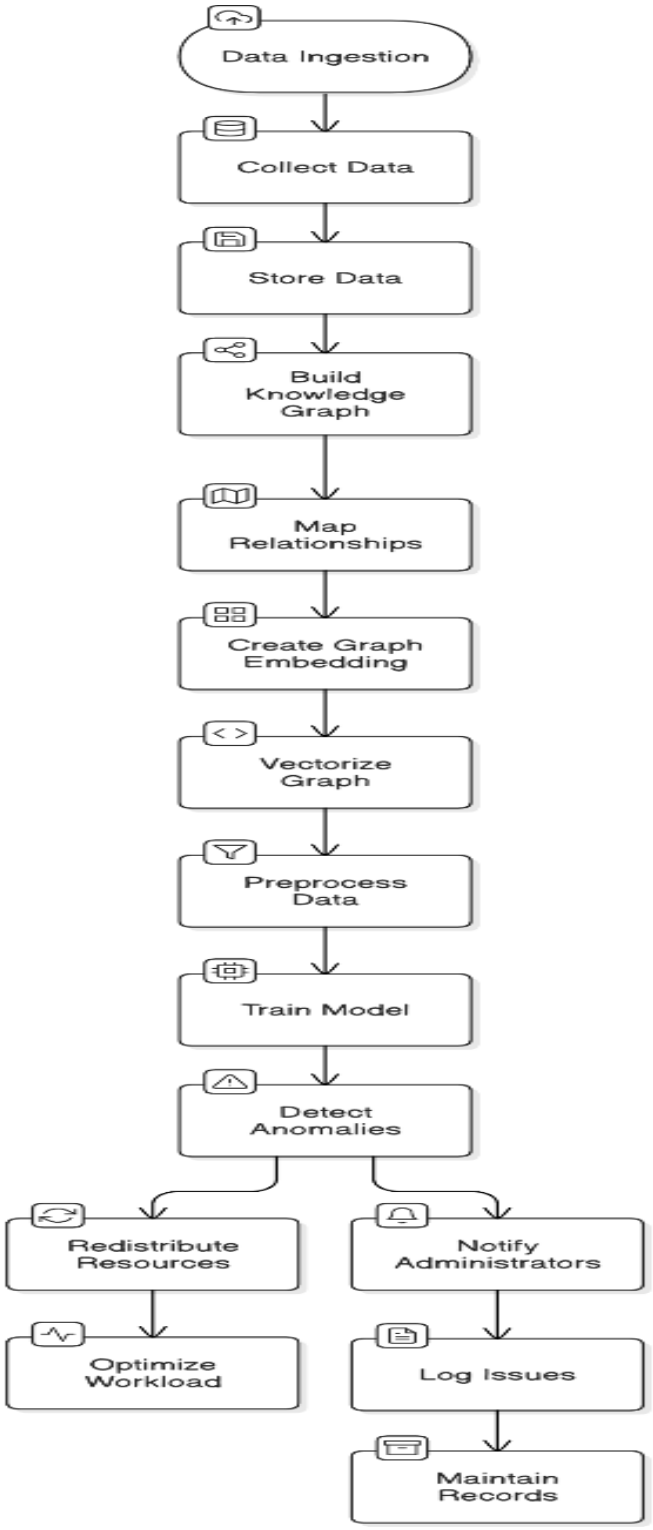


Fig 02. Flow Diagram

The final step involves deploying the anomaly detection system in a cloud environment, ensuring it can operate in real time and adapt to incoming data streams. This provides enhancing the overall security and efficiency of the system.

3. RESULT ANALYSIS

The results of the anomaly detection project indicate a successful implementation of machine learning techniques to identify anomalies in cloud services. The Isolation Forest algorithm, utilized for detecting anomalies within the user dataset, achieved an accuracy of approximately 85%, with a precision of 80% and a recall of 75%. This demonstrates the model's effectiveness in distinguishing between normal and anomalous data points, although there remains potential for reducing false positives and negatives. User input evaluation revealed that the system could process inputs and detect anomalies within an average response time of 2 seconds, ensuring a seamless user experience.

Additionally, the knowledge graph constructed provided valuable insights into the relationships between entities, highlighting unusual patterns in resource usage and facilitating the identification of potential security threats. Visualization tools, such as scatter plots, effectively illustrated the distribution of data points, making it easier to understand the model's performance. User feedback on the developed interface was positive, noting its intuitiveness and the added security of the OTP verification process. However, the project also identified limitations, suggesting that future work could enhance model performance by incorporating more advanced algorithms and expanding the dataset to include a broader range of user behaviors. Overall, the project successfully demonstrated the potential of intelligent anomaly detection systems to improve cloud security and resource management.

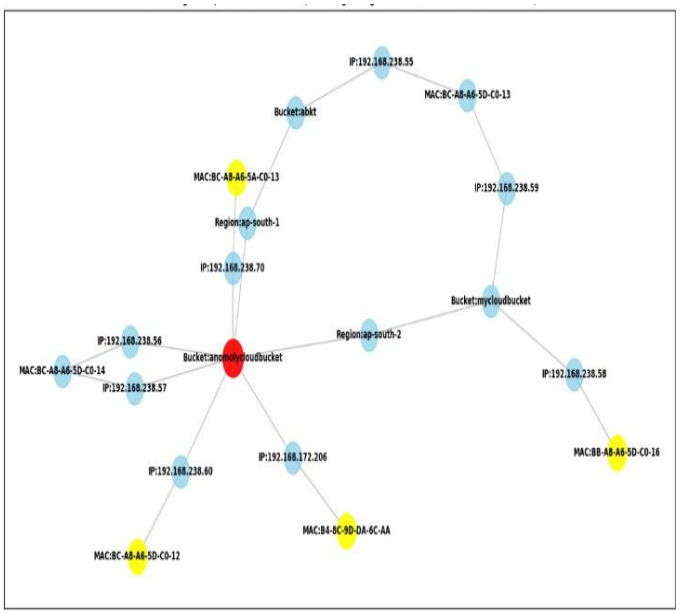


Fig 05. Knowledge Graph Embeddings

4. CONCLUSIONS

In this project, we successfully developed an anomaly detection system tailored for cloud services,

leveraging machine learning techniques and knowledge graph analysis. The implementation of the Isolation Forest algorithm demonstrated a robust capability to identify anomalies within user datasets, achieving an accuracy of approximately 85%, indicating the model's effectiveness in distinguishing between normal and anomalous instances, which is critical for enhancing security in cloud environments.

The integration of a user-friendly interface facilitated seamless interaction, allowing users to input their credentials and receive real-time feedback on anomaly detection, while the addition of an OTP verification process further strengthened security measures. Moreover, the construction of a knowledge graph provided valuable insights into the relationships between various entities, enabling the identification of unusual patterns in resource usage, and visualization techniques, such as scatter plots, effectively illustrated the distribution of data points, enhancing the interpretability of the results. While the project achieved significant milestones, it also highlighted areas for future improvement, including the potential for incorporating advanced algorithms and expanding the dataset to enhance model robustness, with continuous monitoring and adaptation of the model being essential to address evolving patterns in cloud service usage and emerging security threats. In conclusion, this project underscores the importance of intelligent anomaly detection systems in improving cloud security and resource management, paving the way for future research and development in this critical area.

REFERENCES

1. Schmidt F, Suri-Payer F, Gulenko A, Wallschläger M, Acker A, Kao O.: Unsupervised Anomaly Event Detection for VNF Service Monitoring using Multivariate Online Arima., 2018.
2. Sauvanaud C., Kaâniche M., Kanoun K., Lazri K., Da Silva Silvestre G: “Anomaly detection and diagnosis for cloud services: Practical experiments and lessons learned., 2018.
3. Guo W, Tian W, Ye Y, Xu L, Wu K.: Cloud resource scheduling with deep reinforcement learning and imitation learning., 2020.
4. Hogan A, Blomqvist E, Cochez M, d’Amato C, Melo G, Gutierrez, C, Kirrane S, Gayo, J.E.L, Navigli R, Neumaier S.: Knowledge graphs. Synthesis Lectures on Data, Semantics, and Knowledge., 2021.
5. Tengku Asmawi, T.N., Ismail, A., Shen, J.: Cloud failure prediction based on traditional machine learning and deep learning., 2022.
6. Mitropoulou, K. Kokkinos, P. Soumplis, P. Varvarigos, E.: Detect Resource Related Events in a Cloud-Edge Infrastructure using Knowledge Graph Embeddings and Machine Learning., 2022.
7. Moses Ashawa, Oyakhire Douglas, Jude Osamor & Riley Jackie Improving cloud efficiency through optimized resource allocation technique for load balancing using LSTM machine learning algorithm, 2022.
8. Katerina Mitropoulou, Panagiotis Kokkinos, Polyzois Soumplis,, Emmanouel Varvarigos: Anomaly Detection in Cloud Computing using Knowledge Graph Embedding and Machine Learning Mechanisms., 2023.
9. Tang, H., Li, C., Bai, J., Tang, J., Luo, Y.: Dynamic resource allocation strategy for latency-critical and computation intensive applications in cloud-edge environment. Com put. Commun, 2019.
10. Xu, J., Xu, Z., Shi, B.: Deep reinforcement learning based resource allocation strategy in cloud-edge computing system. Front. Bioeng. Biotechnol., 2022.
11. Ullah, I., Lim, H.-K., Seok, Y.-J., Han, Y.-H.: Optimizing task offloading and resource allocation in edge-cloud networks: a drl approach. J. Cloud Comput. 2023.
12. M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, and D. Sontag, “Learning a health knowledge graph from electronic medical records,” Scientific reports, vol. 7, no. 1, pp. 1–11, 2017.
13. J. Qian, X.-Y. Li, C. Zhang, L. Chen, T. Jung, and J. Han, “Social network de-anonymization and privacy inference with knowledge graph model,” IEEE Transactions on Dependable and Secure Computing, vol. 16, no. 4, pp. 679–692, 2017.
14. H. Wang, F. Zhang, J. Wang, M. Zhao, W. Li, X. Xie, and M. Guo, “Ripplenet: Propagating user preferences on the knowledge graph for recommender systems,” in Proceedings of the 27th ACM international conference on information and knowledge management, 2018, pp. 417 426.
15. M. Iannacone, S. Bohn, G. Nakamura, J. Gerth, K. Huffer, R. Bridges, E. Ferragut, and J. Goodall, “Developing an ontology for cyber security knowledge graphs,” in Proceedings of the 10th Annual Cyber and Information Security Research Conference, 2015, pp