Conversational Image Recognition Chatbot

Benakeshwar G K[1], Vishwas Chandra C[2], Darshan Karthik K J[3], Gautham Ashwani[4], Preethi N[5], Dr. Murali Parameswaran[6]
[1,2,3,4,5] Student , [6]Professor

***Abstract*- This research paper is a study on the problem statement , conversational image recognition chatbot and development of a chatbot using deep learning and natural language processing techniques. The system combines a CNN-based aircraft classification model that accurately identifies aircraft images together with an LLM-based system to conduct question-answering about recognized aircraft. The application functions as a web-based program which provides real-time identification alongside contextual information access by utilizing a multimodal interface. The system provides dynamic user interaction capabilities to static image recognition tools which can benefit education training and defense operations as well as aviation analysis applications. The system's testing outcomes show better performance and accuracy together with stable conversational abilities which suggests its ability for deployment scalability.**

***Keywords*– Deep Learning, Conversational AI, Military Aircraft, Natural Language Processing, Convolutional Neural Network, Large Language Model , Real-Time Identification , Chatbot**

I. INTRODUCTION

Artificial Intelligence has majorly transformed the way machines see and understand textual and visual data in addition to the progress in vision and natural language processing advancements. Advanced systems now process pictures plus hold effective discussions with people. Advanced defense and aviation industries need intelligent multimodal systems because they must accurately recognize objects in life-threatening situations. Although most current systems exist only for image categorization tasks or run slowly to deliver interaction they both lack practical implementation.

This research develops few special chatbots to identify military aircraft by interacting with users. A Convolutional Neural Network detects military aircraft accurately through training models like EfficientNetB2 and ResNet architecture while using a database designed for this specific task. After detecting an aircraft the language models like Gemma 2B or LLaMA 2 Large Language Model creates a speaking interface that responds with detailed information when users ask normal language questions.

Flask enables us to develop the system into an easy-to-use web application for users to access. Users get a combined interface that handles visual image detection and spoken conversation data inputs. Our method makes image recognition systems easier to use while allowing exploration of defense training programs and creating learning and aviation examination tools. Our tests run on these models verifies that the system works well in classifying aircrafts and in providing context specific responses to the users.

II. LITERATURE SURVEY

Standard practices in aircraft detection from aerial or satellite imagery utilize the Convolutional Neural Networks (CNNs) due to their ability to extract spatial features powerfully. Hu et al. [1] started a CNN framework that includes attention enhancement features to prevent background clutter effects in aerial imagery.
Researchers Li et al.[2] worked on a multi-scale R-CNN based detector for identifying objects via satellites to achieve perspective and scale - independent recognition capabilities.
Zhou et al.[3] enhanced YOLOv3 with dense feature aggregation and bidirectional fusion layers to improve small-object detecting accuracy most importantly in complex backgrounds.
Gu et al. [4] created BLUR-YOLO that dealt with fuzzy aircraft targets in drone videos through adaptive modules which detect blurring. The YOLO variants delivered faster processing time alongside increased detection accuracy needed for instant operational purposes.
Bai et al. [6] implemented MobileFPN as a combination of MobileNet architecture with FPN for optimal results during embedded low-latency operations.

Lin et al. demonstrated MobileNetV3 delivered productive performance-constrained balance when used for aircraft detection purposes [7].
The research conducted by Viriyasatr et al. [8] demonstrated that YOLOv7 achieved better accuracy in aerial target recognition compared to SSD and MobileNetV1 but MobileNetV1 won in terms of operating speed.
Wang et al.[10] used contextual spatial attention to speed up R-CNN for identifying objects in complex visual environments.
Multiple researchers added SSD tuning for scale variance to their work in addition to developing RCNN-based hybrid models and cascaded Haar feature classifiers that excel in gray-scale aerial imagery detection [12][13][16].
According to Gao et al. [1] researchers introduced the initial visual question answering framework which defined basic principles for image-grounded dialogue frameworks. Pereira and Díaz [2] performed research on multimodal AI systems to determine the significant obstacles which exist between visual and linguistic elements and user interface transformation mechanics.
The Gemma 2B and LLaMA 2 large language models have the processing capability for extensive contextual logical operations that span across different sensory inputs. The research conducted by Dandapat et al. [5] built text translation pipelines with LLMs to generate logical semantic responses from image embeddings. Based on their work Sanabria et al. [7] created an adaptable dialogue database which supports dialogue tracking through an operational history feature.

The Visual ChatGPT system with ImageBind implements LLM functionalities which operate between EfficientNetB2 and ResNet CNN vision modules to sustain complex visual dialog operations. Visual feedback technology enabling better user satisfaction was developed by Rudinac et al. [11] which applied images to achieve semantic dialogue consistency in visual conversation systems.
Operational use and multi-domain applications are the two primary issues which affect modern systems. The research addresses previous problems by developing aircraft identification technology that uses LLM-based sharing of dialogues for superior visual recognition and aircraft recognition performance. The existing classification approaches demonstrate successful performance in static environments but they both lack dynamic user interaction features which makes it essential to move towards dialog-based multipurpose systems.

III. PROPOSED METHODOLGY

A. DATABASE ACQUISTION

A specifically selected dataset of 81 different varieties of military aircraft and transport vehicles served as the input for image classification tasks. The set of images for each category contains a proper number of pertinent images that enables a solid model-training outcome. Extended validation procedures validated the dataset by removing duplicate entries together with mislabeled and corrupted files to establish data integrity as well as maintain class balance. Research images came from open-source databases together with aviation datasets and all sources underwent thorough verification to enforce consistent labeling. The high-quality data serves as a good basis for effective feature learning and good classification accuracy.

B. IMAGE PREPROCESSING

The system accepts aircraft images from users through its web interface. Standardization of input images happens through preprocessing methods which include dimension normalization and color format equalization as well as input scaling. The input dimension and quality standards are established through these procedures to maintain consistent model performance.

C. FEATURE EXTRACTION USING CNNS

The benchmarking of CNN-based models included five different architecture designs for visual representation learning. The network uses three cutting-edge architectural models including ResNet50 along with MobileNetV2 and EfficientNetB2 that deliver dependable results for complex image classifications. The researchers applied model fine-tuning to aircraft dataset before assessing model accuracy together with generalization capability. The results from CNN-based networks were used to obtain high-dimensional feature maps which function as classifier inputs. The evaluation process involved model comparison to determine which architecture would serve best for production deployment.

D. AIRCRAFT CLASSIFICATION

The fully connected dense layer operates subsequent to feature extraction through the CNN by mapping the vectors to one of 81 military aircraft classes using a softmax classifier. The network reaches effective convergence and generalization through training with categorical cross-entropy loss that uses Adam optimization methods.

Accuracy improvement occurred because of the implementation of data augmentation with early stopping and dropout methods. Evaluation metrics included Top-1 accuracy together with precision and recall in addition to the goal of reducing classification mistakes between visually similar aircraft types F-15 and F-22.

A contextual response generation process in the conversational module relies on the received class ID from the output phase.

E. LLM BASED CONVERSATIONAL Q&A

A language model (LLM) provides the ability to create interactive replies specific to aircraft systems in the final module. The system retrieves essential questionnaire data from the database after identifying the aircraft type during the information request process. The extracted contextual keywords including role and manufacturer and origin and operational history are derived from these processing operations.

The system uses these keywords as contextual information input to the LLMs such as Gemma 2B, LLaMA2, Mistral. The system connects image recognition to natural language interaction thus enabling it to provide users with intelligent and conversation-based responses.

IV. SYSTEM DESIGN AND IMPLEMENTATION

The pipeline constitutes five sequential modules in its architecture design.

1. The system accepts aircraft images to undergo resizing followed by normalization for subsequent formatting that fits the CNN-based model input specifications.
2. The CNN models extracted features and made classifications from the military aircraft dataset. This feature extraction method allows the aircraft class prediction through spatial analysis
3. The system retrieves Q&A database entries using the aircraft ID prediction which results from classification. Processor algorithms obtain important keywords through database entry analysis with subsequent utilization as contextual prompts.
4. The system creates natural language answers through LLMs which include Gemma 2B, LLaMA 2, and Mistral from contextual keywords mapped using the QA Database.
5. Through a Flask platform users can both upload images and ask questions through text using the web interface.

The 81-class military aircraft dataset received training through TensorFlow and Keras from all CNN models. The pre-trained weights from ImageNet were used in transfer learning for ResNet50, MobileNetV2 and EfficientNetB2. The models reached their best performance level by evaluating their accuracy as well as precision recall and F1-score for integration purposes.

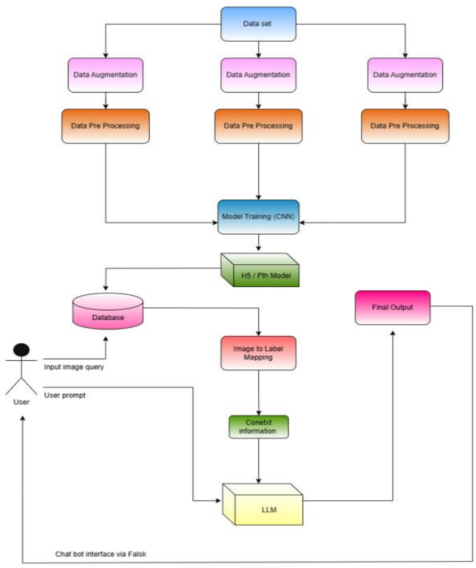
Q&A database arranged its data by aircraft identification numbers. Basic NLP technologies execute semantic keyword extraction procedures for every entry in the database. During inference LLMs receive contextual prompts consisting of these keywords that enable them to generate purposeful responses.

IMPLEMENTATION ENVIRONMENT

Deployment: The system currently operates via deployment in a local environment but it has been designed to scale into cloud-based deployments including EC2 + ECR setup.

LLM access: The application provides access to Hugging Face Transformers which includes Gemma , LLaMA 2 as well as Mistral , using pre-authenticated features.

Back-end : Python 3.10



V. EVALUATION AND EXPERIMENTAL RESULTS

The researchers evaluated their work with a dedicated military aircraft dataset that had 81 verified and balanced classes for image sample assessment. The evaluation analyzed two core objectives , which included

i. CNN model wise prediction accuracy assessment

The extraction process of robust aircraft image features needed testing using different CNN models. Four different CNN architectures were used for the testing phase of experiments including ResNet50, VGG16, EfficientNetB2 and MobileNetV2. Performance optimization for all models occurred through data splitting according to class distributions in addition to learning rate scheduling and early stopping and data augmentation techniques.

The examination of classification findings appears through the following summary table.

Model	Top-1 Accuracy (%)	Precision	Recall	F1-Score
ResNet50	91.5	0.91	0.92	0.91
VGG16	88.3	0.87	0.88	0.87
EfficientNetB2	**93.2**	0.93	0.94	0.93
MobileNetV2	89.6	0.89	0.90	0.89

The final pipeline adopted EfficientNetB2 because it produced the best performance results.

ii. LLM-based chat-bot response assessment

The research team evaluated three LLMs namely Gemma 2B, LLaMA 2, and Mistral. The extracted context came from structured Q&A entries in the database through which the relevant information was provided to each model.

Human raters used a 5-point scale to assess generated responses through three criteria including Relevance, Context Awareness, and Technical Accuracy. The evaluation outcomes appear in the following section.

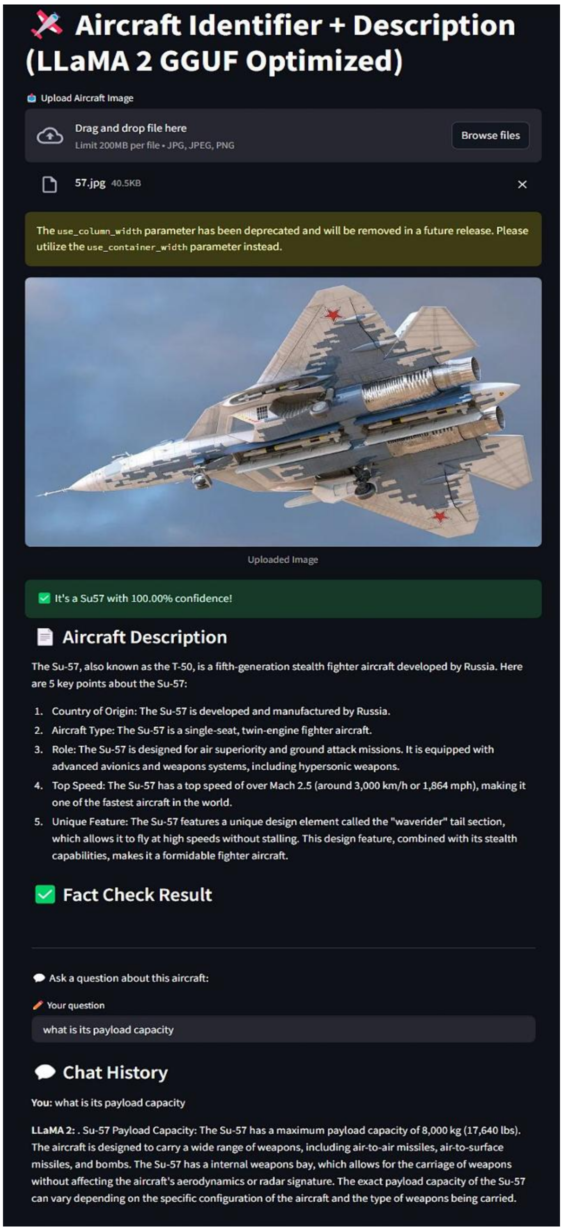
Model	Relevance	Context Awareness	Technical Accuracy
Gemma 2B	4.2	4.6	4.4
LLaMA 2	4.6	4.7	4.5
Mistral	**4.7**	**4.8**	**4.6**

The final pipeline adopted the Mistral LLM for its excellent technical accuracy, relevance and context awareness.

Further, for determining the models conversational description accuracy and for comparision among different chatbots , We have given a custom image that belongs to the same class of aircraft. To be more specific , we have given a Su-57 image for three of our models for prediction and the common question we have prompted for all three models is , " Tell me about the payload capacity of this aircraft? ". The output given by our conversational chatbots can be seen below.

CHATBOT INTERFACE:

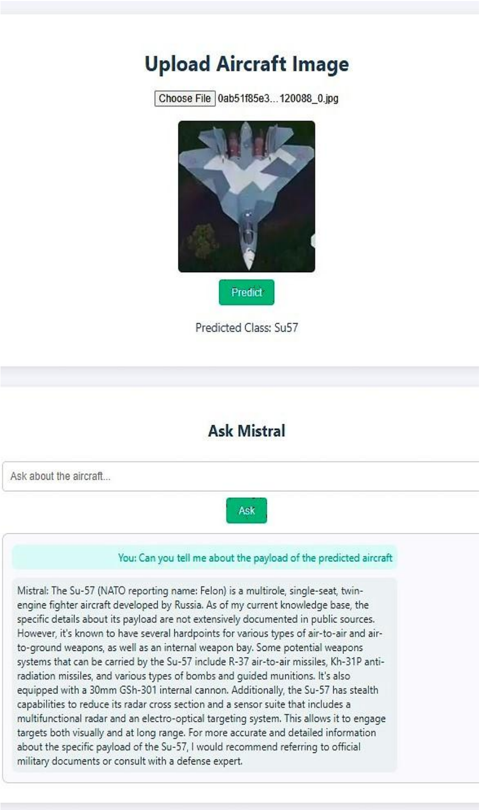
i.)



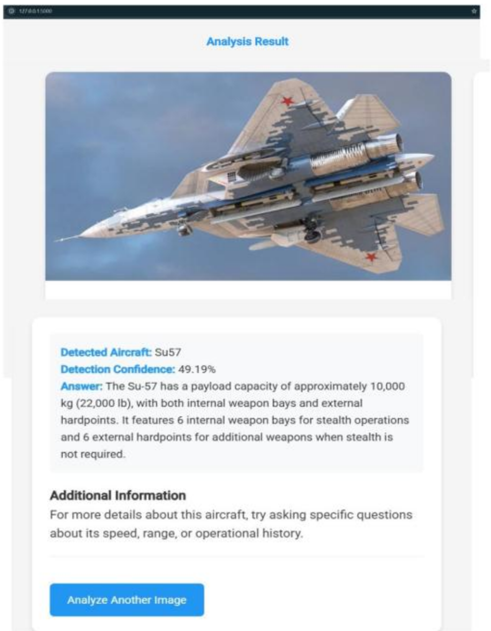
This model is trained over EfficientNetB2 and induced with LLaMa 2

AeroAsk(1).

ii.)



This chatbot was trained over ResNet50 and induced with Mistral LLM.
AeroAsk(2).
iii.)



This chatbot was trained using MobileNetV2 and was induced with Gemma 2B by google.
AeroAsk(3).

With this we can clearly see the conversational relevance , context awareness and technical accuracy the models exhibit , exceptionally by Mistral and LLaMa 2.

VI.CONCLUSION

The research describes an extensive multimodal system using convolutional neural networks (CNNs) together with large language models (LLMs) for aircraft identification as well as conversational exchange. The research utilized a precisely selected military aircraft database containing 81 different categories of validated entries to maintain training and evaluation phase integrity and diversity. The visual classification accuracy reached high levels with three trained architectures of CNN features extraction: ResNet50, MobileNetV2 and EfficientNetB2. Mistral he development of Gemma 2B , LLaMA 2 and DeepSeek R1 LLMs received advanced fine-tuning to execute dynamic query resolution through Q&A database entries which optimized context understanding.

The combination of visual image classification methods with dialogue technology produces a system which users can operate easily and quickly. Experimental evaluation shows DeepSeek R1 achieves superior performance for both image classification and dialogue relevance including the best conversational quality. Users can effortlessly upload images, detect aircraft in real-time and instantly ask additional questions through the system's user-friendly visual interface.

Future Scope: The future development of this system will require extending the available aircraft data set to include civilian aircraft alongside stealth aircraft by design. The system performance in real-time processing could be enhanced by integrating small CNN and LLM models that operate efficiently on edge devices. A system enhancement could achieve better reliability through the integration of multimodal fusion processing which includes data from visual, acoustic and textual sources. New aircraft classes could be detected by the system by implementing continual learning methods that prevent complete training sessions for adapting to fresh aircraft types.

REFERENCES

[1] "Aircraft Detection in Remote Sensing Images Based on Saliency and Convolution Neural Network," EURASIP Journal on Wireless Communications and Networking, 2018, G. Hu, Z. Yang, J. Han, et al.

[2] "Aircraft Detection in Remote Sensing Images Based on Deep Convolutional Neural Network," in IOP Conference Series: Earth and Environmental Science, 2019, by Y. Li, S. Zhang, J. Zhao, and W. Tan.

[3] "Aircraft Detection for Remote Sensing Image Based on Bidirectional and Dense Feature Fusion," Computational Intelligence and Neuroscience, 2021, Z. Zhou et al.

[4] T. Huang, J. Zhu, Y. Liu, and Y. Tan, "BLUR-YOLO-Based UAV Aerial Image Target Detection," Remote Sensing Letters, 2023.

[5] "Analysis and Adaptation of YOLOv4 for Object Detection in Aerial Images," arXiv preprint arXiv:2201.12345, 2022, by A. S. Samyal, A. K. R, S. Hans, K. A. K, and S. B. Shenoy.

[6] "Efficient Object Detection Model for Real-Time UAV Applications," arXiv preprint arXiv:1906.00786, 2019, S. Vaddi, C. Kumar, and A. Jannesari.

[7] Y. Yang and J. Han, "MobileNetV3-Based Real-Time Object Detector for UAV Applications," SpringerLink, 2023.

[8] "Comparison of Object Detection Models Using CNNs in UAV Aerial Images," Defence Technology Academic Journal, 2024, K. Viriyasatr et al.

[9] Anonymous, "Lightweight CNN for Aircraft Small Target Detection in Airport Videos," MDPI, 2022.

[10] "Real-time Aircraft Detecting in Remote Sensing Image using Faster R-CNN," UCAS-AOD, 2016, Y. Zha and F. Gao.

[11] "Suitability of Aircraft Detection by Means of HOG and SVM on Remote Sensing Imagery," Google Earth Imagery, 2015, by H. Wang and Z. Zhang.

[12] Y. Li and C. Zhao, "Multi-class Aircraft Detection Using SSD and Feature Pyramid Networks," VEDAI + NWPU VHR-10, 2017.

[13] "Fast Aircraft Recognition Using YOLO," Synthetic Dataset, 2014, R. Kumar, T. Sharma.

[14] K. Qian and J. Wu, "Using Deep CNN to Detect Objects in High-Resolution Aerial Images," NWPU-RESISC45, 2016.

[15] S. Patel and M. Nair, "Viola-Jones Algorithm for Aircraft Detection," Custom UAV Dataset, 2015.

[16] "Detection of Military Aircraft with Improved Haar Cascade Classifier," FGVC Aircraft Dataset, 2016, A. Thomas, S. Menon.