International Journal of Advanced Trends in Engineering and Management
(ISSN 2583-7052)

Vol. IV, No. 09, pp. 23-32, September

AI-Driven Autonomous Image Classification Using Agentic Deep Learning in a Cloud Computing Environment

Sushil Prabhu Prabhakaran[1]

[1]Full Stack Lead, Tata Consultancy Services, USA.

Corresponding Author E-mail: [1]sushilprab@gmail.com

ABSTRACT: The rapid increase in large-scale visual data in cloud computing systems appeals for intelligent, autonomous and highly scalable AI- driven image classification systems. This work proposes a novelty of agentic-based DL framework applied to image classification in cloud computing system and increases accuracy through enhanced operational efficiency with a new agent-driven AI system. The DL architecture is adaptable and closely connected to an autonomous AI agent. The new proposed architecture includes an agent driven by hybrid attention guided Convolutional Neural Network (CNN) that adapts during pre-processing and tuning of model to real-time feedback and performance that is capable of controlling image pre-processing inference procedures. With deployment on a cloud-native architecture, the proposed framework increased capability for distributed training, elastic resource allocations and large-scale inferences. A series of experiments conducted on standard image datasets proved the proposed agentic DL framework outperformed the existing methods, this validates the effectiveness of agentic intelligence in the development of next generation intelligent and autonomous cloud-based image classification systems.

Keywords: Image classification, Agentic- based DL, Cloud computing environment, hybrid attention guided CNN.

1. Introduction

Through a combination of technological advances and increased investment in infrastructure to support the expansion of digital imaging technologies, AI-based automatic classification of digital images has developed into a key area of expertise that are leveraged across a variety of industries, including the healthcare industry, where this ability is used to enhance the clinical care provided by physicians to their patients through the application of advanced imaging technologies that assist doctors with the diagnoses of disease and with the coordination of treatment for those patients, diagnosed with a specific illness [1].

Autonomous vehicles utilize AI-enabled automated identification and classification of images in order to create an intellectual image of their surroundings and to identify people, traffic signals, and other vehicles so that they make intelligent decisions that allow for safe navigation [2]. Additionally, video surveillance systems that use automated AI-powered image identification technology are able to filter through large amounts of recorded video footage in real time to identify people or objects that are of interest in order to prevent harm or theft [3]. Finally, visual recognition is used in the personalization of marketing by allowing marketers to analyze their consumers' images and preferences so that they serve targeted marketing messages more effectively and improve their overall marketing effectiveness [4]. AI-assisted image recognition capabilities has become increasingly complicated for traditional computer systems due to overwhelming amounts of visual information growing day-by-day. Due to the sheer size and variety within images generated by the industry each day, conventional systems struggle to handle those challenges [5-6]. The advent of cloud computing represents one of the greatest advancements in technology. Through the ability to provide flexible resource allocation as demands change, cloud computing represents one of the greatest opportunities for organizations to utilize their computing resources more efficiently. In order to maximize efficiency of processing images, particularly regarding their use for training and model creation, organizations need the capability to scale up the amount of processing power used as needed in order to handle the extreme workloads required for these tasks [7].

Cloud computing also enables organizations without large amounts of physical space or financial resources to have access to sufficient storage, processing power, and other necessary resources to appropriately deploy advanced image recognition systems. By utilizing the cloud, organizations take advantage of the power of computing resources located within data centers rather than being restricted by anything which is physically available to them [8]. The advanced AI technology known as DL provides a method to classify images when used with image processing. DL uses the design of an Artificial Neural Network that imitates the technique that our intelligences are designed and work. Due to this similarity to human intellects, CNNs perform calculations using artificial neurons that are traversed through certain layers of the CNN [9]. Therefore, AI-enhanced CNNs are one of the most potent branches of DL and it is very useful for many image classification ventures. However, as the image data transferred over the internet into offsite cloud servers for processing, there is a large amount of delay in transmitting this information and a significant amount of bandwidth being used, resulting in lower responsiveness and making it challenging to do any type of classification during real time [10]. To overcome this issue the proposed system uses an agentic-based DL which is used to increase robustness by automatically starting remedial learning procedures, spotting anomalies, and recognizing model drift.

1.1 Related Works

Godwin Olaoye et al [2025] [11] have developed a Cloud-Based Self-Learning Neural Networks for Autonomous AI Systems. They learn from new information automatically and adapt in real-time to changing situations, thus enabling organizations and businesses to use these systems in the most effective way for dynamic, large scale applications. With the ability to run in the cloud, self-learning neural networks take advantage of the computing power and elasticity of the cloud, which means that they train faster and process very large datasets Nevertheless, continuous cloud-based learning could create additional vulnerabilities in terms of privacy and security, since there is continuous communication of sensitive or proprietary information between networks.

Shantanu Kumar et al [2025] [12] have proposed a utilizing cloud computing to facilitate AI-accelerated distributed training of residual CNNs for aerial image classification. With virtually limitless computing ability, cloud computing enables a significantly accelerated training period for highly depth and skip-connection architecture based Deep Residual Neural Network (DRNN). Furthermore, cloud-based dispersed training eliminates the need for pricey local gear and improves collaboration between institutions. However, the heavy dependency on large cloud-based infrastructures imposes much higher operational costs associated with training large models across large clusters of nodes.

Cheng Ji et al [2025] [13] have established an autonomous self-healing through intelligent fault detection using cloud-based AI systems and Large Language Models (LLMs). LLMs are designed specifically to read and analyze massive amounts of log info, telemetry or other forms of unstructured operational data, resulting in very precise fault predictions and capabilities of identifying anomalies at extremely high rates compared to other technologies. Nonetheless, LLM create a higher level of complexity and less transparency, thus increasing the potential difficulty to provide rationale for systems failures to the user or allowing for proper auditing and establishing a sense of trust.

Taiwo Joseph Akinbolaji et al [2024] [14] have implemented the globalization of artificial intelligence and machine learning for the real-time detection of AI-enabled threats within cloud-based environments. As a consequence of developing capabilities associated with cloud-native technologies, the ability to perform distributed monitoring on multiple-cloud infrastructures has allowed for enhanced adaptive capabilities and increased proactive measures with respect to identifying threats. However, due to the complexity of these systems, there is potential for challenges with respect to understanding the security professionals interpret model explanations and verify that automated activities have taken place based upon the aforementioned models.

Biswanath Saha et al [2025] [15] have suggested a cloud-based technology for synthesizing images using GANs. It is increasingly popular due to its ability to reduce bandwidth usage by allowing only the required data to be sent from edge devices to the cloud. In addition, using both edge and cloud resources allows for distributed learning and collaborative model improvement, and provide increased privacy because sensitive raw data remain on the local device rather than being sent to a remote location. However, coordinating computations between edge and cloud devices creates additional complexity to the whole system, and thus, there needs to be strong orchestration, synchronization, and fault tolerance mechanisms in place for effective operation.

1.2 Contributions

The main contributions of this research includes,

- To design an AI-agentic deep learning system that autonomously manages picture preprocessing, model tweaking, and inference in cloud environments.
- To implement into practice an AI-powered hybrid Attention-Guided CNN agent for adaptive, real-time picture categorization.
- To provide elastic cloud resource allocation and distributed training possible for scalable large-scale inference.
- To improve robustness, accuracy, and convergence speed over traditional CNN and static cloud classifiers.
- To validate the efficiency of agentic intelligence for autonomous, intelligent cloud-based image classification systems.

2. Proposed System Description

The proposed system is an AI-intelligent and autonomous image classification framework that responds to the growing prevalence of large-scale visual data within cloud environments by taking in and storing image data. This data is managed through an AI-agentic controller, which employs a hybrid Attention-Guided CNN. The proposed block diagram is shown in figure 1.

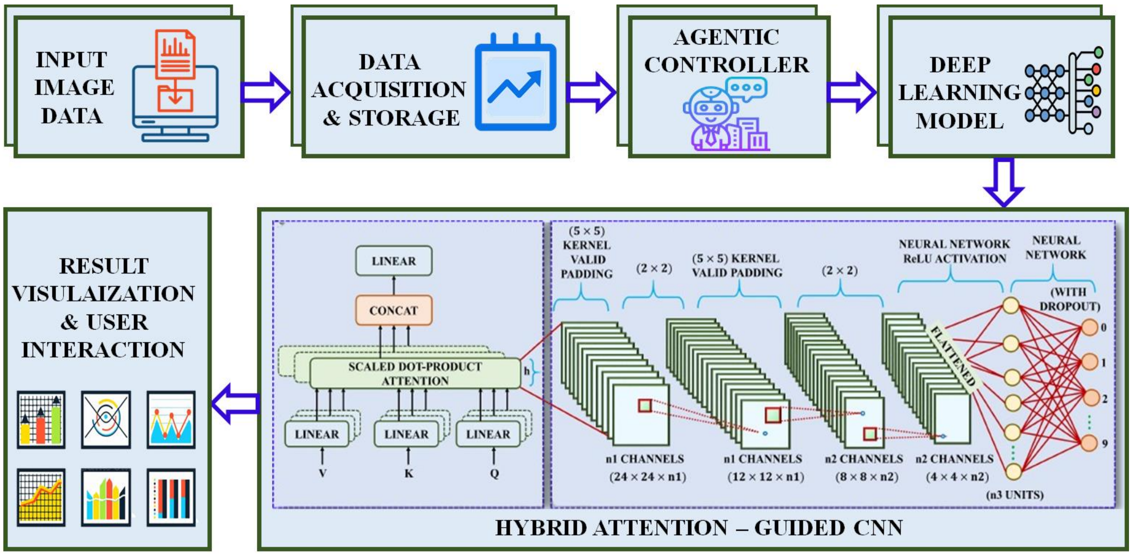

Figure 1: *Proposed block diagram*

By using an agentic controller, image preprocessing, model parameter tuning, and inference are dynamically adjusted based on real-time feedback from the data, and through the implementation of distributed training and elastic resource allocation, the system easily scale up or down and adjust accordingly based on changing workloads. A dedicated layer that is closely related to the AI-agentic controller classifies or predicts each image, thus improving accuracy and allowing for more correct decisions to be made. Finally, the system provides visual information, allowing for further analysis and user interaction with actionable insight. Therefore, the integration of agentic intelligence and cloud scalability greatly improves accuracy, convergence time and overall robustness compared to existing CNN and static cloud-based image classifiers, thus creating the potential for a compelling next generation autonomous image classification solution.

3. Proposed System Modelling

3.1 Cloud Computing Infrastructure

In recent times, cloud computing has overturned the traditional means of acquiring and working with one's own computer equipment. Instead of relying on physical servers to house the data. The software used to run a business, computer resources like storage, processing power, and applications are hosted on the internet. Some of the dominant cloud service providers today are Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure. Through these cloud service platforms, customers are able to meet a broad array of computing needs through numerous offerings, ranging from the most fundamental storage and computing capabilities through to more complex technologies such as Machine Learning (ML) and big data AI-analytics. The following list outlines the various categories of cloud services that are available, cloud service models of image classification is shown in figure 2.

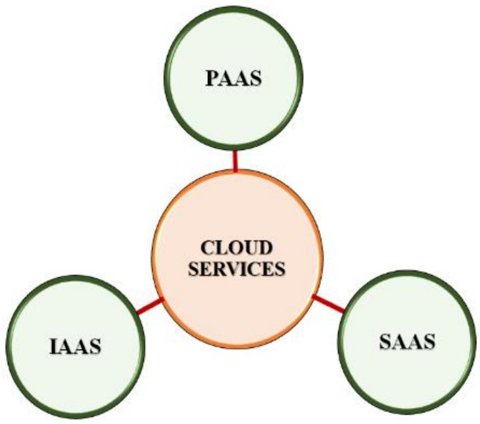

Figure 2: *Cloud service models for image classification*

3.1.1 Infrastructure as a Service (IAAS)

IAAS Services provide the ability to use simulated computer equipment over the Internet. It gives customers complete control of their computer system. Users manage their own IT equipment, and work collaboratively with other customers. Users must pay for all IAAS equipment they use on the Internet. Diagram of IAAS is depicts in figure 3. Examples of key components of IAAS include,

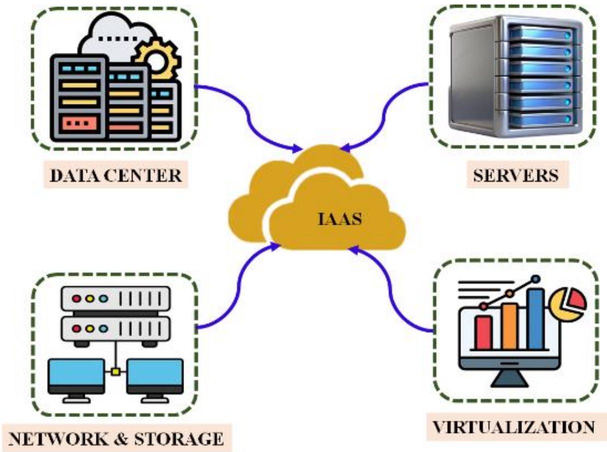

Figure 3: *IAAS cloud computing*

1. Virtual Machines- Users can use their own applications and Operating Systems (OSs) on virtual machines provided by IAAS providers, using a specified number of CPUs, memory and storage. Google Compute Engine (GCE) and Amazon Web Services Elastic Compute Cloud (AWS EC2) are two examples.

2. Scalability- It allows users to easily adjust the amount of resources they use, and only charge them for what they used during the billing period. This is particularly beneficial for companies whose business needs vary from month to month.

3. Storage- IAAS includes several types of storage (Block Storage, Object Storage, and File Storage), each designed to work with specific types of data and data access patterns.

3.1.2 Platform as a Service (PAAS)

With PAAS, customers create, operate, and manage their applications without needing to maintain the underlying technologies. Therefore, it allows to avoid dealing with system administration and focus instead on developing code and creating applications. Some key features and examples include:

1) Development Tools- PAAS provides a set of AI-development tools, frameworks, and middleware that simplifies the app creation process. Sample PAAS offerings include Google App Engine and AWS Elastic Beanstalk.

2) Managed Services- PAAS services manage the infrastructure that supports the application, including server, storage, network, and security. PAAS services manage development tools, databases, and tools for application creation. Therefore, application developers concentrate on writing code without worrying about completing administrative tasks.

3) Automated Scaling.-PAAS systems automatically scale their applications based on need, ensuring optimal performance while maximising the use of available resources.

3.1.3 Software as a Service (SAAS)

SAAS refers to software applications that are accessible over the internet and are paid for based on a user's payment plan. As shown in figure 4, these types of apps allow users to access them directly from an online computer, and ensure that they don't have to perform any updates or keep track of them themselves. This makes the SAAS model very convenient and typically inexpensive for users who want to use applications for routine purposes. Some of the key points to remember with regards to SAAS are,

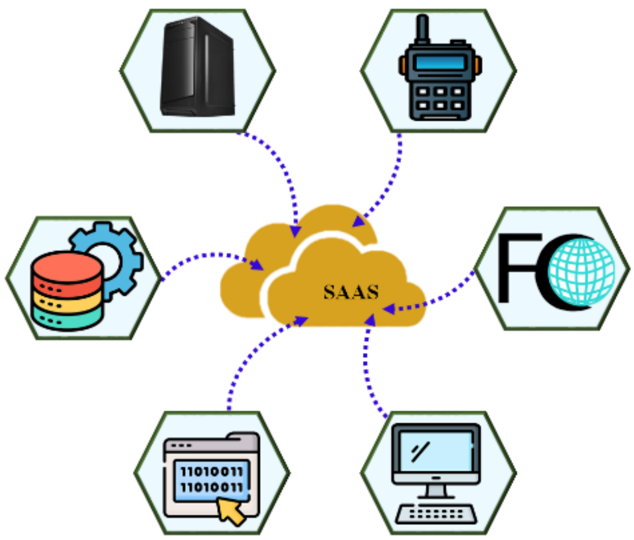

Figure 4: *SAAS cloud computing*

1) Accessibility-With SAAS apps, users have access to their applications from any device capable of connecting to the internet; however, they do not need to worry about transferring applications to a different device to continue using them. Google Workspace and Microsoft Office 365 are two prominent examples.

2) Automatic Updates-Companies providing SAAS applications perform all maintenance, as well as updates, to their users' applications. Therefore, users do not have to maintain/upkeep their software and can always access new capabilities as well as feature enhancements.

3) Subscription Model-Most SAAS offerings follow a subscription-based payment model, whereby users only pay for features they require. Additionally, pricing plans are usually variable, based on volume of activity or quantity of users.

3.2 Agentic Image Classification System

For a fully autonomous image classification solution, multiple AI-DL models are required. Due to factors such as the quality of images, variety of inputs, and fluctuating computing resources in cloud environments, it is essential that a solution dynamically adjust AI-adaptive pre-processing, model selection, and inference methodologies. Agentic Intelligence provides this capability by enabling the system to reason about its current state of operation, make modifications as necessary, and continually optimize the performance of its various aspects rather than relying on predefined fixed pipeline configurations. Levels of agentic capability in image processing system is shown in figure 5.

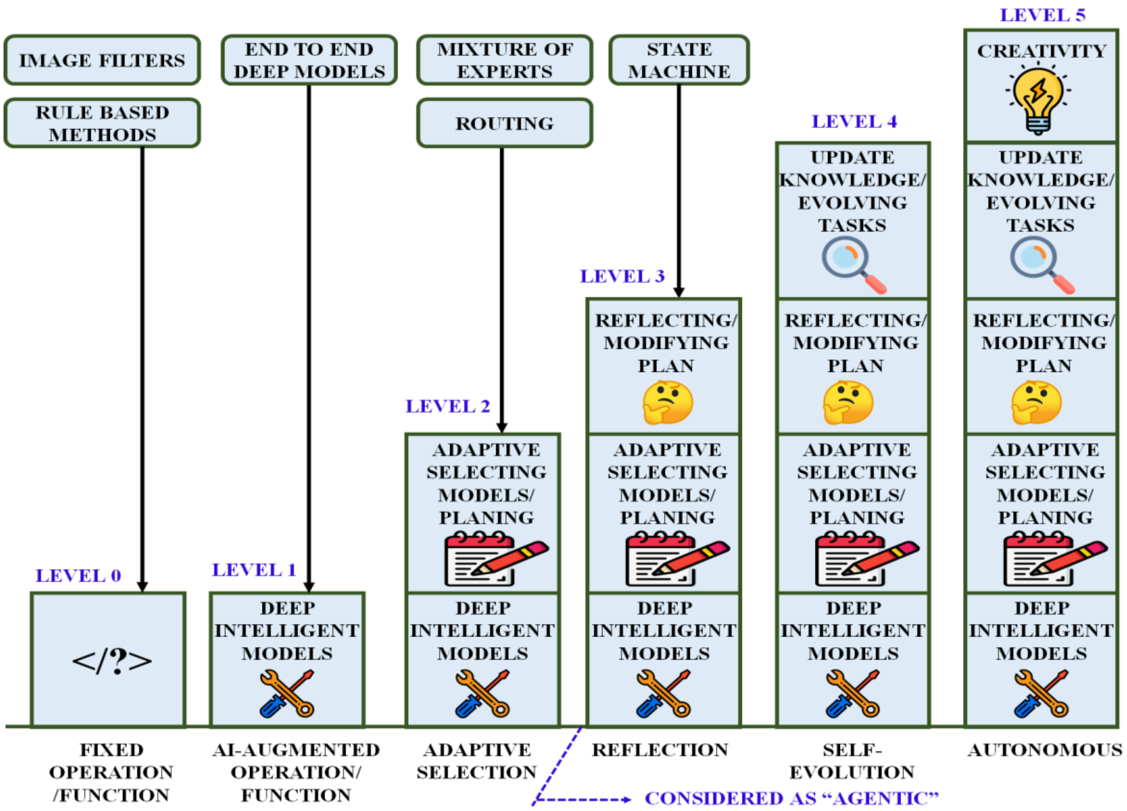

Figure 5: *Levels of agentic capability in image processing system*

3.2.1 Paradigms of Current Classification System

End-to- End

Conventional CNN or Transformer-based classifiers produce a label straight from an image. They are strong yet inflexible, unable to modify preprocessing, fine-tune settings, or modify behavior in response to changes in cloud resources or picture complexity.

Pipeline

In contrast, pipelines decompose the classification process into stages, which offer a modular method to attain better algorithms from a variety of sources, but all pipelines still require humans to build and configure them and do not have the ability to automatically reconfigure when handling new image types.

Mixture-of- Experts (MoE)

MoE approach provide some coverage of classifications by using multiple classifiers and gathering the best expert or classifier to use for each classification task. MoE models have some coverage but can't automatically select classifications, as they are based on either static routing decision logic or limited routing decision processes.

Routing based systems

Routing take the MoE concept and provide the flexibility of creating many potential paths and models for classification, but once the path is chosen, routing has no ability to re-evaluate the selection of prior processors in the routing.

State Machine

Although state machine models allow for a decision process in a step-by-step format, allow for revisiting the prior state to either modify or refine the previous steps of the workflow, provide for the ability to select the best model to use, and provide for the iterative workflow process to achieve a definitive classification. They allow for an infinitely adaptive and self-correcting process closely aligned with the DL framework as proposed.

3.3 Hybrid Attention-Guided CNN

AI-Self-directed image classification conducted in a cloud-based computing atmosphere, the self-attention function enables the image classification model to designate importance to particular regions of an image by computing attention scores related to different parts (patches) of an image. To create these attention scores, a combination of three vector types: a query (Q), a key (K), and a value (V), is used and the output of the attention score calculated by the following formula,

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

AI-attention mechanism helps to capture the global relationships within an image, improving classification accuracy. The dot product of query vectors with key vectors, after being scaled, is passed through a softmax function that produces an attention weight that will be multiplied by the value vector. By doing this, the model is able to learn the relationships between the parts of an image and create a structured representation of that image. The ability of this attention mechanism to scale in the cloud enables it to efficiently classify images based on large amounts of data and provide real-time results for applications. When classifying images in the cloud, CNNs use a series of convolutional layers to extract features from an image. This process is performed by applying a convolutional operation to the input image, mathematically expressed as,

$$O = (W * X) + b \quad (2)$$

Where, O is the resulting output image, W is the weight matrix (or filter), X is the input, and b is the bias. The size of the output image then be calculated using the following formula,

$$Output\ Size = \left(\frac{Input\ Size - Filter\ Size}{Stride}\right) + 1 \quad (3)$$

With same padding, the size of the output image will remain the same as that of the input image and with valid padding, the size of the output image will be less than that of the input image. When CNN's are deployed in the cloud, they leverage the power of the cloud to efficiently process large amounts of data and provide real-time decisions on classification actions for applications such as detecting objects.

The Hybrid Attention-Guided CNN combines convolutional layers to extract features from images with a combined attention mechanism to focus on important areas of an image. The attention mechanism calculates attention scores from the query, key and value vectors which provide global relationships while the CNN provides local features. The diagram of hybrid attention-guided CNN is shown in figure 6.

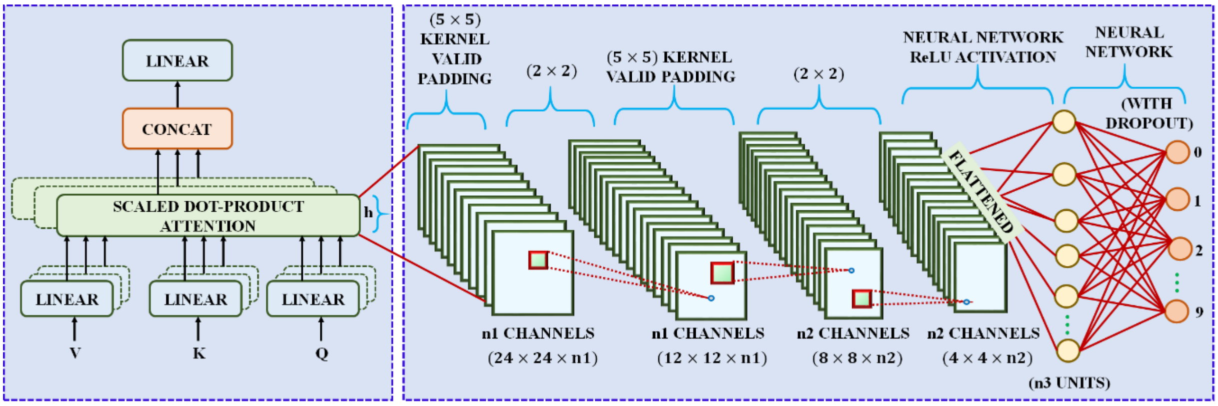

Figure 6: *Hybrid attention-guided CNN*

This combination of local and global context enhances Image classification and supports scalable real-time processing of large amounts of data in the cloud for applications such as object detection and medical imaging.

4. Result and Discussion

By providing a framework of the newly developed AI-agentic DL system for cloud-based image classification along with an instant improvements made over traditional methods of execution such an analysis, this paper illustrates the capabilities of the new cloud-based system. The agentic DL framework consists of autonomous AI agents working together to execute an attention-guided CNN and improve the accuracy of the classification process through real-time dynamic responses.

Table 1: Comparative analysis of accuracy

Image classification method	Accuracy (%)
PPDM [16]	90
PROPOSED	99

In table 1, according to results from a recent study comparing methods for image classification, PPDM [16] achieved an accuracy rate of 90%, while the new method proposed by this paper achieved an accuracy rate of 99%. This demonstrates that the proposed method outperforms PPDM [16] when measuring accuracy.

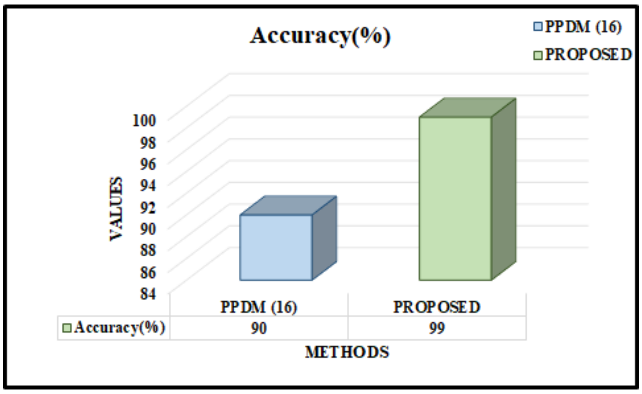

Figure 7: *Accuracy comparison*

Figure 7 depicts a bar graph showing the results of both approaches, with the percentage of accuracy of PPDM (16) being 90% and the PROPOSED method's being 99%. The graphical representation indicates that the proposed approach produces improved accuracy compared to PPDM.

Table 2: Comparative analysis of data hiding time

Data hiding methods	Data hiding time
PPDM [16]	4000
PROPOSED	2000

In table 2, the PPDM method for data hiding takes 4000 units of time. This new technique only takes about 2000 units. Therefore, it is possible to perform data hiding in less time than with the PPDM method.

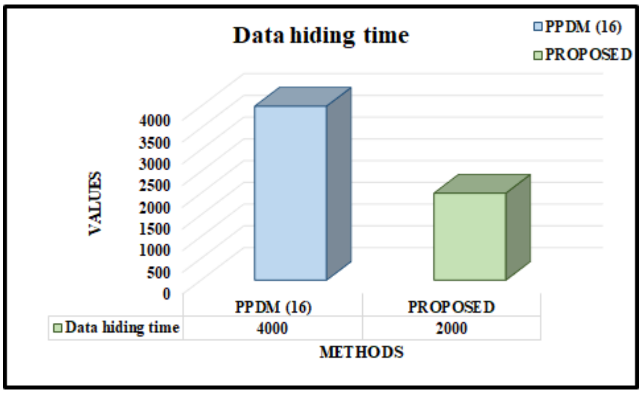

Figure 8: *Data hiding time comparison*

Figure 8 provided an overview of PPDM and the new proposed method in terms of duration. PPDM has a very long hide duration of 4000 units while the proposed method shows a significant time savings of 2000 units. This means that the new proposed method is twice as efficient in relation to processing.

5. Conclusion

An innovative AI-empowered agentic DL framework has been developed for image classification within the emerging restraint of cloud computing. The framework uses an AI-driven hybrid attentional CNN architecture combined with an autonomous AI agent, thus adapting in real-time throughout both the model's preprocessing and tuning phases. This ensures that at all times the system is utilizing the best available data and is responding appropriately to performance feedback received. With the AI-cloud-native deployment of this framework, the system take full improvement of cloud infrastructures, employing distributed training, elastic resource allocation, and extensive inference capabilities. As such, the agentic framework is able to provide highly efficient and adaptable solutions for processing large amounts of image data in real-time, enabling higher-quality image classification applications. Experimental results showed that when applied to standard image datasets, the agentic framework significantly outperformed the existing methods. The successful integration of agentic intelligence into cloud-based image classification solutions demonstrates the potential application of intelligent, agent-based cloud systems and their capability to effectively process high-volume image data in real time.

References

1. Zheshu Jia, Year: 2025, “Research on Image Recognition and Classification Algorithms in Cloud Computing Environment Based on Deep Neural Networks”, IEEE Access.
2. Natthasak Vechprasit; Annop Monsakul; Pramuk Boonsieng, Year: 2024, “Performance Analysis of Image Classification between Edge and Cloud Computing”, In 2024 International Symposium on Parallel Computing and Distributed Systems (PCDS), pp. 1-8. IEEE.
3. P. D. N. K. Kommisetty; Avula Nishanth, Year: 2024, “AI-Driven Enhancements in Cloud Computing: Exploring the Synergies of Machine Learning and Generative AI,” Iarjset, Vol. 9, No. 10.
4. Jeya Bright Pankiraj; Vishnuvarthanan Govindaraj; Yudong Zhang; Pallikonda Rajasekaran Murugan; Anisha Milton, Year: 2023, “Development of scalable coding for the encryption and decryption of images using Modified Diagonal Min Max Block Truncation Code,” Multimedia Tools and Application, Vol. 82, No. 8, pp. 12263-12277.
5. Yehia Ibrahim Alzoubi; Alok Mishra; Ahmet Ercan Topcu, Year: 2024, “Research trends in deep learning and machine learning for cloud computing security,” Artificial intelligence review, Vol. 57, No. 5, pp. 132.
6. Wa’ad H. Aljuaid; Sultan S. Alshamrani, Year: 2024, “A deep learning approach for intrusion detection systems in cloud computing environments,” Applied Sciences, Vol. 14, No. 13, pp. 5381.

7. Jeya Bright Pankiraj; Vishnuvarthanan Govindaraj; Yudong Zhang; Pallikonda Rajasekaran Murugan; R. Kottaimalai; Paras Jain; Arunprasath Thiagarajan, Year: 2024, "Development of Scalable Coding of encrypted Images Using Modified Absolute Moment Block Truncation Code", Journal of Signal, Image and Video Processing, Vol. 18, No. 5, pp. 3973-3981.
8. Yaqian Qi; Yuan Feng; Xiangxiang Wang; Hanzhe Li; Jingxiao Tian, Year: 2024, "Leveraging federated learning and edge computing for recommendation systems within cloud computing networks," arXiv preprint arXiv: 2403.03165.
9. P. D. N. K. Kommisetty; Avula Nishanth, Year: 2024, "AI-Driven Enhancements in Cloud Computing: Exploring the Synergies of Machine Learning and Generative AI," Iarjset, Vol. 9, No. 10.
10. Katerina Mitropoulou; Panagiotis Kokkinos; Polyzois Soumplis; Emmanouel Varvarigos, Year: 2024, "Anomaly detection in cloud computing using knowledge graph embedding and machine learning mechanisms," Journal of grid computing, Vol. 22, No. 1, pp. 6.
11. Godwin Olaoye, Year: 2025, "Self-Learning Neural Networks in the Cloud: Towards Autonomous AI Systems," Available at SSRN 5129553.
12. Shantanu Kumar; Shruti Singh; Narendra Kumar Dewangan, Year: 2025, "Leveraging Cloud Resources for Distributed Training of Residual CNNs in Aerial Image Classification," IEEE Access.
13. Cheng Ji, Huaiying Luo, Year: 2025, "Cloud-based AI systems: Leveraging large language models for intelligent fault detection and autonomous self-healing," arXiv preprint arXiv: 2505.11743.
14. Taiwo Joseph Akinbolaji, Year: 2024, "Advanced integration of artificial intelligence and machine learning for real-time threat detection in cloud computing environments," Iconic Research and Engineering Journals, Vol. 6, No. 10, pp. 980-991.
15. Biswanath Saha, Year: 2025, "The role of edge computing and cloud integration in GAN-based image synthesis," Available at SSRN 5224776.
16. Mehmet Guzel; Muruvvet Kalkan; Erkan Bostanci; Koray Acici; Tunc Asuroglu, Year: 2024, "Cloud type classification using deep learning with cloud images," PeerJ Computer Science, Vol. 10, pp. e1779.