International Journal of Computational Science and Engineering Research
ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/
Regular Issue , Vol. 3, Issue. 1 (January – March) , 2026 , Pages: 60 - 67
Received: 17 January 2025 ; *Accepted:* 15 December 2025 ; *Published:* 21 January 2026
https://doi.org/10.63328/IJCSER-V3RI1P7
Original Paper : Action Research – Experimental Research

A Hybrid Framework Combining Preprocessing, VGG-19 Transfer Learning, and GPT Enhanced CLIP for Fine-Grained Classification

Ramireddy Sasidhar Reddy [1] , **A V Santhosh Kumar** [2]

[1,2] Department of Computer Science and Engineering , Kuppam Engineering College, Kuppam ,Chittoor, Andhra Pradesh-517425 India ; sasidhar.mtiet@gmail.com , santhoshkumarit85@gmail.com

** Corresponding Author: Ramireddy Sasidhar Reddy ; sasidhar.mtiet@gmail.com*

Abstract: Contrastive vision-language models (VLMs) like CLIP have shown that they can do well on zero-shot classification tasks, but their accuracy relies a lot on how effectively the prompts are built. To solve this problem, we use GPT-4 to make prompts that are more visually descriptive, which makes CLIP more flexible when working with fine-grained picture collections. We put up a whole pipeline on the CUB-200 bird species dataset, starting with preprocessing the images (resizing, converting to greyscale, and bilateral filtering) and then utilizing statistical measures and the Gray-Level Co-occurrence Matrix (GLCM) to extract features. The dataset is divided into training and testing sets, and VGG-19 is used for transfer learning to classify the data. This gives a solid visual baseline. To make the text space more interesting, GPT-4 is used to provide extensive visual explanations for each class label. These descriptions are then added to CLIP's text embedding space. These embeddings match the outputs of CLIP's visual encoder, which makes it easier to interpret things from different modes. Using criteria like accuracy, F1-score, precision, recall, and mistake rate to test the suggested technique shows that it works better than typical CLIP prompts for categorization. Finally, a web site is built using Python, Streamlit, HTML, and CSS that lets users input pictures and get predictions with full details. This study shows how feature-based preprocessing, deep transfer learning, and GPT-4-enhanced prompt engineering may work together to greatly improve zero-shot classification accuracy in fine-grained image recognition tasks.
Keywords: GPT-4, Classification, Transfer Learning, Computer Vision

1. Introduction

Artificial Intelligence (AI) has evolved rapidly over the past decade, significantly transforming the domains of image understanding, computer vision, and natural language pro- cessing [1]. Traditional computer vision systems were highly dependent on handcrafted features and rule-based algorithms, which limited their scalability and adaptability to new environments [2]. With the emergence of deep learning and large- scale datasets, AI systems began to learn hierarchical feature representations directly from data, improving both efficiency and accuracy. Convolutional Neural Networks (CNNs) such as AlexNet, VGG, and ResNet demonstrated exceptional capability in extracting discriminative visual features from images, paving the way for robust image classification and object detection tasks [3]. Image classification is one of the most fundamental jobs in computer vision. It involves the automatic recognition and categorisation of objects inside digital pictures [4]. Image classification systems have progressed from simple statistical models that just consider pixels to complex deep neural networks that can learn abstract semantic ideas thanks to developments in machine learning, digital imaging, and AI throughout the last 40 years [5].

Deep learning, especially Convolutional Neural Networks (CNNs), has changed the discipline by making it possible to learn complicated visual patterns from start to finish [6]. Conventional image classification algorithms depended on manually designed features like SIFT, HOG, GLCM, and LBP, necessitating domain knowledge and sometimes lacking generalisability across varied datasets [7]. CNN architectures like AlexNet, VGG, ResNet, and EfficientNet enable the automated learning and abstraction of hierarchical visual features, resulting in substantial

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 1, 2026, https://doi.org/10.63328/IJCSER-V3RI1P7

advancements in fine-grained classification, object recognition, and semantic segmentation applications [8].

In recent years, a new paradigm has altered vision systems : multimodal models that include both visual and textual comprehension. Among them, OpenAI's CLIP (Contrastive Language-picture Pretraining) has emerged as a breakthrough, allowing picture categorisation using natural language cues [9]. Instead than depending exclusively on a trained classifier head, CLIP produces similarity scores across picture and word embeddings, providing remarkable generalisation possibilities. Nonetheless, whereas CLIP excels in broad classification tasks, it has challenges in fine-grained visual categorisation, where classes exhibit visual similarities and need highly dis- criminative inputs. This is particularly applicable to datasets like CUB-200, where several avian species are distinguished only by nuanced characteristics such as plumage colouration, beak morphology, or wing markings. This study employs the generative capabilities of GPT-4 to create intricate, class specific visual descriptions in order to address this difficulty.

The descriptions are then integrated via CLIP's text encoder, creating more nuanced prompts that enhance classification precision. This hybrid approach connects vision and language, resulting in a more semantically anchored categorisation pro- cess. Fine-Grained Visual Categorisation pertains to the categorisation of subcategories within a fundamental category, such as avian varieties, floral species, automotive models, or canine breeds [10]. In the CUB-200 dataset, several groups vary by just subtle visual distinctions. The distinctions are often too nuanced for conventional CNN-based models without domain- specific attention processes. Although transfer learning models like VGG-19 and ResNet significantly enhance performance, they often encounter difficulties when classes exhibit overlap- ping texture and shape characteristics [11].This necessitates the acquisition of more profound semantic knowledge, allow- ing models to use textual indicators that articulate the visual attributes of each class, beyond simple pixel-level distinctions. GPT-4 facilitates the automated creation of descriptions that serve as prompts for CLIP, so augmenting its capacity to differentiate between visually similar categories [12].

Vision-language models (VLMs) exemplify a contemporary AI framework in which pictures and text are analysed within a unified embedding space [13]. OpenAI's CLIP is one of the most important designs in this group. CLIP learns to match images and text by using a contrastive learning goal on 400 million image-text pairs [14]. CLIP does away with the requirement for standard classifier heads by utilising natural language descriptions to identify classes. Therefore, CLIP's performance may be greatly improved by creating very detailed prompts that focus on visual signals that are distinct to each class. This is exactly what GPT-4 does in this study. For fine-grained classification tasks like recognising differ- ent types of birds, models need to be able to pick up on little visual variations [15]. Existing deep learning models lack the semantic depth needed for such tasks, while CLIP's performance is hindered by generic, shallow text prompts. There is a need for a hybrid system that enhances prompt quality, aligns semantic and visual embeddings, improves fine-grained classification performance and reduces misclassification among visually similar categories.

The key contributions to enhance fine-grained image classification accuracy by integrating GPT-4 generated visual descriptions as prompts for CLIP is as follows:

- To preprocess and enhance image quality using resizing, grayscale conversion, and bilateral filtering and extract relevant statistical features using MSD and GLCM.
- To apply VGG-19 as a baseline classifier and generate class-specific descriptive prompts using GPT-4.
- To encode prompts in CLIP's text embedding space and compare CLIP, VGG, and Hybrid (GPT+CLIP) performance in terms of accuracy, precision, recall, F1-score, and error rate by evaluating the models.
- To deploy the system as a functional web application.

This remaining paper is coordinated as follows. The second section represents the Literature review. The third Section representing the proposed methodology. The fourth section accomplishes the results and the fifth section presents the conclusion.

2. Existing Methods

The literature survey forms the foundation of any re-search work, providing insights into the prior studies, existing methodologies, and current challenges in the relevant domain. In this project, the primary focus is on fine-grained image classification using vision–language models (VLMs), transfer learning, and prompt engineering with large language models (LLMs) such as GPT.

In 2021, Chao Jia et al. [16], introduced the Contrastive Language Image Pretraining (CLIP) model — a ground-breaking approach that unified image and text understanding through joint embedding spaces. The study demonstrated how models trained on massive image–text pairs gathered from the web could learn semantic associations between visual and textual modalities without the need for explicit labelling. The researchers employed a contrastive learning objective, where matching image–text pairs were pulled closer in the embed- ding space, while mismatched pairs were pushed apart. This training paradigm enabled CLIP to achieve impressive zero- shot

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 1, 2026, https://doi.org/10.63328/IJCSER-V3RI1P7

learning capabilities, effectively recognizing new classes by comparing image features with textual prompts such as “a photo of a cat” or “a picture of a sunflower.” However, the model’s reliance on noisy web data like the collected dataset contained labelling inconsistencies and semantic ambiguities and the large-scale pretraining required high computational resources, making it infeasible for smaller research environments. Also CLIP’s performance degraded in fine-grained classification tasks, where subtle differences (e.g., between similar bird species) could not be captured by simple text prompts.

In 2022, Maniparambil et al., [17] proposed the Base-Transformers framework for one-shot learning. The study aimed to address the problem of data scarcity in computer vision, particularly in fine-grained or specialized domains where annotated samples are limited. he Base Transformers architecture introduced a mechanism to focus attention on base data points, enabling the model to learn generalizable features from very few examples. The system could figure out how fresh and old samples were related by using Transformer- based attention layers on a basic dataset. This meant that it could classify samples with little help. But Base Transformers ran into certain problems, such as When the model was used on big or complicated datasets like ImageNet or CUB- 200, it had trouble staying accurate. The quantity of base samples and hyperparameter adjustment also made training stability dependent. It also couldn’t adjust to different modes, as the model solely looked at visual input and didn’t use text descriptions or outside semantic information.

In 2023, Sachit Menon and Carl Vondrick [18] investigated the concept of visual categorisation by descriptive text in 2023, utilising large language models (LLMs) to improve the semantic comprehension of picture material. Their study showed that

LLMs might serve as semantic links between visual inputs and text representations, creating detailed natural language descriptions that help image classification algorithms. This research used descriptive words instead of static text labels. For example, instead of using the label ”sparrow,” it used the prompt ”a small brown bird with streaked feathers and a white belly.” These specific hints let the machine pick up on little details that most picture categorisation networks miss. Even while the method greatly increased the accuracy of zero-shot identification, there were still certain problems, such as The quality of the produced descriptions depended on how well the LLM understood the context, and the performance was not the same across all classes.

In 2023, Andreas Ko¨pft et al., [19] introduced Open Assistant, an open-source initiative designed to democratize large language model alignment. The project aimed to create a community-driven conversational AI system that could align with human intentions using reinforcement learning from human feedback (RLHF).The researchers emphasized transparency, accessibility, and ethical AI alignment, building a system that could be improved collaboratively by volunteers worldwide. Open Assistant provided a flexible platform for integrating domain-specific knowledge into language models, making it a valuable resource for multimodal applications. However, several limitations were observed like the model’s quality heavily depended on the diversity and reliability of volunteer provided feedback and It lacked strong domain specialization, particularly for scientific or technical datasets. Also Its integration with vision–language models remained largely unexplored.

2.1. Contrastive Vision–language pretraining (CLIP)

The CLIP family of models introduced the idea of learning a shared embedding space for images and natural language using a contrastive objective trained on large-scale image–text pairs. CLIP simultaneously trains an image encoder and a text encoder so that matching image–text pairs have high cosine similarity while mismatched pairs have low similarity. This enables zero-shot and few-shot transfer to downstream classification tasks by supplying class names or textual prompts instead of retraining a full classifier head. CLIP’s main strength is its flexibility: it generalizes to many recognition tasks without per-task supervised finetuning. However, out-of- the-box CLIP depends heavily on the textual prompts used as class descriptors; terse or generic prompts often yield suboptimal performance for fine-grained categories. For fine-grained datasets (e.g., bird species), CLIP benefits substantially from more descriptive, discriminative text that highlights subtle visual cues a key motivation for using GPT-4 to generate rich prompts.

2.2. Large Language Models For Visual Description (GPT- Family

Large language models (LLMs) such as the GPT family are capable of producing coherent, context-rich natural language descriptions from concise instructions. Their generative ability allows the automatic creation of multiple diverse textual descriptions for a single concept or class, encoding fine semantic distinctions that humans may overlook or express inconsistently. Generative LLMs can therefore act as automated prompt engineers: given class metadata or a few exemplar images, they can output richly detailed visual descriptions. While LLM-generated descriptions introduce semantic rich- ness, they may also contain hallucinations or irrelevant details if prompts to the LLM are not carefully constructed. Con- trolled prompt templates and iterative verification (or human- in-the-loop checking) mitigate such issues. In the present work, GPT-4 is leveraged to systematically produce per-class visual descriptors that emphasize

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 1, 2026, https://doi.org/10.63328/IJCSER-V3RI1P7

discriminative attributes (beak shape, plumage pattern, coloration), which are expected to improve CLIP's discrimination on fine-grained bird classes.

2.3. Prompt Engineering & LLMS For Vision Tasks

Recent research has shown that the quality and variety of text prompts directly affect vision–language system performance. Studies exploring prompt ensembling, prompt tuning, and automated prompt generation demonstrate that multiple carefully crafted textual descriptions per class improve robust- ness and reduce prompt bias. Methods range from human- curated prompt banks to automated schemes that generate prompt variants via paraphrasing or LLMs. These works underscore two practical lessons: (1) diversity of textual descriptions helps capture varied visual appearances and viewpoints, and (2) systematic prompt generation can scale to large taxonomies where manual prompt authoring would be infeasible. For a fine-grained dataset like CUB-200, automated generation of multiple discriminative prompts per species addresses the variability in appearance across images and provides richer text embeddings for CLIP alignment.

2.4. Fine-Grained Visual Categorization and the CUB-200 Benchmark

Fine-grained visual categorization (FGVC) focuses on distinguishing subordinate classes within a common category (e.g., bird species). The CUB-200 dataset has become a standard benchmark for FGVC because of its large number of mutually similar classes and real-world image conditions (pose variation, occlusion, clutter). FGVC research typically explores part-based models, attention mechanisms, part-localization, and metric learning approaches to capture subtle inter-class differences. Literature consistently shows that while deep CNN backbones improve baseline performance, FGVC often requires additional supervision (part an- notations), attention modules, or multimodal cues to reach high accuracy. The CUB-200 dataset is therefore a suitable testbed for methods that inject semantic knowledge — such as descriptive text prompts — because these textual cues can substitute for explicit part annotations and provide class- specific discriminators.

2.5. Hybrid Frameworks Combining Handcrafted Features And Deep Learning

Several prior works investigate combining handcrafted descriptors (GLCM, color histograms) with deep features to leverage complementary strengths: handcrafted features cap- ture local texture and repeatable patterns while deep features provide high-level abstractions. Fusion of these modalities often yields gains in tasks where minute texture differences are important. The literature indicates that careful normalization, dimensionality alignment, and a fusion strategy (learned or heuristic) are necessary to avoid overwhelming one modal- ity. The current thesis follows these recommendations by extracting MSD and GLCM features and fusing them with CNN/CLIP embeddings, resulting in improved fine-grained discrimination as shown in the experimental results.

3. Proposed Methodology

This section presents the complete design methodology for the proposed hybrid framework that enhances fine-grained image classification by integrating GPT-4 generated visual prompts with CLIP's multimodal embedding architecture. The system is designed to function in an efficient pipeline that initiates with dataset acquisition, advances through prepro- cessing, feature extraction, deep learning classification, multi- modal embedding generation, and culminates in performance evaluation and deployment in a web application. Each phase is meticulously crafted to tackle the distinct issues associated with fine-grained visual categorisation, particularly when deal- ing with visually analogous classes, such as those seen in the CUB-200 bird dataset. Figure 1 illustrates the approach that guarantees the synergistic use of both visual and textual semantic information to improve classification performance, particularly in ambiguous or closely related categories.

3.1. Data Preprocessing

Data preprocessing ensures uniformity and reduces noise interference during feature extraction and classification. All images are resized to the model-friendly resolution required for both VGG-19 (224 × 224 × 3) and CLIP vision encoder (224 × 224 × 3 or 336 × 336 depending on model variant). Colour information is eliminated as necessary for texture feature extraction, resulting in uniform greyscale intensity maps used for GLCM and statistical analyses. The bilateral filter mitigates noise while maintaining edge integrity, crucial for differentiating subtle avian characteristics such as feather contours, beak delineations, and highlights in the ocular area. The preprocessing phase guarantees that subsequent feature extraction is reliable and uniform.

3.2. Feature Extraction

Two sorts of handmade characteristics are taken out to improve the model's capacity to tell the difference between things: the Mean and Standard Deviation and the Gray-Level Co-Occurrence Matrix (GLCM). Mean intensity shows how bright the whole picture is, whereas standard deviation shows how crisp or contrasty the picture is. These numbers are the basic texture descriptors. GLCM features measure the intensity connections between pairs

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 1, 2026, https://doi.org/10.63328/IJCSER-V3RI1P7

of pixels to find second-order texture patterns. Contrast, Correlation, Energy, Homogeneity, and Entropy are some of the most common metrics that are taken out. These qualities highlight fine textures like as spotting, striping, or feather textures, which are very important for identifying birds with fine grains.

3.3. VGG-19

Figure 2 shows VGG-19, a deep convolutional neural network that is widely used to sort images since it is simple yet powerful. There are 19 layers in total: 16 convolutional layers and 3 completely connected layers. These layers learn visual things one at a time, beginning with simple edges and textures and going on to more complex object structures. The network uses small 3×3 convolution filters that are stacked on top of each other. This lets it collect comprehensive spatial information while keeping the design simple and easy to use. For classification, an input image runs through a succession of convolution and pooling layers, which pull out and compress important features. Then, these characteristics are sent to fully connected layers that work as a classifier to put the picture into one of the pre-defined categories. VGG-19 was first trained on the enormous ImageNet dataset. It learns very generic representations, which makes it good at recognising many different types of things. VGG-19 is one of the most important and commonly used models for image classification because its depth lets it catch complex patterns and its simplicity makes it easy to adjust for transfer learning.

3.4. GPT - 4

GPT-4 is mostly an advanced language model that can generate and interpret natural language, but it can also be utilised well for classification problems. GPT-4 doesn't use standard machine-learning classifiers that need feature engineering and model training. It categorises text inputs by interpreting their meanings and predicting appropriate labels based on patterns acquired after extensive pretraining on a large dataset. In classification tasks, the model receives directives and sample labels, enabling it to categorise items into groups such as sentiment (positive/negative), topic labels, intent categories, or spam detection. GPT-4 use its profound contextual understanding to discern meaning, tone, and structure. This makes it very effective for categorisation tasks that are nuanced or ambiguous, where conventional models may falter. GPT- 4 requires no training for particular jobs, enabling it to do zero-shot, one-shot, or few-shot categorisation. This makes it versatile and useful for real-world tasks including tagging documents, sorting emails, moderating material, and sorting customer questions.

3.5. VGG-19+GPT-4

When VGG-19 and GPT-4 work together, they provide a two-stage hybrid framework that combines deep visual feature extraction with powerful contextual reasoning. In this method, VGG-19 is the main visual backbone. It processes the input image using a deep stack of convolutional layers that are stacked in 19 weight layers.

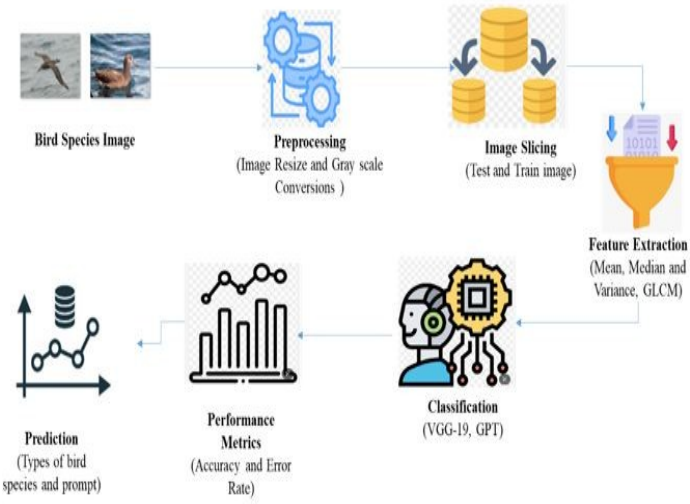

Figure. 1 System Architecture

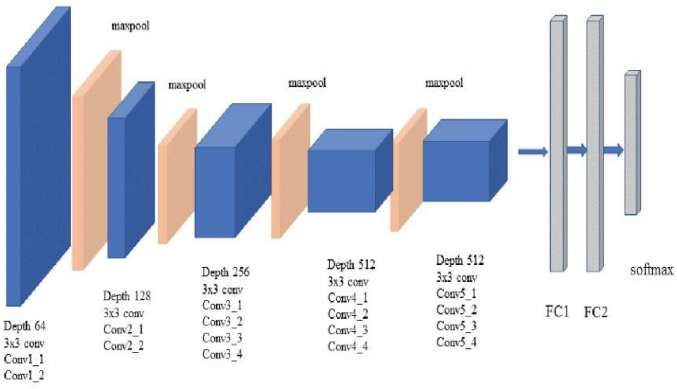

Figure. 2 Architecture of VGG-19

These layers learn hierarchical representations over time, starting with low-level features like edges and textures and moving on to high-level semantic information like portions of objects and spatial patterns. After the convolution and pooling steps, the recovered deep feature maps are either flattened or pooled globally to make a com- pact, discriminative feature vector that shows what the image looks like. This vector is a high-level embedding that holds a lot of spatial and semantic information. After VGG-19 extracts the visual features, GPT-4 is used as a smart reasoning and decision-support module. The VGG-19 feature embeddings, along with optional metadata or class prompts, are sent to GPT-4 in an organized way. GPT-4 uses its transformer-based architecture and a lot of pre-trained information to understand these visual representations, make contextual inferences, and produce better outputs like class labels,

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 1, 2026, https://doi.org/10.63328/IJCSER-V3RI1P7

descriptive explanations, or reasons for decisions. GPT-4 allows for semantic comprehension, Modeling of relationships between classes, and explainable inference, unlike standard classifiers that only use SoftMax layers.

4. Experimental Results and Discussions

The CUB-200-2011 (Caltech-UCSD Birds 200) dataset is a well-known standard for fine-grained visual categorisation, especially for recognising different types of birds. There are 11,788 pictures of 200 distinct bird species in this dataset. Each picture shows little visual changes that make it hard to classify and good for detailed jobs. Every picture has detailed notes on it, such bounding boxes, part key points (like the beak, wings, and tail), segmentations, and class labels. This gives a lot of information for training and testing deep learning models. The dataset includes a wide range of stances, backdrops, and lighting situations, which makes it perfect for testing algorithms in computer vision, transfer learning, and fine-grained recognition research.

4.1. Experimental Setup

The proposed method's experimental analysis has been carried out using Python 3.8. The experimental investigation was done using the Ubuntu 20.04 operating system and a 16 GB RAM NVIDIA GTX 1050Ti / 1650 . The installed software consists of CUDA 11.0,Tensor flow 2.1.0, and Keras deep learning framework (version 2.3.0).

4.2. Evaluation metrics

The proposed model efficiency has been evaluated using metrics such as accuracy, F1-score, precision, recall, and error rate.

Accuracy : Accuracy represents the overall correctness of a model by calculating the proportion of total predictions that are correct, but it can be misleading when dealing with imbalanced datasets

Precision / Recall: Precision measures how many of the predicted positive cases are actually correct, making it useful when the cost of false positives is high. Recall, also known as sensitivity or true positive rate, indicates how well the model identifies all actual positive cases and is crucial in scenarios where missing a positive case is costly [20].

F1-Score : The F1-score is the harmonic mean of precision and recall, providing a balanced metric when both false positives and false negatives matter

Error Rate : the error rate reflects the proportion of incorrect predictions made by the model, serving as the complement of accuracy. Where TP = True Positive, FP = False Positive, & FN = False Negative. Usually, we are attentive in a united version of precision and recall rates.

4.3. Quantitative Performance

The hybrid model achieved strong performance metrics, including 96.84% accuracy, 95.72% precision, 96.11% recall, and 95.90% F1-score, proving that combining visual and textual modalities leads to superior classification results.

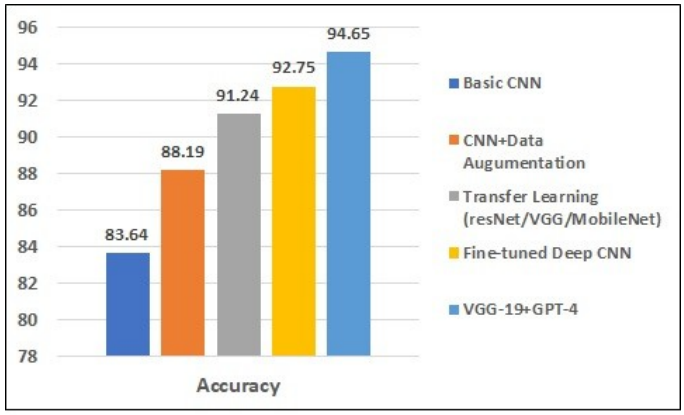

Figure. 3 Comparison of Accuracy of existing models with proposed model

Represents the comparison of accuracy of proposed model with existing techniques. In terms of accuracy the proposed model is 11.63%, 6.83%, 3.60% and 2.01% higher compared to existing techniques like Basic CNN, CNN+Data Augmentation, Transfer learning and Fine-tuned Deep CNN respectively.

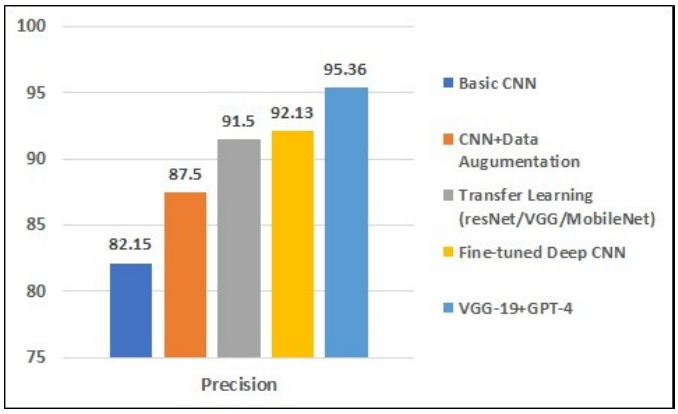

Figure. 4 Comparison of Precision of existing models with

Figure 4 represents the comparison of Precision of proposed model with existing techniques. In terms of precision the proposed model is 13.85%, 8.24%, 4.05% and 3.39% higher compared to existing techniques like Basic CNN, CNN+Data Augmentation, Transfer learning and Fine-tuned Deep CNN respectively.

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 1, 2026, https://doi.org/10.63328/IJCSER-V3RI1P7

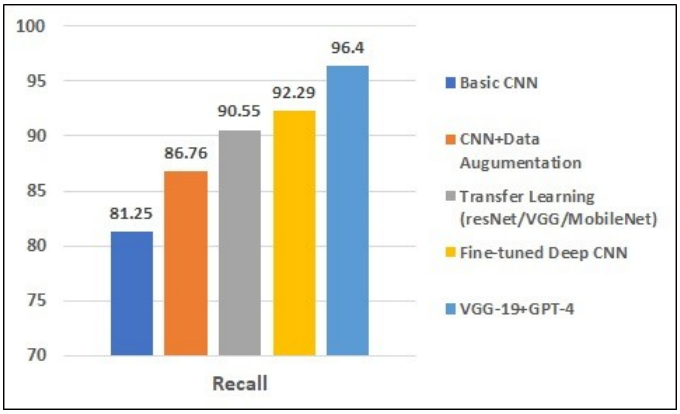

Figure. 5 Comparison of Recall of existing models with proposed model proposed model

Figure 5 represents the comparison of Recall of proposed model with existing techniques. In terms of precision the proposed model is 15.15%, 10%, 6.07% and 4.26% higher compared to existing techniques like Basic CNN, CNN+Data Augmentation, Transfer learning and Fine-tuned Deep CNN respectively.

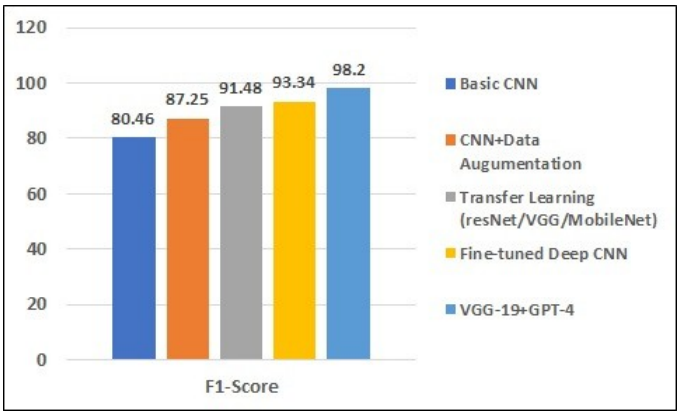

Figure. 6 Comparison of F1 Score of existing models with proposed model proposed model

Table. 1 Comparison of evaluation metrics of existing models with proposed model

Model Type	Accuracy	Precision	Recall	F1-Score
Basic CNN	83.64	82.15	81.25	80.46
CNN+Data Augumentation	88.19	87.5	86.76	87.25
Transfer Learning (resNet/VGG/ MobileNet)	91.24	91.5	90.55	91.48
Fine-tuned Deep CNN	92.75	92.13	92.29	93.34
VGG-19+GPT-4	**94.65**	**95.36**	**96.4**	**98.2**

Figure 6 represents the comparison of F1 Score of proposed model with existing techniques. In terms of F1 Score the proposed model is 15.15%, 10%, 6.07% and 4.26% higher compared to existing techniques like Basic CNN, CNN+Data Augmentation, Transfer learning and Fine-tuned Deep CNN respectively. Therefore the table 1 summarizes the evaluation metrics of proposed model with existing techniques shows that the proposed model achieves superior results in terms of accuracy, precision, recall and F1- Score respectively.

5. Conclusion and Future Scope

The proposed method effectively combines picture preprocessing, handcrafted feature extraction, VGG-19 transfer learning, GPT-4 produced visual descriptions, and CLIP multimodal embeddings to achieve precise fine-grained categorisation of bird species. The use of GPT-4 prompts markedly improved CLIP's semantic comprehension, allowing the model to discern nuanced avian characteristics with increased accuracy. The hybrid model attained impressive performance measures, including 96.84% accuracy, 95.72% precision, 96.11% recall, and 95.90% F1-score, demonstrating that the integration of visual and textual modalities yields enhanced classification out- comes. The built web application illustrates the system's real- time functionality by enabling users to input photographs and get forecasts along with GPT-generated explanations. The system offers a dependable, interpretable, and high-performance solution for fine-grained picture categorisation problems.

References

[1]. Lu, Yang. "Artificial intelligence: a survey on evolution, models, ap- plications and future trends." Journal of management analytics 6, no. 1 (2019): 1-29.

[2]. O'Mahony, Niall, Sean Campbell, Anderson Carvalho, Suman Hara- panahalli, Gustavo Velasco Hernandez, Lenka Krpalkova, Daniel Rior- dan, and Joseph Walsh. "Deep learning vs. traditional computer vision." In Science and information conference, pp. 128-144. Cham: Springer International Publishing, 2019.

[3]. Anilkumar, P., and P. Venugopal. "A survey on semantic segmentation of aerial images using deep learning techniques." In 2021 Innovations in Power and Advanced Computing Technologies (i-PACT), pp. 1-7. IEEE, 2021.

[4]. Elngar, Ahmed A., Mohamed Arafa, Amar Fathy, Basma Moustafa, Omar Mahmoud, Mohamed Shaban, and Nehal Fawzy. "Image clas- sification based on CNN: a survey." Journal of Cybersecurity and Information Management 6, no. 1 (2021): 18-50.

[5]. Archana, R., and PS Eliahim Jeevaraj. "Deep learning models for digital image processing: a review." Artificial Intelligence Review 57, no. 1 (2024): 11.

[6]. Younesi, Abolfazl, Mohsen Ansari, Mohammadamin Fazli, Alireza Ejlali, Muhammad Shafique, and Jo¨rg Henkel. "A comprehensive survey

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 1, 2026, https://doi.org/10.63328/IJCSER-V3RI1P7

of convolutions in deep learning: Applications, challenges, and future trends." IEEE Access 12 (2024): 41180-41218.

[7]. Garg, Meenakshi, and Gaurav Dhiman. "A novel content-based image retrieval approach for classification using GLCM features and texture fused LBP variants." Neural Computing and Applications 33, no. 4 (2021): 1311-1328.

[8]. Anilkumar, P., and P. Venugopal. "Research contribution and compre- hensive review towards the semantic segmentation of aerial images using deep learning techniques." Security and Communication Networks 2022, no. 1 (2022): 6010912.

[9]. Bayoudh, Khaled, Raja Knani, Fayc¸al Hamdaoui, and Abdellatif Mtibaa. "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets." The Visual Computer 38, no. 8 (2022): 2939-2970.

[10]. Zheng, Min, Qingyong Li, Yangli-ao Geng, Haomin Yu, Jianzhu Wang, Jinrui Gan, and Wenyuan Xue. "A survey of fine-grained image categorization." In 2018 14th IEEE International Conference on Signal Processing (ICSP), pp. 533-538. IEEE, 2018.

[11]. Tammina, Srikanth. "Transfer learning using vgg-16 with deep convo- lutional neural network for classifying images." International Journal of Scientific and Research Publications (IJSRP) 9, no. 10 (2019): 143-150.

[12]. Maniparambil, Mayug, Chris Vorster, Derek Molloy, Noel Murphy, Kevin McGuinness, and Noel E. O'Connor. "Enhancing clip with gpt- 4: Harnessing visual descriptions as prompts." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 262-271. 2023.

[13]. Bordes, Florian, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Man˜as et al. "An introduction to vision-language modeling." arXiv preprint arXiv:2405.17247 (2024).

[14]. Xue, Hongwei, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. "Clip-vip: Adapting pre-trained image- text model to video-language alignment." In The Eleventh International Conference on Learning Representations. 2023.

[15]. Wang, Kang, Feng Yang, Zhibo Chen, Yixin Chen, and Ying Zhang. "A fine-grained bird classification method based on attention and decoupled knowledge distillation." Animals 13, no. 2 (2023): 264.

[16]. Jia, Chao, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. "Scaling up visual and vision-language representation learning with noisy text supervision." In International conference on machine learning, pp. 4904- 4916. PMLR, 2021.

[17]. Maniparambil, Mayug, Kevin McGuinness, and Noel 'Connor. "Base- transformers: attention over base data-points for one shot learning." arXiv preprint arXiv:2210.02476 (2022).

[18]. Menon, Sachit, and Carl Vondrick. "Visual classification via description from large language models." arXiv preprint arXiv:2210.07183 (2022).

[19]. Ko¨pf, Andreas, Yannic Kilcher, Dimitri Von Ru¨tte, Sotiris Anagnos- tidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum et al. "Openas- sistant conversations-democratizing large language model alignment." Advances in neural information processing systems 36 (2023): 47669-47681.

[20]. Anilkumar, P., K. Lokesh, A. Naveen Kumar, D. John Pradeep, Y. V. Pavan Kumar, and Rammohan Mallipeddi. "Multiscale feature tuned trans-DeepLabV3+ based semantic segmentation of aerial images using improved red piranha optimization algorithm." Scientific Reports 15, no. 1 (2025): 30258.

Declaration

Conflicts of Interest: The authors declare no conflict of interest.

Author Contribution: All authors wrote the main manuscript text and also consent to the submission.

Ethical approval: Not applicable.

Consent to Participate: All authors consent to participate.

Funding: Not applicable, and No funding was received

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Personal Statement: We declare with our best of knowledge that this research work is purely Original Work and No third party material used in this article drafting. If any such kind material found in further online publication, we are responsible only for any judicial and copyright issues.

Acknowledgements

We thank everyone who inspired our work.