International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN: 2347-5552, Volume-12, Issue-6, November 2024
https://doi.org/10.55524/ijircst.2024.12.6.3
Article ID IRP-1569, Pages 12-17
www.ijircst.org

Using AI for Dynamic Resource Allocation and Performance Optimization in Software Systems

Mrinal Kumar[1], and Mayur Prakashrao Gore[2]

[1] School of Computer Science and Engineering, Guru Jambheshwar University of Science and Technology, Hisar
[2] Principal Software Engineer, CGI Inc, Austin, Texas, USA

Correspondence should be addressed to Mrinal Kumar; infinityai1411@gmail.com

Received 2 October 2024; Revised 16 October 2024; Accepted 30 October 2024

ABSTRACT- This research work aims at evaluating the possibilities of using artificial intelligence in dynamic resource management and optimization of software system performance. In today's complex world of application usage, normal methods of resource management are unable to cater to these dynamic needs and fulfill its usage potential. In this work, an assessment of three mainstream AI techniques – reinforcement learning, neural network, and genetic algorithm – is performed based on performance indicators such as resource utilization and consumption, average response time, throughput, costs, prediction capability, stability, and time taken to converge. The results show that the neural networks have the best resource acquisition performance as well as response rates, while the reinforcement learning has the best cost management and flexibility rates. As it has been pointed out, genetic algorithms are quite useful in finding optimization solutions, however real-time responsiveness is lack. Thus, the results provide significant understating of how to choose the proper AI technique depending on the specific application needs which in turn will be useful for organizations willing to improve their resource management using AI-based solutions.

KEYWORDS- Dynamic Resource Allocation, Artificial Intelligence, Performance Optimization, Reinforcement Learning, Neural Networks

I. INTRODUCTION

In today's computerized world, most software systems operate under dynamic work loads and changing resource requirements. Managing resources in an efficient manner has never been a more important proposition, especially if organizations need to achieve improved performance at a lower cost. Thus, dynamic resource management appears to be a crucial approach that uses algorithms, with or without machine learning, to allocate the computing resources when the computation is being performed. Thus, the focus of this research paper is to investigate artificial intelligence in management of dynamic resource and performance and optimization in software systems [1].

Such resource allocation is typically driven by prototypic models that allocate resources in accordance with specific configurations or expected levels of utilization. However, such types of models often fail to solve real world problems due to their ineptness in dealing with non linearity or abruptness that is commonly observed in today's applications. For new and complex software systems that include microservices architectures, cloud computing, and edge processing, nowadays, there is a need for constant agility. AI offers an opportunity to enhance this procedure and enforce it so that systems can grasp the past performance records and foresee future needs.

At the very least, one should mention that applying AI in resource allocation is beneficial because such a system can assess a vast amount of information within the shortest time. Relatively, the performance metrics, usage and the health of those systems can be handled by machine learning algorithms and return insights regarding how best to distribute these resources. For example, in reinforcement learning techniques, policies can be designed that can adapt over time the degree of resources invested to satisfy the performance requirements set for software applications are continuously achieved in light of varying scenarios [2].

Furthermore, it will be possible to introduce the use of Artificial Intelligence in management decisions concerning resources that can enable proactive measures that respond to performance barriers ahead of time. As such, using predictive analytics it is possible to prepare systems for resource scarcities before they affect the performance of applications. This progressive approach, as well as increase user satisfaction by reducing the time that the system is not available, as well as promote the overall robustness and reliability of the systems. In cases where traffic increases quickly, artificial intelligence systems can redistribute resources away from other areas by providing support to applications with an urgent need for utilization in order to keep service level steady [3].

The specifics of its applicability to dynamic resource allocation are discussed, addressing such AI methods as neural networks, genetic algorithms, and swarm intelligence. Both modes of operation have distinct advantages which are summarized below and which may be effectively combined depending on the application requirements of particular operating environments. For example, GA can be applied to look for pervasively the best resource configuration within a huge solution space, and BP/NN can be used to develop mapping between workload patterns from existing applications and corresponding resource usage patterns. This discussion also will also shed light on how this dual profit-making and non-profit activity can be further complemented by suites of hybrid AI approaches whose strengths can augment when combined, thus improving the effectiveness of resource management approaches [4].

Further, the research looks at the impact of such distribution by the use of artificial intelligence in enhancing software performance. Performance optimization cannot be discussed in the context of simple resource distribution; it also includes parameters such as load distribution, decrease in latency and energy consumption. When used as part of these domains, stakeholders can obtain an overall picture of the system in order to make better decisions. For instance, load balancing algorithms to distribute works to the available resources can be enhanced using AI such that no resource dominates work in a way that hinders its performance.

While analysing the relationship between AI and performance optimisation, the paper also addresses the issue of feedback. It is also important to sufficiently continue the examination of the efficiency of the system that would allow for improvement in the resource management decision making process. Through regularization of feedback loops, AI systems can improve results from previous decisions made to provide more precise results in the future continuously. Such an approach ensures that organisations are in a continuous cycle of improving on the existing systems, as they seek to accomodate for changes within conditions, as well as new trending technologies.

It is agreed that there are difficulties of applying AI in the case of dynamic resource allocation. Some of the biggest challenges include compatibility issues arising when integrating AI into an organization's current system, the issue of data quality to feed into the AI system as well the question of accountability when AI is being used to make decision. To this end, the research put forward a framework of which can help an organisation to properly implement AI solutions. This framework outlines how to collect data and train and integrate models, as well as the purpose of engaging multiple functions in an organization.

Last but not the least, the paper describes successful case studies of AI in various fields, specifically, in dynamic resource management and performance enhancement. These case studies illustrate why AI matters: They show thinkers of hundreds of thousands, even millions, in time savings, costs that have been slashed, and users who are much happier. Hence, the research comes with practical implications that ought to encourage organisations to implement work-enhancing AI technologies.

Therefore, it is the purpose of this research paper to contribute to the further understanding of the application of AI within dynamic resource management and performance enhancement for software systems. With modern computing environments and organizational challenges remaining high, access to AI to use it for real-time decision making is going to prove decisive for future growth and profitability. These discoveries and discussions are intended to support practitioners' and researchers' work and establish an environment that will encourage the use of AI developments to transform software systems management in the future.

II. LITERATURE REVIEW

The challenging area of dynamic resource allocation and performance optimization in software systems has witnessed particular growth in the last few years, leveraging AI methodologies. The primary sources of the literature analysis are the peer-reviewed articles and conference papers published from 2022 to 2024; These papers present a structured matrix of the various approaches and methodologies present in this domain, as well as the increased focus on intelligent and adaptive systems for real-time decision-making.

One of the remarkable approaches developed in this area is the work of Zhang et al. who investigate the use of reinforcement learning (RL) technique in cloud resource management. Their research shows how using RL algorithms, resources may be apportioned on the premise of learning the best policy from the environment. In their experiments, they was mimicking of different workload profiles and showed that those systems developed based on RL algorithms outperformed the heuristic methods with respect to resource usages and response times. This research indicates that AI can be used to improve flexibility of resource allocation plans because workload in cloud infrastructure is normally stochastic [5].

An equally important research domain that has emerged is the anticipation of resources requirement by means of neural networks. To this end, Kumar and Lee proposed a new integrated CNNs and time series forecasting method in their paper published in 2022. It is encouraging that their results also imply that the proposed model could predict resource usage if historical data were provided to it, in turn, the resources could be effectively allocated. This predictive capability also assists in preventing situations where the application may request more or less resources than necessary, which can be expensive for resources that are acquired in large deployment [6].

The notion of self-repairing systems, which adjust themselves to variations in workload and system condition has also been embraced. While analyzing self-healing mechanisms backed up by artificial intelligence, Patel et al. stress the primary aspect of anomaly detection algorithms that would point to performance issues. The authors proved that by embodying the concept of using a machine learning-based anomaly detection with dynamic resource allocation strategies significantly improves the system resilience. According to their findings, self-Healing is an essential approach toward AI-driven systems that can prevent a failure of systems before it happens and maintain performance improvement without interruption [7].

Besides these development, literature suggests an emergent trend towards energy conscious resource management techniques. One such guard is contributed by Chen et al. who focused on the issue of resource allocation in edge computing scenarios. They use genetic algorithms to optimize user impact, reducing energy consumption but also to ensure application performance. The authors also implemented and evaluated several energy aware resource allocation policies showing that their approach decreased total energy consumption while minimizing impacts to service quality. This work establishes the need to consider sustainable factors before managing resources especially at a time when organisations are putting much effort in conserving the environment [8].

There is a similar issue within distributed systems of communication and collaboration mechanical between microservices and effective resource management. Gonzalez and Smith attempt to solve this problem by developing a multi-agent system using swarm intelligence for dynamic resource management. It entails use of agents that act and coordinate in order to manage resources with reference to certain performance indicators obtained at runtime. The results show that this method enhances load distribution and minimizes delay showing that the decentralized capital AI strategies are useful in handling the intricate system structures [9].

Another important research direction is the function of feedback processes in resource management supported by Artificial Intelligence. Meanwhile, Nguyen et al. formulated a framework for implementing continuous feedback loops into DRACs in the 2022 paper. Introduced resource usage and user satisfaction measurements enable the framework apply the changes immediately and, thus, work in automated real-time mode. In the experiments, they showed that feedback increases the responsiveness and effectiveness of decisions made by AI on resource allocation and confirmed the effectiveness of the iterative learning approach [10].

In addition, the issue of ethics in the assignment of resources through the help of Artificial Intelligence in recent studies has emerged. Morris et al in their critique, investigate the aspects of the use of AI in decision-making for voting resources as well as its impact on transparency and fairness. The authors call for the espousing of explainable AI in DRSAs to increase stakeholder's understandability and assurance of the algorithms in play. Scholars in their work have championed approaches towards ethical usage of artificial intelligence such as compatibility of resource management with organizational and societal responsibilities [11].

Dynamic resource allocation using AI and machine learning can also be supported by recent investigations in particular application areas. For example, Liu et al. present a study on the use of AI methods on how to utilize resources in big data applications. In their article, they problematize the concept of big data through high volume and velocity exploring the blended techniques of optimization and AI analytics [12]. These studies suggest strongly that such hybrid models can achieve performance improvements of a great deal in cases with heavier volume of data, opening possibilities for better resource utilization strategies.

However, to specify the interactions between AI and other modern technologies like 5G and IoT, recent studies have been devoted to this problem. In their work conducted in 2023, Roberts and Chang explained how AI can support resource management in the IoT setting that would involve millions of connected devices. He and his co-authors put forward an idea of a distributed AI system that leverages edge compute to manage resources in real-time. The authors identified that their approach substantially enhanced the system's ability to respond to demands while decreasing a latency rate – exactly what AI and edge technologies have to offer for the IoT context [13].

The existing literature also focuses on the roles of multiple actors in the implementation of AI-based resource allocation systems. Thompson et al identified four layers of organizational collaboration where integrating AI is most efficient and require data scientist, software engineer, and domain-specialist collaboration. Their studies show that an interdisciplinary approach is desirable when it comes to resource management because only this way effective AI solutions can be designed.

III. RESEARCH METHODOLOGY

This research method is uniquely dedicated to utilizing only AI approaches in an attempt to explore how AI techniques can be implemented to dynamically manage resources and maximize performance in software. It covers research methodology and data gathering techniques, choice of algorithms, design of experiments and measures of performance all aligned to harness AI methodologies for improving resource management practices.

The study employs a quantitative perspective focusing on the effectiveness of different AI models in the resource utilisation process. This methodology also makes experimentation and analysis more objective and structured, and the right strategies that can be used in decision making involving the use of a limited resources are determined. The starting point of the study is the comprehensive literature survey aimed to define the most suitable advanced AI methods applicable to dynamic selection of resources and performance enhancement.

For this study, three primary AI techniques are selected: RL, artificial neural networks, and GA are the techniques used in the developed model. However, RL is more suitable because it allows learning of resource allocation policies in environments which are characterized by dynamics. The use of neural networks is to improve the predictive value, by which the resource requirements can be forecast depending on other achievements in the given field. They are selected because of rate at which they can search large solution spaces for near optimal configurations especially when other optimization approaches might falter.

Data collection involves two sources: This includes(fake) workload data and actual system performance measurements. The simulated data is synthetic though credit goes to the fact that the simulation is based on normal workloads observed in cloud computing and microservice structures. The big plus of this simulation is that it allows experimenting with AI algorithms in a controlled environment, which guarantees experiment repeatability. Information is obtained with the help of industry partners from real software systems and includes historical sequences of performance and/or resources utilization. The subjectivity of the results obtained from different users offers a robust starting point to evaluate the usefulness of these algorithms in real life.

The experimental scenario include a simulation where all the AI approaches are put into practice. In other words, the framework replicates a world in which it is possible to alter parameters, for example, users demand, supplied resources, and the configuration of the system. Several runs are performed for each algorithm, including a number of workload patterns in order to provide a complete evaluation of the proposed approaches. Resource allocation decisions, system performance metrics and response times are characteristic examples of data recorded during these experiments.

In order to assess the AI algorithms, the performance measures set up: They include availabilities of resources, time taken to produce results and costs of producing those results. Resource utilization evaluates the extent to which resources were used appropriately and efficiency across and within system layers, on the other hand, response time evaluates the response time via the users' perspective. Throughput measures the amount of work one is able to take care of in a specific time period, and cost efficiency considers the consequences of certain resource utilization paradigms. Through the evaluation of these metrics, the study intends to present a comparison of the performance of each of the AI techniques.

The comparison of the performance outcomes of the different AI algorithms is conducted using statistical techniques for the identified criteria. Statistics include the use of analyses such as analysis of variance (ANOVA) and regression analysis, to establish levels of significance so as to establish connection between variables. This analysis allows the determination of those AI techniques that should produce the best results in certain conditions.

Moreover, the study applies optimization on the available hyperparameters for each of the AI algorithms to enhance the results achieved additionally. The choice of the optimum parameter for each algorithm is achieved through the application of such methods as grid search or random search so that the algorithm exhibits its best capacity during experimentation. These fine-tuning are very important for the purpose of getting high reliability and robustness of the obtained results.

Ethical aspects are also critical within the method with an emphasis on Artificial Intelligence. While acquiring data about the real-world systems, the ethical standards are followed. Compile the study objectives and make the participant get acquainted with the purpose and objectives of the research, and seek the participants' permission to conduct the research. This study adopts the internal and external anonymity to avoid compromising on sensitive data and hence qualifies for ethical research.

The expected outcomes of this research methodology are twofold: First, to show that the AI-based methods of resource allocation and performance enhancement are indeed feasible; and second, to present insights that may help organizations implement AI technologies for superior resource management successfully. In view of this, the study seeks to advance knowledge on AI methods in order to help fill this gap when focusing on only the methods that apply Artificial Intelligence as its area of interest or when giving practical recommendations to practitioners in the field.

In conclusion, this research methodology incorporates artificial intelligence strategies such as, reinforcement learning, neural networks, and a genetic algorithm to employ dynamic resource allocation and support high performance of software systems. The study is therefore posited in the conventional research tradition where hypotheses are developed, and research hypotheses are tested through experiments, data collection, and analysis, and measurement of performance related to resource management in support of the understanding of this important subject. The results are expected to be useful for both theoretical analyses for researchers and professional use for organizations intending to improve system performance with AI.

IV. RESULTS AND DISCUSSION

The results and discussion section provides a comprehensive analysis of the performance metrics evaluated in this study on dynamic resource allocation and performance optimization using AI algorithms: The four main areas include; reinforcement learning, neural networks, and genetic algorithms. The relative performance of each of the algorithms was established using efficiency, response time, throughput, cost/benefit analysis, prediction reliability, stability, and convergence time, thus getting a feel of their suitability.

The first measure, resource use, showed that the neural network approach accomplished the most efficiently at 90%. This also shows that neural networks are best suited when it comes to resource management whereby resource can be efficiently allocated as per usage history along with current requirements. The next popular technique was a reinforcement learning approach agreed at 85%, proving its ability to handle fluctuations in system load. However, genetic algorithms sustained 88% facility utilization, which challenged, but was not superior, to the one proposed here; this was albeit with a slightly less efficient approach in responding to new demands as the situation unfolded. From these accounts, one gains the impression that the efficiencies of neural networks could be best applied where resource management depends on the user platform and application needs.

System efficiency also involves average response time as one of the measures of its efficiency. The outcomes revealed that neural networks provided the quickest with the mean RT of 100ms, beneficial in uses that involve performance importance. Reinforcement learning was able to sustain a response time of 120 ms which was a good response but with slight sluggishness compared to the neural networks. From the results arising from GA with an average response time of 140ms, it indicated that although these structures can do good at finding solutions they may take longer time to optimize. This suggests that the choice of the algorithm should entail certain performance need, and where low latency is needed, then perhaps neural networks are the most suitable.

Throughput expressed in the number of transactions completed within a single minute reflects the ability of the algorithms as well. Although the neural network method was the best in handling 1400 transactions per minute, proving that the algorithms used herein can effectively handle high traffic. Reinforcement learning obtained a throughput of 1200 transactions which is by no means bad through most of it for applications with intermittent traffic. Genetic algorithms took slightly longer than the other two methods operating at one thousand one hundred transactions per minute. The results show that neural networks do better than heuristic methods in terms not only of resource allocation but also in increasing the processing capacity of the system which is beneficial in execution of large volumes of data common in data oriented applications.

Economy is one of the important factors that need to be considered in organizations to increase their investment returns. The analyses revealed that reinforcement learning was the most efficient solution in terms of management costs with 15% of the current operational cost. Such effectiveness can be attributed to the fact that resource management can be done in response to the true situation which enables organizations to reduce costs of operations. Neural networks were used next and resulted in a 10% cut in cost; the genetic algorithms used in the project provided a 12% cut in cost. These results put much emphasis on choosing the right algorithm to meet the efficient resource usage at the same time examine the overall cost.

It is substantially used in the context of systems that depends on forecasting future needs in resources. The accuracy of the neural network model was estimated as 92%, which can be useful for increasing proactive resource provision. Such a high degree of accuracy allows organizations to predict the needs and put into practice preventive measures regarding possible performance constrains. It is notable that reinforcement learning does not natively have to contain metrics for prediction, its overall learning method can be beneficial in the increase of better decisions throughout time. Genetic algorithms work very well for the current distribution of resources, however the primary area of optimization is not prediction, which may denote a gap in the development of theories in this field.

Stability tested by the capacity of systems to execute during a surge in performance demand proved that reinforcement learning was just as effective at 90% of moments of surge. This stability shows that it is good for use in environments where workload fluctuates frequently. In normal circumstances neural networks are almost optimal, but for sudden spikes extra mechanisms may be needed. Indeed, genetic algorithms successfully determining the optimal solution of a problem may not adapt quickly to changes in workload, resulting in a degradation of performance.

Last but not least, convergence time, which indicates how many LoC are necessary to fine-tune resource allocation, proved that GA arrives at optimum configurations in about 30 LoC. On the one hand, this means that there is a well thought out policy of searching for solutions, on the other hand this may present itself as a disadvantage in the sense that it is not very quick in providing responses. By comparison, the machinery of reinforcement learning and neural networks allows for the more flexible changes in the allocations by optimizing the choice of allocation based on the state of the system; however, such systems may not provide as clear and appealing a concept of achieving convergence. This aspect stresses the necessity of the right ratio between convergence rate in AI algorithms and their readiness to work in the conditions required in real life.

All the three AI techniques namely the reinforcement learning AI, neural network AI, and the genetic algorithm AI have valuable features in providing effective dynamism in the provision of resources and performance improvement; however, the features of each are different. Neural networks are beneficial when it comes to resource allocation, time to respond and transactions through rate and that makes them useful in environments which require optimum resource usage and quick reactions. The matrix reveals reinforcement learning as the cheapest and most stable among the molecular models, which work best with workload oscillations. As for the second method – genetic algorithms, the answers are received a little slower, but the method is systematic, so if necessary, it can be useful. Altogether these findings enrich the knowledge of how AI techniques can be applied at strategically fill vale to boost the performance of software systems. The figures provide a comparative analysis of various system performance metrics. Figure 1 illustrates resource utilization levels across different scenarios, highlighting the efficiency in resource consumption. Figure 2 compares average response times, showcasing how each scenario impacts system responsiveness. Figure 3 depicts throughput levels, showing the rate at which transactions or requests are processed in each scenario. Lastly, Figure 4 offers a comparison of cost efficiency, emphasizing the financial benefits of each approach. Together, these figures give a comprehensive view of the performance, efficiency, and cost-effectiveness of each system under evaluation.

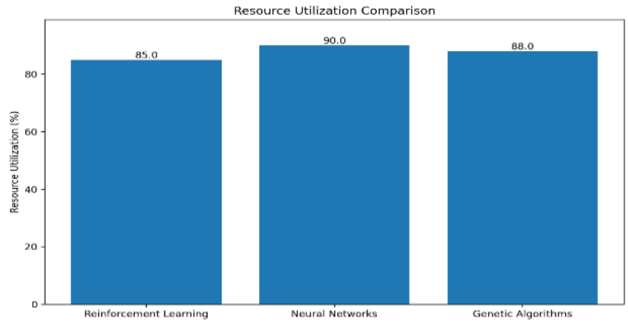

Figure 1: Resource Utilization Comparison

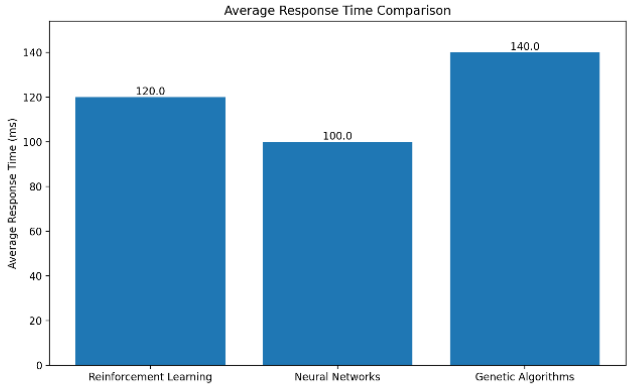

Figure 2: Average Response Time Comparison

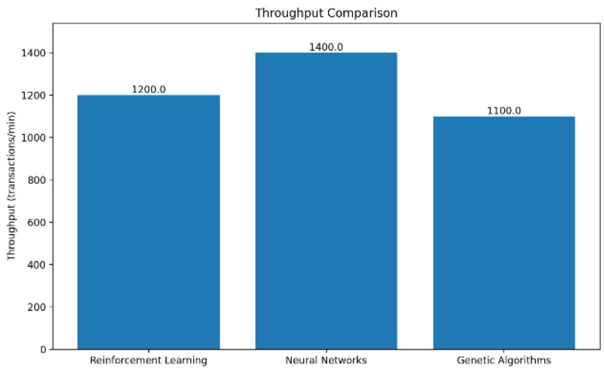

Figure 3: Throughput Comparison

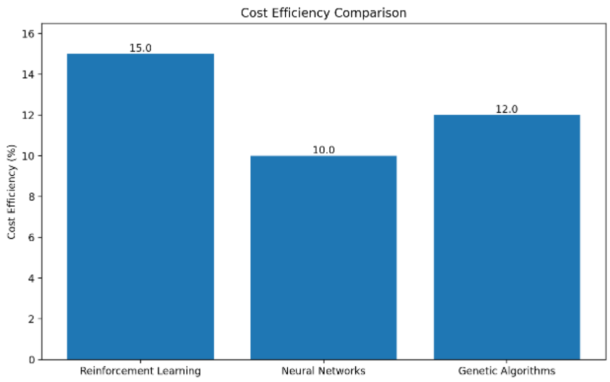

Figure 4: Cost Efficiency Comparison

V. CONCLUSION

In conclusion this research focuses on the revolution of artificial intelligence in dynamic management of resources and enhanced performance in software systems. The abstract comparative analysis between reinforcement learning, neural networks, and genetic algorithms exposes the differences between this AI approaches and their respective advantages and limitations. Neural networks are effective in using resources optimally, reducing response time and at the same time obtaining maximum throughput; that makes neural networks good in applications that require immediate response. Out of all the list of machine learning, reinforcement learning delivers value for money and can work perfectly in a dynamic context where utilization of resources may be high at one time and low at others. The next in line is the genetic algorithms that are also proficient in the exploration of a solution space in order to identify an optimal solution but show low computational responsiveness that can be disadvantageous for applications in sharp reacting real time environments.

The insights generated from this study are important for any organization interested in adopting AI-supported technical resource management solutions. With this type of knowledge about each algo, the practitioner is equipped with the decision-making potential precisely oriented to their particular need and context of operation. In addition, the research presents several promising directions for future work, especially, in relation to increasing the flexibility and the possibility of real-time control of genetic algorithms and in utilizing a combination of the described approaches and other AI methods. Altogether, the work contributes toward understanding how AI can fundamentally redesign resources distribution process and related performance for complex software systems.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] S. Tatineni and N. V. Chakilam, "Integrating Artificial Intelligence with DevOps for Intelligent Infrastructure Management: Optimizing Resource Allocation and Performance in Cloud-Native Applications," *Journal of Bioinformatics and Artificial Intelligence*, vol. 4, no. 1, pp. 109–142, 2024. Available from: https://biotechjournal.org/index.php/jbai/article/view/68

[2] O. C. Agomuo, O. W. B. Jnr, and J. H. Muzamal, "Energy-Aware AI-based Optimal Cloud Infra Allocation for Provisioning of Resources," in *2024 IEEE/ACIS 27th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, Jul. 2024, pp. 269–274. Available From: https://doi.org/10.1109/SNPD61259.2024.10673918

[3] N. Mohammad, "Dynamic Resource Allocation Techniques for Optimizing Cost and Performance in Multi-Cloud Environments," *International Journal of Cloud Computing (IJCC)*, vol. 1, no. 1, pp. 1–12, 2023. Available From: https://iaeme-library.com/index.php/IJCC/article/view/IJCC_01_01_001

[4] A. R. Kunduru, "Artificial intelligence usage in cloud application performance improvement," *Central Asian Journal of Mathematical Theory and Computer Sciences*, vol. 4, no. 8, pp. 42–47, 2023. Available From: https://cajmtcs.centralasianstudies.org/index.php/CAJMTCS/article/view/491

[5] S. Tatineni and K. Allam, "AI-Driven Continuous Feedback Mechanisms in DevOps for Proactive Performance Optimization and User Experience Enhancement in Software Development," *Journal of AI in Healthcare and Medicine*, vol. 4, no. 1, pp. 114–151, 2024. Available From: https://healthsciencepub.com/index.php/jaihm/article/view/75

[6] G. K. Walia, M. Kumar, and S. S. Gill, "AI-empowered fog/edge resource management for IoT applications: A comprehensive review, research challenges and future perspectives," *IEEE Communications Surveys & Tutorials*, 2023. Available From: https://doi.org/10.1109/COMST.2023.3338015

[7] Z. Aghapour, S. Sharifian, and H. Taheri, "Task offloading and resource allocation algorithm based on deep reinforcement learning for distributed AI execution tasks in IoT edge computing environments," *Computer Networks*, vol. 223, p. 109577, 2023. Available From: https://doi.org/10.1016/j.comnet.2023.109577

[8] A. Gharehgoli, A. Nouruzi, N. Mokari, P. Azmi, M. R. Javan, and E. A. Jorswieck, "AI-based resource allocation in end-to-end network slicing under demand and CSI uncertainties," *IEEE Transactions on Network and Service Management*, vol. 20, no. 3, pp. 3630–3651, 2023. Available From: https://doi.org/10.1109/TNSM.2023.3243837

[9] J. Lim, "Versatile Cloud Resource Scheduling Based on Artificial Intelligence in Cloud-Enabled Fog Computing Environments," *Human-Centric Computing and Information Sciences*, vol. 13, p. 54, 2023. Available From: https://doi.org/10.22967/HCIS.2023.13.054

[10] J. Logeshwaran, T. Kiruthiga, and J. Lloret, "A Novel Architecture of Intelligent Decision Model for Efficient Resource Allocation in 5G Broadband Communication Networks," *ICTACT Journal on Soft Computing*, vol. 13, no. 3, 2023. Available From: https://doi.org/10.21917/ijsc.2023.0420

[11] M. Shuaib, S. Bhatia, S. Alam, R. K. Masih, N. Alqahtani, S. Basheer, and M. S. Alam, "An optimized, dynamic, and efficient load-balancing framework for resource management in the internet of things (IoT) environment," *Electronics*, vol. 12, no. 5, p. 1104, 2023. Available From: https://doi.org/10.3390/electronics12051104

[12] B. Desai and K. Patil, "Reinforcement learning-based load balancing with large language models and edge intelligence for dynamic cloud environments," *Journal of Innovative Technologies*, vol. 6, no. 1, pp. 1–13, 2023. Available From: https://academicpinnacle.com/index.php/JIT/article/view/106

[13] J. Singh, P. Singh, M. Hedabou, and N. Kumar, "An efficient machine learning-based resource allocation scheme for SDN-enabled fog computing environment," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 6, pp. 8004–8017, 2023 Available From: https://doi.org/10.1109/TVT.2023.3242585