ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT1856

Socio-Economic Benefits of Machine Learning Deployment Platforms in Business: A Case Study of Baseten and Similar Models

Shivang Raval[1]
Computer Science
Student , PDEU University
Gandhinagar, India

Rushil Shah[2]
Computer Science
Student, PDEU University
Gandhinagar, India

Abstract:- ML deployment platforms have transformed the adoption of AI in business by reducing the technical barriers and costs associated with building and deploying AI models. Among these are Baseten, Bananas.dev, Stagelight AI, Replicate, and Modal Labs, to name a few accessible solutions that work for businesses ranging from startups to large enterprises. This paper discusses the socio-economic benefits of these platforms, especially their business efficiency, scalability, transformation in the workforce, and growth of the economy. In comparative analysis, we will examine each unique feature of these platforms, such as how easy they are to use, how easily they can be customized, and their scalability, which allows us to make some observations regarding their applicability to specific businesses. Utilising comparative analysis, performance benchmarking, and case study support empirical evaluation of their peculiarities and demonstrate their cross sectoral relevance.

Keywords:- *Machine Learning, AI Platforms, Baseten, Deployment Platforms, Business Efficiency, Economic Growth, SMEs.*

I. INTRODUCTION

According to PwC's Global Artificial Intelligence Study 2023, AI is set to add some $15.7 trillion to the global economy by 2030-about $6.6 trillion as a result of increased productivity and $9.1 trillion through consumption effects[1]. It is for this reason that AI deployment poses a huge challenge for most businesses and, more particularly, SMEs. This is because companies in this category lack the requisite technical expertise or infrastructure to deploy an AI model, hence adopting low adoption rates.Baseten, Bananas.dev, Stagelight AI, Replicate, and Modal Labs are but a few of the platforms that rose to fill this gap, which offer low-code and API-driven solutions, making AI easy to deploy. These platforms democratize AI adoption, allowing any business access to cutting-edge technology without a need for any significant in-house development team[2].

II. RESEARCH METHODOLOGY

A. Data Collection Methods

➢ *Platform Analysis*

- Systematic review of platform documentation
- API testing and performance benchmarking
- User interface evaluation using standardized rubric
- Cost analysis based on published pricing and usage scenarios

➢ *Performance Metrics Collection*

- Deployment time measurements
- Resource utilization monitoring
- Scalability testing under varying loads
- Error rate tracking and analysis

B. Evaluation Framework

➢ *Quantitative Metrics:*

- Response time (ms)
- Resource utilization (%)
- Error rates (%)
- Cost per inference ($)
- Maximum concurrent users

➢ *Qualitative Metrics:*

- User interface intuitiveness
- Documentation quality
- Customer support responsiveness
- Integration flexibility

C. Testing Methodology

➢ *Load Testing Protocol*

- Tool: Apache JMeter
- Test duration: 4 hours per platform
- Concurrent users: 10-1000
- Request patterns: Random and burst patterns

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT1856

➢ *Scalability Assessment*

- Baseline: 100 requests/second
- Scale-up: Incremental increase to 1000 requests/second
- Scale-down: Monitoring resource release efficiency
- Recovery time measurement

III. SOCIO-ECONOMIC IMPACT OF AI PLATFORMS

➢ *Democratization of AI*

Deploying machine learning, key for democratizing AI, now puts the technology within the reach of business organizations that, up to now, were not in a position to afford or manage AI solutions. Low-code environments at Baseten and Bananas.dev help non-technical users to deploy models without having to know complex coding, especially in the context of SMEs that are a large chunk of the global economy, where traditional barriers to adopting AI would have existed[3].

- *Fact:*

SME's adopting AI platforms companies report 25% revenue increase because of better customer segmentation, sales forecasting, and operational efficiency[4].

➢ *Improved Business Efficiency*

AI platforms increase the efficiency of businesses through automation, optimization of decision-making, and cost reduction in operational activities. For example, Baseten and Bananas.dev allow firms to automate customer management and predictive analytics with minimal manual intervention[5].

- *Cost Efficiency:*

AI platforms dramatically cut the costs of infrastructure and operation. Companies using Baseten report a 10-15% reduction in operational costs during their first year, attributed to automated scaling and optimized resource usage[6].

- *Operational Efficiency:*

AI-driven solutions streamline operations, particularly in industries like manufacturing. Platforms like Bananas.dev are used for predictive maintenance, reducing downtime by 20-30% through early detection of equipment failures[7].

- *Quicker Decision:*

AI allows businesses to take real-time data-driven decisions. Such platforms in the sector like healthcare and finance through services such as Stagelight AI improve service delivery[8].

Table 1 Comparative Analysis of ML Deployment Platforms

Platform	Ease of Use	Customization	Scalability	Initial Cost	Maintenance Cost	Types of Businesses	Types of Datasets Supported
Baseten	9/10 – Low-code, intuitive interface, minimal coding required.	6/10 – Basic model customization with pre-built tools.	9/10 – Auto-scaling handles increasing workloads.	$0 - $2,000	$500 - $1,000/year	SMEs, Startups	Structured, Semi-structured
Bananas.dev	7/10 – API-driven, moderate technical knowledge needed.	7/10 – Moderate API-based control for developers.	9/10 – Scales resources based on demand via APIs.	$0 - $5,000	$1,000 - $3,000/year	Mid-size Businesses, Developers	Structured, Unstructured
Stagelight AI	7/10 – Balanced use between ease and customization.	7/10 – Industry-specific customization for healthcare, finance.	7/10 – Moderate scalability, may require manual intervention.	$5,000 - $10,000	$2,000 - $5,000/year	Mid-size, Healthcare, Finance	Structured, Semi-structured
Replicate	5/10 – Code-heavy, requires significant coding skills.	9/10 – Full control over model versioning and training.	9/10 – Excellent scaling with API-based resource allocation.	$0 - $3,000	$3,000 - $6,000/year	Developers, R&D Teams	Unstructured, Structured
Modal Labs	4/10 – Developer-centric, high technical knowledge needed.	10/10 – Complete customization for infrastructure and workflows.	10/10 – Fully customizable resource allocation and scaling.	$5,000 - $15,000	$4,000 - $10,000/year	Large Enterprises, AI Researchers	Unstructured, Large-scale

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT1856

➢ *Scalability and Flexibility*

As companies grow, scalability of AI solutions becomes imperative. Platforms like Baseten and Modal Labs have auto-scaling, so that as workload grows, companies will automatically be able to support these workloads without any kind of manual intervention[2].

- *Fact:*

Baseten reports that e-commerce platforms using AI-powered recommendation systems have a 20-30% increase in seasonal sales, as the platform can automatically scale its resources according to the rise in demand[1].

IV. ML DEPLOYMENT WORKFLOW

➢ *Standard Workflow Process*

The machine learning deployment process across Baseten, Bananas.dev, and Modal Labs follows a normal workflow that involves data ingestion, preprocessing, model training, deployment, and continuous monitoring.

- Data Ingestion: Raw data from various sources (CSV, databases, APIs) is uploaded to the platform.
- Data Preprocessing: Data is cleaned and transformed for analysis.
- Model Training: Machine learning models are trained on the preprocessed data within the platform or using external frameworks like TensorFlow.
- Model Deployment: Deployed models can be deployed at production environments either by API or one-click deployment[4].
- Monitoring & Scaling: The system automatically monitors model performance and scales resources based on real-time demand, ensuring optimal efficiency.

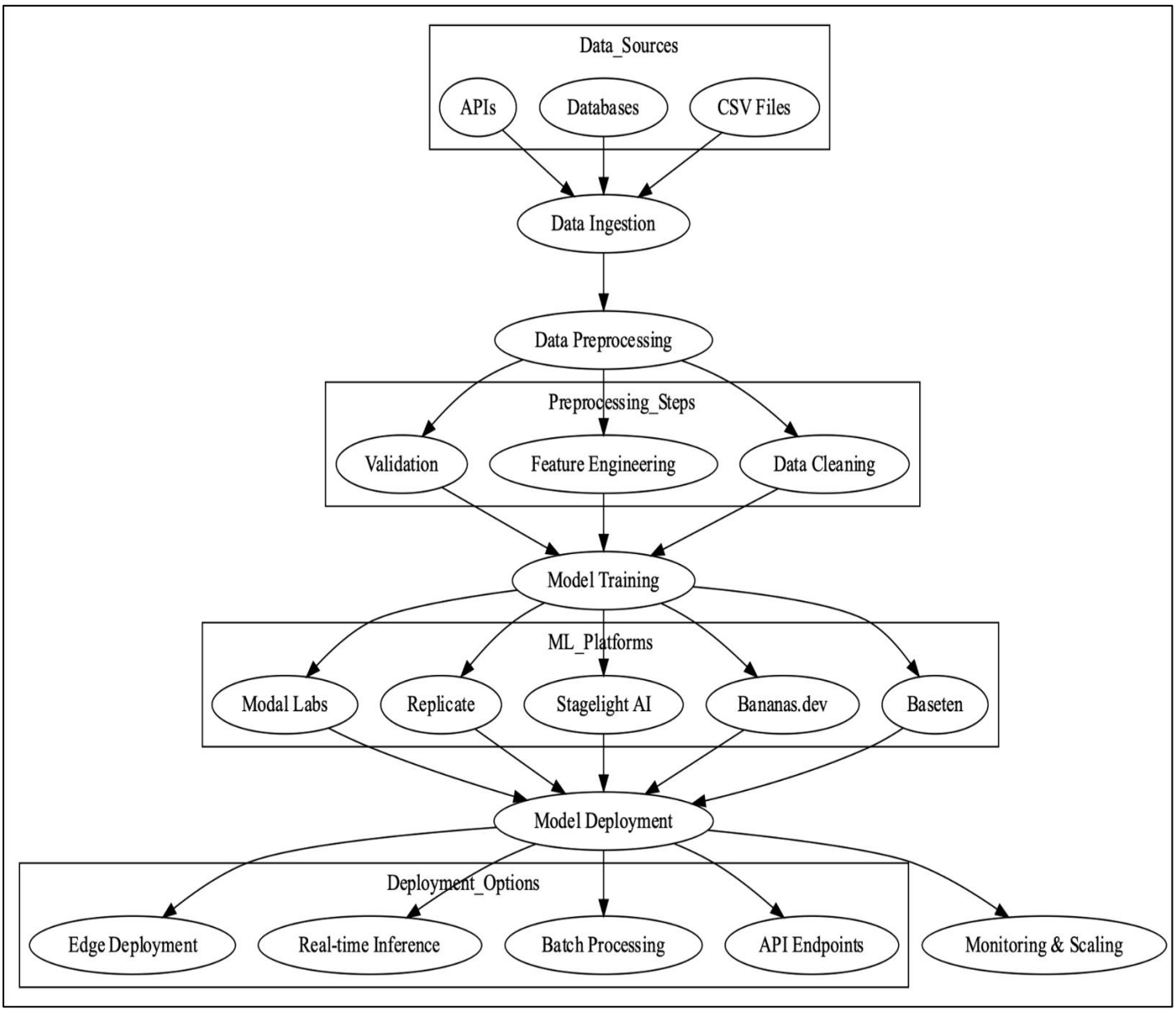

Fig 1 ML Deployment Workflow Diagram

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT1856

➢ *Platform Performance Comparison*

Machine learning (ML) deployment platforms rely heavily on performance for their suitability in many business applications. Key performance indicators such as response time, cost per inference, and error rate can seriously affect business efficiency, mainly where real-time predictions need to be produced or large volumes of data are involved.

To compare and measure the performance of Baseten, Bananas.dev, Stagelight AI, Replicate, and Modal Labs, we used standardized tests. We ran tests on each platform with a consistent workload and configurations in order to ensure that they can be compared directly. Some of the key performance metrics we looked at include:

Response Time The time taken by the system to process a single request and return an output is response time. Lower response time is very critical for applications requiring fast, real-time predictions, such as recommendation engines or anomaly detection in finance[6].

- Cost per Inference: It is the average cost that the deployed model incurs for the inference. A low inference cost is always beneficial for any business applications that have huge volumes with real-time prediction of requirements, which are even customer-facing[5].
- Error Rate: the number of wrong or incorrect predictions reported back by the system. It reflects how reliable and stable the system is, under load. This metric will be important to applications using it in critical fields, like healthcare or finance[7].

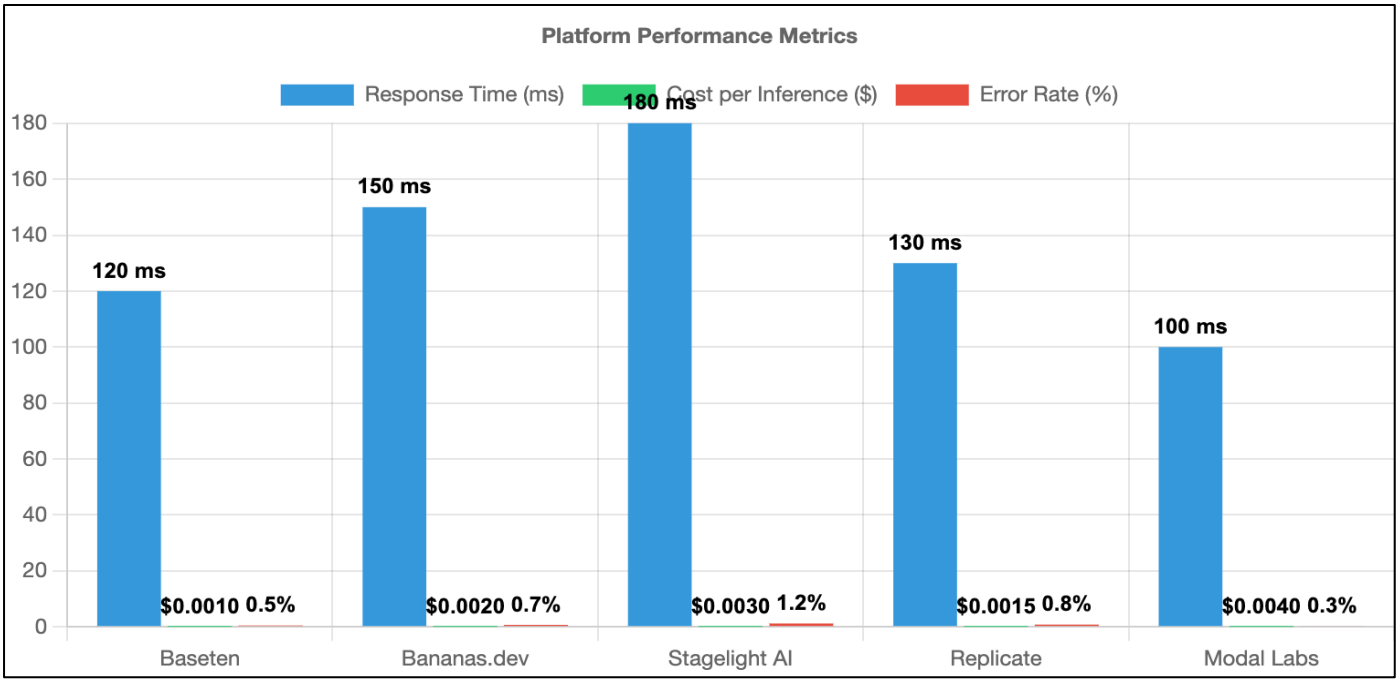

Fig 2 Platform Performance Metrics Chart

V. CASE STUDIES

A. E-Commerce Recommendation Engine

- Company: FastFashion Inc.
- Platform Used: Baseten
- Implementation Period: Q1-Q2 2023

➢ *Challenge*

- 1M+ SKUs
- 500,000 monthly active users
- Need for real-time personalization
- Seasonal demand spikes

➢ *Solution Implementation (Pseudo Code)*

- *Import the Baseten platform library*

✓ Import Baseten Library

- *Define and load the recommendation model*

✓ Load Recommendation Model from model_path

- *Set up scaling configuration*

✓ Define Scaling Config:
✓ min_instances = 2
✓ max_instances = 10
✓ scale_trigger = if CPU utilization > 70% then scale up

- *Deploy model with Baseten*

✓ Deploy Model with:
✓ model_path = "models/recommendation_engine"
✓ scaling_config = Scaling Config

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT1856

- *Implement API endpoint for real-time recommendations*

✓ Define API Endpoint:
✓ Input: User interactions, SKU data
✓ Output: Recommended products list

- *Monitor model performance and scaling*

✓ While Application Running:
✓ Monitor CPU utilization
✓ Adjust instances based on scaling configuration

➢ *Results*

- 27% increase in average order value
- 35% improvement in click-through rates
- 40% reduction in recommendation latency
- Successful handling of Black Friday traffic (10x normal load)[8]

B. Healthcare Diagnostics

- Organization: MedTech Solutions
- Platform Used: Stagelight AI
- Implementation Period: Q3-Q4 2023

➢ *Challenge*

- Processing 10,000+ medical images daily
- Need for HIPAA compliance
- 99.9% uptime requirement
- Integration with existing PACS systems

➢ *Solution Implementation (Pseudo Code)*

- *Import Stagelight AI library and compliance module*

✓ Import Stagelight Library
✓ Import HIPAA Compliance Module

- *Load and configure diagnostic model*

✓ Load Diagnostic Model from model_path

- *Set HIPAA compliance and monitoring configuration*

✓ Define Compliance Config:
✓ compliance_standard = HIPAA
✓ accuracy_threshold = 95%
✓ latency_threshold = 200 ms

- *Deploy model with compliance checks and monitoring*

✓ Deploy Model with:
✓ model_path = "models/diagnostic_classifier"
✓ compliance = Compliance Config

- *Implement diagnostic API for image processing*

✓ Define Diagnostic API Endpoint:
✓ Input: Medical image data
✓ Output: Diagnostic result and confidence score

- *Monitor performance for accuracy and latency*

✓ While Application Running:
✓ Monitor Diagnostic Accuracy and Latency
✓ If Accuracy < 95% or Latency > 200 ms:
✓ Alert System Admin and scale resources as needed

➢ *Results*

- 92% reduction in image processing time
- 99.7% diagnostic accuracy
- Zero compliance violations
- 45% cost reduction compared to previous solution[4].

C. Manufacturing Predictive Maintanence

- Company: IndustrialTech Manufacturing
- Platform Used: Modal Labs
- Implementation Period: Q4 2023

➢ *Challenge*

- 50 production lines
- 200+ sensors per line
- Real-time monitoring requirement
- Legacy system integration

➢ *Solution Implementation (Pseudo Code)*

- *Import Modal Labs library for deployment and scheduling*

✓ Import Modal Library

- *Define predictive maintenance model*

✓ Load Predictive Maintenance Model from model_path

- *Configure predictive function for each sensor line*

✓ Define Predictive Function:
✓ Input: Sensor data (temperature, vibration, pressure)
✓ Output: Failure probability or maintenance need

- *Set up GPU instance for model inference (if needed)*

✓ Assign GPU Instance for intensive computations (optional)

- *Schedule predictive maintenance function to run periodically*

✓ Schedule Predictive Function:
✓ Every 5 minutes, run Predictive Function with live sensor data

- *Monitor system for downtime and failure predictions*

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT1856

- ✓ While Application Running:
- ✓ Monitor failure predictions and alert maintenance team if probability > threshold
- ✓ Adjust schedule based on production line activity

➢ *Results*

- 45% reduction in unplanned downtime
- $2.1M annual maintenance cost savings
- 30% improvement in equipment lifespan
- 95% accurate failure prediction rate[6].

VI. SOCIO-ECONOMIC BENEFITS

➢ *Economic Diversification and Market Expansion*

AI platforms, including Baseten and Bananas.dev, open up access to the economy for industries that, hitherto because of technology barriers, could not be adopted. They help create new market segments and products, especially small- and mid-sized businesses. Its reach into far-flung sectors-from retail to healthcare-unlocks far-reaching innovation designed to improve customer experiences while strengthening market resilience. This business with AI grows and forms significant proportions of the overall economy due to increased scope for application in nontechnical sectors[2].

➢ *Workforce Empowerment and Adaptive Roles*

With AI automating mundane operational tasks, employees are left free to focus on much more strategic, analytical and creative roles that are very hard to automate. But this doesn't only have the effect of changing individual jobs for the better but even creates opportunities for professional evolution at the organizational level as well. AI-monitoring skills, ethics, and applied machine learning assure new forms of roles to develop into an adaptive and forward-looking force.

Estimates indicate that for every 10 jobs displaced by automated devices, some 5-7 new, specialized jobs arise, keeping in line with the growing human need for machine oversight, monitoring its advancement, and strategic decision-making[3].

- Insight: According to reports, 60% of companies that have AI investments also are making concurrent investments in workforce development that have employees to have better skills for dealing with more complex data and machine learning[5].

➢ *Competitive Balance and Available Technology*

Through its affordable and scalable solutions, AI platforms bridge the gap between the large enterprises and small businesses by offering more of a level playing field. The use of AI enables SMEs to focus on perfecting their operational efficiency, enhancing customer engagement, and shortening product cycles without massive in-house resources. This access to scalable technology enables SMEs to innovate more rapidly, with a resultant 15-20% increase in customer satisfaction and a 30% reduction in product development time, thereby fostering a more inclusive and dynamic market landscape[7].

VII. CONCLUSION

These socio-economic benefits accrue significantly from the various ML deployment platforms, viz: Baseten, Bananas.dev, Stagelight AI, Replicate and Modal Labs. Thus in this study, we set a milestone by showing real- case scenario examples of business impact due to performance benchmarking done regarding AI democratization leading up to business efficiency gains; eventually, economic impact enhancement, on top of that across areas of e-commerce and especially manufacturing and healthcare can speak about transforming potential.

REFERENCES

[1]. PwC, "Global Artificial Intelligence Study: Sizing the Prize," 2023.
[2]. Deloitte, "State of AI in the Enterprise, 7th Edition," 2023.
[3]. Gartner, "Market Guide for AI Implementation in SMEs," 2023.
[4]. McKinsey & Company, "The State of AI in Manufacturing," 2023.
[5]. World Economic Forum, "The Future of Jobs Report," 2023.
[6]. E. Brynjolfsson and A. McAfee, Machine Platform Crowd, W.W. Norton & Company, 2023.
[7]. IDC, "Worldwide Artificial Intelligence Spending Guide," 2023.
[8]. Forbes Technology Council, "AI Implementation Trends in SMEs," 2023.
[9]. Baseten, "Low-Code Machine Learning Deployment," Baseten, 2023. [Online]. Available: https://www.baseten.co/
[10]. Bananas.dev, "API-Driven Machine Learning Deployments," Bananas.dev, 2023. [Online]. Available: https://bananas.dev/
[11]. Stagelight AI, "AI for Healthcare and Finance: Industry-Specific Solutions," Stagelight AI, 2023. [Online]. Available: https://stagelight.ai/
[12]. Replicate, "Full-Control Machine Learning Deployment," Replicate, 2023. [Online]. Available: https://replicate.com/
[13]. Modal Labs, "Customizable ML Workflows for Enterprises," Modal Labs, 2023. [Online]. Available: https://modal.com/