# ENHANCING THE MLOPS DEPLOYMENT PROCESS USING GEN AI

Pankaj Pilaniwala[1], Girish Chhabra[2] and Ameya Naik[3]

[1]University of Arizona, Tucson, Arizona, USA
[2] San Jose State University, California, USA
[3] Stony Brook University, New York, USA

***ABSTRACT***

*This paper examines the current state of MLOps and emphasizes the importance of developing ML and AI-powered applications. The study commences by defining MLOps and conducting a Literature Review to highlight critical studies in this area. It then delves into the current state of MLOps, providing examples of how major tech companies implement MLOps across their organizations. Additionally, the paper discusses the benefits of utilizing MLOps and addresses deployment challenges in the current process. Furthermore, it proposes an innovative solution to improve existing practices. It presents a technical architecture system design for implementing this novel approach using GenAI to enhance and streamline the MLOps deployment process.*

***KEYWORDS***

*MLOps, Machine Learning, GenAI, Deployment*

## 1. INTRODUCTION

We live in an exciting time when Artificial Intelligence has become a household term. Everyone uses some form of AI application. These AI-powered apps are built by analyzing and learning from a vast amount of data the world produces every second. Companies collect these data and create Machine Learning models to provide intelligent solutions to their users. AI has influenced several industries – Finance [1], Commerce Platforms [2], Gaming [3], etc. NFTs [4][35][36] are upcoming tech and AI can have profound impact to it. To provide intelligent and innovative products/services, ML algorithms must be productionized, and the process of productionizing the ML algorithms is known as MLOps. It stands for a collection of tools and techniques to productionize ML algorithms [5]. ML can be categorized into two states - Supervised Learning and Unsupervised Learning. As per a report by Deloitte, "MLOps can encourage experimentation and rapid delivery, helping enterprises industrialize machine learning" [6].

## 2. LITERATURE REVIEW

MLOps is a new and upcoming field with limited existing research, as the scientific community and practiconers have recently started exploring its potential [7]. This section will discuss some of the field's critical developments and ongoing evolution. Notably, the scientific community demonstrated the pivotal role of ML in assisting the government during the Covid crisis [8]. Additionally, Sculley et al. emphasized the complexity and substantial technical debt associated with maintaining ML systems [9]. At the same time, Makinen et al. conducted a global survey of data scientists to confirm the global importance and criticality of MLOps [10]. Ruf et al.

David C. Wyld et al. (Eds): NLMLT, SEAU– 2024
pp. 13-21, 2024. - CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.142002

examined and presented their findings on selecting the best open-source tools that are available for various operational and execution phases of MLOps [11]. Furthermore, Schelter et al. delved into the criticality of data quality in ML applications and proposed APIs for large-scale automated quality verification [12]. Finally, researchers have underscored the importance of continuous delivery for Machine Learning systems [13][14]. But none of the existing study talks about an advanced automated GenAI powered MLOps deployment tech. the paper bridges the gap in the current literature by proposing a new concept in the following sections.

## 3. OVERVIEW OF MLOPS

This section will provide an overview of the MLOps. It will go in details about the current state of the industry, MLOps benefits, discuss the lifecycle and will review how MLOps is being practiced in Big Tech.

### 3.1. Current State

The market size of MLOps in 2024 is estimated to be around $3 Billion. It is expected to reach a whopping $60 Billion by 2033, growing at a CAGR of 40% [15]. The market will be captured majorly by Large Enterprise - 71%. Some prominent companies in this space are Amazon, IBM, Microsoft, Databrick, and Google. Google Trends [Fig. 1] shows a steep rise in interest in MLOps starting in 2022, proving that more companies are trying to adopt MLOps for their ML-based application development and deployment. However, according to a survey, only 47% of these models are going into production [16]. The financial sector presents a massive opportunity for MLOps learners and practitioners as the sector is bound to leverage data and use ML to automate many back-office and front-office processes [17].

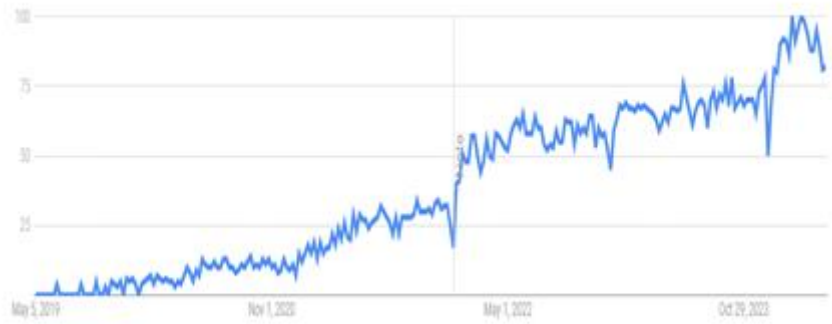

Figure 1. MLOps Google Keyword Trend

### 3.2. Benefits of MLops

There are benefits to doing MLOps, such as making it a widespread practice across big techs and startups. Some of the critical benefits of MLOps are:

- It builds trust by automating the deployment process through testing and validation
- Helps organizations maintain and scale several ML-based applications on production
- MLOps helps organizations better manage and maintain big data. It helps organizations to deploy solutions faster on production
- MLOps helps teams and organizations be at the forefront of Compliance law by automating tasks and better-managing data to maintain the laws in different regions.

### 3.3. MLOps Lifecycle

As seen in Fig. 2, the MLOps lifecycle begins with Data gathering and cleaning, and it is an iterative process where ML Engineers, Data Scientists, Data Engineers, and Software Engineers work together to build, train, deploy, and measure ML Models. Different companies have different versions of this model, but at a high level, this is how the life cycle for any ML-based application is managed. To effectively implement MLOps processes, companies often build core capabilities powered by in-house ML platform technologies or built by integrating vendor capabilities [18].

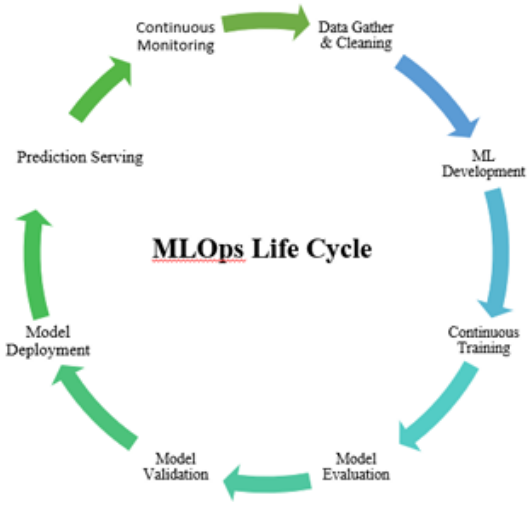


Figure 2. MLOps Process Lifecycle

### 3.4. MLOps in Big-Tech

Meta has released its white paper to showcase how it has built the infra that powers its ecosystem, fueled by ML and AI algorithms built on the massive dataset produced by Meta Platforms [19]. Similarly, Google has released a couple of papers around MLOps, discussing the technical debt [9] and describing the MLOPS lifecycle, processes, and infra requirements [18]. Nvidia is another big-tech company that's at the forefront of the deployment of AI and ML models. Nvidia blogs provide a sneak peek into its processes for handling the MLOps pipeline [20][21]. Spotify, a global music-streaming app company, has written about its ML Infra and its workflow to build, deploy, and maintain ML Algorithms to power its app worldwide in its engineering blog post [22][23]. Uber, another global tech company, has shared in its blog how it has built the infra to scale and operate MLOps processes [24][25]. Most companies globally have a strategy to implement MLOps to elevate their technical processes and build a robust platform.

### 3.5. Tools & Technologies

Some of the end-to-end popular MLOps platforms are DataRobot, Azure ML, Databricks, Domino, Amazon SageMaker, Metaflow, Vertex AI, Weights & Biases, Valohai, Kubeflow, TrueFoundry. Of these Kubeflow and Metaflow are open-source platforms.

# 4. CHALLENGES IN ML DEPLOYMENT

MLOps comes with quite a few challenges and it makes the overall experience of the developers frustrating. It involves too much manual work which makes it time consuming and intricate. Also, this makes MLOps error prone, putting the business systems at risks. The following sub-section will highlight the challenges in the current ML Deployment process.

## 4.1. Model Versioning and Management

Managing multiple versions of ML models and tracking changes over time can become complex during deployment [26].

## 4.2. Scalability and Performance

Ensuring deployed ML models can handle varying workloads and scale effectively is critical. Resource constraints, such as limited memory or processing power, can impact the performance and responsiveness of deployed models, especially in high-demand scenarios [27][28].

## 4.3. Dependency Management

ML models often rely on complex dependencies, including libraries, frameworks, and external services. Managing these dependencies and ensuring compatibility across different environments can be difficult while deploying.

## 4.4. Data Drift and Model Decay

ML models are trained on historical data, which may become outdated or no longer represent the current environment. Addressing data drift and model decay requires continuous monitoring and retraining strategies to maintain model accuracy and effectiveness in production [29].

## 4.5. Monitoring and Maintenance

Once deployed, ML models require ongoing monitoring and maintenance to detect performance degradation, address issues promptly, and incorporate feedback for continuous improvement.

# 5. LEVERAGING GEN AI IN MLOPS

Recent developments in GenAI can ease the deployment of models at various levels. One of the premises was that the packaging of models results in ease of deployment and maintains traffic well; GenAI can play a significant role in that. We tried a methodology where models could be packaged based on the size of the group, correlated workloads, and Machine Type (CPU/GPU types) rather than a manually trying to create them. By automating the packaging process with GenAI, organizations can streamline deployment workflows, reduce manual intervention, and ensure that models are efficiently deployed and maintained to handle multiple workloads and resource limitations.

- Automated Workload Analysis: GenAI can be used to do a lot of tasks and in this paper, authors propose a new model to analyze MLOps workloads using advanced pattern recognition and data analysis techniques, analyzing workload size, complexity, and resource requirements.

- Contextual Understanding: GenAI understands the context of the workload, considering multiple factors like correlated workloads and machine types (CPU/GPU types). It analyzes and structures how these factors influence the optimal packaging configuration for deployment.
- Recommendation Generation: The proposed system generates recommendations for the most suitable packaging configuration. It considers available CPU/GPU types, memory constraints, and expected traffic patterns, ensuring efficient resource utilization and optimal performance.
- Continuous Learning: The GenAI model keeps on learning from the past deployments and their performance metrics and with new existing processes. It keeps on updating and revising its recommendations over time, adapting to changes in workload pipelines and deployment environments.

Streamlined Deployment Workflows: By automating the packaging process with GenAI, organizations streamline deployment workflows. They reduce manual intervention and eliminate the need for human experts to manually analyze and configure deployment settings.

## 6. SYSTEM ARCHITECTURE DESIGN

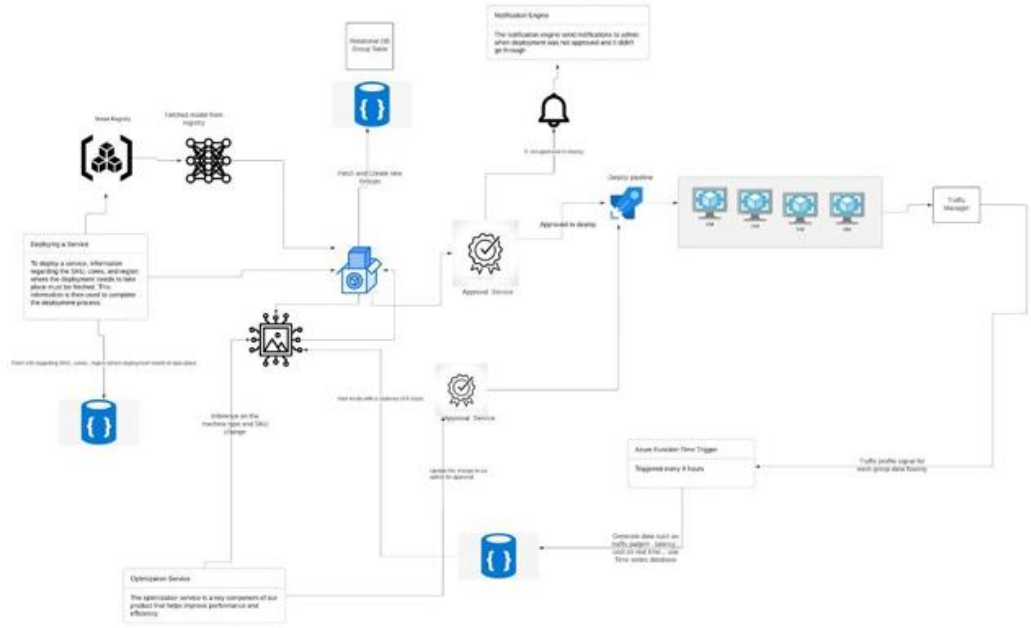

Figure 3. System Design

Steps:

- Identification of New Workload: The process begins when a new workload is introduced to the existing group.
- Metric Understanding: The model collects essential data, including workload size, memory capacity requirements, traffic utilization patterns, and the cost associated with the machine's SKU, and comes up with all-around recommendations that align with the task.
- Model evaluation/inference: The model analyzes the collected data in detailed and determines the optimal deployment processes. This analysis considers various factors such as workload, available resources, and cost efficiency.

- Prediction: The model analyzes several factors and parameters to predicts the right grouping for the deployment pipeline.
- Recommendations: The model provides recommendations on potential adjustments to the machine's SKU to effectively accommodate the new workload. Sometimes, it may suggest creating a new group with specific machine specifications, including type, cores, and thread count.
- Presentation to Tech Team: Three recommendations are presented to teach teams responsible for deployment. These recommendations outline deployment options customized to the specific use case and overall business objectives.
- Recommendation Evaluation: The team evaluates the recommendations. The team considers several factors such as performance, scalability, and cost-effectiveness when deciding on which recommendation to choose.
- Decision Making: The team selects the most suitable deployment option from the recommendations provided by the model from their dashboard. The decision is based on careful consideration of all relevant factors.
- Deployment: Post decision making, the deployment process is initiated according to the chosen option. This may involve deploying the model to an existing group or creating a new group with the recommended machine specifications.
- Monitoring and Optimization: The system is continuously monitored to ensure optimal performance after deployment. Adjustments may be made based on real-time data and feedback to optimize resource utilization and cost efficiency.

## 6.1. Deployment using a grouping of workloads

Deploying hundreds of models at scale poses a significant challenge [30], particularly when managing workloads across diverse geographical regions within a DevOps pipeline [31]. Suppose an organization is tasked with deploying hundreds of models across various geo-regions [32]; the complexity arises in precisely tracking each model to its respective endpoint deployment while ensuring real-time deployment without disrupting production traffic.

To address this challenge, our study explored a grouping of models based on common characteristics. Each group is assigned a dedicated endpoint, with slight variations in the additional query pattern. This approach streamlines the deployment process, reducing complexity and minimizing monitoring overhead.

For example, an e-commerce platform operating globally, might want to deploy recommendation models targeted to regional preferences. Each group will be associated with a distinct endpoint by categorizing models based on product categories, user demographics, or purchasing behavior. Categorization allows for efficient deployment management, enabling seamless updates or additions to the model portfolio without causing disruptions or compromising performance. Also, this grouping strategy supports simplified monitoring [33], as administrators can focus on monitoring performance and health metrics at the group level rather than individually tracking each model. This enhances operational efficiency across the entire deployment infrastructure.

## 6.2. Example

Let us consider an example scenario involving an e-commerce platform [34] deploying recommendation models across different regions, as discussed above. Here is a hypothetical dataset showing how models could be grouped based on common characteristics, Table 1:

Table 1. Endpoints for Grouping ML Models by SKUs

| Model ID | Product Category | Region | Endpoint URL |
|---|---|---|---|
| 001 | Electronics | USA | https://us-recommendations.com/model1 |
| 002 | Fashion | USA | https://us-recommendations.com/model2 |
| 003 | Electronics | Europe | https://eu-recommendations.com/model1 |
| 004 | Fashion | Europe | https://eu-recommendations.com/model2 |
| 005 | Electronics | Asia | https://asia-recommendations.com/model1 |
| 006 | Fashion | Asia | https://asia-recommendations.com/model2 |

Product categories are used to create distinct Models, such as "Electronics" and "Fashion.". Each category has models deployed in different geographical regions: USA, Europe, and Asia. Models within the same category and region share a common endpoint URL with a slight variation in the model identifier (e.g., "model1" and "model2").

This grouping strategy allows for efficient management and deployment of recommendation models. For example, if there is an update to the "Electronics" recommendation model, it can be rolled out simultaneously across all regions where electronics are sold, using the standard endpoint URLs associated with the respective model groups.

The deployment process becomes simpler and we can focus on monitoring and tracking performance and health metrics at the group-level by categorizing the models as discussed above.

## 7. CONCLUSION

In conclusion, utilizing recent advancements in AI facilitates streamlining machine learning model deployment through automated packaging methodologies. By utilizing GenAI's capabilities, organizations can optimize their deployment workflows, reduce manual workloads, and effectively manage several workloads and resource challenges. GenAI empowers organizations to make informed packaging decisions by using the proposed system design and implementing GenAI powered MLOps deployment process. Using this advanced deployment strategy, organizations can truly harness the power of LLMs and enhance their efficiency and serve their users faster and better.

This paper along with others is just the start of utilizing GenAI to advance the deployment process of MLOps systems. Researchers should continue to explore and study further to develop more advanced and nuanced GenAI use cases and algorithms to further the field and introduce advanced AI algorithms and techniques that’ll disrupt the changing landscape for good.

### ACKNOWLEDGEMENTS

The authors would like to thank everyone, just everyone!

## REFERENCES


[1] Pankaj Pilaniwala. How AI is Changing the Financial Industry Landscape and Its Ethical Consideration. IEEE FeedForward Magazine. Vol 3 Issue 3. July - Sept 2024. P 7 -20. Available: https://r6.ieee.org/scv-cs/wp-content/uploads/sites/81/2024/07/FeedForwardV3I3-updated.pdf

[2] Pankaj Pilaniwala. April 2024. How AI is changing Commerce Platforms globally. Retrieved From: https://aijourn.com/how-ai-is-changing-commerce-platforms-globally/

[3] Pankaj Pilaniwala. 2023. How Games Aid in Teaching Real-World Complex Skills: Positive Impact of Gaming. IJRASET Vol 11 Issue II. DOI: https://doi.org/10.22214/ijraset.2023.48987

[4] Pankaj Pilaniwala. March 2023. How NFT is Changing Video-Games and Sports Industry: A Review. IJRASET Volume 11 Issue 3. DOI: https://doi.org/10.22214/ijraset.2023.49426

[5] S. Alla and S. K. Adari, "What is mlops?" in Beginning MLOps with MLFlow. Springer, 2021, pp. 79–124

[6] Deloitte, Tech Trends 2021, [Online] Available: https://www2.deloitte.com/content/dam/insights/articles/6730_TT-Landing-page/DI_2021-Tech-Trends.pdf

[7] Y. Zhao, Machine learning in production: A literature re-view, 2021, [online] Available: https://staff.fnwi.uva.nl/a.s.z.belloum/LiteratureStudies/Reports/2021-LiteratureStudy-report-Yizhen.pdf

[8] Y. Zoabi, S. Deri-Rozov, and N. Shomron, Machine learning-based prediction of covid-19 diagnosis based on symptoms, npj Digital Medicine, vol. 4, Dec. 2021

[9] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison, Hidden technical debt in machine learning systems, in Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS'15, (Cambridge, MA, USA), p. 2503–2511, MIT Press, 2015

[10] Sasu Mäkinen, Henrik Skogström, Eero Laaksonen, Tommi Mikkonen, Who Needs MLOps: What Data Scientists Seek to Accomplish and How Can MLOps Help?, Arxiv, DOI: https://doi.org/10.48550/arXiv.2103.08942

[11] Ruf P, Madan M, Reich C, Ould-Abdeslam D. Demystifying MLOps and Presenting a Recipe for the Selection of Open-Source Tools. Applied Sciences. 2021; 11(19):8861, DOI: https://doi.org/10.3390/app11198861

[12] Sebastian Schelter, Dustin Lange, Philipp Schmidt, Meltem Celikel, Felix Biessmann, and Andreas Grafberger, Automating large-scale data quality verification, Proceedings of the VLDB EndowmentVolume 11Issue 12pp 1781–1794, DOI: https://doi.org/10.14778/3229863.3229867

[13] Lwakatare, L.E., Crnkovic, I., Rånge, E., Bosch, J. (2020). From a Data Science Driven Process to a Continuous Delivery Process for Machine Learning Systems. In: Morisio, M., Torchiano, M., Jedlitschka, A. (eds) Product-Focused Software Process Improvement. PROFES 2020. Lecture Notes in Computer Science(), vol 12562. Springer, Cham. https://doi.org/10.1007/978-3-030-64148-1_12

[14] S. Garg, P. Pundir, G. Rathee, P. K. Gupta, S. Garg and S. Ahlawat, "On Continuous Integration / Continuous Delivery for Automated Deployment of Machine Learning Models using MLOps," 2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), Laguna Hills, CA, USA, 2021, pp. 25-28, doi: 10.1109/AIKE52691.2021.00010

[15] Market Research Future, Mlops Market Research Report, [Online] Available: https://www.marketresearchfuture.com/reports/mlops-market-18849

[16] Information Week, Getting Machine Learning into Production: MLOps, [Online], Available https://www.informationweek.com/machine-learning-ai/getting-machine-learning-into-production-mlops

[17] Dr. Joseph N. Kozhaya, Souva Majumder, Anushree Bhattacharjee, A Framework For Integrating Governed MLOps In Financial Services, IBM, [Online], Available: https://community.ibm.com/community/user/power/viewdocument/a-framework-for-integrating-governe?CommunityKey=273aac8b-92d1-4fbe-a228-018776b76a19&tab=librarydocuments

[18] Khalid Salama, Jarek Kazmierczak, Donna Schut, Practitioners guide to MLOps: A framework for continuous delivery and automation of machine learning, Google Cloud, White Paper, May 2021, [Online] Available: https://services.google.com/fh/files/misc/practitioners_guide_to_mlops_whitepaper.pdf

[19] Kim Hazelwood, Sarah Bird, David Brooks, et al., Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective, Meta Inc., [Online], Available:

https://research.facebook.com/publications/applied-machine-learning-at-facebook-a-datacenter-infrastructure-perspective/
[20] Nvidia, What Is MLOps?, [Online], Available: https://blogs.nvidia.com/blog/what-is-mlops/
[21] Nvidia, Demystifying Enterprise MLOps, [Online], Available: https://developer.nvidia.com/blog/demystifying-enterprise-mlops/
[22] Spotify, The Winding Road to Better Machine Learning Infrastructure Through Tensorflow Extended and Kubeflow, [Online], Available: https://engineering.atspotify.com/2019/12/the-winding-road-to-better-machine-learning-infrastructure-through-tensorflow-extended-and-kubeflow/
[23] Spotify, Unleashing ML Innovation at Spotify with Ray, [Online], Available: https://engineering.atspotify.com/2023/02/unleashing-ml-innovation-at-spotify-with-ray/
[24] Uber, Continuous Integration and Deployment for Machine Learning Online Serving and Models, [Online], Available: https://www.uber.com/blog/continuous-integration-deployment-ml/
[25] Uber, Scaling AI/ML Infrastructure at Uber, [Online], Available: https://www.uber.com/blog/scaling-ai-ml-infrastructure-at-uber/
[26] N. Nahar, S. Zhou, G. Lewis and C. Kästner, Collaboration Challenges in Building ML-Enabled Systems: Communication, Documentation, Engineering, and Process, 2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE), Pittsburgh, PA, USA, 2022, pp. 413-425, DOI: 10.1145/3510003.3510209
[27] Akkiraju, R. et al. (2020). Characterizing Machine Learning Processes: A Maturity Framework. In: Fahland, D., Ghidini, C., Becker, J., Dumas, M. (eds) Business Process Management. BPM 2020. Lecture Notes in Computer Science(), vol 12168. Springer, Cham. DOI: https://doi.org/10.1007/978-3-030-58666-9_2
[28] R. V. Kulkarni, A. Thakur, S. Nalbalwar, S. Shah and S. Chordia, Exploring Scalable and Efficient Deployment of Machine Learning Models: A Comparative Analysis of Amazon SageMaker and Heroku, 2023 International Conference on Information Technology (ICIT), Amman, Jordan, 2023, pp. 746-751, DOI: 10.1109/ICIT58056.2023.10225793
[29] Shreya Shankar, Rolando Garcia, Joseph M. Hellerstein, Aditya G. Parameswaran, Operationalizing Machine Learning: An Interview Study, Arxiv, DOI: https://doi.org/10.48550/arXiv.2209.09125
[30] L. Savu, "Cloud Computing: Deployment Models, Delivery Models, Risks and Research Challenges," 2011 International Conference on Computer and Management (CAMAN), Wuhan, China, 2011, pp. 1-4, DOI: 10.1109/CAMAN.2011.5778816
[31] Henrik Heymann, Alexander D. Kies, Maik Frye, Robert H. Schmitt, Andrés Boza, Guideline for Deployment of Machine Learning Models for Predictive Quality in Production, Procedia CIRP, Volume 107, 2022, Pages 815-820, ISSN 2212-8271, DOI: https://doi.org/10.1016/j.procir.2022.05.068
[32] M. Villari, A. Celesti, G. Tricomi, A. Galletta and M. Fazio, "Deployment orchestration of microservices with geographical constraints for Edge computing," 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, Greece, 2017, pp. 633-638, DOI: 10.1109/ISCC.2017.8024599
[33] Nvidia , A Guide to Monitoring Machine Learning Models in Production , [Online] , Available: A Guide to Monitoring Machine Learning Models in Production | NVIDIA Technical Blog
[34] G. Chang, "The deployment of customer requirements of e-commerce website," 2008 7th World Congress on Intelligent Control and Automation, Chongqing, China, 2008, pp. 6636-6641, doi: 10.1109/WCICA.2008.4593930
[35] Pankaj Pilaniwala. Systemic Review & Analysis of the Current and Future State of NFT as an Artwork. Jan 2023. IJRASET Volume 11 Issue 1. DOI: https://doi.org/10.22214/ijraset.2023.48613
[36] Pankaj Pilaniwala. How NFT is Changing Fashion Industry: A Systemic Review and Analysis. Jan 2023. IJRASET Volume 11 Issue 1. DOI: https://doi.org/10.22214/ijraset.2023.48480.