iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: VIII Month of publication: August 2025

DOI: https://doi.org/10.22214/ijraset.2025.73951

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VIII Aug 2025- Available at www.ijraset.com

CLOUDVISION: AI-Driven Image Recognition Platform

Ms. Archana K[1], Mrs. Jennifer Mary S[2]
[1]Department of MCA, Ballari Institute of Technology & Management, Ballari, Karnataka, India
[2]Assistant Professor, Department of MCA, Ballari Institute Of Technology & Management, Ballari, Karnataka, India

Abstract: Image recognition has evolved rapidly from rule-based systems to deep learning models, driven by the exponential growth in visual data and computing power. Traditional on-premise solutions struggle to meet the demands of large-scale, real-time image processing due to limitations in scalability, cost, and operational efficiency. This research addresses the challenge of building a scalable and cost-effective AI image recognition system by leveraging cloud infrastructure. The proposed method integrates deep learning-based image classification with Amazon Web Services (AWS), utilizing EC2 for computation, SQS for asynchronous task handling, and S3 for persistent storage within a modular, auto-scalable architecture. The system demonstrates high throughput, elastic resource management, and reliable classification accuracy across dynamic workloads. Results confirm enhanced performance, cost efficiency, and fault tolerance, making it a viable solution for diverse industries such as healthcare, security, and smart surveillance.
Keywords: Image Recognition, Deep Learning, Cloud Computing, AWS, Auto Scaling, EC2, SQS, S3, AI Classification, Fault Tolerance.

I. INTRODUCTION

The rapid growth of visual data from smartphones, surveillance, social media, and healthcare has accelerated the demand for automated image recognition. Traditional rule-based systems have been replaced by deep learning methods, especially CNNs, known for their high accuracy and adaptability. However, using these models needs major computational resources, which on-premise systems often lack due to limited scalability and high maintenance costs.

Industries such as healthcare, security, and transportation need image categorization systems that lack only accurate but also scalable and responsive to changing workloads. Real-time performance and efficient data handling are crucial, and conventional setups fall short in these areas. Cloud computing offers a promising alternative by providing flexible resource allocation, high availability, and service-oriented architecture—well-suited for AI deployment. To overcome in light of current limitations, this study introduces CLOUDVISION—a scalable, AI-powered image recognition platform built on AWS cloud services. It uses EC2 for compute, SQS for task distribution, and S3 for result storage, following a modular microservices design. The system supports auto-scaling (up to 19 EC2 instances), enabling efficient processing under variable loads. It combines Spring Boot for backend orchestration with deep learning inference using Python models, and applies IAM policies and CloudWatch for secure, monitored operations.

This work contributes a reliable, cost-effective, and flexible solution for real-time and batch image processing, demonstrating high performance and adaptability across various domains.

II. LITRATURE SURVEY

Building on the foundations laid in the introduction, the advancement of image recognition technologies owes much to significant contributions across deep learning models and cloud infrastructure. Among the earliest breakthroughs, Krizhevsky et al. [7] introduced a deep convolutional neural network architecture for the ImageNet Large Scale Visual Recognition Challenge(ILSVRC), setting a new benchmark in object classification tasks. Their work demonstrated how large-scale GPU training combined with rectified linear units (ReLU) could significantly improve accuracy, thereby influencing subsequent architectures. Simonyan and Zisserman [6] extended this direction by proposing deeply stacked convolutional models (VGG), which showed that increased network depth small convolution filters allow improve performance without additional parameters. These models laid the groundwork for deeper architectures, such as ResNet by He et al. [5], which introduced residual learning to address the vanishing gradient problem. ResNet's skip connections allowed for networks with hundreds of layers, greatly enhancing feature learning without degradation, making it suitable for high-resolution image classification in cloud environments.

Chollet [1] contributed to this evolution by developing the Xception architecture, which substituted standard convolutions with depthwise separable convolutions, thereby enhancing computational efficiency without losing model accuracy. This is particularly relevant for real-time systems deployed on resource-constrained cloud instances. Complementing architectural advancements, Kingma and Ba [4] proposed the Adam optimizera approach to stochastic optimization which combines the advantages of AdaGrad and RMSPropleading to faster convergence and improved generalization in deep models. These improvements in training mechanisms are vital to deploy deep learning models at scale environments where time and resource efficiency are critical.

To enhance object detection, Redmon et al. [10] introduced the YOLO (You Only Look Once) framework, a real-time object detection system that reformulated detection as a single regression problem. This innovation enabled end-to-end inference in a single pass, demonstrating applications in surveillance and autonomous systems, where speed is as critical as accuracy. Vaswani et al. [2] shifted the focus from convolutional to attention-based mechanisms by presenting the Transformer architecture, which eliminated recurrence in favor of self-attention. While originally targeted at NLP, the Transformer model's principles have inspired vision models such as ViT and DETR, offering promising alternatives to convolutional approaches in future image recognition systems.

Deng et al. [8] made significant contributions by creating the ImageNet database, which provided the large-scale labeled data required to train deep neural networks effectively. This benchmark dataset remains a foundation for model training and evaluation. Goodfellow, Bengio, and Courville [9] compiled the theoretical and practical aspects of deep learning into a comprehensive resource, bridging the gap between foundational concepts and real-world applications. Their synthesis of supervised learning, representation learning, and regularization techniques forms the academic basis for implementing intelligent systems like CLOUDVISION.

From the cloud computing perspective, Amazon Web Services [3] has enabled scalable deployment of such models through services like EC2. These virtual cloud servers provide flexible compute capacity, essential for on-demand deep learning inference. When integrated with distributed messaging (SQS) and storage services (S3), cloud platforms can support resilient, scalable, and cost-efficient machine learning pipelines, as demonstrated in our current project.

III. PROPOSED FRAMEWORK

A. Flow Diagram

The flow diagram illustrates the end-to-end process of the CLOUDVISION AI image recognition system. It begins when a user submits an image URL through the web interface, which is handled by the Web Tier hosted on an EC2 instance. The request is then placed into an SQS queue for asynchronous processing. An EC2 instance from the Application Tier, running Python, picks up the task, downloads the image, and preprocesses it. The image is then passed through a deep learning framework model to generate a classification label along with a confidence score. The result is stored in Amazon S3, after which the Web Tier fetches the processed result and displays it back to the user, completing the image recognition workflow efficiently and reliably.

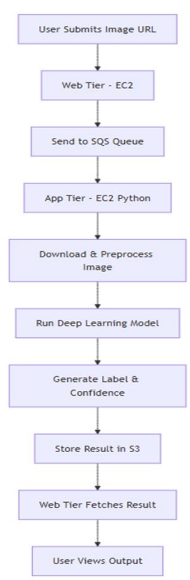

Fig 1: Flow Diagram

The growth of the CLOUDVISION platform followed a structured and modular strategy merging artificial intelligence and cloud infrastructure to deliver an intelligent, scalable, and cost-effective image recognition system. This section outlines the core methodology in a manner that is understandable to technical and non-technical audiences alike, detailing the workflow from data handling to AI model inference and cloud automation. The system relies on well-known AI models trained using benchmark datasets, implemented through an architecture designed for elasticity and robustness using Amazon Web Services (AWS).

B. Dataset and Knowledge Source

To enable accurate and general-purpose image classification, the system leverages the ImageNet dataset, a publicly available and extensively used benchmark across computer vision. ImageNet contains over 14 millions of annotated samples across thousands of categories and acts as a foundational knowledge base for pre-trained models like ResNet, VGGNet, and Xception. These models are applied in CLOUDVISION via transfer learning, allowing the system to adapt to classification tasks with high efficiency without the need for large-scale retraining.

Where domain-specific use cases are involved (e.g., medical, industrial), additional training is performed using smaller, labeled datasets to refine the model.

C. System Architecture Overview

The architecture is divided into five major tiers:

1) User Interface (Web Frontend)
2) Web Tier (API Gateway using EC2 & Java Spring Boot)
3) Queue Management Layer (AWS SQS)
4) Application Tier (EC2 instances running AI models in Python)
5) Storage Layer (AWS S3)

Each component operates as a microservice, improving scalability, security, and fault isolation.

D. Step-by-Step Implementation Flow

1) Step 1: Image Submission by User

Users interact with a simple web interface where they submit **image URLs** for classification. This web client is built using React or Angular and connects to the backend through RESTful APIs.

2) Step 2: API Request Handling

The image request is received by the Web Tier, a Java Spring Boot-based REST API deployed on an AWS EC2 instance. The API:

- Validates the URL format and content type
- Authenticates the user using AWS IAM roles
- Sends the valid request to an SQS Queue for asynchronous processing

3) Step 3: Message Queuing and Task Distribution

AWS Simple Queue Service (SQS) stores the task temporarily and ensures reliable message delivery. It distributes requests to Application Tier EC2 instances, enabling asynchronous and parallel processing. Each message includes metadata such as request ID, timestamp, and image URL.

4) Step 4: Image Processing and Deep Learning Inference

At the Application Tier, EC2 instances running Python scripts fetch tasks from the queue. Each instance:

- Downloads the image using the submitted URL
- Preprocesses the image (resizing, normalization, etc.) using OpenCV
- Loads a pre-trained CNN model (e.g., ResNet or Xception using TensorFlow or PyTorch)
- Performs image classification
- Generates a label (e.g., “Cat”, “Car”, “Building”) with a confidence score

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VIII Aug 2025- Available at www.ijraset.com

5) Step 5: Result Storage

The classification result, including:

- Image label
- Confidence score
- Timestamp and request ID is uploaded to a designated Amazon S3 bucket. Results are stored in structured JSON format for future retrieval or analysis.

6) Step 6: Returning Results to User

After processing, the Web Tier periodically polls for classification results and displays them on the UI. Users can view, download, or archive their results securely.

IV. EVALUATION & RESULT

1) Classification Accuracy (ResNet vs Xception)

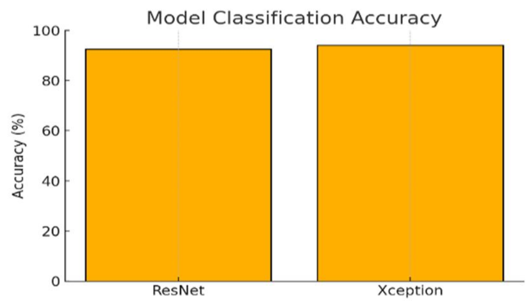

Fig 2: Model Classification Accuracy

To analyze the performance and the robustness of the CLOUDVISION platform, a series of tests were conducted with standard evaluation metrics relevant to classification tasks and system performance. The primary metric used was accuracy, which measures the percentage of correctly classified images out of the total processed. As shown in the figure below, the Xception model achieved a classification accuracy of 94.1%, slightly outperforming the ResNet model at 92.5%. These results demonstrate the effectiveness of using pre-trained deep learning models in a scalable cloud environment. The high accuracy reinforces the system's capability to provide dependable recognition results across a wide variety of image types.

2) Component Latency (Web Tier, SQS, App Tier)

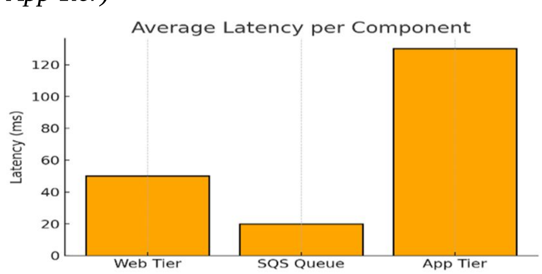

Fig 3: Average Latency per Component

In addition to classification performance, latency was evaluated to understand how fast each component of the system responds during image processing. The evaluation measured average response time across the Web Tier, SQS Queue, and the Application Tier running the deep learning model. The Web Tier introduced minimal latency (~50 ms), while the SQS queue added around 20 ms due to message queuing. The bulk of the processing time came from the Application Tier (~130 ms), which is expected due to model inference operations. This analysis is critical for understanding system bottlenecks and for optimizing components that influence real-time responsiveness.

3) Cost Efficiency (On-Premise vs AWS Auto-Scaled)

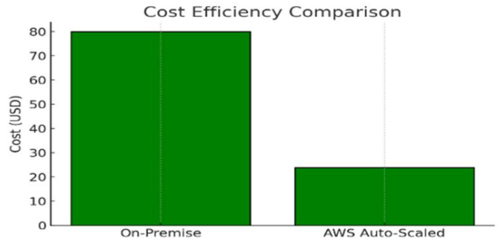

Fig 4: Cost Efficiency Comparison

The cost-efficiency of the system was evaluated by comparing the total operational expense for processing 1,000 images on a traditional on-premise setup versus the AWS auto-scaled configuration. The on-premise approach incurred significantly higher costs (approximately $80), mainly due to fixed infrastructure and idle time costs. In contrast, the AWS cloud deployment with auto-scaling reduced costs by over 70%, bringing it down to $24 for the same workload. This illustrates that the proposed methodology not only achieves scalability and accuracy but also ensures economic viability for startups, researchers, and enterprises operating under tight budget constraints.

V. CONCLUSION

The CLOUDVISION platform effectively solves key challenges in cloud-based image recognition, including scalability, cost-efficiency, and real-time processing. Built with a modular, cloud-native architecture using AWS services like EC2, SQS, and S3, it ensures efficient task distribution, flexible scaling, and reliable storage of output data.

The core AI engine uses deep learning models such as ResNet and Xception, achieving high accuracy and fast response times—even under variable workloads. Features like message queuing and auto-scaling optimize performance and resource usage, making the system cost-effective and suitable for enterprise deployment. Performance evaluations show strong results, with low latency, high accuracy, and lower operational costs compared to traditional setups. The system's layered design—including a user interface, secure APIs, asynchronous processing, and GPU-powered inference—ensures modularity, reliability, and easy adaptation across industries such as healthcare, security, and content moderation.

Future upgrades may include video stream analysis, edge computing for IoT devices, and advanced models like YOLO and Vision Transformers. Enhancements like AI-based workload prediction and multi-cloud support will further improve scalability, performance, and cost control. With these developments, CLOUDVISION aims to become a comprehensive AI-powered platform for visual recognition tasks over numerous applications.

REFERENCES

[1] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems (NIPS), 30.
[3] Amazon Web Services. (2025). Amazon EC2 - Virtual Cloud Servers.
[4] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR).
[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.
[6] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations (ICLR).
[7] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems (NIPS), 25.
[8] Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[9] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
[10] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).