A STUDY ON DEEP LEARNING AND TRANSFORMER BASED MODELS FOR HAND GESTURE AND ACTION RECOGNITION

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY
IN
INFORMATION TECHNOLOGY

Submitted by

SAHIL SUTTY (23/ITY/02)

Under the supervision of

DR. VIRENDER RANGA

INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

JUNE, 2025

DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, SAHIL SUTTY, Roll No's – 23/ITY/02 student of M.Tech (INFORMATION TECHNOLOGY),hereby declare that the project Dissertation titled "A Study on Deep Learning and Transformer Based Models for Hand Gesture and Action Recognition" which is submitted by me to the Information Technology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi

Sahil Sutty

Date: 30.05.2025

DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled " Study on Deep Learning and Transformer Based Models for Hand Gesture and Action Recognition" which is submitted by Sahil Sutty, Roll No's – 23/ITY/02, Inlformation Technology ,Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Dr. Virender Ranga

Date: 30.05.2025

SUPERVISOR

DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I wish to express my sincerest gratitude to Dr Virender Ranga for his continuous guidance and mentorship that he provided me during the project. He showed me the path to achieve my targets by explaining all the tasks to be done and explained to me the importance of this project as well as its industrial relevance. He was always ready to help me and clear my doubts regarding any hurdles in this project. Without his constant support and motivation, this project would not have been successful.

Place: Delhi Sahil Sutty

Date: 30.05.2025

Abstract

Fundamental technologies in the evolution of human-computer interaction (HCI), hand gestures and human action recognition enable more natural, intuitive, and accessible interfaces across sectors including assistive technologies, robotics, virtual reality, and surveillance. Using the MSRA Hand Gesture Dataset and the UCF101 Dataset, this paper presents a thorough comparative analysis of state-of- the-art deep learning and transformer-based models for hand gesture recognition and for human action recognition.

Comprising 76,500 depth images distributed over 17 gesture classes, the MSRA Hand Gesture Dataset offers a strong basis for spatial feature extraction. ResNet101 obtained the highest F1-score (0.9978) among all architectures; closely followed by DenseNet 169 (0.9919) and DenseNet 201 (0.9901). MobileNetV2 demonstrated a good balance between computational efficiency and accuracy with an F1-score of 0.9847; VGG variants lagged since they lacked sophisticated architectural elements.

Human action recognition using the UCF101 dataset—which consists of over 13,000 video clips in 101 action categories—was driven with an eye toward the 50 most frequent classes to guarantee computational feasibility and class balance.With F1-score 0.9997, transformer-based models especially ViT Tiny Patch surpassed even the deepest CNNs. While MobileNetV2 once shown efficiency in settings with limited resources, VGG16bn's performance revealed the limits of older CNN architectures for demanding tasks.

The results underline how architectural innovations including residual connections, dense connectivity, and attention mechanisms help to raise recognition accuracy and computational efficiency. The paper claims that transformer-based models are redefining benchmarks even if deep CNNs continue to be strong candidates. More particularly, considering hybrid CNN-transformer designs, explicit temporal modeling, and advanced augmentation techniques helps to increase recognition capacities in pragmatic settings.

Contents

List of Tables

List of Figures

List of Symbols

HCI	Human-Computer Interaction
GUI	Graphical User Interface
WIMP	Windows, Icons, Menus, Pointers
UX	User Experience
NUI	Natural User Interface
VR	Virtual Reality
AR	Augmented Reality
CNN	Convolutional Neural Network
NLP	Natural Language Processing
ViT	Vision Transformer
DeiT	Data-efficient Image Transformer
MSRA	Microsoft Research Asia (Hand Gesture Dataset)
RGB	Red, Green, Blue
SVM	Support Vector Machine
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GNN	Graph Neural Network
TSN	Temporal Segment Network
GCN	Graph Convolutional Network
HMDB	Human Motion Database
DHG	Dynamic Hand Gesture
IoT	Internet of Things
ONNX	Open Neural Network Exchange

Chapter 1

INTRODUCTION

1.1 The Evolving Landscape of Human-Computer Interaction (HCI)

Human-Computer Interaction (HCI) stands as a dynamic and multidisciplinary field dedicated to the design, evaluation, and implementation of interactive computing systems for human use and the study of major phenomena surrounding them.Fundamentally, HCI aims to comprehend and maximize the interaction between humans and computers, so improving the usability, accessibility, and efficiency of technology in so enhancing human capabilities and experiences. From command-line interfaces and basic graphical user interfaces (GUIs) to the sophisticated, intuitive, and progressively natural interaction paradigms we see today, HCI has experienced a tremendous evolution over the past few decades. Fast developments in computing power, sensor technology, artificial intelligence, and a better knowledge of human cognition and behavior have propelled this evolution.

Early HCI was mostly system-centric, stressing computer technical capabilities with less regard for user needs or cognitive load. Often rigid, interactions demanded users to learn difficult commands and adjust to the machine's limitations. A major change towards user-centric design came with the arrival of the personal computer and later development of GUIs pioneered by companies like Xerox PARC and popularized by Apple and Microsoft. Leveraging human visual processing and spatial reasoning, these interfaces—which feature windows, icons, menus, and pointers (WIMP)—made computers more approachable to a larger audience.

The spread of the internet and mobile devices in the late 20th and early 21st centuries transformed HCI still more. New modalities of interaction made possible by touchscreens, voice assistants, and pervasive connectivity led to more direct manipulation and context-aware computing based on considering elements like engagement, emotion, and general satisfaction. The emphasis moved beyond simple usability to include user experience (UX). Natural user interfaces (NUIs) first emerged in this age and seek to make interactions feel as simple and easy as interacting with the physical world or another human being. By using human capacities including speech, touch, gesture, and even gaze, NUIs reduce the need for explicit instruction of abstract commands.

Among this changing terrain, hand gestures and human action recognition have become especially interesting and transforming tools for HCI. In human-machine communication, these technologies seem to open hitherto unattainable degrees of naturalism and expressiveness. Fundamental to human communication, hand gestures can transmit a wide range of meaning, feeling, and knowledge. Including gesture recognition into computing systems enables touchless control, which is not only handy but also essential in places like public kiosks or sterile medical environments where physical contact is either unacceptable or unworkable.

Applications range from controlling smart home appliances with a wave of the hand, navigating complex menus in virtual reality (VR) and augmented reality (AR) environments, to facilitating communication for individuals with speech or hearing impairments using sign language translating systems, and allowing sophisticated control of robotic systems in manufacturing or exploration.

Likewise, a wide range of applications requiring situational awareness and behavioral understanding depend on human action recognition—that is, the capacity of a system to identify and understand human actions from video or sensor data. Automated action recognition can support public safety enhancement, crowd monitoring, and detection of suspicious activity in surveillance and security. Content-based video retrieval lets users search enormous video archives for particular events or actions. Action recognition in patient monitoring, senior fall detection, and rehabilitation assessment all help healthcare. Using this technology, sports analytics tracks game statistics, examines player performance, and offers tactical analysis. Action recognition is also used in interactive entertainment and gaming systems to produce more realistic and interesting interactions.

Still, there are many difficult obstacles in the way strong and accurate gesture and action recognition systems develop. Generalization is challenging because of the natural variability in human gesture and action performance influenced by personal style, cultural background, emotional state, and physical context. Further complicating the recognition challenge are environmental elements including changing illumination, background clutter, and occlusions—that is, hiding of parts of the hand or body from view. Capturing and interpreting the intricate spatio-temporal patterns of motion adds still another level of difficulty for dynamic gestures and actions. Classical machine learning techniques and handcrafted elements were common components of conventional methods of recognition. Although these techniques showed some success in limited settings, they usually lacked the scalability and resilience needed for practical implementation.

1.2 Motivation: The Quest for Advanced Recognition Systems

The limits of conventional recognition methods and the growing need for more intelligent and natural HCI systems have given research on more potent and flexible recognition models a great incentive.

Deep learning, especially convolutional neural networks (CNNs), signaled a paradigm change in the field of computer vision including gesture and action recognition. CNNs have the amazing capacity to automatically learn hierarchical feature representations straight from raw pixel data, so saving the time required for careful hand feature engineering. This capacity has produced innovations in many image and video understanding projects. CNNs can efficiently learn discriminative spatial features from either individual video frames or stationary images for hand gesture detection. Architectures like 3D CNNs and two-stream networks—combining spatial and temporal information—have shown notable advancements over previous approaches for human action recognition.

Even with CNNs' success, though, there are still difficulties particularly in accurately capturing global contextual information and modeling long-range temporal dependencies. Originally designed for natural language processing (NLP), transformer-based models have lately become rather important in computer vision. Treating image patches as sequences, Vision Transformers (ViTs) and their variants apply self-attention techniques to balance the importance of various areas and model global relationships. For tasks requiring an awareness of complicated interactions and long-range dependencies, such those inherent in dynamic human actions, this method has shown to be especially successful.

Notwithstanding these developments, thorough comparative studies evaluating the performance, computational efficiency, and practical feasibility of several deep learning and transformer architectures across standardized benchmarks remain much needed.

Such studies are crucial for several reasons:

- Guiding Model Selection: Practitioners and researchers need clear guidance on which models are best suited for specific applications, considering trade-offs between accuracy, speed, and resource requirements.

- Identifying Architectural Strengths and Weaknesses: A systematic comparison can illuminate the specific architectural innovations (e.g., residual connections, dense blocks, attention mechanisms) that contribute most to performance gains.

- Benchmarking Progress: Standardized evaluations help track the progress of the field and identify areas where further research and development are needed.

- Fostering Innovation: Understanding the current state-of-the-art can inspire the development of novel hybrid architectures or further refinements to existing models.

This work attempts to provide a comprehensive view on the possibilities of modern deep learning models by concentrating on two different but related tasks: human action recognition from RGB videos (UCF101 dataset) and hand gesture recognition using depth images (MSRA dataset).

While the UCF101 dataset presents the complexity of action recognition in unconstrained "in-the-wild" videos, the MSRA dataset with its depth information enables an evaluation targeted on spatial feature extraction in a rather controlled environment. From established CNNs like VGG, ResNet, and DenseNet to efficient networks like MobileNetV2 and innovative transformers like ViT and DeiT, the choice of a varied range of models guarantees a complete evaluation of the present architectural scene.

Moreover, a fair and rigorous comparison depends on the emphasis on a consistent training and evaluation framework including consistent preprocessing, data augmentation, and cross-validation techniques. The results of this study are meant to be insightful for the academic community and industry practitioners, so helping to shape more strong, clever, and intelligent HCI systems. The ultimate aim is to progress the state of the art in gesture and action recognition, so opening the path for a future whereby human-machine interaction is really seamless and natural.

1.3 Research Objectives and Contributions

The main goal of this work is to perform a thorough and exact comparison of well-known deep learning and transformer-based architectures for hand gesture recognition and human action detection. This main objective is divided into several particular research goals:

1. Comprehensive Performance Evaluation:
    (a) To meticulously evaluate the performance of selected CNN architectures (VGG16bn, VGG19bn, DenseNet169, DenseNet201, ResNet101, ResNet152, MobileNetV2) and transformer-based architectures (DeiT3 Base Patch, ViT Base Patch, ViT Small Patch, ViT Tiny Patch).

    (b) To similarly evaluate these architectures (where applicable, with appropriate modifications for video data) on a significant subset of the UCF101 Dataset for human action recognition.

    (c) To use a consistent set of evaluation metrics, with a primary focus on the F1-score, complemented by accuracy, precision, recall, and training/validation loss, to provide a multi-faceted view of model performance.

2. Analysis of Architectural Efficacy and Trade-offs:
    (a) To analyze and compare the effectiveness of different architectural paradigms (e.g., deep sequential layers in VGG, residual connections in ResNet, dense connectivity in DenseNet, depthwise separable convolutions in MobileNetV2, self-attention in Transformers) in extracting salient features and achieving high recognition accuracy for both static depth images and dynamic video sequences.

    (b) To investigate the trade-offs between model accuracy, computational complexity and potential for real-time feasibility across the evaluated architectures. This is crucial for understanding the practical deployability of these models, especially in resource-constrained environments such as mobile devices or embedded systems.

3. Investigation of Training Dynamics and Generalization:
    (a) To study the training dynamics of each model, including convergence speed, stability of learning (as observed from loss and F1-score curves over epochs), and susceptibility to overfitting, particularly in the context of the 10-fold cross-validation strategy.

    (b) To assess the generalization capabilities of the models by evaluating their performance on unseen test data and across different folds of the cross-validation process.

4. Derivation of Insights and Guidelines:
    (a) To synthesize the empirical findings into a set of actionable insights and guidelines that can assist researchers and practitioners in selecting appropriate deep learning architectures for specific gesture and action recognition tasks.

    (b) To highlight the strengths and limitations of current state-of-the-art models and identify promising avenues for future research and development in these domains.

The contributions of this report are anticipated to be manifold:

- A Unified Comparative Benchmark: This work provides a side-by-side comparison of a wide range of influential deep learning models for two key HCI tasks, using consistent evaluation protocols.
- In-depth Analysis of Transformer Models for Action/Gesture Recognition: While transformers are increasingly popular, their comparative performance against well-established CNNs, especially for hand gesture recognition from depth data and action recognition within a unified framework, is a valuable contribution. The report specifically explores different sizes of ViT models (Base, Small, Tiny), providing insights into the impact of model scale.
- Elucidation of Architectural Advantages: The study will clarify which particular architectural elements—e.g., skip connections, dense blocks, attention layers—are most advantageous for the nuances of hand gesture patterns versus dynamic human actions and how these components help to overcome obstacles including feature repetition, vanishing gradients, and modeling long-range dependencies.
- Practical Considerations for Model Deployment: Examining not only accuracy but also elements like computational efficiency (implied by model choice like MobileNetV2 or ViT Tiny) and parameter count helps the work offers insights relevant to the pragmatic implementation of these recognition systems.
- Identification of Research Gaps and Future Directions: The thorough review and empirical analysis will help to identify present constraints and motivate future research paths, so guiding perhaps even more effective and efficient recognition models.

This work intends to advance the knowledge of how several deep learning paradigms can be efficiently used to address the complexity of hand gesture and human action recognition, contributing to the more general objective of building more intelligent, simple, and human-centric computing systems.

Chapter 2

LITERATURE REVIEW

2.1 Foundations and Evolution

2.1.1 Handcrafted Features to Classical Machine Learning

Driven by the necessity for more natural and expressive human-computer interaction (HCI), hand gestures and human action recognition have been main subjects in computer vision for decades. Early studies concentrated on extracting handcrafted features from images or video sequences including optical flow, contours, skin colour segmentation, and motion trajectories [1, 2, 3].

Hidden Markov Models (HMM), Support Vector Machines (SVM), and decision trees among other classical machine learning algorithms—were fed these elements to classify gestures and actions [1, 3]. Although these methods attained some success, their resilience was limited by high sensitivity to background clutter, illumination changes, occlusions, and the variability of human appearance and motion [2, 3].

For hand gesture recognition, for instance, techniques based on skin color segmentation or edge detection frequently failed in complex backgrounds or under changing illumination conditions [1]. In action recognition, too, handcrafted elements found it difficult to depict the temporal dynamics and minute variations between like actions, especially in free environments [2]. Consequently, the recognition rates frequently traded off computational power with temporal resolution [3].

2.1.2 Depth Sensors and Multi-Modal Data

The depth sensors of Microsoft Kinect indicated a major progress in gesture recognition [4, 5]. Depth-based feature extraction helped to more precisely segment hands, improve spatial representation, and lower background noise effect. Structured depth images using 3D skeletal data from sources such as the MSRA Hand Gesture [4] and SHREC17 [6] allow one to evaluate both conventional and deep learning techniques in more controlled surroundings.

Combining RGB, depth, and sometimes electromyography (EMG) signals or skeleton data, multi-modal techniques started to show up in action recognition [7, 8, 9]. By means of the complementing strengths of several modalities, these approaches enhanced the system's capacity to control occlusions [10, 11], several points of view, and complex scenes [7].

2.1.3 Deep Learning Approach

By letting automatic learning of hierarchical spatial features from raw data, deep learning—especially convolutional neural networks—revolutionized gesture and action recognition [12, 13, 14]. CNNs obtained state-of- the-art performance in stationary gesture recognition and image-based action classification [12], hence they eliminated the need for hand feature engineering.

Especially Long Short-Term Memory (LSTM) networks, researchers integrated CNNs with Recurrent Neural Networks (RNNs) for dynamic gestures and actions to capture temporal dependencies [15, 16, 7]. Because hybrid CNN-RNN models could learn both spatial and temporal patterns, they shown enhanced accuracy in sequential datasets. By separating the modeling of spatial features (from still frames) and temporal features (from optical flow), two-stream CNN architectures further advanced action recognition and produced state-of-the-art results on benchmarks including UCF101 and HMDB [17, 18, 19].

2.1.4 Transformer-Based and Hybrid Architectures

More lately, transformer-based models have become rather popular in gesture and action recognition. Originally designed for natural language processing, transformers capture global dependencies and parallel process sequences using self-attention mechanisms. Particularly shining in modeling long-range spatial and temporal relationships, Vision Transformers (ViT) and Data-efficient Image Transformers (DeiT) have set new benchmarks in video understanding [9, 8, 20].

Transformers have been used on EMG signals [9, 8] for hand motions. Transformers-based methods including CLIP have shown promise for action recognition [21, 19]. Leveraging the local feature extracting capabilities of CNNs with the global context modeling of transformers [22, 20], hybrid architectures combining CNNs and transformers have also emerged [22, 20]. Particularly on large-scale databases such as Kinetics-400 and UCF101 [17, 19], these models have shown improved generalization and convergence speed.Human action detection has also been accomplished with Graph Neural Networks (GNNs [13, 23, 20].

2.2 Hand Gesture Recognition: Techniques, Datasets, and Challenges

2.2.1 Taxonomy of Hand Gesture Recognition Methods

In general, hand gesture recognition methods fall into sensor-based and vision-based systems. Using EMG sensors, accelerometers, or data gloves, sensor-based techniques track hand motions. They are less fit for natural HCI [2, 1] and intrusive even if their accuracy is rather high. More natural but struggle with segmentation, background clutter, and lighting changes vision-based methods—which rely on cameras to record hand images or videos [3, 24].

Vision-based methods also differ in dynamic gesture recognition—that is, in recognizing sequences of movements—from static gesture recognition—that is, in recognition of fixed hand poses [4, 15]. Whereas dynamic gestures demand temporal modeling using video sequences or frame stacks [16, 25], static gestures are usually categorized using image-based models.

2.2.2 Key Datasets for Hand Gesture Recognition

The development and benchmarking of hand gesture recognition algorithms have been facilitated by several publicly available datasets:

- MSRA Hand Gesture Dataset: Nine subjects complete 76,500 depth images spread over 17 gesture classes. Depth images improve feature extraction and help to lower background noise, hence the dataset is perfect for assessing CNN-based architectures on spatial aspects [4].
- Dynamic Hand Gesture (DHG-14/28) Dataset: Designed for testing temporal models such as RNNs and transformers [5], it offers frame sequences of dynamic hand gestures using a depth sensor.
- SHREC17 Dataset: Provides 3D skeletal models and depth images of hand gestures, so supporting hybrid and multi-modal recognition techniques.

These datasets vary in the number of classes, types of gestures (static vs. dynamic), and modalities (RGB, depth, skeleton), providing comprehensive benchmarks for different recognition tasks [3, 2, 1].

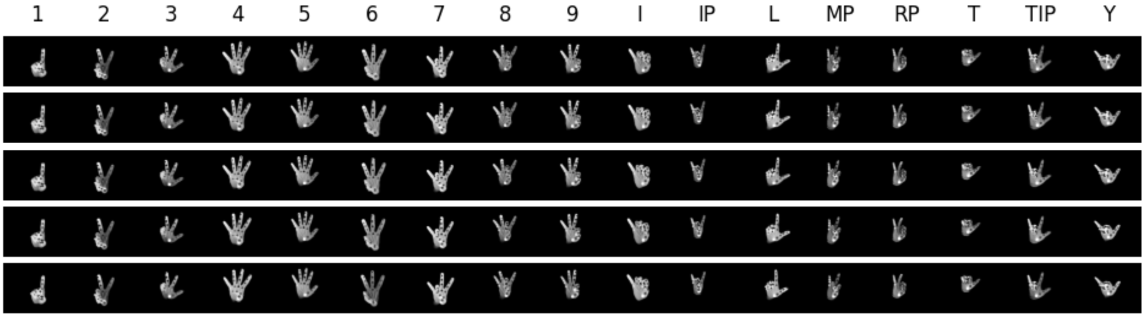


Figure 2.1: MSRA Dataset

2.2.3 Evolution of Deep Learning in Hand Gesture Recognition

Early vision-based approaches for hand gesture recognition drew on Haar-like features, edge detection, and histogram analysis among other feature extraction methods. Particularly in the presence of noise or complicated backgrounds [3, 2], these characteristics were sometimes limited in their capacity to differentiate between like gestures. CNNs let direct learning of discriminative features from raw images, so greatly enhancing accuracy and robustness [26, 24, 25]. Recent research has explored various CNN architectures, each with unique strengths:

- VGG16/19: Deep sequential layers with small convolutional filters are computationally costly but effective for hierarchical feature learning [2].
- DenseNet169/201: Effective gradient flow and feature reuse made possible by dense connectivity help to lower parameter redundancy and enhance performance [3].

- ResNet101: Residual connections solve vanishing gradients, so allowing the training of very deep networks [2, 26].
- MobileNetV2: Edge [27, 24] and real-time compatible lightweight architecture using depthwise separable convolutions.

Particularly for dynamic gestures [15, 7, 16], hybrid models such CNN-RNN and CNN-LSTM architectures have been suggested to capture both spatial and temporal aspects. By using self-attention mechanisms to model long-range dependencies [8, 9], transformer-based models have lately shown promise especially for sequential data such as EMG signals. Furthermore useful for efficiently modelling hand structure and dynamics are graph-based models [10].

2.2.4 Challenges and Open Issues in Hand Gesture Recognition

Despite significant progress, several challenges persist in hand gesture recognition:

- Variability in Gestures: Generalization is challenging because of variations in hand shapes, orientations, and motion patterns across people and settings.cite 3,12,17.
- Environmental Factors: Changes in background clutter, lighting, and occlusions can all compromise recognition performance [24, 28].
- Real-Time Processing: Many deep learning models call for large computational resources, which presents difficulties for real-time applications on embedded or mobile devices [16, 27, 7].
- Data Scarcity and Annotation: Training deep models requires large annotated datasets, but gathering and labeling such data is labor-intensive and expensive [2, 3].

Recent studies have concentrated on tackling these difficulties by means of lightweight architectures, self-supervised learning, and sophisticated data augmentation [3, 2, 9].

2.3 Human Action Recognition: From Still Images to Videos

2.3.1 Early Methods and the Shift to Deep Learning

Human action recognition (HAR) seeks to automatically find and categorize human actions from still images or video sequences. Early approaches often combined classical classifiers like SVMs and HMMs with handcrafted features including motion trajectories, spatio-temporal interest points, and semantic attributes [3, 29]. These methods battled to adequately capture complicated temporal dynamics and the great variability of actual video material.

Especially with convolutional neural networks (CNNs), the shift to deep learning allowed end-to- end learning of hierarchical feature representations directly from raw video frames [4, 8]. Two-stream architectures—which separately handle spatial (RGB) and temporal (optical flow)—achieve notable improvements [4, 12] by capturing both appearance and motion cues. Expanding this idea, temporal segment networks (TSN) segmented videos to replicate long-range temporal structures [8].

2.3.2 Datasets for Human Action Recognition

Several benchmark datasets have driven progress in HAR:

- UCF101: Comprising 13,320 video clips across 101 action categories, UCF101 introduces substantial variations in camera motion, background, and lighting, making it a standard for evaluating deep learning models [5, 6].
- HMDB51: Includes 6,849 video clips across 51 diverse action categories, providing a more challenging dataset due to higher intra-class variation and realistic scenarios [5, 17].
- Kinetics (400/600/700): These large-scale datasets contain hundreds of thousands of annotated video clips spanning up to 700 action categories. Their scale and diversity make them suitable for training and benchmarking deep CNNs and transformer-based models [9, 21, 22].

These datasets provide diverse and realistic scenarios, enabling the evaluation of models under challenging conditions.

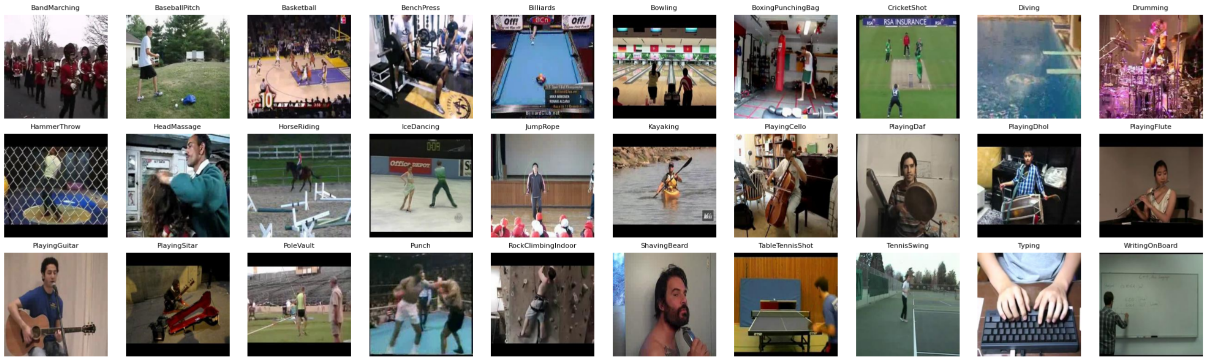


Figure 2.2: UCF101 Dataset

2.3.3 Deep Learning Architectures for Action Recognition

Key deep learning architectures for HAR include:

- CNNs are used to extract spatial features from individual frames or short clips [4, 8, 30].
- 3D CNNs process video volumes, enabling simultaneous modeling of spatial and short-range temporal features [3, 12].
- RNNs and LSTMs are employed to capture temporal dependencies across frames and are often integrated with CNNs for feature extraction [31, 10].
- Two-Stream Networks explicitly model spatial and temporal components using RGB and optical flow inputs [4, 12].
- Graph Convolutional Networks (GCNs) represent skeletal or object-based interactions, improving the recognition of complex and fine-grained actions [1, 11, 32].

- Transformer-Based Models, such as Vision Transformers (ViT), DeiT, and hybrid CNN-Transformer combinations, utilize self-attention mechanisms to capture long-range dependencies in both space and time [9, 20, 21, 22].

Hybrid models combining CNNs for localized feature extraction with transformers for global contextual modeling are especially effective on large-scale datasets [8, 20]. Additionally, skeleton-based recognition is a prominent subfield, leveraging joint coordinates and their dynamics to enhance model robustness [11, 32].

2.3.4 Challenges and Future Directions in Action Recognition

Despite advances, HAR faces ongoing challenges:

- Complexity of Real-World Videos: Recognition accuracy is hampered by camera motion, occlusions, cluttered backgrounds, and diverse action execution styles [3, 5, 33].

- Temporal Modeling: Especially in untrimmed or multi-action videos, long-range temporal dependencies and action progression remain challenging to adequately model [8, 9, 21].

- Computational Efficiency: Deep CNNs and transformers are limited in use in real-time or edge applications due to their often high computational demand [7, 16, 26, 27].

- Data Annotation and Generalization: Manual labeling is labor-intensive; current models find it difficult to extend over domains or to new, unseen actions [13, 19, 34].

Recent work has concentrated on knowledge distillation, self-supervised learning, multi-modal fusion, and the development of lightweight, efficient architectures [7, 8, 13, 21]. These challenges are addressed here. Particularly interesting for efficiently capturing local and global features in complex video data are hybrid models including CNNs, transformers, and GCNs [8, 20, 22, 32].

2.4 Comparative Analysis and Synthesis

2.4.1 Architectural Innovations and Their Impact

The literature consistently highlights the importance of architectural innovations in advancing the state of the art in gesture and action recognition:

- Residual Connections (ResNet) reduce vanishing gradients, so enhancing both accuracy and convergence speed [12, 35, 36, 37] and enabling the training of very deep networks.

- By means of effective gradient flow and feature reuse, Dense Connectivity (DenseNet) lowers parameter redundancy and improves discriminative power [12, 37].

- Crucially for modeling complex actions and gestures in sequential data, self-attention mechanisms (transformers) capture global dependencies and long-range relationships [8, 9, 21].

- Excellent performance and generalizing capacity have been shown by hybrid architectures combining the strengths of convolutional neural networks (CNNs) for local feature extraction and transformers for global context modeling [8, 20, 22].

2.4.2 Trade-Offs and Practical Considerations

In gesture and action recognition, model choice entails compromises between feasibility for real-time applications, computational economy, and accuracy. For settings with limited resources, such mobile or embedded devices [26, 9, 27], lightweight models such MobileNetV2 and ViT Tiny Patch provide a reasonable mix. On the other hand, although requiring more computational resources, deeper architectures like ResNet 152 or ViT Base Patch generate better accuracy [35, 22, 38]. By means of techniques including transfer learning and data augmentation, extensive application helps to improve model generalization and reduce dependency on large annotated datasets [17, 24, 13].

2.4.3 Open Issues and Future Research Directions

Key open issues identified in the literature include:

- Generalization to Unseen Scenarios: Many models struggle in novel settings, subjects, or action classes not observed in training [3, 5, 29].

- Data Efficiency: Using self-supervised or semi-supervised learning techniques [13, 19, 21, 39] will help to strongly push reliance on big labeled datasets towards reduction.

- Real-Time and Edge Deployment: Key issues still remain ensuring that high-accuracy models satisfy the latency and resource constraints of real-time or edge deployment [7, 16, 27].

- Multi-Modal Fusion: Combining several data modalities (e.g., RGB, depth, skeleton, radar, EMG) is progressively seen as a path to improve recognition robustness and accuracy.Also wrist-wearable electromyography (EMG)-based hand gesture recognition systems.[10, 8, 9, 32, 40].

Chapter 3

METHODOLOGY

This section describes the complete approach used in the comparison between deep learning and transformer models for hand gesture recognition (using the MSRA dataset) and human action recognition (using the UCF101 dataset). The approach is set up to guarantee consistent, fair, and exact evaluation for both projects. Dataset selection and justification, data preparation, augmentation, partitioning techniques, model architecture selection and adaptation, training protocols, and evaluation metrics comprise the process.

3.1 Dataset Selection and Justification

3.1.1 Hand Gesture Recognition: MSRA Hand Gesture Dataset

Especially in depth-based analysis [41, 18], the MSRA Hand Gesture Dataset is a major benchmark in vision-based gesture recognition. It features 76,500 depth images spread over 17 gesture classes executed by nine people. Gathered with a depth camera, the dataset minimizes lighting and background clutter—limitations typical of RGB datasets—while introducing variability in hand shape, motion, and orientation.

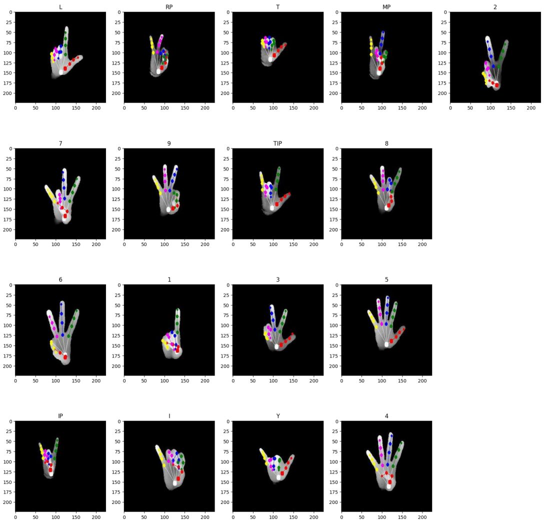


Figure 3.1: Different Classes in MSRA DataSet

The structure of the dataset helps to guarantee strong spatial feature extraction capability of models. Its thorough coverage of gesture variants and clean, noise-free depth images makes it perfect for hand gesture recognition benchmarking CNN and transformer architectures. Recent polls and studies also routinely cite the dataset as the main tool for assessing new algorithms in this field [41, 18, 42].

3.1.2 Human Action Recognition: UCF101 Dataset

The UCF101 dataset is chosen for human action recognition because of its size, variety, and standing as a field standard benchmark. Inspired from YouTube, it features 13,320 video clips spanning 101 action categories covering a wide range of real-world events including sports, musical performances, human-object interactions, and more [43]. Variations in camera motion, object appearance, background clutter, and illumination conditions define the complexity of the dataset, so reflecting real-world challenges [5, 6]. This work uses the top 50 most frequent classes in computational manageability and to address class imbalance [9, 5, 21]. This subset lets effective training and evaluation of deep learning models while maintaining the variety of human activities.

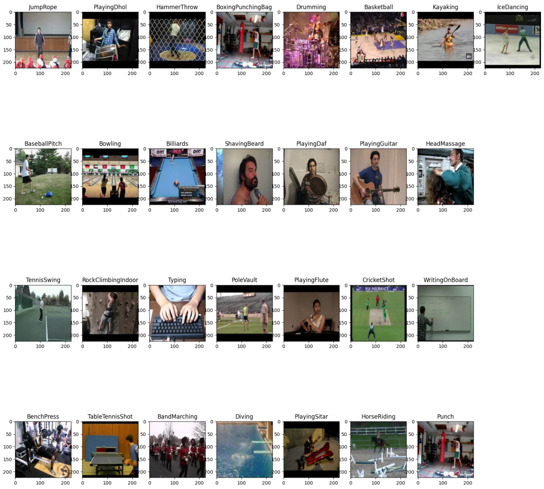


Figure 3.2: Top Classes with highest occurrence in UCF101 Dataset

3.2 Data Preprocessing

3.2.1 Image and Video Frame Preparation

- Image Resizing: Every image and video frame is cropped to 224 by 224 pixels. Typically expecting inputs of this dimension, this size is selected for compatibility with standard deep learning architectures (including VGG, ResNet, DenseNet, MobileNet, and ViT)[31].
- Normalization: To stabilize training and avoid bias resulting from different intensity scales, pixel values are scaled to a consistent range usually 1 or [-1, 1].
- Depth Data Handling: Depth images for the MSRA dataset are kept grayscale, preserving the spatial structure vital for gesture. recognition[41].

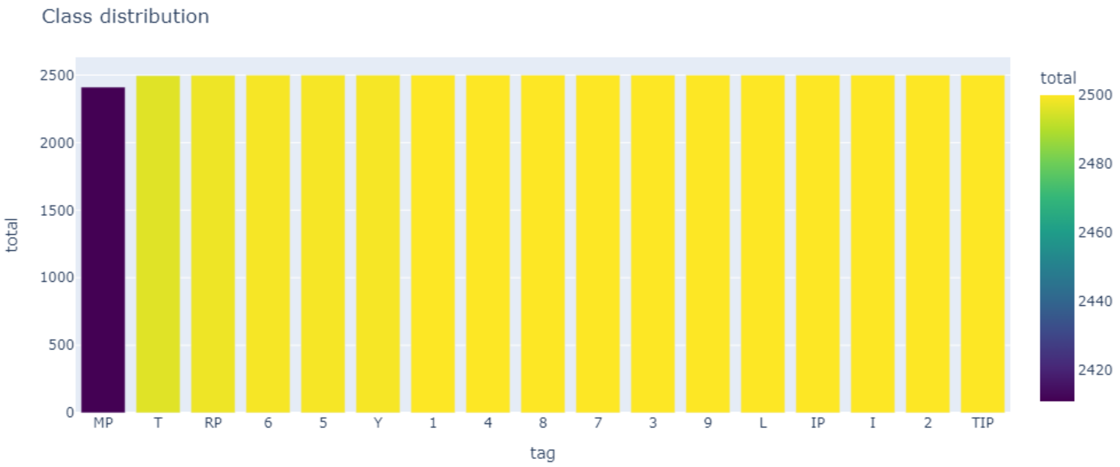


Figure 3.3: Class Distribution in MSRA Dataset

3.2.2 Data Augmentation

To enhance model robustness and generalization, extensive data augmentation is applied:

- Hand Gesture Recognition (MSRA):
  - Rotation: Random rotations simulate various hand orientations.
  - Flipping: Horizontal flips account for left- and right-handed gestures.
  - Contrast Adjustment: Varies illumination to make models invariant to lighting changes.
  - Gaussian Noise: Simulates sensor noise, improving robustness.
- Human Action Recognition (UCF101):
  - Temporal Sampling: From every video, ten equally spaced frames are taken to preserve computational efficiency while capturing the temporal evolution of motions.
  - Rotation and Flipping: To accommodate camera angle and action direction variation, frames are rotated randomly (within ±15 degrees) and flip horizontally.
  - Spatial Cropping: Random crops reduce background clutter by concentrating the model on action-relevant areas.

3.2.3 Dataset Partitioning

- MSRA: The dataset comprises 33,989 testing images and 42,402 training images. Ten-fold cross-validation is used to minimise overfitting and guarantee strong performance assessment. Ten percent of the data is used for validation and ninety percent for training across each fold.
- UCF101: Training and validation follow an 80:20 stratified split, so maintaining class distribution. Furthermore used for thorough evaluation is ten-fold cross-valuation.

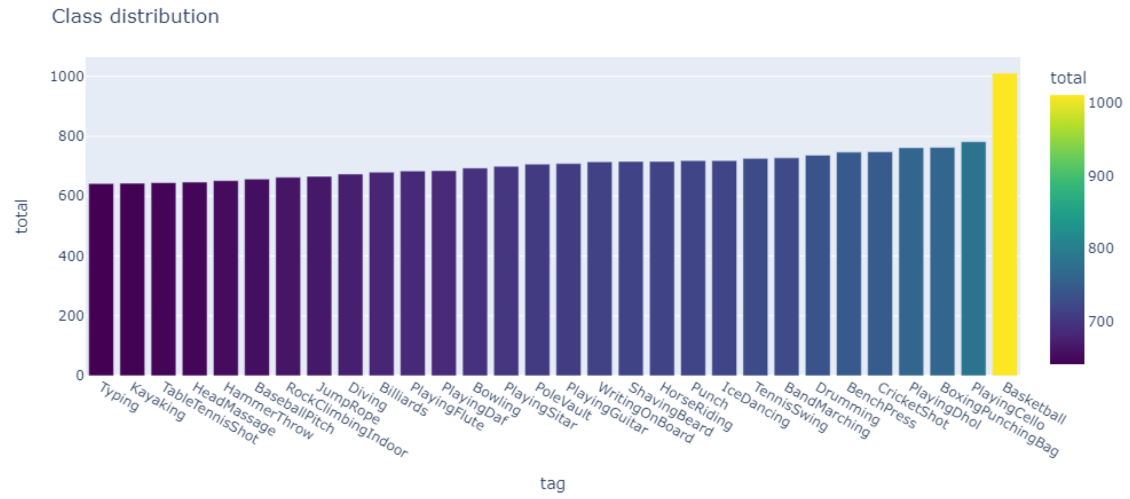


Figure 3.4: Class Distribution of top 30 classes in UCF101 Dataset

3.3 Model Architecture Selection and Adaptation

3.3.1 CNN-Based Architectures

- VGG16bn and VGG19bn: Deep sequential CNNs with batch normalisation, applied as spatial feature extraction baselines.

- DenseNet169 and DenseNet201: Feature-dense designs that encourage feature reuse and gradient flow help to lower parameter redundancy.

- ResNet101 and ResNet152: Deep residual networks enabling the training of very deep models by using skip connections to fight vanishing gradients.

- MobileNetV2: Designed for efficiency and fit in edge computing situations, a lightweight network using depthwise separable convolutions and inverted residuals.

3.3.2 Transformer-Based Architectures

- DeiT3 Base Patch: Particularly useful for video-based tasks, a data-efficient Vision Transformer model makes advantage of self-attention for spatial-temporal modeling.

- ViT Base Patch, ViT Small Patch, ViT Tiny Patch: Variations of the Vision Transformer architecture with different model size and patch resolution. Excellently capturing global dependencies, these models divide input images into non-overlapping patches and treat them as sequences.

3.3.3 Model Adaptation for Task-Specific Requirements

- Hand Gesture Recognition: Adapted for single-frame classification, models emphasize spatial feature extraction from depth images.

- Human Action Recognition: Models trained on individual frames for video data generate final predictions for each video by averaging the class probabilities over the sampled frames, so capturing temporal context.

3.4 Training Protocols

3.4.1 Optimization and Regularization

- Optimizer: Adaptive learning rate features of the Adam optimizer enable effective convergence by themselves. Included to regularize weights and stop overfitting is a weight decay value—let say 0.05.

- Learning Rate Scheduling: Starting at 0.0001, a cosine annealing schedule or step decay is used to progressively lower the learning rate during training so guaranteeing steady convergence.

- Dropout: Particularly in fully connected and transformer layers, dropout layers with rates up to 0.3 are included to help to further prevent overfitting.

- Early Stopping: If validation loss does not increase for three straight epochs, training is stopped, so avoiding needless computation and overfitting.

3.4.2 Data Loading and Augmentation During Training

- Dynamic Data Loaders: Custom data loaders dynamically apply augmentations and shuffle data to avoid order bias and improve generalizing capability.

- Mini-Batch Processing: Gradient updates are stabilized and training is accelerated using effective mini-batch processing.

- Temporal Sequence Sampling: Sequence samplers guarantee that frames from the same video are grouped for UCF101 so preserving temporal order for models that can take advantage of it.

3.5 Model Evaluation

3.5.1 Metrics

- F1-Score: Especially appropriate for multi-class classification with imbalanced data, the main metric for model comparison balances recall and accuracy.

- Accuracy, Precision, Recall: Recording complementary metrics helps to give a whole picture of model performance.

- Training and Validation Loss: Monitored across epochs to analyze convergence and detect overfitting.

3.5.2 Cross-Validation and Statistical Robustness

- Ten-Fold Cross-Validation: guarantees that models are tested on several subsets of the data, so generating strong approximations of generalization performance and lowering the overfitting risk to particular data splits.

- Ensemble Predictions: Where relevant, ensemble techniques—such as averaging forecasts from several trained models—can help to increase resilience and lower result variance.

3.5.3 Model Comparison and Visualization

- Validation Curves: Plotting loss and F1-score curves for every model throughout epochs helps one to see training dynamics and convergence behavior.

- Confusion Matrices: Used to examine per-class performance and spot typical misclassifications.

- Parameter Count and Inference Speed: In order to guide trade-offs between accuracy and computational efficiency, where feasible the number of parameters and inference speed are noted.

3.6 Task-Specific Considerations

3.6.1 Hand Gesture Recognition (MSRA)

- Spatial Feature Extraction: Using CNN's strengths and transformer models in learning from structured spatial data, the aim is to extract discriminative spatial features from depth images.

- Data Augmentation: Focus on modeling actual changes in hand orientation, illumination, and noise level.

3.6.2 Human Action Recognition (UCF101)

- Temporal Modeling: Although models learn on individual frames, temporal modeling is approximated by averaging predictions across frames. Explicit temporal modules—e.g., LSTM, temporal transformers—may find place in future work.

- Class Imbalance Handling: Choosing the top 50 most often occurring classes guarantees a balanced dataset, so enhancing the validity of performance evaluation.

Chapter 4

RESULTS and DISCUSSION

The experimental results for human action recognition (using the UCF101 dataset) and hand gesture recognition (using the MSRA dataset) are systematically analyzed in this part. The debate is set to underline the relative performance of all assessed models, examine convergence and generalization, and offer ideas on architectural trade-offs, computational economy, and pragmatic application.

4.1 Hand Gesture Recognition on the MSRA Dataset

4.1.1 Experimental Setup and Evaluation Metrics

Comprising 76,500 depth images across 17 gesture classes, the MSRA Hand Gesture dataset consisted in 42,402 training images and 33,989 testing images. Robust performance estimate and overfitting minimization were guaranteed by means of a 10-fold cross-valuation system. To improve generalization, data augmentation—rotation, flipping, contrast adjustment as well as normalizing was used. The F1-score was the main assessment tool; additionally recorded were accuracy, precision, and recall.

4.1.2 Comparative Model Performance

Key Findings:

Table 4.1: Performance Comparison of Different Models MSRA DataSet

Model	**Training Accuracy (%)**	**Testing Score (%)**	**F1 Score**
DenseNet169	99.94	99.91	0.9919
ResNet101	99.96	99.95	0.9978
DenseNet201	99.93	99.91	0.9901
VGG16_bn	99.58	99.40	0.9404
MobileNetV2	99.89	99.84	0.9847
VGG19_bn	98.67	98.34	0.8495

- With an F1-score of 0.9978 ResNet101 proved to be most adept in differentiating between gesture classes and reducing false positives and negatives. Its deep residual architecture reduces vanishing gradients so that, even at higher depth, stable and accurate learning is enabled.

- Closely following with use of dense connectivity to maximize feature reuse and preserve strong gradient flow were DenseNet 169 (0.9919) and DenseNet 201 (0.9901). The continuous learning curves and great generalization of this architecture mirror their parameter efficiency.

- With an F1-score of 0.9847 MobileNetV2 offered a good compromise between computational efficiency and recognition accuracy. Though with a small accuracy compromise relative to deeper models, its lightweight design using depthwise separable convolutions makes it well-suited for real-time and edge applications.

- Underperforming relative to more modern designs were VGG16bn (0.9404) and VGG19bn (0.8495). The performance decline, especially between VGG16bn and VGG19bn, draws attention to the constraints of raising depth without architectural innovations including residual or dense connections.

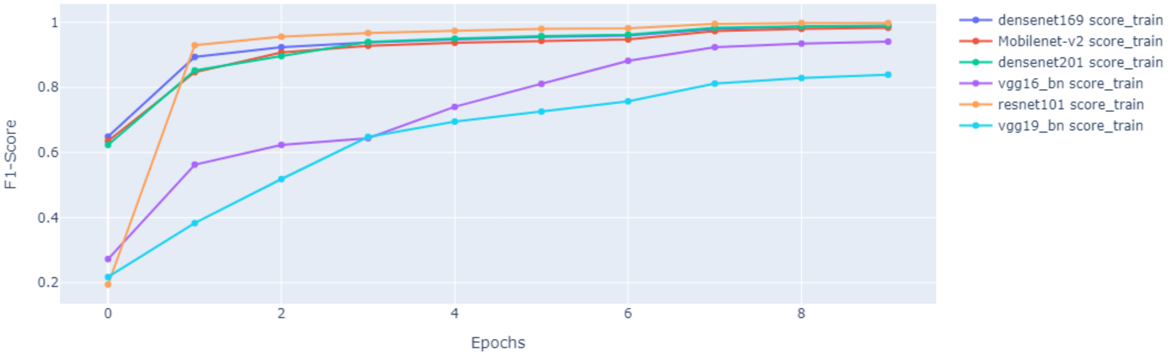


Figure 4.1: F-1 score versus epochs for all the models for MSRA Dataset.

4.1.3 Training Dynamics and Convergence Analysis

- Convergence Speed: Validation loss curves for every model stabilized in six to eight epochs. ResNet101 exhibited fastest initial loss reduction, suggesting quick convergence. While VGG variants showed more fluctuation and overfitting despite batch normalizing, DenseNet models also converged smoothly.

- F1-Score Evolution: ResNet101 kept constant excellence in F1-score throughout training, while DenseNet designs shown stable, high performance. Reflecting its efficiency-oriented architecture, MobileNetV2's F1-score was lower but stable.

- Generalization: Ten-fold cross-valuation guaranteed that models were assessed on several data splits, so lowering the overfitting risk and offering a consistent projection of real-world performance.

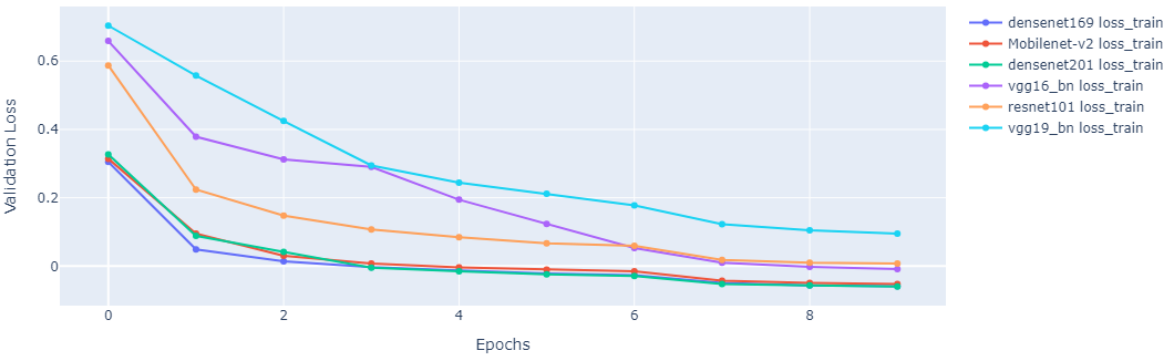


Figure 4.2: Validation loss versus epochs for all the models for MSRA Dataset.

4.1.4 Discussion: Architectural Insights and Practical Implications

- Residual Learning (ResNet101): ResNet101's skip connections let gradients pass more readily through deep networks, so enabling the model to learn intricate spatial features vital for differentiating subtle hand gestures.

- Dense Connectivity (DenseNet): When memory economy is a factor, DenseNet is a good choice since its feature reuse helps to achieve high discriminative power with less parameters.

- Lightweight Design (MobileNetV2): The performance of MobileNetV2 confirms the possibilities of effective models for deployment in settings with limited resources, including embedded systems or mobile devices.

- Sequential Depth Limitations (VGG): The notable performance decline for VGG19bn relative to VGG16bn highlights the declining returns of increasing depth without addressing optimization problems inherent in deep sequential networks.

4.2 Human Action Recognition on the UCF101 Dataset

4.2.1 Experimental Setup and Evaluation Metrics

For computational feasibility, the 13,320 video clips in the UCF101 dataset—which spans 101 action categories—were filtered to the 50 most often occurring classes. From every video, ten consistently sampled frames were resized, cropped, and augmented (rotation, flipping). Applying 10-fold cross-validation, the dataset was split 80:20 for training and validation. F1-score dominated the measurements; accuracy, precision, and recall were also tracked.

4.2.2 Comparative Model Performance

Key Findings:

Table 4.2: Performance Comparison of Different Models UCF101 Dataset

Model	Test Accuracy (%)	Val F1-Score	Val Loss
ResNet152	98.74	0.9976	0.0115
ResNet101	98.46	0.9956	0.0107
MobileNetV2	98.53	0.9895	0.0020
VGG16_bn	94.58	0.8482	0.1148
DeiT3 Base Patch	98.90	0.9996	0.0008
ViT Base Patch	99.03	0.9925	0.0015
ViT Small Patch	99.17	0.9994	0.0005
ViT Tiny Patch	99.47	0.9997	0.0001

- With an F1-score of 0.9997, ViT Tiny Patch outperformed all CNN-based models—including the rather deep ResNet152 and ResNet101. This reveals how effectively transformer-based designs capture long-range spatial-temporal dependencies in video data.

- DeiT3 Base Patch and ViT Small Patch also shown amazing performance with F1-scores above 0.9994. These models' self-attention mechanism helps to enable strong modeling of challenging action sequences.

- Among CNNs, ResNet 152 and ResNet 101 performed rather well; but, the slight improvement ResNet 152 over ResNet 101 points to declining returns with increasing depth for this work.

- Retaining a strong F1-score (0.9895), MobileNetV2 proved suitable for edge or real-time deployment requirements.

- VGG16bn lagged clearly with an F1-score of 0.8482, highlighting the limitations of older designs for demanding recognition systems.

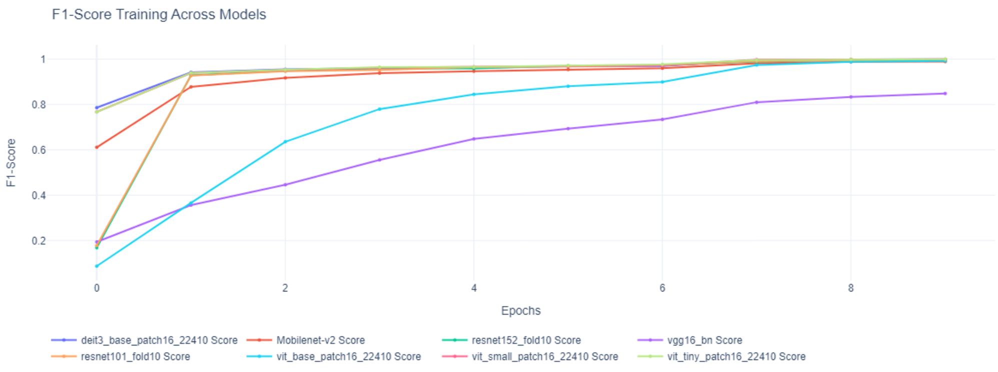


Figure 4.3: F-1 score versus epochs for all the models for UFC101 Dataset.

4.2.3 Training Dynamics and Convergence Analysis

- Transformer Models: ViT Tiny Patch and DeiT3 Base Patch shown fast convergence as validation loss fell sharply in the first few epochs and F1-scores rose rapidly to almost perfect levels. Moreover displaying lower final loss values, these models suggested more consistent and effective learning than CNNs.
- CNN Models: ResNet designs consistently improved loss and F1-score, but peak performance called for more epochs. MobileNetV2 converged rather well, but its final accuracy was rather less than that of the best transformers.
- Generalization: Consistent performance across folds and low variance in F1-scores propose strong generalization for the top-performance models.

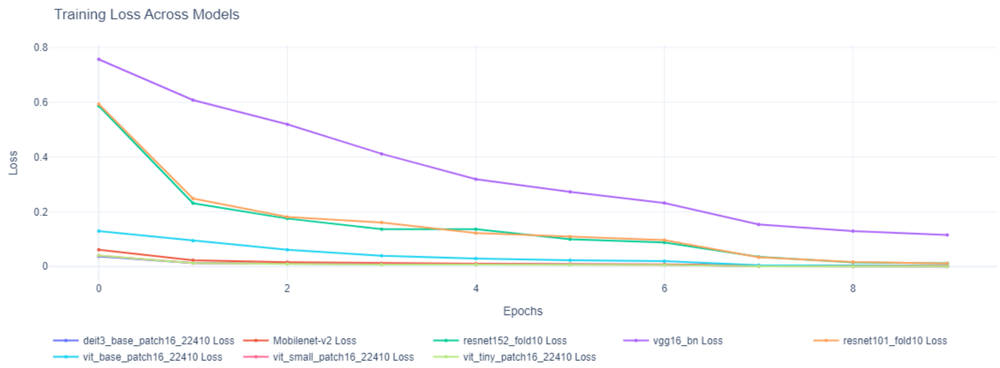


Figure 4.4: Validation loss versus epochs for all the models for UFC101 Dataset.

4.2.4 Discussion: Architectural Insights and Practical Implications

- Self-Attention Dominance: The better performance of transformer-based models stresses the need of self-attention for obtaining global context and temporal relationships in video data even with limited parameter counts (e.g., ViT Tiny Patch).
- Depth vs. Efficiency: The small depth increase in CNNs (ResNet 152 against ResNet 101) indicates that architectural changes—rather than only adding layers—are more crucial for improving performance.
- Resource-Constrained Deployment: Excellent MobileNetV2 performance confirms its suit for uses requiring efficient inference, such mobile video analysis or real-time surveillance.
- VGG Limitations: VGG16bn's underperformance highlights the need of contemporary architectures in video understanding, particularly in view of task complexity rising.

4.3 Cross-Task Synthesis and Broader Implications

4.3.1 Model Selection for Application Scenarios

- Image-Based Tasks (Hand Gesture Recognition): Deep CNNs with dense (DenseNet 169/ 201) or residual (ResNet 101) connectivity shine in extracting spatial features from structured data including depth images. For edge deployment, lightweight models (MobileNetV2) offer a reasonable compromise.

- Video-Based Tasks (Action Recognition): Even with smaller parameter footprints, transformer-based models (ViT, DeiT) are now the gold standard for modeling intricate spatial-temporal patterns. CNNs remain competitive, but as task complexity and data scale rise their benefits fade.

4.3.2 Trade-offs: Accuracy, Efficiency, and Real-Time Feasibility

- Accuracy: While transformers shine in video-based applications, deep CNNs get almost perfect F1-scores on benchmark datasets.

- Efficiency: MobileNetV2 and ViT Tiny Patch offer outstanding low computational cost performance for real-time and embedded systems.

- Generalization: While cross-valuation and ensemble methods boost resilience, future work should look at domain adaptation and self-supervised learning to improve generalization to hitherto unmet conditions.

4.3.3 Limitations and Future Directions

- Temporal Modeling: Explicit temporal modeling—e.g., temporal transformers, LSTM layers—could enhance current frame-based methods for action recognition for even higher accuracy.

- Multi-Modal Fusion: Especially in demanding situations, including other data sources (e.g., skeleton, depth, audio) can help to increase resilience.

- Data Efficiency: By lowering dependence on extensive labeled datasets, advanced augmentation and self-supervised learning help to enable more general application.

4.4 Summary

For both hand gesture and human action recognition, this comparison analysis shows that modern deep learning and transformer architectures have attained astonishing degrees of accuracy and robustness. Important developments in architecture are residual and dense connections in CNNs, self-attention in transformers. While transformer-based methods are redefining video understanding, lightweight models allow practical deployment. These realizations direct model choice for various HCI uses and point up interesting avenues for next studies.

Chapter 5

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

Two important domains of human-computer interaction—hand gesture recognition (using the MSRA dataset) and human action recognition (using the UCF101 dataset)—were thoroughly compared in this paper between efficient deep learning and transformer-based models. From classical convolutional neural networks (CNNs) like VGG16bn, VGG19bn, DenseNet169, DenseNet201, ResNet101, ResNet152, and MobileNetV2, to state-of-the-art transformer models such as DeiT3 Base Patch and Vision Transformer (ViT) variants (Base, Small, Tiny Patch)—under a unified experimental framework with consistent preprocessing, augmentation, and cross-validation strategies. The work methodically evaluated a wide spectrum of architectures. Key Findings:

- Hand Gesture Recognition: ResNet101 obtained the best F1-score (0.9978) on the MSRA dataset, so highlighting the value of deep residual learning for depth image analysis. By means of dense feature reuse, DenseNet 169 (0.9919) and DenseNet 201 (0.9901) also performed well. For real-time use, MobileNetV2 presented a good mix between accuracy (0.9847) and efficiency. Lacking modern additions like skip or dense connections, VGG models trailed behind.

- Human Action Recognition: Transformers (ViT Tiny Patch, DeiT3 Base Patch) led on the UCF101 dataset with F1-scores up to 0.9997, so faithfully modeling spatial-temporal dependencies. Though gains were less with depth, ResNet101 (0.9956) and ResNet 152 (0.9976) also performed well. While VGG16bn lagged (0.8482) reflecting its limitations for complex video tasks, MobileNetV2 balanced accuracy (0.9895) and efficiency.

- Architectural Insights: Accuracy and generalization across both tasks depend critically on developments in residual connections, dense blocks, and self-attention. While deeper architectures excelled where maximum accuracy is crucial, lightweight models (e.g., MobileNetV2, ViT Tiny Patch) offered strong performance with great efficiency, ideal for real-time use.

Implications: While transformer models are becoming the norm for video and sequence modeling, this work emphasizes that deep CNNs are still useful for image-based tasks including stationary hand gesture recognition. Architectural choice should strike a mix between deployment requirements, efficiency, and accuracy.

5.2 Future Scope

While this report provides a comprehensive benchmark and analysis, several promising directions remain for further advancing gesture and action recognition systems:

1. Hybrid Architectures
   (a) CNN-Transformer Fusion: Global context modeling of transformers combined with CNN's local feature extraction generates models that excel in both spatial and temporal tasks. Early CNN and transformer features for sequence modeling could help to boost resilience and accuracy.
2. Explicit Temporal Modeling
   (a) Temporal Transformers and Sequence Models:
   (b) Future work should investigate explicit temporal modeling—such as temporal transformers, LSTM/GRU layers, or 3D CNNs—to better capture action dynamics and increase recognition of complex temporal patterns—this work averaged frame-level predictions for video classification.
3. Edge and On-Device Inference:
   (a) Real-time applications in robotics, AR/VR, and IoT depend on optimizing inference for hardware constraints—e.g., by using TensorRT, ONNX, or custom accelerators.
4. Generalization and Adaptation
   (a) Domain Adaptation: Still a difficult task is developing models that extend across several environments, subjects, and sensor kinds. Further investigation should go into domain adaptation and transfer learning techniques.
   (b) Robustness to Occlusions and Variability: Robust real-world systems will depend critically on methods for managing occlusions, viewpoint changes, and inter-subject variability.
5. Ethical and Societal Considerations
   (a) Bias and Fairness: Especially for uses in healthcare, surveillance, and accessibility, it is imperative that recognition systems function fairly across many populations and avoid unintentionally encoding or amplifying prejudices.
   (b) Privacy and Security: As gesture and action recognition systems spread, safeguarding user privacy and protecting sensitive data will need constant attention.

All things considered, this work provides a strong foundation for pragmatic advances in gesture and action recognition as well as for next investigations. Rigorously benchmarking a broad spectrum of deep learning and transformer models stresses architectural innovations pushing ahead change and provides unambiguous direction on model selection.Along with improving model architectures and training approaches, the road forward addresses more general issues of real-world deployment, generalization, and ethical use. These systems will become ever more important as the field develops in allowing more natural, intelligent, inclusive human-computer interaction.

Appendix A

Dataset Details

A.1 MSRA Hand Gesture Dataset

The MSRA Hand Gesture Dataset is a benchmark collection for depth-based hand gesture recognition, curated by Microsoft Research Asia. It consists of 76,500 depth images, each representing one of 17 distinct gesture classes performed by nine subjects. Each gesture was captured using a depth camera, ensuring diversity in hand shapes, orientations, and motion patterns. The depth images are 320×240 pixels, and each gesture sequence is accompanied by joint location data for 21 hand joints per frame, stored as 3D coordinates. This dataset is particularly advantageous for spatial feature extraction because depth data reduces the impact of background clutter and lighting variability. The dataset was split into 42,402 training images and 33,989 testing images, with 10-fold cross-validation used to ensure robust evaluation and minimize overfitting.

A.2 UCF101 Human Action Recognition Dataset

The UCF101 dataset is a widely used benchmark for human action recognition in videos. It contains 13,320 video clips spanning 101 action categories, sourced from YouTube. The videos are characterized by significant variations in camera motion, object appearance, background clutter, and illumination. For this study, the top 50 most frequent classes were selected to manage computational complexity and ensure class balance. Each video was sampled to extract 10 uniformly spaced frames, resized to 224×224 pixels. The dataset was divided into 100,372 training frames and 17,202 testing frames, with stratified 80:20 split and 10-fold cross-validation for comprehensive evaluation.

Appendix B

Data Preprocessing and Augmentation

B.1 Hand Gesture (MSRA)

- Resizing: All depth images resized from 320×240 to 224×224 pixels.
- Normalization: Pixel intensities normalized to 1 or [-1, 1].
- Augmentation: Random rotation, horizontal flipping, contrast adjustment, and noise was added to increase data diversity and robustness against variations in hand orientation and lighting.

B.2 Human Action (UCF101)

- Frame Extraction: 10 equally spaced frames per video to capture temporal dynamics.
- Resizing: All frames resized to 224×224 pixels.
- Augmentation: Included random rotations (±15°), horizontal flipping, and spatial cropping to simulate camera viewpoint changes and action direction variability.
- Stratification: Ensured class balance in training and validation splits.

Appendix C

Model Training and Evaluation Protocols

- Optimization: Adam optimizer with initial learning rate of 0.0001, reduced via cosine annealing or step decay.
- Regularization: Dropout (up to 0.3) and early stopping (patience of 3 epochs without validation improvement).
- Cross-Validation: 10-fold cross-validation for both datasets to ensure robust and unbiased performance estimates.
- Metrics: F1-score (primary), accuracy, precision, recall, and loss tracked per epoch.

Appendix D

Hardware and Training Environment

- Hardware: Training was conducted on a workstation equipped with an Intel i7-14650HX 2.20 GHz CPU, NVIDIA RTX5060 GPU, and 16 GB RAM.
- Software: Models implemented in PyTorch, with data augmentation and loading handled by custom DataLoader scripts.
- Training Duration: Each model was trained for up to 10 epochs, with early stopping based on validation loss.

List of Publications

1. Sahil Sutty and Virender Ranga, ”A Comparative Study of Efficient Deep Learning Models for Hand Gesture Recognition on MSRA”, *Sixteenth International Conference on Advances in Computing, Control, and Telecommunication Technologies.(ACT-2025)*,2025.(Accepted)

2. Sahil Sutty and Virender Ranga, ”A Comparative Study on Deep Learning and Transformer Architectures for Human Action Recognition on the UCF-101 Dataset”, *International Conference on Data Science and Network Security (ICDSNS)*, 2025.(Minor Review)

Bibliography

[1] S. R. A. B. M. L. I. Khalaf, S. A. Aswad and M. R. Ahmed, “Survey on recognition hand gesture by using data mining algorithms,” in *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, Ankara, Turkey, 2022, pp. 1–4.

[2] M. B. M. N. Mohamed and N. Jomhari, “A review of the hand gesture recognition system: Current progress and future directions,” pp. 157 422–157 436, 2021.

[3] S. Z. M. H. A. Osman Hashi and A. B. Asamah, “A systematic review of hand gesture recognition: An update from 2018 to 2024,” pp. 143 599–143 626, 2024.

[4] S. L. X. T. X. Sun, Y. Wei and J. Sun, “Cascaded hand pose regression,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 824–832.

[5] H. W. Q. De Smedt and J.-P. Vandeborre, “Skeleton-based dynamic hand gesture recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2016, p. 1–9.

[6] J.-P. V. J. G. B. L. S. Q. D. Smedt, H. Wannous and D. Filliat, “Shrec’17 track: 3d hand gesture recognition using a depth and skeletal dataset,” in *Proc. 3DOR-10th Eurographics Workshop 3D Object Retr.*, 2017, p. 1–6.

[7] Z. L. A. M. W. Qi, S. E. Ovur and R. Song, “Multi-sensor guided hand gesture recognition for a teleoperated robot using a recurrent neural network,” in *IEEE Robotics and Automation Letters*, 2021, pp. 6039–6045.

[8] A. A. S. Zabihi, E. Rahimian and A. Mohammadi, “Trahgr: Transformer for hand gesture recognition via electromyography,” pp. 4211–4224, 2023.

[9] E. R. A. M. M. Montazerin, S. Zabihi and F. Naderkhani, “Vit-hgr: Vision transformer-based hand gesture recognition from high density surface emg signals,” in *2022 44th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, Glasgow, Scotland, United Kingdom, 2022, pp. 5115–5119.

[10] M. A. M. H. A. S. M. Miah and J. Shin, “Dynamic hand gesture recognition using multi-branch attention based graph and general deep learning model,” pp. 4703–4716, 2023.

[11] T.-T. N. et al., “A continuous real-time hand gesture recognition method based on skeleton,” in *2022 11th International Conference on Control, Automation and Information Sciences (ICCAIS)*, Hanoi, Vietnam, 2022, pp. 273–278.

[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, Columbus, OH, USA, 2014, pp. 1725–1732.

[13] H. Chang, L. Liang, X. Li, S. Wang, X. Pan, and J. Hu, "A parallelized framework for human action recognition and prediction based on graph neural networks," in *2024 China Automation Congress (CAC)*, Qingdao, China, 2024, pp. 6018–6023.

[14] F. S. Khan, J. Xu, J. van de Weijer, A. D. Bagdanov, R. M. Anwer, and A. M. Lopez, "Recognizing actions through action-specific person detection," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4422–4432, 2015.

[15] K. Lai and S. N. Yanushkevich, "Cnn+rnn depth and skeleton based dynamic hand gesture recognition," in *2018 24th International Conference on Pattern Recognition (ICPR)*, Beijing, China, 2018, pp. 3451–3456.

[16] S.-J. R. J.-W. Choi and J.-H. Kim, "Short-range radar based real-time hand gesture recognition using lstm encoder," pp. 33 610–33 618, 2019.

[17] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[18] M. Ramesh and K. Mahesh, "Sports video classification framework using enhanced threshold based keyframe selection algorithm and customized cnn on ucf101 and sports1-m dataset," *Computational Intelligence and Neuroscience*, 2022.

[19] H. Lee, "Evaluation of lc-ksvd on ucf 101 action dataset," n.d.

[20] G. Wang, J. Guo, J. Zhang, X. Qi, and H. Song, "Design of human action recognition method based on cross attention and 2s-agcn model," in *2024 IEEE ICCASIT*, Hangzhou, China, 2024, pp. 1341–1345.

[21] Y. Ou, X. Shi, J. Chen, R. He, and C. Liu, "From body parts to holistic action: A fine-grained teacher-student clip for action recognition," *IEEE Signal Processing Letters*, vol. 32, pp. 1336–1340, 2025.

[22] P. Parmar, E. Peh, and B. Fernando, "Learning to visually connect actions and their effects," in *WACV 2025*, Tucson, AZ, USA, 2025, pp. 1477–1487.

[23] Z. Li *et al.*, "Object-augmented skeleton-based action recognition," in *2023 IEEE AICAS*, Hangzhou, China, 2023, pp. 1–4.

[24] S. R. B. G. M. S. A. Das, K. Maitra and S. Biswas, "Development of a real time vision-based hand gesture recognition system for human-computer interaction," in *2023 IEEE 3rd Applied Signal Processing Conference (ASPCON)*, India, 2023, pp. 294–299.

[25] J. K. V. N. S. C. D. S. D. A. Reddy, V. E. Jyothi and S. Sindhura, "A pattern recognition model: Hand gestures recognition using convolutional neural networks," in *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, Coimbatore, India, 2023, pp. 460–465.

[26] D. G. L. et al., “Video hand gestures recognition using depth camera and lightweight cnn,” pp. 14 610–14 619, 2022.

[27] Y. M. A. K. S, V. P and W. A. J, “Innovative hand gesture recognition techniques for volume adjustment in real-time,” in *2024 4th International Conference on Artificial Intelligence and Signal Processing (AISP)*, VIJAYAWADA, India, 2024, pp. 1–5.

[28] K. Sachdeva and R. Sachdeva, “A novel technique for hand gesture recognition,” in *2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)*, Faridabad, India, 2023, pp. 556–560.

[29] Y. Mitsuzumi, A. Kimura, G. Irie, and A. Nakazawa, “Cross-action cross-subject skeleton action recognition via simultaneous action-subject learning with two-step feature removal,” in *2024 IEEE ICIP*, Abu Dhabi, UAE, 2024, pp. 2182–2186.

[30] Y.-H. Wu, W.-J. Tsai, and H.-T. Chen, “Temporal action detection based on hierarchical object detection networks,” in *2019 Ubi-Media*, Bali, Indonesia, 2019, pp. 119–123.

[31] L. Jiashan and L. Zhonghua, “Dynamic gesture recognition algorithm combining global gesture motion and local finger motion for interactive teaching,” pp. 128 117–128 128, 2024.

[32] R. Zhang and X. Yan, “Video-language graph convolutional network for human action recognition,” in *ICASSP 2024*, Seoul, South Korea, 2024, pp. 7995–7999.

[33] L. Ting-Long, “Short-term action learning for video action recognition,” *IEEE Access*, vol. 12, pp. 30 867–30 875, 2024.

[34] R. Tanigawa and Y. Ishii, “Hear-your-action: Human action recognition by ultrasound active sensing,” in *ICASSP 2024*, Seoul, South Korea, 2024, pp. 7260–7264.

[35] T. Su, H. Wang, and L. Wang, “Multi-level content-aware boundary detection for temporal action proposal generation,” *IEEE Transactions on Image Processing*, vol. 32, pp. 6090–6101, 2023.

[36] C.-K. Lu, M.-W. Mak, R. Li, Z. Chi, and H. Fu, “Action progression networks for temporal action detection in videos,” *IEEE Access*, vol. 12, pp. 126 829–126 844, 2024.

[37] A. K. Pandey and A. S. Parihar, “A comparative analysis of deep learning based human action recognition algorithms,” in *2023 ICCCNT*, Delhi, India, 2023, pp. 1–7.

[38] V.-D. Le, T.-L. Nghiem, and T.-L. Le, “Accurate continuous action and gesture recognition method based on skeleton and sliding windows techniques,” in *2023 APSIPA ASC*, Taipei, Taiwan, 2023, pp. 284–290.

[39] Y. S. P. Raj, S. Pandey and M. Singh, “Hand sign gesture recognition system,” in *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, Gautam Buddha Nagar, India, 2023, pp. 639–642.

[40] K. S. Prakash and N. Kunju, “An optimized electrode configuration for wrist wearable emg-based hand gesture recognition using machine learning,” *Expert Systems with Applications*, vol. 274, p. 127040, 2025.

[41] S. L. X. T. X. Sun, Y. Wei and J. Sun, “Cascaded hand pose regression,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, p. 824–832.

[42] S. Su and Y. Zhang, “Online hierarchical linking of action tubes for spatio-temporal action detection based on multiple clues,” *IEEE Access*, vol. 12, pp. 54 661–54 672, 2024.

[43] T. P. K. B. S. V. S. C. B. Jaganathan, K. R and A. Mital, “Dynamic hand gesture recognition,” in *2022 IEEE 2nd International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)*, Gunupur, Odisha, India, 2022, pp. 1–6.

List of Publications

1. Sahil Sutty and Virender Ranga, ”A Comparative Study of Efficient Deep Learning Models for Hand Gesture Recognition on MSRA”, *Sixteenth International Conference on Advances in Computing, Control, and Telecommunication Technologies.(ACT-2025)*,2025.(Accepted)

2. Sahil Sutty and Virender Ranga, ”A Comparative Study on Deep Learning and Transformer Architectures for Human Action Recognition on the UCF-101 Dataset”, *International Conference on Data Science and Network Security (ICDSNS)*, 2025.(Minor Review)

SIXTEENTH INTERNATIONAL CONFERENCE ON ADVANCES IN COMPUTING, CONTROL, AND TELECOMMUNICATION TECHNOLOGIES
ACT 2025

June 25-26, 2025; Hyderabad, India.
http://act.theides.org/2025/

GRENZE Scientific Society

Honorary Chair
Dr. Pawel Hitczenko
(Drexel University, USA)
Dr. Jiguo Yu
(Qufu Normal University, China)

Technical Chair
Dr. Mahesh P K (Rajeev Institute of Technology, Karnataka, India)
Dr. H N Prakash (Rajeev Institute of Technology, Karnataka, India)

Technical Co-Chair
Dr. Aravinda C V (NAMNITTE, Nitte, Mangalore, India)
Dr. Arjun B C (Rajeev Institute of Technology, Karnataka, India)

Chief Editor
Dr. Mahesh P K (Rajeev Institute of Technology, Karnataka, India)
Dr. H N Prakash (Rajeev Institute of Technology, Karnataka, India)

General Chair
Dr. Janahanlal Stephen
(Mookambika Tech. Campus, India)

General Co-Chair
Prof. K. U Abraham (Holycross College of Engineering, India)

Finanace Chair
Prof. Ford Lumban Gaol
(University of Indonesia)

Publicity Chair
Dr. Amit Manocha (Maharaja Agrasen Institute of Tech., India)

Poster Chair
Dr. Ashadi Salim (Bina Nusantara University, Indonesia)

National Advisory Committee
Dr. Togar Alam Napitupulu (Bina Nusantara University, Indonesia)

Program Committee Chair
Dr. Raymond Kosala (Bina Nusantara University, Indonesia)
Dr. Richard Kumaradjaja (Bina Nusantara University, Indonesia)

International Advisory Committee
Dr. Marc van Dongen
(University College Cork, Ireland)
Dr. Hooman Mohseni
(Northwestern University, USA)
Dr. Suresh Subramoniam (Prince Sultan University, Saudi Arabia)
Dr. Kabekode V. Bhat (The Pennsylvania State University, USA)

Sixteenth International Conference on Advances in Computing, Control, and Telecommunication Technologies - ACT 2025 will be held during June 25-26, 2025; Hyderabad, India.

ACT 2025, aims at bringing together the researchers, scientists, engineers, and scholar students in all areas of Computer Science, Control Engineering, Electrical and Electronics and Telecommunication Technology, and provides an international forum for the dissemination of original research results, new ideas and practical development experiences which concentrate on both theory and practices. The conference focuses on the frontier topics in the Computer Science, Electrical, Electronics and Telecommunication and Engineering subjects. The conference will be held every year to make it an ideal platform for people to share views and experiences in Information, Telecommunication, Computing Techniques and related areas. The conference is jointly organised by the **IDES** and the **Association of Computer Electrical Electronics and Communication Engineers (ACEECom).**

All the accepted registered papers will be published by the **Grenze Scientific Society** and it will be made available in the **GRENZE International Journal of Engineering and Technology (GIJET)** and will be indexed in **Scopus.**

Computational Mathematics
Data Structures, Algorithms
Artificial Intelligence
Automated Software Engineering
Bioinformatics and Scientific Computing
Compilers and Interpreters
Computational Intelligence
Computer Animation, Computer Games
Computer Architecture, Computer Vision
Computer Architecture and Embedded Systems, Data Mining
Computer Graphics & Virtual Reality
Computer Modeling, Databases
Computer Networks, Watermarking
Computer Security, Data Encryption
Computer Simulation, Image Processing
Computer-aided Design/Manufacturing
Computing Ethics, Data Compression
Computing Practices & Applications
Data Communications
Software Engineering & CASE

Digital Signal Processing
Distributed Systems, Robotics
Event Driven Programming
Expert Systems, Field Theory
High Performance Computing
Information Retrieval
Telecommunication Technologies
Mobile Computing, Digital Security
Multimedia Applications
Natural Language Processing
Parallel and Distributed Computing
Performance Evaluation
Programming Languages
Reconfigurable Computing Systems
Security & Cryptography
Applications of Computer Science and Engineering, Information Systems
Electrical Materials and Process
High Voltage Engineering and Insulation Technology
Electronic Materials, Mechatronics

Prospective authors are invited to submit full (original) research papers; which are NOT submitted/published/under consideration anywhere in other conferences/journals; in electronic (Doc or Docx only) format through the email **act.chair@gmail.com**

Important Dates

Last Date of Paper Submission: ***22 May 2025***
Paper Acceptance Notification: 25 May 2025
Camera Ready Paper: 20 June 2025
Paper Registration: ***20 June 2025***
Additional Co-author/Non-Author registration: June 22,2025

IDES ICPS IDES Conference Publishing System ACEECom

http://conference.theides.org/

Fwd: ACT2025 :: Individual Review Result

1 message

ACT Chair <act.chair@gmail.com> Thu, May 29, 2025 at 10:19 AM
To: 23/ITY/02 SAHIL SUTTY <sahilsutty_23ity02@dtu.ac.in>

---------- Forwarded message ---------
From: **ACT Chair** <act.chair@gmail.com>
Date: Sun, May 25, 2025 at 10:40 PM
Subject: ACT2025 :: Individual Review Result
To: <sahilsutty23ity02@dtu.ac.in>, <drvirender.ranga@gmail.com>

Dear ACT2025 Authors,

Welcome to **ACT2025.**

Congratulations - Your paper for the **Sixteenth International Conference on Advances in Computing, Control, and Telecommunication Technologies - ACT 2025,** has been accepted.

Paper ID	**ACT2025 -524**
Paper Title	A Comparative Study of Efficient Deep Learning Models for Hand Gesture Recognition on MSRA
Category Accepted	**Full Paper**

The **ACT2025** conference is jointly organized by the **IDES** and **Association of Computer Electrical Electronics and Communication Engineers (ACEECom)** and will be held during **June 25-26, 2025; Hyderabad, India.**

https://act.theides.org/2025/

All the accepted registered papers will be published by the ***Grenze Scientific Society*** and it will be made available in the **GRENZE International Journal of Engineering and Technology (GIJET)**, and will be indexed in **Scopus.** For your reference, to review previously published papers in the **GIJET**, please visit the following link: ***GIJET - Previously Published Papers.***

Recent Scopus Indexed Issues » 2021 | 2022 | 2023

Authors submitting their papers to **GRENZE** may use the **GRENZE standard** template. Failing this will result in rejection. All camera ready papers must be submitted in **MS WORD** file format only (PDF is NOT accepted) not exceeding stipulated pages including text, figures, tables and references.

The registration fee details, Registration form, copyright form and camera ready paper submission are mentioned in the following link.

https://act.theides.org/2025/reg.htm

It is mandatory for at least one author of an accepted paper to register in order for the paper to appear in the Proceedings of the Technical Sessions of the **ACT2025.**

Registration form: https://act.theides.org/2025/reg.htm

Payment option: http://theides.org/payment-in-sib.htm

Best Regards,

--

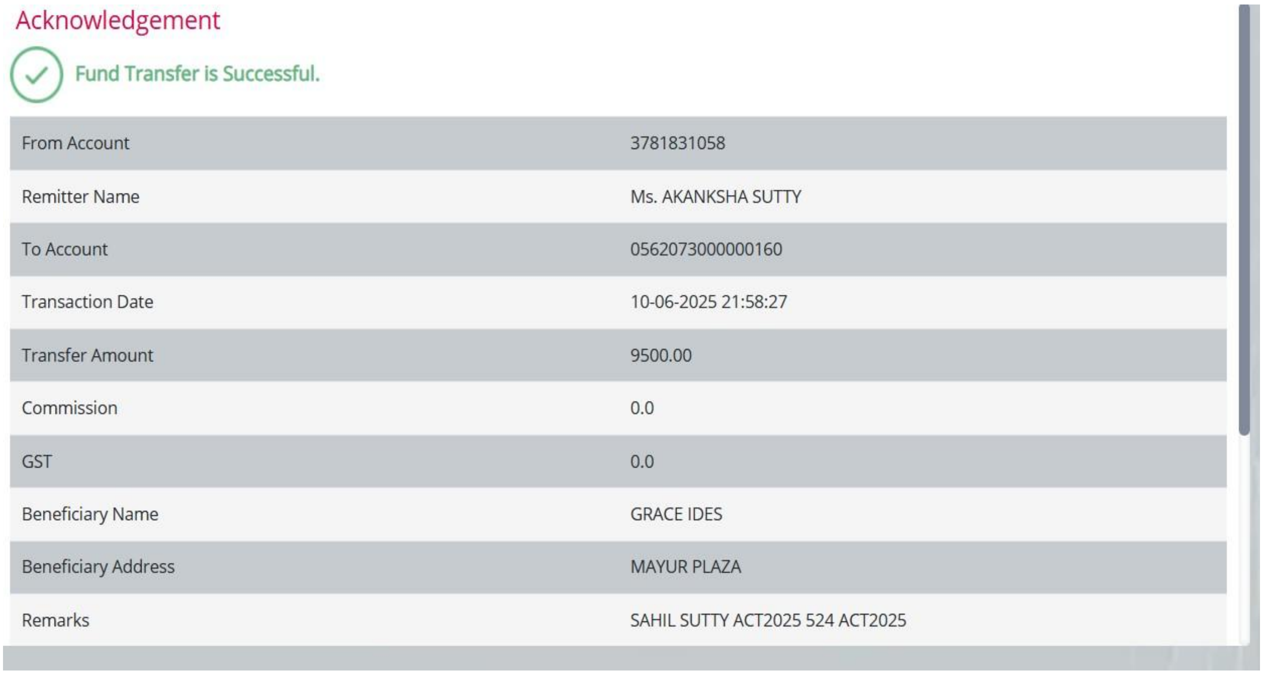
NOTE: Please visit the conference website carefully for any doubt. In all your communications please quote your Paper ID and Category

--
--
===
Sudheer
The IDES
Email:act.chair@gmail.com,Website:http://act.theides.org/

NETHERLANDS | INDIA | INDONESIA | USA
===

--
--
===
Sudheer
The IDES
Email:act.chair@gmail.com,Website:http://act.theides.org/

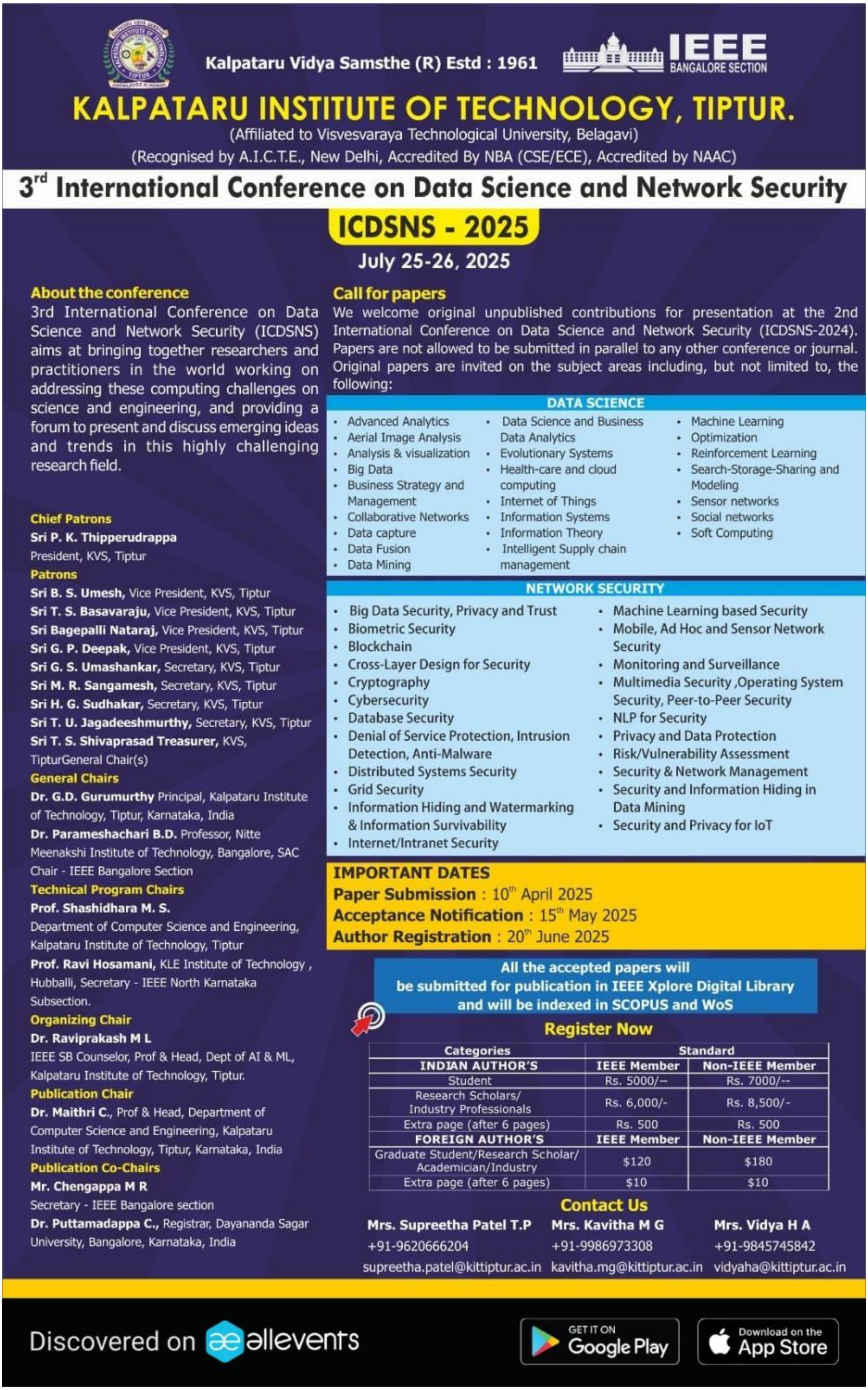
NETHERLANDS | INDIA | INDONESIA | USA
===



Acknowledgement

✓ Fund Transfer is Successful.

From Account	3781831058
Remitter Name	Ms. AKANKSHA SUTTY
To Account	0562073000000160
Transaction Date	10-06-2025 21:58:27
Transfer Amount	9500.00
Commission	0.0
GST	0.0
Beneficiary Name	GRACE IDES
Beneficiary Address	MAYUR PLAZA
Remarks	SAHIL SUTTY ACT2025 524 ACT2025



Kalpataru Vidya Samsthe (R) Estd : 1961

IEEE BANGALORE SECTION

KALPATARU INSTITUTE OF TECHNOLOGY, TIPTUR.

(Affiliated to Visvesvaraya Technological University, Belagavi)
(Recognised by A.I.C.T.E., New Delhi, Accredited By NBA (CSE/ECE), Accredited by NAAC)

3rd International Conference on Data Science and Network Security

ICDSNS - 2025

July 25-26, 2025

About the conference

3rd International Conference on Data Science and Network Security (ICDSNS) aims at bringing together researchers and practitioners in the world working on addressing these computing challenges on science and engineering, and providing a forum to present and discuss emerging ideas and trends in this highly challenging research field.

Chief Patrons

Sri P. K. Thipperudrappa
President, KVS, Tiptur

Patrons

Sri B. S. Umesh, Vice President, KVS, Tiptur
Sri T. S. Basavaraju, Vice President, KVS, Tiptur
Sri Bagepalli Nataraj, Vice President, KVS, Tiptur
Sri G. P. Deepak, Vice President, KVS, Tiptur
Sri G. S. Umashankar, Secretary, KVS, Tiptur
Sri M. R. Sangamesh, Secretary, KVS, Tiptur
Sri H. G. Sudhakar, Secretary, KVS, Tiptur
Sri T. U. Jagadeeshmurthy, Secretary, KVS, Tiptur
Sri T. S. Shivaprasad Treasurer, KVS, TipturGeneral Chair(s)

General Chairs

Dr. G.D. Gurumurthy Principal, Kalpataru Institute of Technology, Tiptur, Karnataka, India
Dr. Parameshachari B.D. Professor, Nitte Meenakshi Institute of Technology, Bangalore, SAC Chair - IEEE Bangalore Section

Technical Program Chairs

Prof. Shashidhara M. S.
Department of Computer Science and Engineering, Kalpataru Institute of Technology, Tiptur
Prof. Ravi Hosamani, KLE Institute of Technology , Hubballi, Secretary - IEEE North Karnataka Subsection.

Organizing Chair

Dr. Raviprakash M L
IEEE SB Counselor, Prof & Head, Dept of AI & ML, Kalpataru Institute of Technology, Tiptur.

Publication Chair

Dr. Maithri C., Prof & Head, Department of Computer Science and Engineering, Kalpataru Institute of Technology, Tiptur, Karnataka, India

Publication Co-Chairs

Mr. Chengappa M R
Secretary - IEEE Bangalore section
Dr. Puttamadappa C., Registrar, Dayananda Sagar University, Bangalore, Karnataka, India

Call for papers

We welcome original unpublished contributions for presentation at the 2nd International Conference on Data Science and Network Security (ICDSNS-2024). Papers are not allowed to be submitted in parallel to any other conference or journal. Original papers are invited on the subject areas including, but not limited to, the following:

DATA SCIENCE

- Advanced Analytics
- Aerial Image Analysis
- Analysis & visualization
- Big Data
- Business Strategy and Management
- Collaborative Networks
- Data capture
- Data Fusion
- Data Mining
- Data Science and Business Data Analytics
- Evolutionary Systems
- Health-care and cloud computing
- Internet of Things
- Information Systems
- Information Theory
- Intelligent Supply chain management
- Machine Learning
- Optimization
- Reinforcement Learning
- Search-Storage-Sharing and Modeling
- Sensor networks
- Social networks
- Soft Computing

NETWORK SECURITY

- Big Data Security, Privacy and Trust
- Biometric Security
- Blockchain
- Cross-Layer Design for Security
- Cryptography
- Cybersecurity
- Database Security
- Denial of Service Protection, Intrusion Detection, Anti-Malware
- Distributed Systems Security
- Grid Security
- Information Hiding and Watermarking & Information Survivability
- Internet/Intranet Security
- Machine Learning based Security
- Mobile, Ad Hoc and Sensor Network Security
- Monitoring and Surveillance
- Multimedia Security ,Operating System Security, Peer-to-Peer Security
- NLP for Security
- Privacy and Data Protection
- Risk/Vulnerability Assessment
- Security & Network Management
- Security and Information Hiding in Data Mining
- Security and Privacy for IoT

IMPORTANT DATES

Paper Submission : 10th April 2025
Acceptance Notification : 15th May 2025
Author Registration : 20th June 2025

All the accepted papers will be submitted for publication in IEEE Xplore Digital Library and will be indexed in SCOPUS and WoS

Register Now

Categories	Standard	
INDIAN AUTHOR'S	**IEEE Member**	**Non-IEEE Member**
Student	Rs. 5000/–	Rs. 7000/--
Research Scholars/ Industry Professionals	Rs. 6,000/-	Rs. 8,500/-
Extra page (after 6 pages)	Rs. 500	Rs. 500
FOREIGN AUTHOR'S	**IEEE Member**	**Non-IEEE Member**
Graduate Student/Research Scholar/ Academician/Industry	$120	$180
Extra page (after 6 pages)	$10	$10

Contact Us

Mrs. Supreetha Patel T.P
+91-9620666204
supreetha.patel@kittiptur.ac.in

Mrs. Kavitha M G
+91-9986973308
kavitha.mg@kittiptur.ac.in

Mrs. Vidya H A
+91-9845745842
vidyaha@kittiptur.ac.in