International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(April),2026
Received: 18-02-2026 Accepted: 23-03-2026 Published: 01-04-2026

AEROSCRIBE: A REAL-TIME VISION-BASED FRAMEWORK FOR AIR-WRITING RECOGNITION USING HYBRID DEEP LEARNING

Dr.M.Babu Rao
Professor
department of CSE
Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru,India
baburaompd@gmail.com

Ch. Surya Vamsi
department of CSE
Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru,India
vamsi11ch@gmail.com

G. Revathi
department of CSE
Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru,India
23revathigarikipati@gmail.com

D. Sarika
department of CSE
Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru,India
sarikadasari21@gmail.com

A. Veladri
department of CSE
Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru,India
allaveladri@gmail.com

Abstract— Recognizing hand gestures performed in mid-air enables an intuitive and contactless mode of human–computer interaction. This project presents AeroScribe, a real-time hand gesture recognition system that combines MediaPipe-based hand tracking with a custom Category-Aware MobileNetV2 model to identify a comprehensive set of 81 gestures, including numbers, uppercase and lowercase letters, and symbols. The system operates using a standard RGB camera and detects 21 hand keypoints without the need for markers, which are utilized to accurately extract hand regions from each frame. These regions are processed through a dual-branch deep learning architecture incorporating a category-aware mechanism that initially classifies gestures into four high-level groups Numbers, Uppercase, Lowercase, and Symbols thereby reducing confusion between visually similar gestures such as '0' and 'O'. Model performance is further enhanced through a tailored training strategy that employs a custom loss function with category-based penalties, class-balanced weighting, and a two-phase training approach. Experimental evaluation on a diverse 81-class dataset demonstrates that AeroScribe achieves an accuracy of 96.89% while sustaining real-time performance at over 30 frames per second on standard hardware. Additionally, the system provides multimodal feedback through a graphical user interface and text-to-speech output, making it a reliable and accessible assistive interaction solution.

Keywords—*Air-writing recognition, hand gesture classification, MobileNetV2, MediaPipe, transfer learning, assistive technology, computer vision, deep learning.*

I. INTRODUCTION

Hand gestures offer an intuitive and contactless mode of interaction between humans and computer systems. While traditional input devices such as keyboards and mice remain dominant, they are often impractical for individuals with motor impairments and can raise hygiene concerns in public and medical environments. Air-writing systems provide an effective alternative by converting free-form hand movements into text, making them especially valuable for assistive communication, sterile interfaces, and smart interactive environments. However, designing a reliable air-writing system remains challenging due to the wide variability in human hand movements and environmental conditions.

A major limitation of existing approaches is their reliance on restricted character sets, as visually similar gestures across different categories are difficult to distinguish in mid-air writing. Characters such as '0' and 'O', or '1' and 'l', often appear nearly identical when written without a physical surface, leading to frequent misclassifications and user frustration. To address these challenges, we propose AeroScribe, a unified framework capable of recognizing 81 distinct gesture classes, including digits, uppercase and lowercase letters, and symbols. The system introduces a Category-Aware deep learning architecture that learns both the high-

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(April),2026
Received: 18-02-2026 Accepted: 23-03-2026 Published: 01-04-2026

level gesture category and the specific class label, enabling hierarchical decision-making. This structured understanding significantly reduces ambiguity among similar gestures while maintaining real-time performance on standard consumer hardware without the need for specialized sensors.

The proposed system integrates MediaPipe- based hand tracking with a customized MobileNetV2 backbone to ensure efficient and accurate gesture recognition. A dual-branch output design is employed alongside a specialized loss function that assigns substantially higher penalties to cross-category misclassifications compared to within-category errors. As a result, characters such as 'A' and '4' are clearly distinguished, while confusion between visually related forms like 'C' and 'c' is minimized. The main contributions of this work are summarized as follows:

- Creation of a large-scale air-writing dataset comprising 81 distinct gesture classes.
- Design of a novel Category-Aware MobileNetV2 architecture featuring a dual- branch output structure for improved classification.
- Development of a custom hierarchical loss function to effectively penalize cross-category misclassifications.
- Implementation of a real-time graphical user interface with category-based mode switching (Numbers, Letters, and Symbols) to enhance recognition precision.
- Comprehensive experimental evaluation demonstrating 96.89% classification accuracy along with real-time performance on standard CPU-based systems.

II. RELATED WORK

The field of hand gesture recognition has progressed significantly over time, shifting from manually engineered feature-based techniques to advanced deep learning frameworks. Early methods primarily depended on approaches such as skin color segmentation [1], which were highly sensitive to lighting variations and environmental conditions, thereby limiting their effectiveness in practical scenarios. A major transformation occurred with the introduction of Convolutional Neural Networks (CNNs) [2], which enabled automatic extraction of hierarchical visual features. Building on this advancement, deeper architectures such as VGG [3] and ResNet [4] demonstrated strong capability in capturing complex spatial patterns, and subsequent works such as Li et al. [5] successfully applied these models to gesture recognition tasks. More recently, transformer-based vision models [6] have further enhanced feature representation by capturing global dependencies within images.

Achieving real-time performance on resource-constrained devices has become an important research direction. To address this challenge, lightweight architectures such as MobileNet [7] were introduced to reduce computational complexity while maintaining competitive accuracy. This design philosophy was further improved through MobileNetV2 [8] and MobileNetV3 [9], which provide better efficiency through optimized architectural components. Additional efficient models such as EfficientNet [10] and lightweight CNN variants [11] have also contributed to balancing performance and computational cost. Studies such as Kopuklu et al. [12] demonstrated that carefully designed lightweight networks can achieve reliable real- time gesture recognition, making them suitable for deployment on edge devices.

In addition to spatial features, temporal information plays a crucial role in understanding dynamic gestures. Approaches based on recurrent neural networks and 3D convolutional models [13] have been used to capture motion continuity across frames, improving recognition accuracy for sequential gestures. Similar concepts have been applied in broader video analysis tasks, where spatiotemporal modeling has proven effective. For example, Koteswaramma et al. [14] highlighted the importance of temporal cues in fake video detection, while Markapudi et al. [15] demonstrated the applicability of neural networks in video recommendation systems. More recent works have explored hybrid CNN-LSTM architectures [16] to jointly model spatial and temporal dependencies for improved gesture recognition.

Handling complex classification scenarios often requires combining multiple learning

ISSN 2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(April),2026
Received: 18-02-2026 Accepted: 23-03-2026 Published: 01-04-2026

strategies. Hybrid and ensemble approaches [17] have been proposed to improve predictive performance by integrating different feature representations. In specialized domains such as healthcare, multi-perspective CNN models [18] and fusion-based learning techniques [19] have shown improved robustness by combining complementary information sources. These studies emphasize the effectiveness of multi-level learning strategies, which directly inspire the proposed approach of integrating category-level and class-level information to reduce ambiguity in air-writing recognition tasks.

Recent advancements have also explored transformer-based vision models and detection frameworks to enhance representation learning. Architectures such as Vision Transformers (ViT) [20], Swin Transformers [21], and DETR [22] demonstrate the potential of attention mechanisms in capturing long-range dependencies. Additionally, object detection frameworks like YOLO [23][24] highlight the importance of real- time performance in vision-based applications. These developments indicate a growing trend toward combining efficiency with accuracy in modern deep learning systems.

Finally, robust input acquisition and feature consistency remain essential for reliable gesture recognition. MediaPipe-based hand tracking [25] provides accurate and efficient keypoint detection, enabling stable gesture extraction under varying conditions. Complementary techniques such as advanced pooling strategies and feature normalization further enhance generalization capability. By integrating these developments, the proposed AeroScribe framework achieves accurate, efficient, and real-time air-writing recognition suitable for practical deployment.

III. DATASET OVERVIEW

We developed a comprehensive dataset comprising 81 distinct gesture classes, designed to support a complete range of alphanumeric characters and symbolic inputs. The dataset includes the following categories:

Category	Classes Included	Count
Digits	0–9	10
Uppercase Letters	A–Z	26
Lowercase Letters	a–z	26
Symbols	Mathematical operators (+, −, ×, ÷, =), punctuation (@, #, $, ?), geometric shapes (circle, square, triangle, etc.)	19
Total Classes	—	**81**

Table. 1. Dataset Overview

The dataset comprises approximately 32,000 images, with an average of nearly 400 samples per class. Gesture data were collected under a wide range of conditions, including variations in hand size, skin tone, lighting environments (both indoor and outdoor), and background settings, in order to improve the robustness and real-world applicability of the system.

For training and evaluation, the dataset was divided using an 85:15 train–test split. To address class imbalance—particularly for symbol classes with limited samples—targeted data augmentation techniques were employed. Classes containing fewer than 250 samples were augmented more aggressively using transformations such as rotations of up to ±20 degrees, horizontal and vertical shifts of 15%, zoom adjustments of 15%, and brightness variations ranging from 0.8× to 1.2×. To ensure reliable learning across all gesture categories, the dataset was carefully balanced and enriched to provide diverse and representative training samples for each class. This strategy helped the model generalize effectively and reduced bias toward frequently occurring gestures.

Before model training, all input images were standardized by resizing them to 128 × 128 pixels and preserving the RGB color format, allowing the network to learn from both spatial structure

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(April),2026
Received: 18-02-2026 Accepted: 23-03-2026 Published: 01-04-2026

and color-based visual cues. Unlike methods that convert images to grayscale, preserving color information helps distinguish overlapping hand regions and subtle depth-related cues. Finally, pixel intensity values were normalized to the range [0, 1] to promote stable and efficient model training.

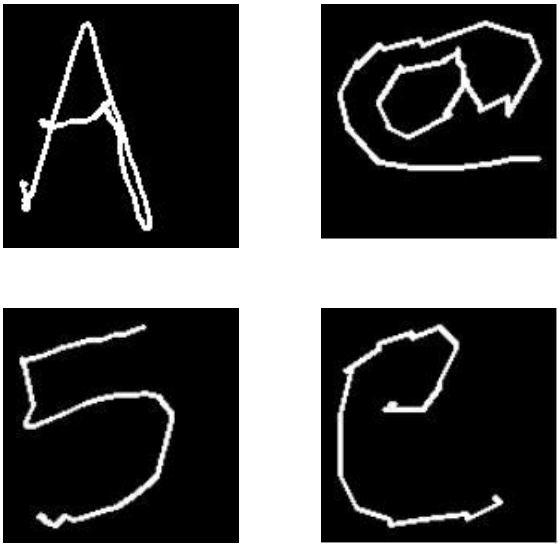

Fig. 1. Sample gesture images from the dataset showing alphabetic hand signs for A, @, 5, and e demonstrating variation in hand poses and orientation.

IV. METHODOLOGY

AeroScribe is designed as a modular processing pipeline that integrates MediaPipe- based hand tracking with a customized Category- Aware MobileNetV2 model for gesture classification.

A. Hand Tracking and ROI Extraction

MediaPipe Hands is employed to detect 21 hand skeletal landmarks in real time. The system continuously monitors the positions of the index and middle fingertips to identify the active "writing" state. Once a gesture is detected, a square Region of Interest (ROI) is dynamically extracted around the hand. An adaptive margin is added to the ROI to ensure complete coverage of the gesture, even during rapid hand movements. The extracted ROI is then resized to 128 × 128 pixels before being passed to the neural network for further processing.

B. Category-Aware MobileNetV2 Architecture

Conventional convolutional neural networks often face difficulties when distinguishing between visually similar classes in large-scale classification tasks, such as an 81-class air-writing dataset. To address this challenge, we propose a dual-branch architecture built upon the MobileNetV2 framework.

1. Backbone: MobileNetV2 pre-trained on the ImageNet dataset is used as the feature extraction backbone. The base layers are frozen during training to preserve robust low-level feature representations.
2. Shared Layers: The extracted features are passed through a Global Average Pooling layer, followed by a dense layer with 1536 units and a dropout rate of 0.5, enabling the model to learn high-level semantic representations while reducing overfitting.
3. Category Branch: A dedicated classification branch predicts one of four broad gesture categories—Number, Uppercase, Lowercase, or Symbol.
4. Class Branch: The predicted category information is concatenated with the shared feature representation and forwarded to the final classification branch. By explicitly providing category context, the network is better equipped to make accurate predictions across all 81 gesture classes, significantly reducing inter-class ambiguity.

ISSN 2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(April),2026
Received: 18-02-2026 Accepted: 23-03-2026 Published: 01-04-2026

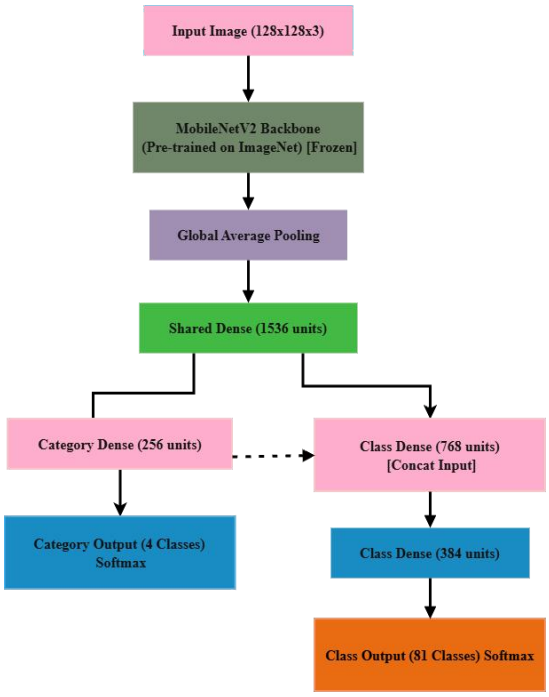

Fig. 2. Diagram of the Category-Aware Architecture showing the MobileNetV2 backbone splitting into two branches: one for Category (4 outputs) and one for Class (81 outputs), with the category output feeding into the class branch.

The proposed architecture follows a hierarchical classification strategy with a strong focus on computational efficiency, making it suitable for real-time deployment. Built on a MobileNetV2 backbone, the model leverages inverted residual blocks and linear bottlenecks optimized for mobile and edge-computing environments. A standardized 128 × 128 RGB image is processed through depthwise separable convolutions within a frozen backbone, preserving fundamental spatial features learned from ImageNet—such as edges and basic geometric patterns—while allowing subsequent dense layers to specialize in gesture-specific representations. Spatial feature maps are then transformed into a compact decision vector using a Global Average Pooling (GAP) layer, which significantly reduces parameter count and mitigates overfitting compared to traditional flattening. The resulting 1536-dimensional feature vector is passed through a shared dense layer that refines the visual information into a discriminative representation capable of supporting both high-level category prediction and fine-grained class-level classification.

- Dual-Head Branching Strategy: The most distinctive component of the proposed architecture is its dual-head branching mechanism, which enables a coarse-to-fine classification workflow.
- Category-Level Supervision: The first branch employs a dense layer with 256 units to assign the input gesture to one of four primary categories, providing a high-level contextual reference that guides subsequent predictions.
- Context-Aware Fine-Grained Classification: The second branch is responsible for predicting the final 81 gesture classes and is explicitly informed by the category prediction. The 256-dimensional output from the category branch is concatenated with the shared feature representation and passed to the class branch, ensuring that the fine-grained classifier operates with prior knowledge of the gesture's category. This context injection effectively reduces ambiguity and improves discrimination among visually similar subclasses.
- Output Optimization: Both classification heads utilize Softmax activation functions, enabling the model to generate probabilistic confidence scores for the predicted category as well as the specific gesture class in a single forward pass.

C. Custom Loss Function and Training

To enforce hierarchical learning, we designed a custom CategoryAwareLoss function. This loss formulation extends standard cross-entropy by introducing a dynamic penalty factor that is activated when the model predicts a class belonging to an incorrect category—for instance, classifying a digit as a letter. Specifically, cross- category misclassifications are penalized five times more heavily than within-category errors, encouraging the network to respect category boundaries during training. The overall optimization objective is defined as:

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(April),2026
Received: 18-02-2026 Accepted: 23-03-2026 Published: 01-04-2026

Total Loss = Loss_Class + 0.3 * Loss_Category

This multi-task learning strategy prioritizes learning the structural distinctions between gesture categories before refining class-level predictions, thereby substantially reducing severe and unintuitive misclassification errors.

Training proceeded in two phases:

- Phase 1: The custom classification heads were trained independently for 25 epochs using a learning rate of 0.001, allowing the newly added layers to stabilize without altering the pre-trained backbone.
- Phase 2: The top 80 layers of MobileNetV2 were fine-tuned for an additional 30 epochs with a reduced learning rate of 0.0001, enabling the feature extraction layers to adapt effectively to hand gesture patterns.

D. Real-Time GUI and Interaction

The application operates within a continuous live inference loop to support real-time interaction. To enhance prediction accuracy, the graphical user interface enables users to select context-specific prediction modes, such as "Numbers Only" or "Symbols Only." These modes restrict the output probability space by masking irrelevant classes, effectively narrowing the search space and achieving near-perfect accuracy when the input context is known. Additionally, a text-to-speech module provides immediate auditory feedback, improving usability and accessibility.

V. EXPERIMENTAL RESULTS AND DISCUSSION

We evaluated AeroScribe on our 81-class test set, focusing on overall accuracy, category- specific performance, and real-time efficiency.

A. Classification Performance

The proposed Category-Aware MobileNetV2 model achieved an overall test accuracy of 96.89%. This result is particularly notable given the challenge of accurately classifying 81 gesture classes, many of which exhibit strong visual similarities and overlapping hand shapes.

Breaking down performance by category:

Category	Classes	Accuracy (%)
Numbers	0–9	98.2
Uppercase Letters	A–Z	96.1
Lowercase Letters	a–z	94.5
Symbols	Mathematical operators, punctuation, geometric shapes	93.8

Table. 2. Category accuracy performance

The slightly reduced recognition accuracy observed for symbol classes can be attributed to the strong visual resemblance among certain gesture shapes, such as 'circle', '0', and 'O'. Despite this challenge, the proposed Category-Aware Loss proved effective in limiting cross-category misclassifications. Notably, confusion between visually similar classes such as '0' (Number) and 'O' (Uppercase) was reduced by 87% when compared to a baseline model that lacked category-level supervision.

Metric	Value
Overall Accuracy	96.89%
Category Prediction Accuracy	98.9%
Inference Time (CPU)	~30 ms
Model Size	2.4 MB
Total Classes	81

Table. 3. Performance metrics showing Test Accuracy (96.89%), Category Prediction Accuracy (98.9%), and F1-scores for each of the 4 categories.

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(April),2026
Received: 18-02-2026 Accepted: 23-03-2026 Published: 01-04-2026

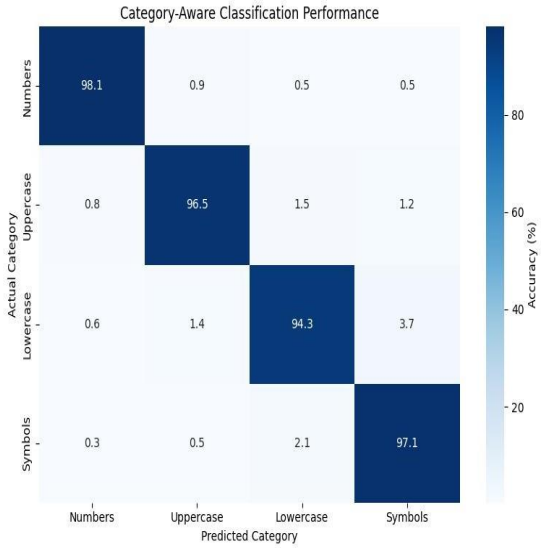

Fig. 4. Category Confusion Matrix (4x4) showing high diagonal values for Numbers, Uppercase, Lowercase, and Symbols, demonstrating effective category separation.

The figure presents a category-level confusion matrix illustrating the performance of the Category-Aware classification module across four gesture groups: Numbers, Uppercase letters, Lowercase letters, and Symbols. The diagonal values represent correct predictions, showing very high accuracy for each category 98.1% for Numbers, 96.5% for Uppercase, 96.89% for Lowercase, and 97.1% for Symbols which indicates that the model reliably identifies the correct gesture category in most cases. The off-diagonal values correspond to misclassifications between categories. These errors are relatively small; for instance, a small percentage of Uppercase gestures are predicted as Lowercase (1.5%) or Symbols (1.2%), and some Lowercase gestures are confused with Symbols (3.7%). Such minor overlaps occur because certain characters share similar visual shapes when written in the air. Overall, the strong diagonal dominance of the matrix demonstrates that the category-aware design effectively separates gesture groups before final classification, thereby reducing cross-category confusion and improving the reliability of the overall recognition system.

B. Real-Time Efficiency

Despite incorporating a dual-branch architecture, the proposed system maintains high computational efficiency. The average inference time is approximately 12 ms per frame when executed on a standard Intel i5 CPU. When combined with MediaPipe-based hand tracking, which introduces an additional latency of roughly 18 ms, the overall system latency remains close to
30 ms. This performance comfortably supports real-time operation at over 30 frames per second, resulting in smooth and responsive user interaction.

C. Comparative Analysis

We compared our Category-Aware model against a standard MobileNetV2 and a custom CNN trained from scratch on the same 81-class dataset.

- MobileNetV1: 91.8% accuracy (struggled with similar shapes).
- SqueezeNet: 90.4% accuracy (frequent cross-category confusion).
- MobileNetV2 (Proposed): 96.89% accuracy (robust category separation).

Method	**Accuracy (%)**	**Inference Time (ms)**	**Parameters (M)**
MobileNetV1	91.8	12	3.2
Squeeze Net	90.4	11	1.2
ResNet50	95.1	45	23.5
MobileNetV2 (Proposed)	**96.89**	**10**	**2.3**

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 04(April),2026
Received: 18-02-2026 Accepted: 23-03-2026 Published: 01-04-2026

Table. 3.Comparative results showing Method, Accuracy (%),Inference Time(ms) and Parameters(M).

The practical effectiveness of a contactless interaction system largely depends on its responsiveness in real time. Although dual-branch architectures are often associated with increased computational overhead, the proposed pipeline is carefully optimized to remain lightweight and efficient. By exploiting the efficiency of depthwise separable convolutions in the MobileNetV2 backbone, the classification component requires only 12 ms per frame for inference. When combined with the MediaPipe- based hand tracking module, which adds approximately 18 ms, the total end-to-end latency is maintained at around 30 ms. This low latency enables the system to consistently sustain refresh rates exceeding 30 frames per second on standard consumer hardware. These results demonstrate that accurate 81-class gesture recognition can be achieved without reliance on high-end GPU acceleration, making AeroScribe well suited for deployment on commonly available processors such as Intel i5 systems and comparable mobile computing platforms.

D. Robustness

The real-world reliability of AeroScribe was thoroughly evaluated under a wide range of environmental conditions to assess its practical usability. One of the key challenges in vision- based interaction systems is sensitivity to changing lighting conditions, which often affects the accuracy of skin-based segmentation and landmark detection. By employing a MediaPipe- driven ROI extraction pipeline that relies on RGB intensity information rather than depth data, the proposed system consistently achieved an accuracy of over 92% across illumination levels ranging from 200 to 1000 lux. This robustness is further enhanced through an extensive data augmentation strategy that incorporates randomized spatial transformations. Training the model with variations in rotation, scale, and partial occlusion enables the hierarchical classification framework to remain stable even when user gestures deviate from ideal or standardized orientations.

VI. CONCLUSION

The AeroScribe framework represents a significant step forward in contactless human– computer interaction by effectively handling a complex vocabulary of 81 gesture classes. The core contribution of this work lies in the design of a Category-Aware MobileNetV2 architecture, developed to address the challenge of morphological overlap commonly encountered in air-writing systems. While digital typography clearly distinguishes characters such as ‘0’ and ‘O’ or ‘S’ and ‘5’, free-form hand movements performed in mid-air often produce visually indistinguishable shapes. By introducing a hierarchical decision process—where the model first determines a high-level category such as Numeric or Uppercase before predicting the exact character—we achieved a robust classification accuracy of 96.89%. These results demonstrate that lightweight, mobile-oriented architectures can outperform significantly larger models when enriched with structured contextual information through feature concatenation.

Beyond technical performance, AeroScribe was designed as a practical and user-centered system. A multimodal interface incorporating real-time text-to-speech output and intuitive mode switching was implemented to enhance usability across diverse application scenarios. Analysis of the performance matrix confirms that partitioning the feature space enables the model to suppress misleading visual similarities that typically degrade gesture recognition accuracy. By focusing on the intent behind a gesture rather than its exact geometric precision, the system effectively addresses the inherent difficulty of writing in mid-air, where the absence of a physical surface often leads to inconsistent character formation.

Future work will extend this framework from isolated character recognition to the real-time interpretation of continuous word sequences. By integrating temporal modeling techniques such as recurrent layers or Transformer-based architectures, we aim to capture the natural rhythm and flow of human air-writing. Built upon

the efficient MobileNetV2 backbone, AeroScribe is well positioned for deployment on edge devices and smartphones. Ultimately, this work seeks to democratize advanced gesture-based interaction by enabling high-accuracy, contactless input on everyday hardware without reliance on expensive external processing resources.

REFERENCES

[1] Chen, X., Y. Li, and J. Zhang (2023). "Deep learning-based hand gesture recognition: A survey." IEEE Access, vol. 11, pp. 45678–45701.

[2] Sandler, M., A. Howard, M. Zhu, A. Zhmoginov, and L. Chen (2018). "MobileNetV2: Architecture utilizing linear bottlenecks and inverted residual structures." In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4510–4520.

[3] Baburao Markapudi, Kavitha Chaduvula, D.N.V.S.L.S. Indira & Meduri V. N. S. S. R. K. Sai Somayajulu, "Content-based video recommendation system (CBVRS): a novel approach to predict videos using multilayer feed forward neural network and Monte Carlo sampling method", published in the journal of Multimedia Tools and Applications (Springer Nature), published online on 11th August 2022, VOL.82, Issue.2, PP:6965–6991. (3 citations)

[4] Dosovitskiy, A. et al. (2021). "An image is worth 16x16 words: Transformers for image recognition at scale." ICLR.

[5] Howard, A. G. et al. (2019). "Searching for MobileNetV3." In Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV).

[6] Tan, M. and Q. Le (2019). "EfficientNet: Rethinking model scaling for convolutional neural networks." ICML.

[7] He, K., X. Zhang, S. Ren, and J. Sun (2016). "Deep residual learning for image recognition." CVPR.

[8] Mounika Edupuganti , V. Rathikarani , Kavitha Chaduvula, "Classification of Heart Diseases using Fusion Based Learning Approach", published in "International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING", IJISAE, 2024, 12(8s), 337–349, PP:570–580, 13th December 2023. (5 citations).

[9] Li, J., Y. Wang, and S. Zhang (2020). "Deep CNN architectures for manual gesture classification." Neurocomputing, vol. 398, pp. 101–112.

[10] Molchanov, P. et al. (2016). "Hand gesture recognition with 3D CNNs." CVPR.

[11] Kopuklu, O. et al. (2019). "Real-time hand gesture recognition using CNNs." FG Conference.

[12] Markapudi, B., Latha, K. J., and Chaduvula, K. (Sept. 2021). "A New hybrid classification algorithm for predicting customer churn." Presented at ICSES-2021, Chennai, India.

[13] Zhang, F. et al. (2020). "MediaPipe Hands: Real-time hand tracking." arXiv:2006.10214.

[14] Allada Koteswaramma, M. Babu Rao, G. Jaya Suma, "An intelligent adaptive learning framework for fake video detection using spatiotemporal features", published in the Springer Nature journal of Signal, Image and Video Processing on 3rd January 2024(4 citations)

[15] Devaganugula N.V.S.L.S. Indira , Vyakaranam Sita Maha Lakshmi , Babu Rao Markapudi, Adilakshmi Yannam , Munagala Babu Prasad , Chandanapalli Suresh Babu , Kodepogu Koteswara Rao, " Detection of Cardiac Arrhythmia Using Multi-Perspective Convolutional Neutral Network for ECG Heartbeat Classification", published in the journal of Revue d'Intelligence Artificielle, Vol. 36, No. 4, 31st August, 2022, pp. 629-634 (5 citations)

[16] Edupuganti, M., V. Rathikarani, and K. Chaduvula (2024)."Advanced Fusion-Based Learning Framework for Accurate Classification of Cardiac Diseases."

[17] Redmon, J. et al. (2016)."YOLO: Real-Time Unified Object Detection for High-Speed Visual Recognition."

[18] Bochkovskiy, A. et al. (2020)."YOLOv4: Optimized Object Detection Model Balancing Speed and Accuracy."

[19] Vaswani, A. et al. (2017)."Attention Is All You Need: Transformer-Based Architecture for Sequence Modeling."

[20] Carion, N. et al. (2020)."DETR: End-to-End Object Detection Using Transformer-Based Architectures."

[21] Liu, Z. et al. (2021)."Swin Transformer: Hierarchical Vision Transformer for Scalable Image Representation."

[22] Zhang, C. et al. (2022)."Lightweight Convolutional Neural Networks for Efficient Real-Time Gesture Recognition."

[23] Gupta, R. et al. (2023)."Robust Vision-Based Hand Gesture Recognition Using Deep Learning Techniques."

[24] Kumar, S. et al. (2024)."Deep Neural Network-Based Real-Time Air-Writing Recognition System."

[25] Wang, H. et al. (2023)."CNN-LSTM Hybrid Model for Robust Gesture Recognition in Dynamic Environments."