# EXPLAINABLESTAGE-AWARE BHARATANATYAM MUDRA RECOGNITION BY INTEGRATING GEOMETRIC HAND LANDMARK ENCODING AND DOUBLE TRANSFER LEARNING

*by* Ganesh Gutti

**Submission date:** 27-May-2026 02:17PM (UTC+0530)
**Submission ID:** 2970465579
**File name:** 24AFI25_GaneshGutti_Thesis_1.pdf (23.11M)
**Word count:** 9699
**Character count:** 63890

## ABSTRACT

Bharatanatyam mudras constitute an essential component of Indian classical dance, serving as a medium for storytelling, emotional expression, and semantic communication. Automatic recognition of these hand gestures is a challenging task due to subtle inter-class variations, complex finger articulations, viewpoint differences, and limited annotated datasets with existing gesture recognition approaches often relyinng solely on visual appearance features and lack robustness, interpretability, and cross-domain adaptability and in addition, limited research has explored the integration of multimodal learning and explainable artificial intelligence for culturally significant gesture recognition tasks such as Bharatanatyam mudra analysis.

This thesis proposes an explainable multimodal dual-stage transfer learning framework for Bharatanatyam mudra recognition by integrating transformer-based visual learning with geometric hand landmark representations which employs a Swin Transformer Tiny backbone for RGB image feature extraction and a dedicated landmark-processing branch utilizing MediaPipe hand keypoints to capture structural hand articulation and thus extracted features from both modalities are fused to enhance discriminative representation learning and improve classification performance. To address data scarcity and improve transferability, a dual-stage transfer learning strategy is introduced, where the model is initially pretrained on an Indian Sign Language (ISL) gesture dataset and subsequently fine-tuned on Bharatanatyam mudras. To add more, explainable artificial intelligence techniques such as GradCAM-based visualization are incorporated to help with the process of the interpreting model predictions and identification of anatomically significant hand regions influencing recognition decisions.

Extensive experiments are conducted on Bharatanatyam mudra datasets along with cross-domain evaluation using external gesture datasets to assess robustness and generalization capability and the proposed framework is evaluated using multiple performance metrics including accuracy, precision, recall, and F1-score, along with cross-dataset experiments to further analyze the effects of domain shift and transferability and the experimental results demonstrate that the proposed multimodal dual-stage transfer learning framework achieves robust and highly accurate Bharatanatyam mudra recognition while providing improved interpretability and cross-domain adaptability thereby contributing towards the development of culturally aware artificial intelligence systems for digital heritage preservation, intelligent dance analysis, and human-centered gesture understanding.

# CHAPTER 1

# INTRODUCTION

## 1.1 Background

Indian classical dance forms represent a significant component of the cultural and artistic heritage of India. Among these, Bharatanatyam is one of the oldest and most structured classical dance traditions, characterized by intricate body movements, rhythmic footwork, facial expressions, and symbolic hand gestures known as *mudras*. Mudras as shown in Fig 1.1 are crucal for the conveying emotions, narratives, and semantic meaning during performances which emphasise on the fact that accurate understanding and interpretation of these gestures are essential for preserving the expressive richness and communicative depth of Bharatanatyam as a dance form[1].

With the advancing technology in areas of artificial intelligence and computer vision technologies, automated gesture recognition systems have gained an increasing attention in the areas including but not limited to such as human-computer interaction, sign language interpretation, surveillance, virtual reality, and healthcare and recent developpments in deep learning, particularly convolutional neurl networks and transformer-based architectures, have significantly improved visual recognition performance across various applications[2] but however, recognizing Bharatanatyam mudras remains a challenging task due to subtle inter-class variations, complex finger articulations, occlusions, varying illumination conditions, and limited availability of annotated datasets[2].

In the recent years, the frameworks for the hand landmark detection like the MediaPipe have enabled the extraction of detailed geometric representations of the hand structures offering additional information beyond raw visual appearance[3, 4] and in the similar fashion, transfer learning techniques have also emerged as effective solutions for performance improvement in the data-constrained environments by making use of the knowledge learned from related domains[5, 6] and despite all these advancements, existing Bharatanatyam mudra recognition systems as shown in Fig 1.2 very often rely primarily on image-based features, lack explainability, and rarely investigate cross-domain generalization or multimodal fusion strategies.


**Figure 1.1:** Sample Bharatanatyam mudras used in the study

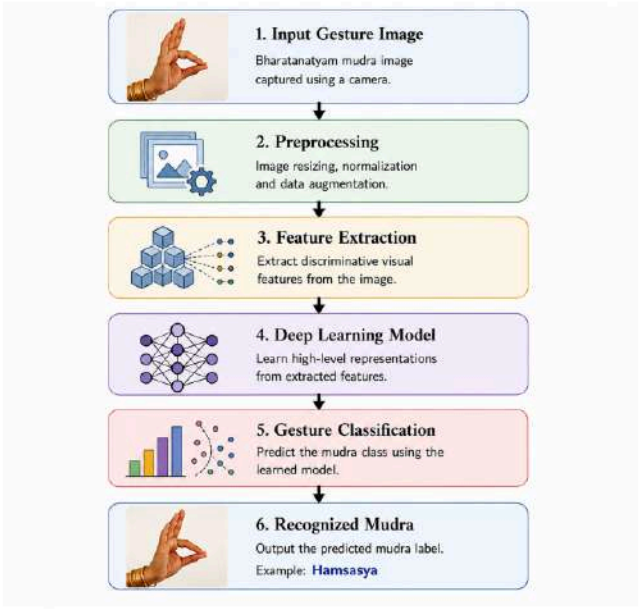



**Figure 1.2:** General workflow of an AI-based gesture recognition system

## 1.2 Problem Statement

Automatic recognition of Bharatanatyam mudras presents several technical and research challenges as presented in Table 1.1 : Traditional image-based classification approaches often struggle to capture the fine-grained geometric distinctions between visually similar mudras; Variations in hand orientation, performer-specific articulation styles, background conditions, and

lighting further complicate the recognition process;the limited size of available Bharatanatyam datasets restricts the ability of deep leaning models to generalise effectively.

Most existing approaches focus primarily on achieving classification accuracy without addressing interpretability and semantic understanding of model decisions and then, very limited research has explored the integration of multimodal representations combining visual and geometric hand features, so the lack of robust cross-domain evaluation also raises concerns regarding the adaptability and generalization capability of current systems when exposed to unseen gesture distributions or datasets.

Thus, there is a need for an explainable and robust Bharatanatyam mudra recognition framework. This should be capable of effectively integrating visual and geometric information. This should also be improving transferability across domains. In addition this should be providing interpretable predictions which are to be aligned with meaningful hand articulation patterns.

**Table 1.1:** Major challenges in Bharatanatyam mudra recognition

| Challenge | Description |
|---|---|
| Inter-class similarity | Many mudras possess highly similar finger configurations |
| Intra-class variation | Differences in performer styles and hand articulation |
| Background complexity | Variations in lighting and environmental conditions |
| Limited datasets | Scarcity of large annotated Bharatanatyam datasets |
| Interpretability limitations | Difficulty in understanding model decision mechanisms |

## 1.3 Aim of the Study

The primary aim of this research is to develop an explainable multimodal dual-stage transfer learning framework for robust Bharatanatyam mudra recognition using transformer-based visual learning and hand landmark fusion and the poposed architecture is as in Fig 1.3.

## 1.4 Research Gaps

After reviewing the existing literature, the following major research gaps were identified:

- Existing Bharatanatyam mudra recognition frameworks suffer from limited cross-dataset generalization due to strong studio bias, performer variability, and acquisition condition differences.

- Most existing approaches rely primarily on either visual appearance features or geometric landmark representations independently, with limited exploration of robust multimodal fusion frameworks for fine-grained mudra recognition.
- The adoption of explainable transformer-based architectures and progressive transfer learning strategies remains limited in Bharatanatyam mudra recognition and cultural gesture understanding systems.

These identified research gaps motivated the development of the proposed explainable multimodal dual-stage transfer learning framework presented in this thesis.

## 1.5 Objectives

The major objectives of the proposed work are as follows:

1. To develop an explainable multimodal Bharatanatyam mudra recognition framework integrating visual gesture representations and geometric hand landmark features,
2. To implement a dual-stage transfer learning strategy for improving generalized gesture representation learning and cross-dataset adaptability,
3. To incorporate explainable artificial intelligence techniques for interpreting model predictions and analyzing semantically meaningful hand articulation regions.

## 1.6 Research Contributions

The major contributions of the proposed research are summarized as follows:

- Development of an explainable multimodal Bharatanatyam mudra recognition framework integrating Swin Transformer-based visual learning and MediaPipe-based geometric hand landmark representations.
- Introduction of a dual-stage transfer learning strategy involving generalized gesture pretraining using Indian Sign Language data followed by Bharatanatyam-specific fine-tuning and cross-dataset evaluation.
- Incorporation of explainable artificial intelligence techniques including GradCAM and landmark sensitivity analysis for interpreting semantically meaningful gesture regions and supporting intelligent cultural heritage preservation.

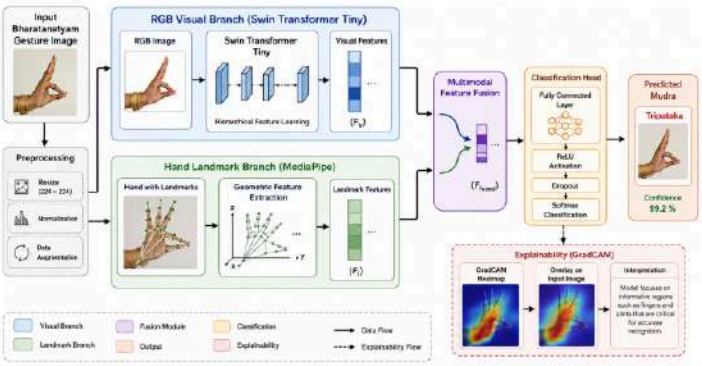



**Figure 1.3:** Overall architecture of the proposed multimodal dual-stage transfer learning framework

## 1.7 Organization of the Thesis

The remaining of this presented thesis is organized as given below:

- Chapter 2 presenting the literature review related to gesture recognition, transfer learning, transformer architectures, hand landmark analysis, and explainable artificial intelligence techniques,
- Chapter 3 describing the proposed methodology, including the tasks of dataset preparation, preprocessing, landmark extraction, multimodal fusion architecture, dual-stage transfer learning, and explainability framework.
- Chapter 4 discusses the experimental setup, implementation details, training configuration, evaluation metrics, and hardware specifications,
- Chapter 5 presents the experimental results, comparative analysis, explainability evaluation, robustness analysis, and cross-domain performance assessment, and
- Chapter 6 concluding the thesis by summarising the major findings, contributions, limitations, and possible future research directions.

# CHAPTER 2

# LITERATURE REVIEW

The literature surrounding automated Indian Classical Dance recognition and particularly Bharatanatyam mudra classification, reveals a significant evolution from the traditional approaches of machine learning towards advanced kind of deep learning and multimodal frameworks where existing research demonstrates continuous progress in visual representation learning, geometric hand analysis, transfer learning methodologies, and explainable artificial intelligence techniques for gesture understanding and despite these advancements, several important research gaps remain unresolved, particularly in terms of robustness, interpretability, and cross-domain generalization.

## 2.1 Evolution from using Hand-Crafted Features to techniques of Deep Learning

Early approaches to Bharatanatyam mudra recognition as shown in Fig 2.1 primarily relied on tradditional image processing and hancrafted feature extraction techniques in which researchers made use of the descriptors such as Histogram of Oriented Gradients (HOG)[7], Scale-Invariant Feature Transform (SIFT)[8], Sped-Up Robust Features (SURF)[9], Hu Moments, and normalized chain codes to represent the hand gestures mathematically and thus extracted features were subsequently classified with the employment of usually used machine learning algorithms including Support Vector Machines (SVM)[10], K-Nearest Neighbors (KNN)[11], Naive Bayes[12], and Random Forest[13] classifiers and although these approaches provided an initial understanding of hand orientation and shape representation, they suffered from several limitations because of the fact that handcrafted features were highly sensitive to illumination variations, background complexity, hand orientation changes, and performer-specific articulation differences and to add more these methods were lacking the ability to capture fine-grained hierarchical spatial relationships between finger joints and hand configurations[14].

In order to address these limitations, researchers hae gradually transitioned towardss deep learning methodologies, particularly Convolutional Neural Networks (CNNs) which automatically learn hierarchical visual representations directly from raw image data[15], eliminating the very need for a manually featureengineering where these Deep learning frameworks demonstrated substantially improved classification performance and robustness compared to traditional handcrafted approaches, partcularly in complex gesture recognition tasks involving subtle structural differences between the mudras[16].

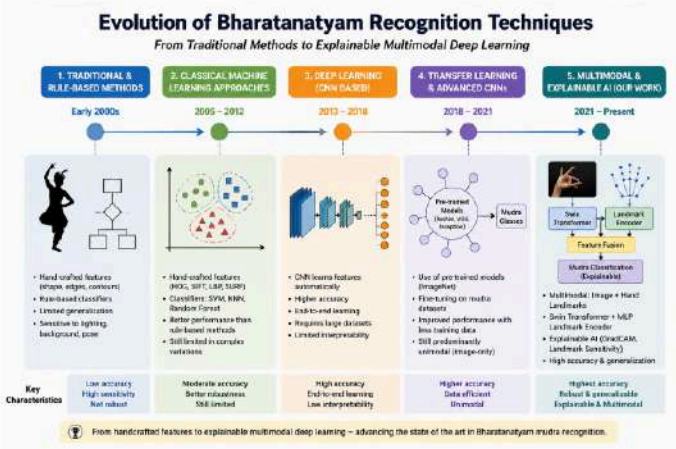



**Figure 2.1:** Evolution of gesture recognition approaches from traditional handcrafted features to deep learning frameworks

## 2.2 Dataset Challenges and Studio Bias

One among the most noticeable limitations identified all throughout the literature is that of the dearth in the case of large-scale, publicly available Bharatanatyam mudra datasets where most of the existing studies have shown the relying on small proprietary datasets collected in the highly controlled studio environments characterized by the presence of the uniform lighting, the plain backgrounds, and also the constrained movement of the dance performer.

Such controlled conditions introduce a phenomenon commonly referred to as *studio bias* and often these models which are trained exclusively on studio-based datasets are prone to fail in generalizing effectively when deployed in real-world stage environments containing cluttered backgrounds, complex costumes, dynamic illumination, overlapping performers, and varying camera viewpoints and so consequently, many existing systems demonstrate high accuracy under laboratory conditions but exhibit poor robustness in unconstrained practical scenarios and in order to mitigate these limitations, recent research stresses on the importance of the use of extensive data augmentation, domain randomization, and cross-domain evaluation strategies so that this simulation of realistic stage conditions through augmentation techniques can improve the generalization capability of the models by reducing overfitting to controlled datasets. These are summarized in Table 2.1.

**Table 2.1:** Common limitations of existing Bharatanatyam mudra datasets

| Limitation | Impact on Recognition Performance |
|---|---|
| Small dataset size | Reduced generalization capability |
| Controlled studio backgrounds | Poor robustness in real-world environments |
| Limited performer diversity | Increased subject-specific bias |
| Uniform lighting conditions | Sensitivity to illumination variations |
| Lack of cross-domain evaluation | Limited adaptability across datasets |

## 2.3 Geometric Landmarking and Spatial Features

To improve robustness against background complexity and visual occlusions, recent studies have explored geometric landmark-based representations for gesture recognition and frameworks such as Google MediaPipe[17] and OpenPose[18] enable real-time extraction of skeletal hand landmarks representing finger joints and palm structures as shown in Fig. 2.2 which provide a geometric abstraction of the hand independent of texture, background, and illumination conditions that helped the researchers to utilize spatial relationships between hand joints, including Euclidean distances, joint angles, and finger articulation patterns, to create structural representations suitable for gesture classification tasks.

Landmark-based approaches offer improved invariance to lighting and environmental variations compared to raw RGB image representations but then however, landmark-only frameworks may also end up losing on the important appearance-based contextual information such as texture, finger contours, and subtle visual semantics, so as a consequence the recent literature is increasingly advocating the integration of visual and geometric representations through multimodal learning frameworks.


**Figure 2.2:** Example of skeletal hand landmark representation using MediaPipe

## 2.4 Transfer Learning and Double Transfer Learning

Pertaining to the limited size of Bharatanatyam mudra datasets, transfer learning has become a widely adopted strategy in gesture recognition research so that instead of training deep learning models from scratch, the researchers are in a position to utilize the architectures that are already pretrained usually on the datasets of large-scale such as ImageNet[19] and subsequently fine-tune them for gesture classification tasks and accordingly a good number of these pretrained architectures including VGG16, ResNet, EfficientNet, and MobileNet have demonstrated strong performance in hand gesture recognition applications and recently some researchers introduced specialized methodologies such as Double Transfer Learning (DTL) to further improve recognition accuracy and representation learning[6].

In Double Transfer Learning, a model pretrained on a large generic dataset undergoes an intermediate training stage using a generalized hand gesture or sign language dataset before final fine-tuning on Bharatanatyam mudras and so this progressive transfer learning strategy enables the network to learn both generalized visual patterns and domain-specific hand articulation structures supported by existing studies reporting that DTL significantly improves classification performance compared to conventional single-stage transfer learning approaches by enhancing feature generalization and reducing domain adaptation difficulty and so this methodology as depicted in Fig 2.3 is particularly effective for culturally specific gesture recognition tasks with not enough training data[6].

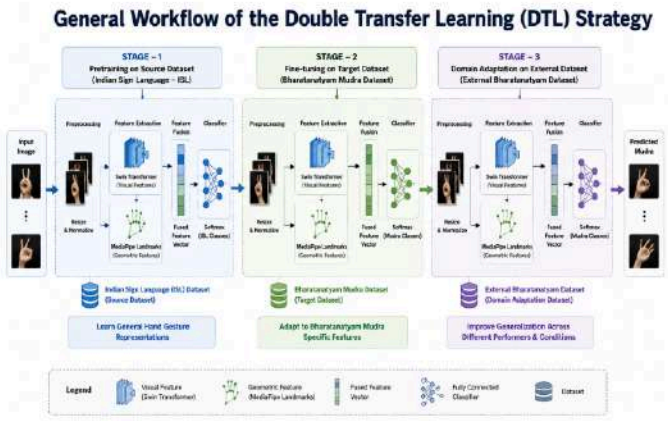



**Figure 2.3:** General workflow of the Double Transfer Learning (DTL) strategy

## 2.5 Explainable Artificial Intelligence in Gesture Recognition

Eevn though the deep learning models achieve remarkable classification performance, they often function as opaque blackbox systems that create difficulty to determine whether the model is learning meaningful gesture semantics or merely memorizing irrelevant background artifacts and this lack of transparency presents significant concerns in cultural heritage applications where interpretability and trustworthiness are essential which prompted reseacrh towards addressing these issues, whererecent literature increasingly incorporates Explainable Artificial Intelligence (XAI) techniques into gesture recognition frameworks and methods such as SHAP (SHapley Additive exPlanations) and Gradient-weighted Class Activation Mapping (GradCAM) as shown in Fig 2.4 generate visual explanations highlighting image regions that are significant in the contribution to that of the model predictions[20].

In Bharatanatyam mudra recognition, explainability techniques help identify critical finger articulations, hand contours, and anatomical regions influencing classification decisions. Such visual interpretations improve transparency and provide insight into the semantic reasoning process learned by deep learning models.

Recent studies suggest that integrating explainable visual representations with geometric landmark analysis and multimodal feature fusion leads to more robust and interpretable gesture recognition systems suitable for real-world deployment.


**Figure 2.4:** GradCAM-based explainability visualization highlighting discriminative gesture regions

**Table 2.2:** Literature Review on Indian Classical Dance Classification

| Author & Year | Methodology | Dataset | Strengths | Limitations |
|---|---|---|---|---|
| K. Thanikasalam et al. (2026)[14] | Ensemble of three EfficientNetV2S models using raw, skeleton, and embedded-landmark images | AHM-UoJ and Jisha Raj Bharatanatyam Mudra Dataset | Achieved high classification accuracy (98.28%) through multimodal feature integration | Computationally expensive due to ensemble learning and dependence on handcrafted feature extraction |
| C. Sarmah and P. Sarma (2024)[16] | CNN and ResNet-50 models using Watershed segmentation | Sattriya-08 Double Handed Mudra Dataset[16] | Effectively captures double-handed gestures with high training accuracy | Small dataset size and segmentation sensitivity under complex backgrounds |
| K. Adalarasu et al. (2025)[20] | Machine learning models using VGRF data with SHAP-based explainability | VGRF force platform dataset containing six dance postures | Provides interpretability and avoids optical occlusion problems | Limited to static poses and affected by class imbalance |
| S. Paul et al. (2025)[21] | YOLOv6-based hand detection with ResNet18 classification and flexion angle descriptors | Oxford Hand Dataset and Jisha Raj Bharatanatyam Dataset | Robust against rotation and scaling variations with real-time implementation capability | Depth estimation inaccuracies and dependency on performer precision |
| J. R. Challapalli et al. (2026)[22] | CNN architectures optimized using Flamingo Search metaheuristic algorithm | ICD dataset augmented using GANs along with CIFAR-10/100 | Efficient hyperparameter optimization and improved computational scalability | Limited generalization due to visually homogeneous datasets |
| S. S. and J. M.V. (2022)[23] | Deep Pose Estimator with GRU, 3D-CNN, and CNN-LSTM architectures | 300 HD dance video clips representing seven classes | Efficient handling of spatio-temporal dynamics for real-time inference | Requires high computational resources and GPU-intensive training |
| A. P. Parameshwaran et al. (2020)[5] | Double Transfer Learning using VGG16 and stacked ensemble models | Custom dataset containing 27 single-hand gestures | Effectively addresses data scarcity through progressive transfer learning | Limited robustness due to controlled environment dataset collection |

| Author & Year | Methodology | Dataset | Strengths | Limitations |
|---|---|---|---|---|
| S. Gupta and S. Singh (2024)[24] | Comparative survey of probabilistic, rule-based, geometric, and neural network models | Multiple gesture and dance datasets | Provides extensive comparative analysis of gesture recognition approaches | Limited experimental validation and dependence on existing literature |

## 2.6 Summary

This chapter is all about the reviewing of existing literature that is related to Bharatanatyam mudra recognition, gesture classification systems, deep learning architectures, transfer learning strategies, geometric landmark representations, and explainable artificial intelligence techniques. The review highlighted the transition from handcrafted feature extraction approaches that are traditional and conventional to modern multimodal deep learning frameworks while identifying major research limitations related to robustness, explainability, and cross-domain adaptability. Based on these observations, the next chapter presents the proposed explainable multimodal dual-stage transfer learning framework designed to address the identified research gaps and improve Bharatanatyam mudra recognition performance.

# CHAPTER 3

# PROPOSED WORK

## 3.1 Introduction

This chapter presents the proposed explainable multimodal dualstage transfer learning framework developed for robust Bharatanatyam mudra recognition. The proposed methodology integrates transformer-based visual learning with geometric hand landmark representations to improve classification performance, interpretability, and cross-dataset generalization capability. The framework combines Swin Transformer Tiny[25] for RGB image feature extraction and MediaPipe-based normalized hand landmark analysis[17] for capturing structural articulation patterns of Bharatanatyam mudras.

To address the limited availability of large-scale Bharatanatyam mudra datasets and improve transferability, a dual-stage transsfer learning strategy is employed. The proposed framework first learns generalized gesture representations from an Indian Sign Language dataset before fine-tuning on Bharatanatyam mudra data. Furthermore, cross-dataset evaluation and domain adaptation experiments are conducted using an external Bharatanatyam mudra dataset[14] containing the same mudras performed by different subjects under varying acquisition conditions.

Adding to the classification performance, the proposed framework as shown in Fig 3.1 incorporates explainable artificial intelligence techniques including GradCAM-based visual explanation and landmark sensitivity analysis for interpreting the decision-making behavior of the multimodal architecture[26].

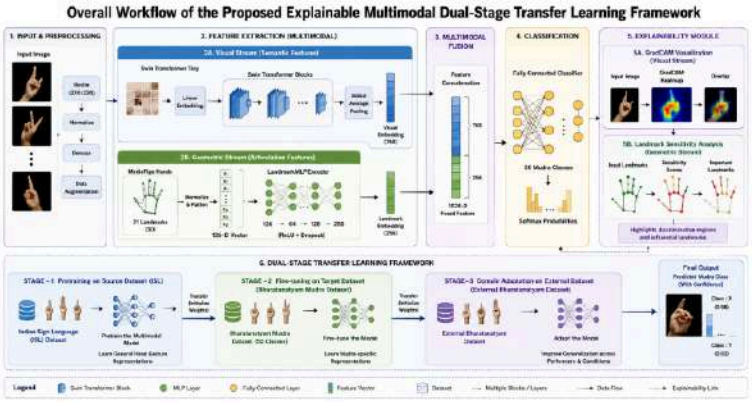



**Figure 3.1:** Overall workflow of the proposed explainable multimodal dual-stage transfer learning framework

## 3.2 Overall Framework

The proposed framework follows a multimodal architecture combining appearance-based visual features and geometric hand articulation features for Bharatanatyam mudra recognition. The full workflow is comprising of the following stages:

1. Dataset preparation
2. Image preprocessing and augmentation
3. Hand landmark extraction using MediaPipe
4. Visual feature extraction using Swin Transformer Tiny
5. Geometric feature encoding using LandmarkMLPEncoder
6. Multimodal feature fusion
7. Gesture classification
8. Explainability analysis

Initially, input Bharatanatyam gesture images are subjected to the preprocessing operations including resizing, normalization, and augmentation. The processed images are then simultaneously forwarded to two parallel branches:

- RGB image branch
- Landmark branch

The RGB branch extracts deep visual representations using Swin Transformer Tiny, while the landmark branch processes normalized three-dimensional hand keypoint coordinates extracted using MediaPipe Hands. The extracted visual and geometric features are fused into a unified multimodal representation, which is subsequently passed through classification layers for final mudra prediction.

The framework additionally incorporates explainability modules to visualize discriminative image regions and identify important hand joints contributing to classification decisions.

## 3.3 Dataset Description

Three datasets as shown in Table 3.1are utilized in the proposed study for transfer learning, Bharatanatyam mudra recognition, and cross-dataset evaluation.

### 3.3.1 Indian Sign Language Dataset

The intermediate transfer learning stage uses an Indian Sign Language (ISL) dataset[27] obtained from Kaggle which contains multiple hand gesture classes representing different ISL signs and is used for learning generalized gesture representations before we do the Bharatanatyam mudra specific fine-tuning using our Bharatantyam Mudra Dataset.

The ISL dataset helps the model to learn the following:

- generalized hand articulation patterns,

- finger configuration structures,
- gesture semantics, and
- visual hand representations.

### 3.3.2 Primary Bharatanatyam Mudra Dataset

The primary Bharatanatyam mudra dataset[28] is used for fine-tuning. This was also used for classification. This was collected as part of doctoral research conducted under the guidance of Dr. Sunil T.T., College of Engineering Attingal, Kerala, India[28]. The dataset contains multiple Bharatanatyam mudra classes represented using static hand gesture images captured under controlled conditions.

The dataset includes variations in:

- hand articulation,
- performer orientation,
- gesture appearance, and
- illumination conditions

### 3.3.3 External Bharatanatyam Mudra Dataset

To evaluate the robustness on cross-dataset and domain adaptability for the same cross-dataset, an external Bharatanatyam mudra dataset[14] is used. This was proposed by Kokul Thanikasalam et al. The external dataset contains the same mudra classes. These are but performed by different subjects under different acquisition conditions.

Unlike the generic gesture transfer evaluation, this cross-dataset analysis performed in our study investigates:

- performer variation,
- acquisition condition variation,
- dataset distribution shift, and
- cross-dataset generalization capability

The external dataset enables evaluation of the proposed framework under non-identical training and testing distributions.

**Table 3.1:** Datasets used in the proposed study

| **Dataset** | **Data Type** | **Purpose** | **Role in Framework** |
|---|---|---|---|
| Indian Sign Language Dataset | Hand gesture images | Intermediate transfer learning | Learning generalized gesture representations |
| Primary Bharatanatyam Mudra Dataset | Mudra images | Final fine-tuning and classification | Primary mudra recognition |
| External Bharatanatyam Mudra Dataset | Mudra images | Cross-dataset evaluation | Robustness and domain adaptation analysis |

## 3.4 Data Preprocessing

Data preprocessing is performed to improve training stability and enhance model generalization capability. The preprocessing pipeline consists of multiple operations applied uniformly across all datasets as shown in Fig 3.2.

The preprocessing operations include:

- Image resizing
- Pixel normalization
- Data augmentation
- Background standardization

All of the imput images are subjected to undergo the resizing to a fixed resolution that is compatible and works well with the Swin Transformer Tiny architecture. Pixel normalization is applied to standardize intensity distributions across samples.

To reduce overfitting and improve robustness against environmental variations, several augmentation techniques are employed, including:

- rotation,
- horizontal flipping,
- translation,
- zoom transformation, and
- brightness adjustment.

The augmentation process enables the model to learn invariant gesture representations under varying acquisition conditions.

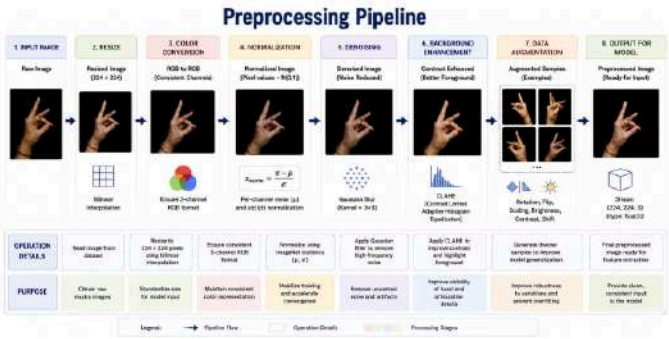



**Figure 3.2:** Image preprocessing and augmentation pipeline

## 3.5 Hand Landmark Extraction

In order to get the information features for geometric hand articulation, MediaPipe Hands is used for the extraction skeletal hand landmarks from input images.

MediaPipe provides 21 hand landmarks. These are then represented using normalized three-dimensional coordinates:

$$(x, y, z)$$

where:

- $x$ and $y$ represent normalized spatial coordinates, and
- $z$ represents relative depth information.

The landmark extraction process consists of:

1. Hand detection,
2. Landmark localization,
3. Coordinate normalization,
4. Feature vector generation.

Each detected hand produces:

$$21 \times 3 = 63$$

features corresponding to the 21 landmarks and their associated three-dimensional coordinates as shown in Fig 3.3.

For two-hand detection scenarios, the coordinates are concatenated to form a fixed-length feature vector of size:

For suppose only one hand is detected, in that case zero-padding is applied so that to preserve dimensional consistency across samples.

The resulting normalized coordinate vector is forwarded to the step/stage of the LandmarkMLPEncoder of our architecture to perform geometric representation learning.

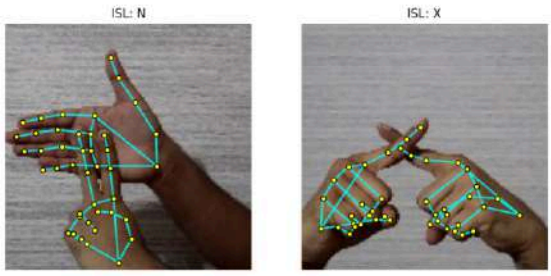



**Figure 3.3:** MediaPipe hand landmark extraction showing 21 skeletal keypoints

## 3.6 Visual Feature Extraction Using Swin Transformer Tiny

The RGB image branch utilizes Swin Transformer Tiny as the primary visual feature extraction backbone, and as discussed earlier this Swin Transformer is a vision transformer that has a hierachical architecture that employs self-attention mechanisms of shifted-window kind for an efficient local as well as that of the global representation learning. Compared to conventional convolutional neural networks, Swin Transformer as shown in Fig 3.4 provides improved capability for modeling long-range dependencies and subtle spatial relationships within gesture images and so this property is particularly important for Bharatanatyam mudra recognition, where fine-grained finger articulation differences significantly influence gesture semantics.

The Swin Transformer Tiny architecture performs:

- patch embedding,
- hierarchical feature extraction,
- window-based self-attention,
- shifted-window attention,and
- feature aggregation.

The extracted visual features capture:

- texture information,
- contour structures,

- finger configurations,and
- visual semantic patterns.

The generated visual embeddings are subsequently passed to the multimodal fusion stage.

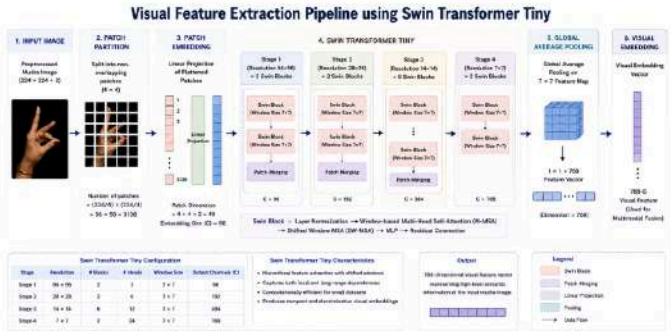



**Figure 3.4:** Visual feature extraction pipeline using Swin Transformer Tiny

## 3.7 LandmarkMLPEncoder

The landmark branch processes normalized geometric landmark coordinates using a multilayer perceptron-based LandmarkMLPEncoder.

The encoder receives the fixed-length 126-dimensional landmark vector generated from MediaPipe extraction and learns compact geometric representations corresponding to hand articulation structures as shown in Fig 3.5.

The LandmarkMLPEncoder enables the framework to learn:

- finger joint relationships,
- spatial articulation patterns,
- geometric gesture structures, and
- relative joint configurations.

The learned geometric embeddings complemented the visual representations extracted by the Swin Transformer branch.

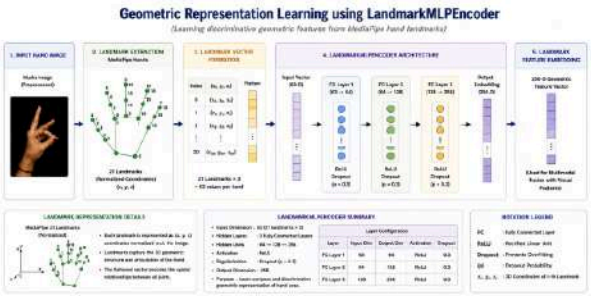



**Figure 3.5:** Geometric representation learning using LandmarkMLPEncoder

## 3.8 Dual-Stage Transfer Learning

To address limited Bharatanatyam dataset availability and improve representation learning, a dual-stage transfer learning strategy as in Fig 3.6 is employed.

Instead of directly fine-tuning an ImageNet-pretrained model on Bharatanatyam mudras, the proposed framework introduces an intermediate gesture adaptation stage using Indian Sign Language data.

The dual-stage transfer learning process consists of:

1. Initial pretraining on ImageNet,
2. Intermediate transfer learning on Indian Sign Language gestures,and
3. Final fine-tuning on Bharatanatyam mudras.

This progressive adaptation strategy enables the model to first learn generalized hand gesture representations before specializing in Bharatanatyam mudra recognition.

The DTL strategy improves:

- feature transferability,
- training stability,
- convergence capability,and
- cross-dataset adaptability.

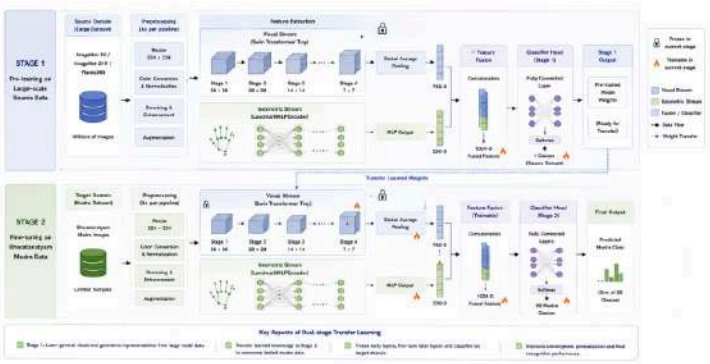


**Figure 3.6:** Dual-stage transfer learning workflow used in the proposed framework

## 3.9 Multimodal Feature Fusion

The proposed framework integrates visual embeddings extracted from Swin Transformer Tiny with geometric embeddings learned using LandmarkMLPEncoder as shown in the Fig 3.7.

This multimodal fusion strategy combines complementary information from:

- appearance-based visual representations
- geometric articulation representations

Visual features capture:

- texture
- contour
- visual semantics

while landmark embeddings capture:

- joint relationships
- spatial articulation
- geometric hand structure

The fused multimodal representation improves robustness against:

- illumination variation
- performer variation
- acquisition differences
- background complexity

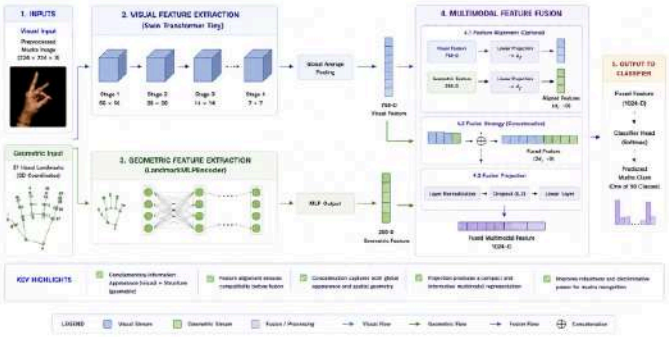



**Figure 3.7:** Multimodal feature fusion between visual and geometric representations

## 3.10 Gesture Classification

The fused multimodal features are processed through fullyconnected classification layers as in Fig 3.8 for final Bharatanatyam mudra prediction.

A softmax activation function is employed to generate probability distributions across all mudra classes.

The framework's optimization is done with categorical cross-entropy loss and gradient-based optimization techniques during training.

The classification stage predicts the most probable Bharatanatyam mudra corresponding to the input gesture image.

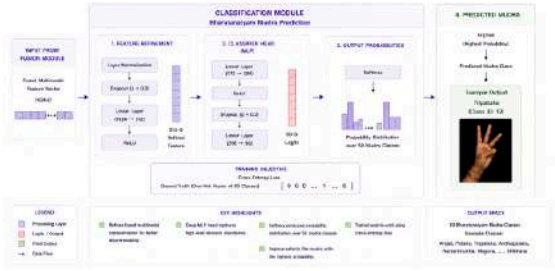



**Figure 3.8:** Classification module for Bharatanatyam mudra prediction

## 3.11 Cross-Dataset Evaluation and Domain Adaptation

To evaluate robustness and generalization capability, cross-dataset experiments are conducted using the external Bharatanatyam mudra dataset.

Two evaluation strategies are employed:

### 3.11.1 Zero-Shot Cross-Dataset Evaluation

The Bharatanatyam-trained model is directly evaluated on the external dataset without additional fine-tuning. This experiment evaluates the inherent cross-dataset generalization capability of the proposed framework under performer and acquisition variation.

### 3.11.2 Domain Adaptation Evaluation

The pretrained Bharatanatyam model is further fine-tuned using the training split of the external dataset and subsequently evaluated on the unseen testing split as in Fig 3.9.

This experiment evaluates the adaptability of the proposed framework under dataset distribution shift conditions.

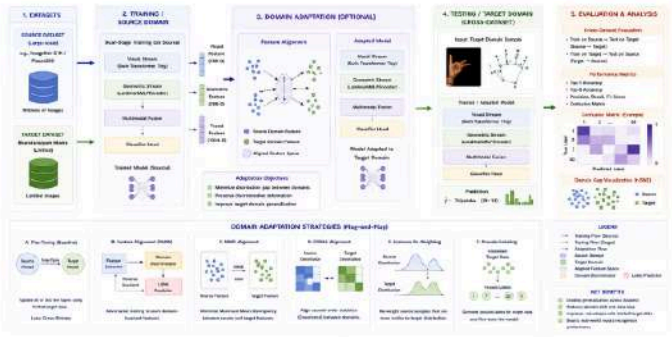


**Figure 3.9:** Cross-dataset evaluation and domain adaptation workflow

## 3.12 Explainability Analysis

To improve transparency and interpretability, the proposed framework incorporates dual-modal explainability analysis.

### 3.12.1 GradCAM-Based Visual Explainability

GradCAM is employed to visualize discriminative image regions contributing significantly to model predictions within the Swin Transformer branch as shown in Fig 3.10.

The generated attention heatmaps help identify:

- important finger regions
- discriminative contours
- gesture-specific visual semantics

### 3.12.2 Landmark Sensitivity Analysis

In addition to GradCAM, landmark sensitivity analysis as in Fig 3.11 is performed using gradient magnitude analysis on the landmark branch.

This analysis identifies:

- influential hand joints
- important articulation regions
- geometric structures affecting classification

The landmark importance patterns are further analyzed with reference to traditional Bharatanatyam Shastra-based gesture semantics.

**Figure 3.10:** GradCAM-based attention visualization highlighting discriminative gesture regions

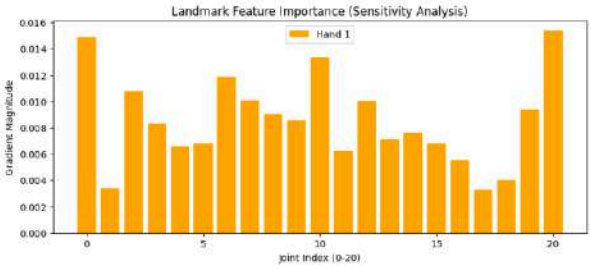



**Figure 3.11:** Landmark sensitivity analysis showing important hand joints influencing classification

## 3.13 Summary

This chapter presented the proposed explainable multimodal dual-stage transfer learning framework developed for Bharatanatyam mudra recognition. The methodology integrates transformer-based visual

learning, geometric landmark representation learning, progressive transfer learning, multimodal feature fusion, cross-dataset evaluation, and explainable artificial intelligence techniques to improve recognition robustness, adaptability, and interpretability.

The next chapter deals on the experimental setup, implementation details, training configuration, and evaluation metrics used for validating the proposed framework.

# CHAPTER 4

# EXPERIMENTAL SETUP

## 4.1 Introduction

This chapter details on the experimental setup and implementation details employed for evaluating the proposed explainable multimodal dual-stage transfer learning framework for Bharatanatyam mudra recognition. The chapter describes the computational environment, dataset preparation, preprocessing pipeline, model architecture, training protocols, evaluation methodologies, and explainability analysis techniques used throughout the study.

The experiments were designed to evaluate:

- Bharatanatyam mudra classification performance
- Cross-dataset generalization capability
- Domain adaptation effectiveness
- Multimodal representation learning
- Explainability and feature interpretability

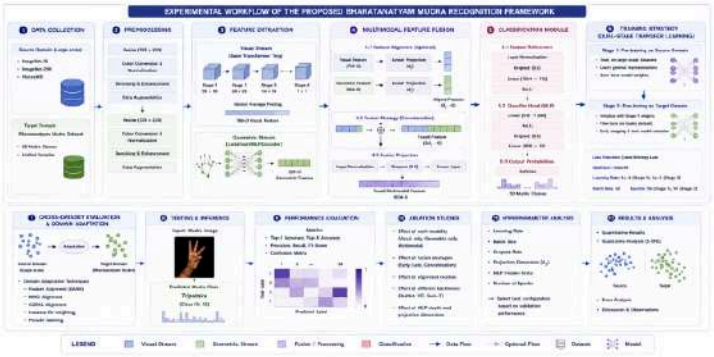



**Figure 4.1:** Experimental workflow of the proposed Bharatanatyam mudra recognition framework

## 4.2 Hardware and Software Environment

All of the experiments were run using the Google Colab environment with GPU acceleration support.

### 4.2.1 Hardware Configuration

The hardware specifications used for model training and evaluation are summarized in Table 4.1.

**Table 4.1:** Hardware configuration used for experimentation

| Component | Specification |
|---|---|
| Platform | Google Colab |
| GPU | NVIDIA T4 GPU |
| GPU Memory | 16 GB |
| Operating System | Linux-based cloud environment |

### 4.2.2 Software Environment

The proposed framework was implemented using Python and the PyTorch deep learning ecosystem. Several computer vision, explainability, and machine learning libraries were utilized throughout the experiments.

The software environment is summarized in Table 4.2.

**Table 4.2:** Software environment and libraries

| Software Component | Version / Library |
|---|---|
| Programming Language | Python 3.9 |
| Deep Learning Framework | PyTorch 2.0.1 + CUDA 11.8 |
| Vision Libraries | Torchvision, OpenCV |
| Transformer Libraries | Transformers, TIMM |
| Landmark Extraction | MediaPipe |
| Machine Learning Utilities | Scikit-learn |
| Visualization Libraries | Matplotlib, Plotly, Seaborn |
| Explainability Libraries | Grad-CAM |
| Feature Analysis Libraries | UMAP-learn, Scikit-image |

## 4.3 Datasets and Data Preparation

Three datasets as shown in Fig 4.3 were utilized in the proposed study for transfer learning, Bharatanatyam mudra classification, and cross-dataset evaluation.

### 4.3.1 Stage-1 Transfer Learning Dataset

The first stage of transfer learning utilized an Indian Sign Language (ISL) dataset obtained from Kaggle. The dataset contains 36 gesture classes representing different ISL hand signs.

The ISL dataset was used to learn generalized hand gesture representations before Bharatanatyam-specific fine-tuning.

The dataset split strategy included:

- 90% training split
- 10% validation split
- Separate held-out test set

The split was performed using `torch.utils.data.random_split()`.

### 4.3.2 Primary Bharatanatyam Mudra Dataset

The primary Bharatanatyam mudra dataset was obtained from the Hugging Face Hub. The dataset consists of 50 Bharatanatyam mudra classes represented using static hand gesture images.

The dataset split strategy was:

- 70% training set
- 15% validation set
- 15% testing set

The split was performed using `Dataset.train_test_split()` with:

```
seed = 42
```

The dataset splitting procedure preserved class distribution across splits through stratified partitioning.

### 4.3.3 External Bharatanatyam Mudra Dataset

To evaluate cross-dataset generalization and domain adaptation capability, an external Bharatanatyam mudra dataset[14] proposed by Kokul Thanikasalam et al. was utilized.

The external dataset contains:

- identical mudra classes
- different performers
- varying acquisition conditions
- different image distributions

This dataset was used for:

- zero-shot cross-dataset evaluation
- domain adaptation experiments

Table 4.3: Summary of datasets used in the proposed study

| Dataset | Classes | Dataset Size | Data Type | Purpose |
|---|---|---|---|---|
| Indian Sign Language Dataset | 36 | 42.7K | Gesture images | Stage-1 transfer learning |
| Bharatanatyam Mudra Dataset | 50 | 28,431 | Mudra images | Primary classification |
| External Bharatanatyam Mudra Dataset | 27 | 3,450 | Mudra images | Cross-dataset evaluation |

## 4.4 Image Preprocessing and Augmentation

All input images were resized to:

$$224 \times 224$$

pixels to match the input requirements of the Swin Transformer Tiny architecture.

### 4.4.1 Training-Time Augmentation

To improve robustness and reduce overfitting, multiple augmentation techniques were applied during training as shown in Fig 4.2:

- Random resizing and cropping
- Horizontal flipping
- Color jitter
- Image normalization

Normalization was performed using ImageNet mean and standard deviation values.

### 4.4.2 Validation and Testing Preprocessing

For validation and testing:

- Images were resized to 256 pixels
- Center cropped to 224 pixels
- Normalized using ImageNet statistics

**Figure 4.2:** Image preprocessing and augmentation examples

## 4.5 Hand Landmark Extraction

Hand landmarks were extracted using the MediaPipe HandLandmarker framework as shown in Fig 4.3.

Each detected hand produced:

$$21$$

skeletal landmarks represented using normalized three-dimensional coordinates:

$$(x, y, z)$$

The landmark coordinates were normalized relative to image dimensions.

For each hand:

$$21 \times 3 = 63$$

features were generated.

For two-hand detection:

$$63 \times 2 = 126$$

features were produced.

If only one hand was detected, zero-padding was applied to preserve fixed-dimensional input representation.

Approximately:

$$18\%$$

of Bharatanatyam training samples resulted in failed landmark detection. These cases were handled using zero-vector landmark representations.

The extracted landmark vectors were directly forwarded to the LandmarkMLPEncoder without additional z-score normalization or feature standardization.

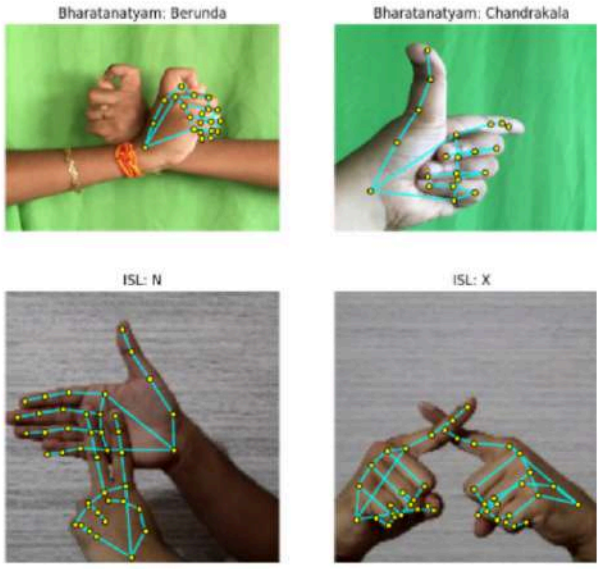



**Figure 4.3:** MediaPipe-based hand landmark extraction process

## 4.6 Model Architecture

The proposed multimodal framework integrates visual and geometric representation learning through two parallel branches as shown in Fig 4.4.

### 4.6.1 Visual Branch

The visual branch utilizes:

`swin_tiny_patch4_window7_224`

initialized using the pretrained Microsoft Swin Transformer[25] Tiny checkpoint.

The Swin Transformer backbone remained fully trainable during all training stages.

The backbone outputs:

768

dimensional visual embeddings obtained after global adaptive average pooling.

### 4.6.2 Landmark Branch

The geometric landmark branch utilizes a lightweight multilayer perceptron called `LandmarkMLPEncoder`.

The architecture follows:

$$126 \rightarrow 64 \rightarrow 128 \rightarrow 256$$

The encoder includes:

- Linear layers,
- ReLU activation, and
- Dropout with probability 0.2

The final output embedding dimension of the landmark branch is:

$$256$$

### 4.6.3 Feature Fusion

The visual and geometric embeddings are concatenated to form a fused multimodal representation:

$$768 + 256 = 1024$$

dimensional fused feature vector.

### 4.6.4 Classifier Head

The fused multimodal representation is directly mapped to the output classes using a single fully connected classification layer:

$$1024 \rightarrow \text{Number of Classes}$$

The final output dimension depends on the active training stage:

- 36 classes for ISL pretraining
- 50 classes for Bharatanatyam mudra recognition

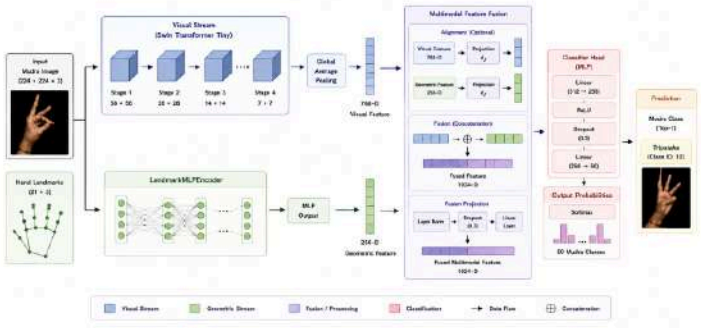


**Figure 4.4:** Architecture of the proposed multimodal Bharatanatyam mudra recognition framework

## 4.7 Training Protocols

### 4.7.1 General Training Configuration

The common training configuration used across all experiments is summarized in Table 4.4.

**Table 4.4:** General training configuration

| Parameter | Value |
|---|---|
| Batch Size | 32 |
| Optimizer | AdamW |
| Loss Function | CrossEntropyLoss |
| Learning Rate Scheduler | CosineAnnealingWarmRestarts |
| Mixed Precision Training | torch.cuda.amp |
| Model Selection Criterion | Highest validation accuracy |
| Gradient Clipping | Not applied |

### 4.7.2 Stage-1 ISL Pretraining

The first training stage focused on generalized gesture representation learning using the ISL dataset.

The configuration included:

- Epochs: 5
- Learning Rate: $1 \times 10^{-4}$
- Weight Decay: 0.01

### 4.7.3 Stage-2 Bharatanatyam Fine-Tuning

The second stage adapted the pretrained model to Bharatanatyam mudra recognition.

The configuration included:

- Epochs: 10
- Learning Rate: $5 \times 10^{-5}$
- Weight Decay: 0.01

Early stopping was employed using validation accuracy monitoring to reduce overfitting.

### 4.7.4 Stage-3 Domain Adaptation

The final stage performed domain adaptation using the external Bharatanatyam mudra dataset as shown in Fig 4.5.

The configuration included:

- Epochs: 5
- Learning Rate: $1 \times 10^{-5}$
- Weight Decay: 0.05

A conservative learning rate was used to preserve previously learned Bharatanatyam representations while adapting to external dataset variations.

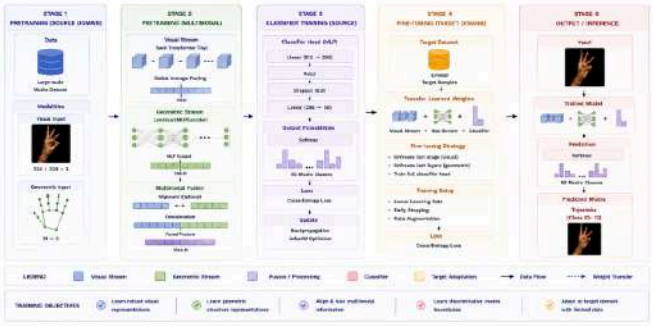


**Figure 4.5:** Multi-stage training pipeline of the proposed framework

## 4.8 Evaluation Metrics

Multiple quantitative metrics were used to evaluate classification performance and feature representation quality.

### 4.8.1 Classification Metrics

The primary evaluation metrics included:

- Accuracy,
- Precision,
- Recall,
- F1-score, and
- Cohen's Kappa Score

Confusion matrices were additionally generated to analyze inter-class misclassification patterns.

### 4.8.2 Cross-Dataset Evaluation

Cross-dataset robustness was evaluated using:

- Zero-shot evaluation
- Domain adaptation evaluation

The evaluation results included:

- Main Bharatanatyam classification accuracy:

  99.31%

- Zero-shot cross-dataset accuracy:

  66.90%

- Domain-adapted cross-dataset accuracy:

  86.11%

### 4.8.3 Feature Space Analysis

To evaluate latent feature separability, clustering metrics were computed on the fused feature embeddings:

- Silhouette Score[43],
- Calinski-Harabasz Index[44],and
- Davies-Bouldin Index[45].

Additionally, t-SNE visualization was employed to analyze feature-space clustering behavior as shown in Fig 4.6 [39].

### 4.8.4 Structural Similarity Analysis

Structural Similarity Index (SSIM)[46] was used to quantify inter-class visual similarity and analyze potential causes of classification confusion.

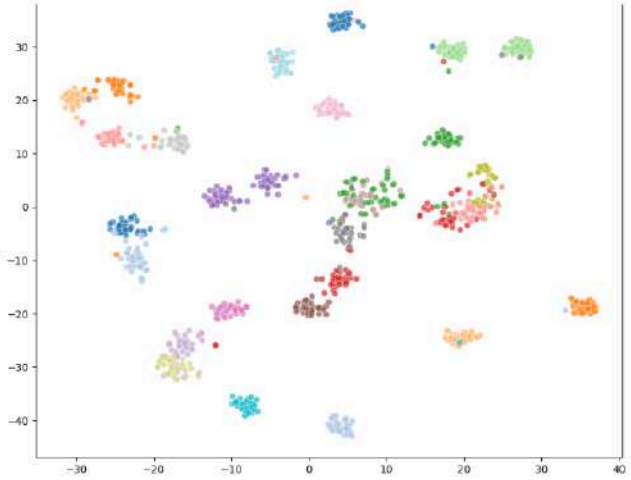



**Figure 4.6:** t-SNE visualization of fused multimodal feature representations

## 4.9 Explainability Analysis

To improve transparency and interpretability, two explainability mechanisms were employed.

### 4.9.1 GradCAM-Based Visual Explainability

GradCAM was applied to the Swin Transformer branch to visualize discriminative image regions contributing significantly to model predictions.

The generated heatmaps highlighted:

- important finger articulations,
- hand contours,and
- gesture-specific visual regions.

### 4.9.2 Landmark Sensitivity Analysis

Gradient-based landmark sensitivity analysis was performed on the landmark branch to identify influential hand joints affecting classification decisions, and the resulting importance patterns were interpreted using Bharatanatyam Shastra-based anatomical terminology for semantic alignment.

## 4.10 Summary

This chapter presented the complete experimental setup used for validating the proposed explainable multimodal dual-stage transfer learning framework for Bharatanatyam mudra recognition. The chapter described the computational environment, dataset preparation strategies, preprocessing pipeline, multimodal architecture, training configurations, evaluation methodologies, and explainability techniques employed throughout the study.

The next chapter presents the experimental results, comparative analysis, cross-dataset evaluation, feature-space analysis, and explainability outcomes obtained using the proposed framework.

# CHAPTER 5

# RESULTS AND DISCUSSION

## 5.1 Introduction

This chapter deals with the experimental results and as well as theperformance analysis of the proposed explainable multimodal dual-stage transfer learning framework for Bharatanatyam mudra recognition. The evaluation focuses on classification performance, cross-dataset generalization capability, domain adaptation effectiveness, latent feature representation quality, and explainability analysis.

The proposed framework was evaluated under multiple experimental settings including:

- Stage-1 gesture pretraining using Indian Sign Language data,
- Bharatanatyam mudra fine-tuning,
- Zero-shot cross-dataset evaluation,
- Domain adaptation evaluation,
- Feature-space clustering analysis,and
- Explainability analysis using GradCAM and landmark sensitivity.

The results demonstrate the effectiveness of multimodal representation learning, progressive transfer learning, and geometric feature integration for robust Bharatanatyam mudra recognition under varying performer and acquisition conditions.

## 5.2 Training Performance Analysis

### 5.2.1 Stage-1 ISL Pretraining Performance

The first stage of training is focused on learning generalized hand gesture representations using the Indian Sign Language dataset where the model demonstrated rapid convergence during the pretraining phase and achieved extremely high validation performance within a limited number of epochs and thus so the final Stage-1 performance metrics included:

- Validation Loss:

  0.0033

- Validation Accuracy:

  99.88%

The results indicate that the proposed multimodal architecture effectively learned generalized hand articulation and gesture representations prior to Bharatanatyam-specific adaptation.

### 5.2.2 Bharatanatyam Fine-Tuning Performance

During Stage-2 fine-tuning, the pretrained model was adapted to Bharatanatyam mudra recognition using the primary Bharatanatyam dataset and the validation performance improved progressively during the initial epochs, indicating successful transfer of generalized gesture knowledge into Bharatanatyam-specific semantic learning as also the highest validation accuracy was achieved during the:

Epoch 7

with:

99.34%

validation accuracy.

The training process demonstrated stable convergence with limited overfitting despite the highly fine-grained nature of Bharatanatyam mudra classification as shown in Table 5.1.

**Table 5.1:** Bharatanatyam fine-tuning performance across epochs

| Epoch | Validation Loss | Validation Accuracy |
|---|---|---|
| 1 | 0.1082 | 97.23% |
| 2 | 0.0669 | 97.93% |
| 3 | 0.0578 | 98.51% |
| 4 | 0.0263 | 99.30% |
| 5 | 0.0375 | 98.55% |
| 6 | 0.0538 | 98.37% |
| 7 | 0.0262 | 99.34% |
| 8 | 0.0721 | 98.11% |
| 9 | 0.0471 | 98.73% |
| 10 | 0.0308 | 99.03% |

The fluctuation observed in later epochs suggests that the model reached convergence relatively early and that extended training introduced mild overfitting behavior as shown in Fig 5.1.

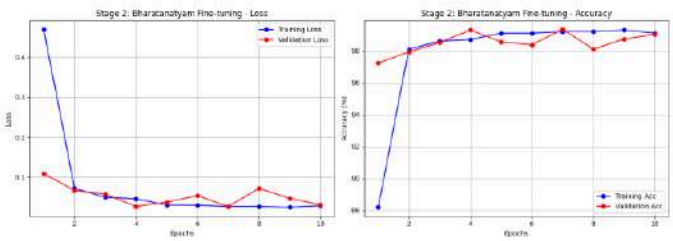



**Figure 5.1:** Training and validation performance during Bharatanatyam fine-tuning

## 5.3 Bharatanatyam Mudra Classification Results

The proposed multimodal framework achieved outstanding classification performance on the Bharatanatyam mudra test dataset.

The final test accuracy achieved was:

99.31%

across:

50

Bharatanatyam mudra classes.

The classification report demonstrated consistently high precision, recall, and F1-scores across most mudra categories as in Table 5.2.

**Table 5.2:** Overall Bharatanatyam mudra classification performance

| Metric | Value |
|---|---|
| Test Accuracy | 99.31% |
| Macro Precision | 0.99 |
| Macro Recall | 0.99 |
| Macro F1-Score | 0.99 |
| Weighted F1-Score | 0.99 |
| Number of Test Samples | 5687 |

Several mudra classes including:

- Anjali
- Garuda
- Kapotham
- Katakavardhana
- Mukulam
- Shanka
- Swastikam

achieved near-perfect classification performance.

The high accuracy as evident in Fig 5.2 demonstrates the effectiveness of combining transformer-based visual representation learning with geometric landmark encoding for fine-grained Bharatanatyam mudra classification.

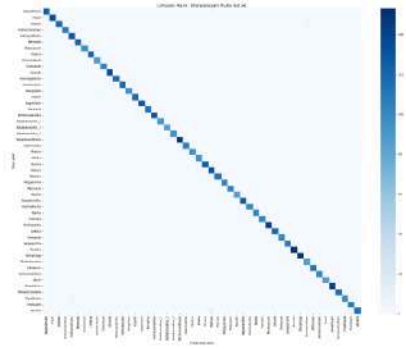


**Figure 5.2:** Confusion matrix for Bharatanatyam mudra classification

## 5.4 Cross-Dataset Generalization Analysis

### 5.4.1 Zero-Shot Cross-Dataset Evaluation

To evaluate robustness under performer and acquisition variation, the trained Bharatanatyam model was directly evaluated on the external Bharatanatyam mudra dataset without additional fine-tuning.

The zero-shot cross-dataset evaluation achieved:

66.90%

classification accuracy.

The significant reduction in performance compared to the in-domain Bharatanatyam test accuracy indicates the presence of:

- dataset distribution shift
- performer variation
- articulation variability
- acquisition condition differences

Despite the performance drop, several mudra classes maintained strong recognition performance, indicating partial transferability of the learned multimodal representations.

However, specific classes including:

- Aralam
- Bramaram
- Chaturam
- Katrimukha
- Trishulam

demonstrated substantial degradation under cross-dataset evaluation conditions.

**Table 5.3:** Zero-shot cross-dataset evaluation performance

| Metric | Value |
|---|---|
| Cross-Dataset Accuracy | 66.90% |
| Macro Precision | 0.75 |
| Macro Recall | 0.67 |
| Macro F1-Score | 0.67 |
| External Test Samples | 3335 |

The results as in Table 5.3 reveal that high in-domain accuracy alone does not guarantee robust generalization under performer and dataset variations.

### 5.4.2 Domain Adaptation Performance

To improve cross-dataset robustness, the pretrained Bharatanatyam model underwent additional domain adaptation using the training split of the external Bharatanatyam dataset.

Following adaptation which is depicted in training curves in Fig 5.3, the model achieved:

86.11%

accuracy on the unseen external testing set.

This represents a substantial improvement over the zero-shot evaluation setting as depicted in Table 5.4.

**Table 5.4:** Performance comparison between zero-shot and domain-adapted evaluation

| Evaluation Strategy | Accuracy |
|---|---|
| Zero-Shot Cross-Dataset Evaluation | 66.90% |
| Domain Adapted Evaluation | 86.11% |

The domain adaptation process enabled the framework to better accommodate:

- performer-specific articulation
- acquisition variability
- lighting differences
- dataset distribution shift

The results strongly validate the effectiveness of transfer learning and multimodal feature fusion for cross-dataset Bharatanatyam mudra recognition.

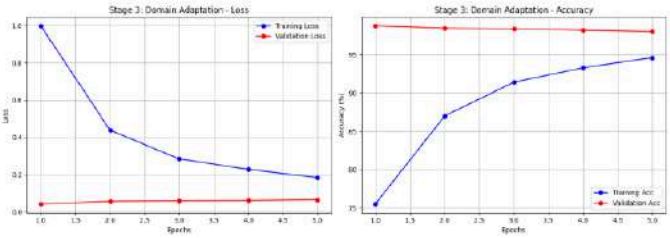



**Figure 5.3:** Training performance during domain adaptation

## 5.5 Feature Space Analysis

To analyze latent representation quality, clustering metrics were computed on the fused multimodal embeddings as shown in table 5.5.

**Table 5.5:** Feature-space clustering analysis

| Stage | Silhouette Score | Calinski-Harabasz Index | Davies-Bouldin Index |
|---|---|---|---|
| Stage-2 Fine-Tuned | 0.3611 | 49.0214 | 1.1793 |
| Stage-3 Domain Adapted | 0.1897 | 53.9176 | 1.9254 |

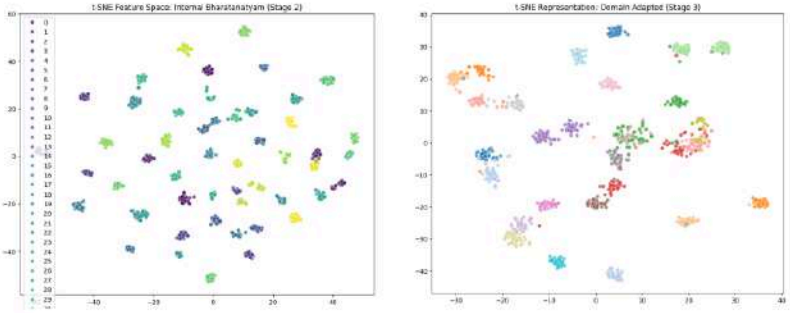
The Stage-2 fine-tuned model demonstrated stronger cluster compactness and separation, indicated by:

- higher silhouette score,and
- lower Davies-Bouldin index.

In contrast, the domain-adapted model exhibited more overlapping feature distributions, suggesting broader feature generalization at the expense of cluster compactness.

This behavior indicates a trade-off between:

- compact in-domain representation learning,and
- robust cross-dataset adaptability.




**(a)** For Primary Dataset **(b)** For Cross Dataset

**Figure 5.4:** t-SNE visualization of feature embeddings for different datasets used in the proposed framework

The t-SNE visualizations as shown in Fig 5.4 further support this observation, where the fine-tuned model exhibits relatively compact and well-separated mudra clusters, while the domain-adapted model demonstrates increased overlap among certain gesture classes.

## 5.6 Structural Similarity Analysis

Structural Similarity Index (SSIM) analysis was performed to investigate the visual similarities between Bharatanatyam mudra classes and the SSIM heatmaps as shown in Fig 5.5 revealed that several visually similar mudras shared overlapping structural characteristics, contributing to cross-class confusion during classification and also the analysis further indicated that subtle finger articulation difference plays a significant role in distinguishing the semantically similar Bharatanatyam mudras.

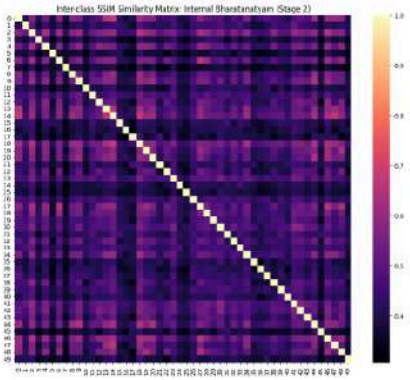



**Figure 5.5:** SSIM heatmap showing structural similarity among Bharatanatyam mudra classes

The SSIM analysis supports the necessity of combining:

- visual semantic learning,and
- geometric articulation encoding

for robust mudra recognition.

## 5.7 Explainability Analysis

### 5.7.1 GradCAM-Based Visual Explainability

GradCAM was employed to visualize the attention regions contributing significantly to the model's predictions.

The generated attention maps revealed that the Swin Transformer branch consistently focused on:

- finger articulation regions,
- fingertip structures,
- palm contours,and
- discriminative gesture boundaries

rather than irrelevant background regions.

This indicates that the proposed framework learned semantically meaningful gesture representations instead of memorizing dataset-specific artifacts.

### 5.7.2 Landmark Sensitivity Analysis

In addition to GradCAM visualization, landmark sensitivity analysis as shown in Fig 5.6 was conducted to identify influential hand joints contributing to classification decisions.

The analysis demonstrated that:

- fingertip landmarks,
- finger articulation joints,
- thumb positioning,and
- inter-finger spatial relationships

played significant roles in distinguishing Bharatanatyam mudras.

The landmark importance patterns aligned closely with traditional Bharatanatyam Shastra-based interpretations of mudra articulation.

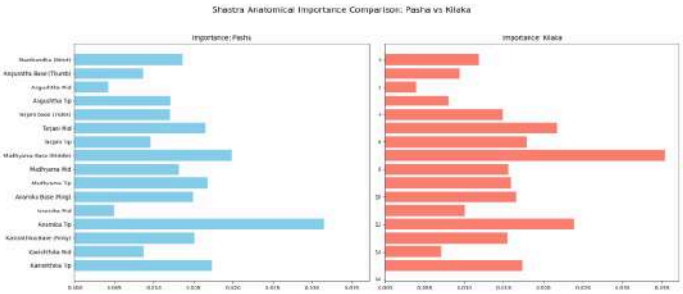



**Figure 5.6:** Landmark sensitivity analysis showing influential hand joints

The explainability analysis validates that the proposed framework learned anatomically and semantically meaningful Bharatanatyam gesture representations.

## 5.8 Discussion

The experimental results demonstrate that the proposed multimodal dual-stage transfer learning framework effectively combines visual and geometric representation learning for Bharatanatyam mudra recognition.

Several important observations can be made from the results:

- The Stage-1 ISL pretraining successfully enabled generalized gesture representation learning,
- Multimodal fusion between Swin Transformer embeddings and geometric landmark representations significantly improved fine-grained gesture discrimination,
- The proposed framework achieved outstanding in-domain Bharatanatyam classification accuracy of 99.31%,
- Zero-shot cross-dataset evaluation revealed substantial dataset distribution and performer variation challenges,
- Domain adaptation substantially improved external dataset generalization performance from 66.90% to 86.11%,
- Feature-space analysis demonstrated a trade-off between compact cluster separation and generalized adaptabilit, and

- Explainability analysis confirmed that the model focused on semantically meaningful Bharatanatyam articulation structures.

The results collectively demonstrate the effectiveness of combining:

- transformer-based visual learning,
- geometric landmark encoding,
- dual-stage transfer learning,and
- explainable artificial intelligence.

for robust Bharatanatyam mudra recognition under varying performer and acquisition conditions.

## 5.9 Summary

This chapter presented the experimental results and performance analysis of the proposed explainable multimodal dual-stage transfer learning framework for Bharatanatyam mudra recognition as the framework achieved strong classification performance, substantial cross-dataset adaptability, meaningful feature-space representations, and interpretable decision-making behavior through multimodal explainability analyssis.

The next chapter presents the conclusion, research contributions, limitations, and future research directions of the proposed study.

# CHAPTER 6

# CONCLUSION AND FUTURE WORK

## 6.1 Conclusion

This thesis presented an explainable multimodal dual-stage transfer learning framework for Bharatanatyam mudra recognition by integrating transformer-based visual representation learning with geometric hand landmark encoding. The proposed framework combined Swin Transformer Tiny for visual feature extraction and MediaPipe-based landmark representations for capturing fine-grained hand articulation semantics.

The study explored a progressive transfer learning strategy in which the model first learned generalized gesture representations using an Indian Sign Language dataset before being adapted for Bharatanatyam mudra recognition. The framework was further evaluated under cross-dataset conditions to analyze robustness against performer and acquisition variations. In addition to classification performance, explainability techniques such as GradCAM and landmark sensitivity analysis were incorporated to improve interpretability and semantic understanding of the learned gesture representations.

The overall findings indicate that combining visual semantic representations with geometric articulation features improves the capability of deep learning models to recognize complex Bharatanatyam mudras. The proposed framework also demonstrated the importance of multimodal representation learning and transfer learning for cultural gesture understanding tasks. Furthermore, the explainability analysis confirmed that the framework focused primarily on meaningful finger articulation and hand structure regions rather than irrelevant background features.

Although the proposed framework achieved strong performance, certain challenges remain. The current work primarily focuses on static mudra recognition and does not explicitly model temporal motion dynamics present in continuous dance sequences. In addition, variations in performer articulation and acquisition conditions continue to influence cross-dataset generalization capability. The framework also depends on the quality of landmark extraction, which may be affected under complex hand poses and partial occlusion conditions.

The proposed work contributes toward the development of robust and explainable artificial intelligence frameworks for Bharatanatyam mudra recognition and demonstrates the potential of multimodal deep learning for cultural heritage preservation and intelligent dance analysis applications.

## 6.2 Future Scope

Future research can extend the proposed framework toward video-based Bharatanatyam analysis incorporating temporal gesture modeling and spatio-temporal transformer architectures for capturing dynamic dance movements and gesture transitions. Additional improvements may be achieved through

advanced domain adaptation strategies, self-supervised representation learning, and lightweight deployment frameworks for real-time cultural AI applications. Furthermore, integrating deeper semantic knowledge from Bharatanatyam Shastra and choreography analysis may contribute toward more culturally grounded, interpretable, and semantically aware Bharatanatyam understanding systems.

# REFERENCES


[1] M. R. Reshma, B. Kannan, V. P. J. Raj, and S. Shailesh, "Cultural heritage preservation through dance digitization: A review," *Digital Applications in Archaeology and Cultural Heritage*, vol. 28, p. e00257, 2023.

[2] R. Amrutha and V. M. Ladwani, "Bharatanatyam hand gesture recognition using normalized chain codes and oriented distances," in *Proc. Int. Conf. Inventive Computation Technologies (ICICT)*, vol. 3, 2016, pp. 1–6.

[3] A. D. Naik and M. Supriya, "Classification of Indian Classical Dance Images using Convolution Neural Network," in *Proc. Int. Conf. Communication and Signal Processing (ICCSP)*, 2020, pp. 1245–1249.

[4] D. P. Akarsha, B. Monisha, K. L. Bhoomika, and D. R. V., "Bharatanatyam Mudra Classification using CNN," in *Proc. 1st Int. Conf. Software, Systems and Information Technology (SSITCON)*, 2024, pp. 1–6.

[5] A. P. Parameshwaran, H. P. Desai, R. Sunderraman, and M. Weeks, "Transfer Learning for Classifying Single Hand Gestures on Comprehensive Bharatanatyam Mudra Dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2019, pp. 1–3.

[6] A. P. Parameshwaran, H. P. Desai, M. Weeks, and R. Sunderraman, "Unravelling of Convolutional Neural Networks through Bharatanatyam Mudra Classification with Limited Data," in *Proc. 10th Annual Computing and Communication Workshop and Conference (CCWC)*, 2020, pp. 342–347.

[7] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2005, pp. 886–893.

[8] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[9] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.

[10] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[11] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[12] P. Langley, W. Iba, and K. Thompson, "An Analysis of Bayesian Classifiers," in *Proc. Nat. Conf. Artificial Intelligence (AAAI)*, 1992, pp. 223–228.

[13] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[14] K. Thanikasalam, A. Ramanan, and P. Kanmanirajah, "A comprehensive review and ensemble CNN approach for Bharatanatyam single-hand gesture classification," *Entertainment Computing*, vol. 56, p. 101069, 2026.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2012, pp. 1097–1105.

[16] C. Sarmah and P. Sarma, "A dataset of Sattriya dance: Classical dance of Assam," *Data in Brief*, vol. 52, p. 109878, 2024.

[17] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "MediaPipe Hands: On-device Real-time Hand Tracking," arXiv:2006.10214, 2020.

[18] Z. Cao, G. Hidalgo Martinez, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, 2021.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 248–255.

[20] K. Adalarasu, R. M. K. Chetty, K. G. Begum, S. Harini, and M. Janardhanan, "An Explainable Machine Learning (XAI) framework for classification of intricate dancing posture among Indian Bharatanatyam dancers," *Applied Soft Computing*, vol. 171, p. 112817, 2025.

[21] S. Paul, G. Sagar, P. P. Das, and K. S. Rao, "Two-stage pipeline based robust hand gesture recognition from Bharatanatyam dance images," *Multimedia Tools Appl.*, vol. 84, pp. 39667–39691, 2025.

[22] J. R. Challapalli, R. Durgam, B. L. Nandipati, and P. Malavath, "Intelligent fine-tuning of convolutional neural networks using flamingo search for traditional dance classification," *Systems and Soft Computing*, vol. 8, p. 200449, 2026.

[23] S. Shailesh and M. V. Judy, "Understanding dance semantics using spatio-temporal features coupled GRU networks," *Entertainment Computing*, vol. 42, p. 100484, 2022.

[24] S. Gupta and S. Singh, "Indian dance classification using machine learning techniques: A survey," *Entertainment Computing*, vol. 50, p. 100639, 2024.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10012–10022.

[26] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.

[27] S. Kumar, "Indian Sign Language Dataset," Kaggle, 2023. [Online]. Available: `https://www.kaggle.com/datasets/saurabh24999/indian-sign-language/data`. [Accessed: May 20, 2026].

[28] R. J. Raj, S. Dharan, and T. T. Sunil, "Optimal feature selection and classification of Indian classical dance hand gesture dataset," *The Visual Computer*, vol. 39, pp. 4049–4064, 2023.

[29] G. K. P. K. Nambiar, N. M. Kumar, V. G., and M. G. Thushara, "AI-Powered Bharatanatyam Mudra Identification and Description System: Preserving Heritage through Technology," in *Proc. 12th Int. Conf. Computing for Sustainable Global Development (INDIACom)*, 2025, pp. 1–7.

[30] S. U. Shetty, R. Puthran, and A. P. Shetty, "Asamyuta Hasta Mudras Recognition Using MediaPipe Keypoint Extraction and Random Forest Classification," in *Proc. Int. Conf. Artificial Intelligence and Data Engineering (AIDE)*, 2025, pp. 173–181.

[31] R. R. Subramanian et al., "Automated Real-Time Hand Gesture Detection for Bharatanatyam Dance," in *Proc. Int. Conf. Computational Robotics, Testing and Engineering Evaluation (ICCRTEE)*, 2025, pp. 1–6.

[32] P. Sadhana, N. Ravishankar, and S. Palaniswamy, "Bharatanatyam Mudra Recognition Using Deep Learning and Meta-Learning Techniques," in *Proc. Int. Conf. Communications and Computer Science (InCCCS)*, 2024, pp. 1–6.

[33] A. S. Nandeppanavar, S. S. Kallur, V. A. Sankannavar, and P. Thotad, "Bharatanatyam hasta mudra categorization using deep learning approaches," in *Proc. IEEE North Karnataka Subsection Flagship Int. Conf. (NKCon)*, 2023, pp. 1–6.

[34] D. Kilari and K. K. Singh, "Comparative study on the effect of HSV Segmentation and ORB Features on Transfer Learning models for recognition of Bharatanatyam Asamyukta Mudras," in *Proc. Int. Conf. Computational Intelligence, Communication Technology and Networking (CICTN)*, 2023, pp. 241–245.

[35] S. Haridas and V. R. Bai, "Detection and Classification of Indian Classical Bharathanatyam Mudras Using Enhanced Deep Learning Technique," in *Proc. Int. Conf. Innovations in Science and Technology for Sustainable Development (ICISTSD)*, 2022, pp. 18–23.

[36] P. Chavan, P. Choudhary, S. Upadhyay, S. Devarshi, S. Sureliya, and A. Kumar, "EfficientNetB0-based Feature Extraction for Single and Double-Hand Mudra Recognition," in *Proc. 5th Int. Conf. Sentiment Analysis and Deep Learning (ICSADL)*, 2026.

[37] V. Amrutha Raj and G. Malu, "EnGesto: An Ensemble Learning Approach for Classification of Hand Gestures," *IEEE Access*, vol. 12, pp. 85709–85723, 2024.

[38] S. Baskar, W. J. Hans, A. V. R., V. S. Solomif, and A. R., "MudraGyaan: A Novel Feature Extraction Algorithm for Machine Learning-Based Bharatanatyam Mudra Classification," in *Proc. Int. Conf. Advancement in Renewable Energy and Intelligent Systems (AREIS)*, 2024.

[39] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[40] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Int. Conf. Learn. Representations (ICLR)*, 2019.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[42] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Int. Conf. Learn. Representations (ICLR)*, 2021.

[43] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[44] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[45] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[46] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[47] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

# EXPLAINABLESTAGE-AWARE BHARATANATYAM MUDRA RECOGNITION BY INTEGRATING GEOMETRIC HAND LANDMARK ENCODING AND DOUBLE TRANSFER LEARNING

ORIGINALITY REPORT

| 7% | 5% | 5% | 3% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | Submitted to University of Ulster<br>Student Paper | <1% |
| 2 | "Pattern Recognition and Machine Intelligence", Springer Science and Business Media LLC, 2026<br>Publication | <1% |
| 3 | www.mdpi.com<br>Internet Source | <1% |
| 4 | Submitted to National Institute of Technology Jamshedpur<br>Student Paper | <1% |
| 5 | Submitted to University of Wales Institute, Cardiff<br>Student Paper | <1% |
| 6 | Submitted to University of Southampton<br>Student Paper | <1% |
| 7 | Submitted to Vellore Institute of Technology<br>Student Paper | <1% |
| 8 | ebin.pub<br>Internet Source | <1% |

| | | |
|---|---|---|
| 9 | repository.nwu.ac.za<br>Internet Source | <1 % |
| 10 | researcharchive.vuw.ac.nz<br>Internet Source | <1 % |
| 11 | Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dhirendra Kumar Shukla. "Recent Trends in Intelligent Computing and Communication", CRC Press, 2025<br>Publication | <1 % |
| 12 | Submitted to UC, Boulder<br>Student Paper | <1 % |
| 13 | ijsret.com<br>Internet Source | <1 % |
| 14 | aclanthology.org<br>Internet Source | <1 % |
| 15 | unige.iris.cineca.it<br>Internet Source | <1 % |
| 16 | www.aimspress.com<br>Internet Source | <1 % |
| 17 | Akshata K, Dharshini K. "An Intelligent AI-Driven Framework for Early Prediction of Heart Disease Using Advanced Machine Learning Techniques", Springer Science and Business Media LLC, 2026<br>Publication | <1 % |
| 18 | Submitted to Indian Institute of Space Science and Technology<br>Student Paper | <1 % |

| | | |
|---|---|---|
| 19 | Submitted to Loughborough University<br>Student Paper | <1 % |
| 20 | Submitted to National Institute of Technology, Rourkela<br>Student Paper | <1 % |
| 21 | journalmsr.com<br>Internet Source | <1 % |
| 22 | Seghers, Estelle. "A Data-Driven Multivariate Process Monitoring Platform for Knowledge Discovery and Model Building in Industrial Applications", Louisiana State University and Agricultural & Mechanical College, 2024<br>Publication | <1 % |
| 23 | docplayer.net<br>Internet Source | <1 % |
| 24 | ir.kabarak.ac.ke<br>Internet Source | <1 % |
| 25 | prod-ms-be.lib.mcmaster.ca<br>Internet Source | <1 % |
| 26 | www.jneonatalsurg.com<br>Internet Source | <1 % |
| 27 | Parikshit N. Mahalle, Nuzhat Faiz Shaikh, Pritibala S. Ingle, Yashwant Sudhakar Ingle. "Skin Cancer Detection Using Artificial Intelligence Techniques", CRC Press, 2026<br>Publication | <1 % |
| 28 | assets-eu.researchsquare.com | |

| | | |
|---|---|---|
| | Internet Source | <1% |
| 29 | library.acadlore.com<br>Internet Source | <1% |
| 30 | reelmind.ai<br>Internet Source | <1% |
| 31 | theses.hal.science<br>Internet Source | <1% |
| 32 | umpir.ump.edu.my<br>Internet Source | <1% |
| 33 | "Proceedings of the International Conference on Computational Intelligence and Sustainable Technologies", Springer Science and Business Media LLC, 2022<br>Publication | <1% |
| 34 | hal.science<br>Internet Source | <1% |
| 35 | mobt3ath.com<br>Internet Source | <1% |
| 36 | ojs.aaai.org<br>Internet Source | <1% |
| 37 | scholarworks.gsu.edu<br>Internet Source | <1% |
| 38 | thesai.org<br>Internet Source | <1% |
| 39 | www.ntp.niehs.nih.gov<br>Internet Source | <1% |

| | | |
|---|---|---|
| 40 | "Intelligent Computing and Technologies", Springer Science and Business Media LLC, 2026<br>Publication | <1 % |
| 41 | K. V. Sambasivarao, Anasuya Sesha Roopa Devi Bhima. "Artificial Intelligence, Computational Intelligence, and Inclusive Technologies - Proceedings of International Conference on Artificial Intelligence, Computational Intelligence, and Inclusive Technologies (ICRAIC2IT – 2025)", CRC Press, 2026<br>Publication | <1 % |
| 42 | Paolo Ferro, Harinadh Vemanaboina, Chander Prakash. "Computational Techniques and Smart Manufacturing", CRC Press, 2026<br>Publication | <1 % |
| 43 | arxiv.org<br>Internet Source | <1 % |
| 44 | bear.buckingham.ac.uk<br>Internet Source | <1 % |
| 45 | dr.ntu.edu.sg<br>Internet Source | <1 % |
| 46 | icsts2025.lbt.ac.in<br>Internet Source | <1 % |
| 47 | jutif.if.unsoed.ac.id<br>Internet Source | <1 % |

| | | |
|---|---|---|
| 48 | Abhilasha Sharma, Vishwas Rathi, Anupam Biswas, Anil Singh, Omer Rana. "Multimodal Artificial Intelligence in Precision Agriculture - Practices, Challenges, and Applications", CRC Press, 2026<br>Publication | <1% |
| 49 | Susovan Pradhan, Prasenjit Mukherjee, Baisakhi Chakraborty. "CVAE-guided triage and modular classifiers for multimodal ASD detection", Computers in Biology and Medicine, 2026<br>Publication | <1% |
| 50 | Uche Onyekpe, Vasile Palade, M. Arif Wani. "Recent Advances in Deep Learning Applications - New Techniques and Practical Examples", CRC Press, 2025<br>Publication | <1% |
| 51 | Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical and Computer Technologies", CRC Press, 2025<br>Publication | <1% |

| | | | |
|---|---|---|---|
| Exclude quotes | Off | Exclude matches | Off |
| Exclude bibliography | On | | |

# Ganesh Gutti

## EXPLAINABLESTAGE-AWARE BHARATANATYAM MUDRA RECOGNITION BY INTEGRATING GEOMETRIC HAND LANDM...

Quick Submit

Quick Submit

Delhi Technological University

### Document Details

**Submission ID**
trn:oid:::1:3580176790

**Submission Date**
May 27, 2026, 2:15 PM GMT+5:30

**Download Date**
May 27, 2026, 2:23 PM GMT+5:30

**File Name**
24AFI25_GaneshGutti_Thesis_1.pdf

**File Size**
23.1 MB

53 Pages

9,699 Words

63,890 Characters

# *% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

**Caution: Review required.**

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

# ABSTRACT

Bharatanatyam mudras constitute an essential component of Indian classical dance, serving as a medium for storytelling, emotional expression, and semantic communication. Automatic recognition of these hand gestures is a challenging task due to subtle inter-class variations, complex finger articulations, viewpoint differences, and limited annotated datasets with existing gesture recognition approaches often relyinng solely on visual appearance features and lack robustness, interpretability, and cross-domain adaptability and in addition, limited research has explored the integration of multimodal learning and explainable artificial intelligence for culturally significant gesture recognition tasks such as Bharatanatyam mudra analysis.

This thesis proposes an explainable multimodal dual-stage transfer learning framework for Bharatanatyam mudra recognition by integrating transformer-based visual learning with geometric hand landmark representations which employs a Swin Transformer Tiny backbone for RGB image feature extraction and a dedicated landmark-processing branch utilizing MediaPipe hand keypoints to capture structural hand articulation and thus extracted features from both modalities are fused to enhance discriminative representation learning and improve classification performance. To address data scarcity and improve transferability, a dual-stage transfer learning strategy is introduced, where the model is initially pretrained on an Indian Sign Language (ISL) gesture dataset and subsequently fine-tuned on Bharatanatyam mudras. To add more, explainable artificial intelligence techniques such as GradCAM-based visualization are incorporated to help with the process of the interpreting model predictions and identification of anatomically significant hand regions influencing recognition decisions.

Extensive experiments are conducted on Bharatanatyam mudra datasets along with cross-domain evaluation using external gesture datasets to assess robustness and generalization capability and the proposed framework is evaluated using multiple performance metrics including accuracy, precision, recall, and F1-score, along with cross-dataset experiments to further analyze the effects of domain shift and transferability and the experimental results demonstrate that the proposed multimodal dual-stage transfer learning framework achieves robust and highly accurate Bharatanatyam mudra recognition while providing improved interpretability and cross-domain adaptability thereby contributing towards the development of culturally aware artificial intelligence systems for digital heritage preservation, intelligent dance analysis, and human-centered gesture understanding.

# CHAPTER 1

# INTRODUCTION

## 1.1 Background

Indian classical dance forms represent a significant component of the cultural and artistic heritage of India. Among these, Bharatanatyam is one of the oldest and most structured classical dance traditions, characterized by intricate body movements, rhythmic footwork, facial expressions, and symbolic hand gestures known as *mudras*. Mudras as shown in Fig 1.1 are crucal for the conveying emotions, narratives, and semantic meaning during performances which emphasise on the fact that accurate understanding and interpretation of these gestures are essential for preserving the expressive richness and communicative depth of Bharatanatyam as a dance form[1].

With the advancing technology in areas of artificial intelligence and computer vision technologies, automated gesture recognition systems have gained an increasing attention in the areas including but not limited to such as human-computer interaction, sign language interpretation, surveillance, virtual reality, and healthcare and recent developpments in deep learning, particularly convolutional neurl networks and transformer-based architectures, have significantly improved visual recognition performance across various applications[2] but however, recognizing Bharatanatyam mudras remains a challenging task due to subtle inter-class variations, complex finger articulations, occlusions, varying illumination conditions, and limited availability of annotated datasets[2].

In the recent years, the frameworks for the hand landmark detection like the MediaPipe have enabled the extraction of detailed geometric representations of the hand structures offering additional information beyond raw visual appearance[3, 4] and in the similar fashion, transfer learning techniques have also emerged as effective solutions for performance improvement in the data-constrained environments by making use of the knowledge learned from related domains[5, 6] and despite all these advancements, existing Bharatanatyam mudra recognition systems as shown in Fig 1.2 very often rely primarily on image-based features, lack explainability, and rarely investigate cross-domain generalization or multimodal fusion strategies.


**Figure 1.1:** Sample Bharatanatyam mudras used in the study

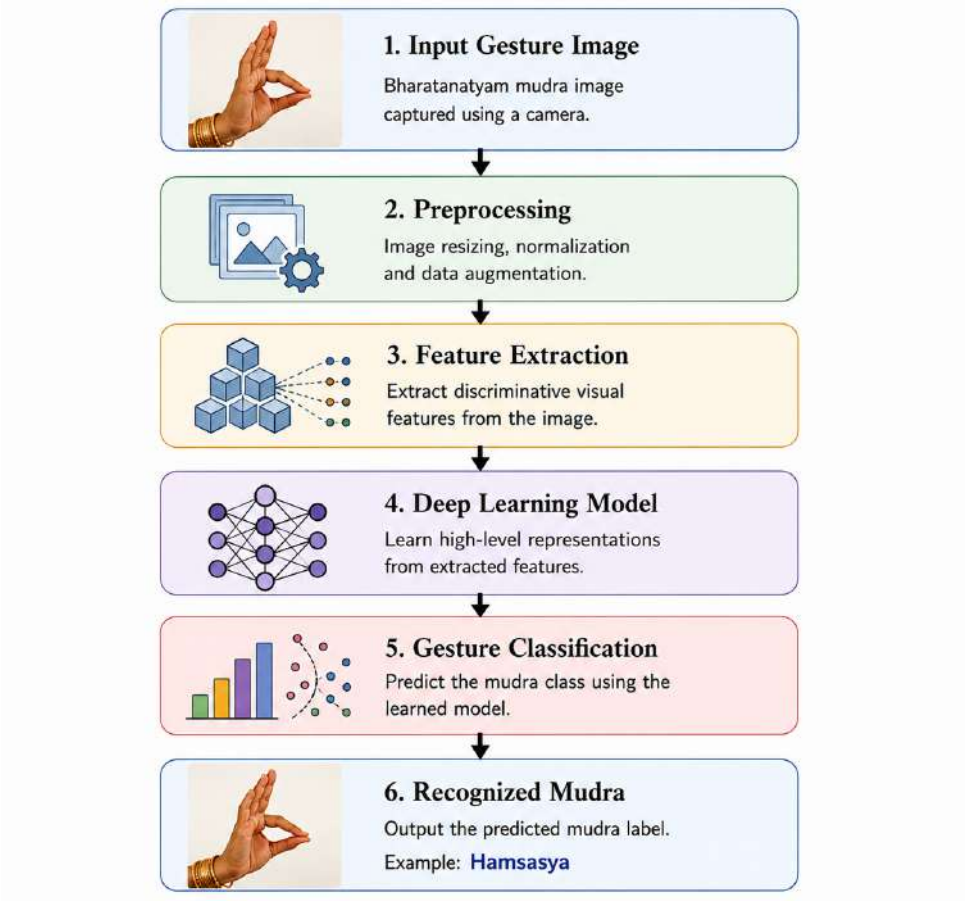



**Figure 1.2:** General workflow of an AI-based gesture recognition system

## 1.2 Problem Statement

Automatic recognition of Bharatanatyam mudras presents several technical and research challenges as presented in Table 1.1 : Traditional image-based classification approaches often struggle to capture the fine-grained geometric distinctions between visually similar mudras; Variations in hand orientation, performer-specific articulation styles, background conditions, and

lighting further complicate the recognition process;the limited size of available Bharatanatyam datasets restricts the ability of deep leaning models to generalise effectively.

Most existing approaches focus primarily on achieving classification accuracy without addressing interpretability and semantic understanding of model decisions and then, very limited research has explored the integration of multimodal representations combining visual and geometric hand features, so the lack of robust cross-domain evaluation also raises concerns regarding the adaptability and generalization capability of current systems when exposed to unseen gesture distributions or datasets.

Thus, there is a need for an explainable and robust Bharatanatyam mudra recognition framework. This should be capable of effectively integrating visual and geometric information. This should also be improving transferability across domains. In addition this should be providing interpretable predictions which are to be aligned with meaningful hand articulation patterns.

**Table 1.1:** Major challenges in Bharatanatyam mudra recognition

| **Challenge** | **Description** |
|---|---|
| Inter-class similarity | Many mudras possess highly similar finger configurations |
| Intra-class variation | Differences in performer styles and hand articulation |
| Background complexity | Variations in lighting and environmental conditions |
| Limited datasets | Scarcity of large annotated Bharatanatyam datasets |
| Interpretability limitations | Difficulty in understanding model decision mechanisms |

## 1.3 Aim of the Study

The primary aim of this research is to develop an explainable multimodal dual-stage transfer learning framework for robust Bharatanatyam mudra recognition using transformer-based visual learning and hand landmark fusion and the poposed architecture is as in Fig 1.3.

## 1.4 Research Gaps

After reviewing the existing literature, the following major research gaps were identified:

- Existing Bharatanatyam mudra recognition frameworks suffer from limited cross-dataset generalization due to strong studio bias, performer variability, and acquisition condition differences.

- Most existing approaches rely primarily on either visual appearance features or geometric landmark representations independently, with limited exploration of robust multimodal fusion frameworks for fine-grained mudra recognition.
- The adoption of explainable transformer-based architectures and progressive transfer learning strategies remains limited in Bharatanatyam mudra recognition and cultural gesture understanding systems.

These identified research gaps motivated the development of the proposed explainable multimodal dual-stage transfer learning framework presented in this thesis.

## 1.5 Objectives

The major objectives of the proposed work are as follows:

1. To develop an explainable multimodal Bharatanatyam mudra recognition framework integrating visual gesture representations and geometric hand landmark features,
2. To implement a dual-stage transfer learning strategy for improving generalized gesture representation learning and cross-dataset adaptability,
3. To incorporate explainable artificial intelligence techniques for interpreting model predictions and analyzing semantically meaningful hand articulation regions.

## 1.6 Research Contributions

The major contributions of the proposed research are summarized as follows:

- Development of an explainable multimodal Bharatanatyam mudra recognition framework integrating Swin Transformer-based visual learning and MediaPipe-based geometric hand landmark representations.
- Introduction of a dual-stage transfer learning strategy involving generalized gesture pretraining using Indian Sign Language data followed by Bharatanatyam-specific fine-tuning and cross-dataset evaluation.
- Incorporation of explainable artificial intelligence techniques including GradCAM and landmark sensitivity analysis for interpreting semantically meaningful gesture regions and supporting intelligent cultural heritage preservation.

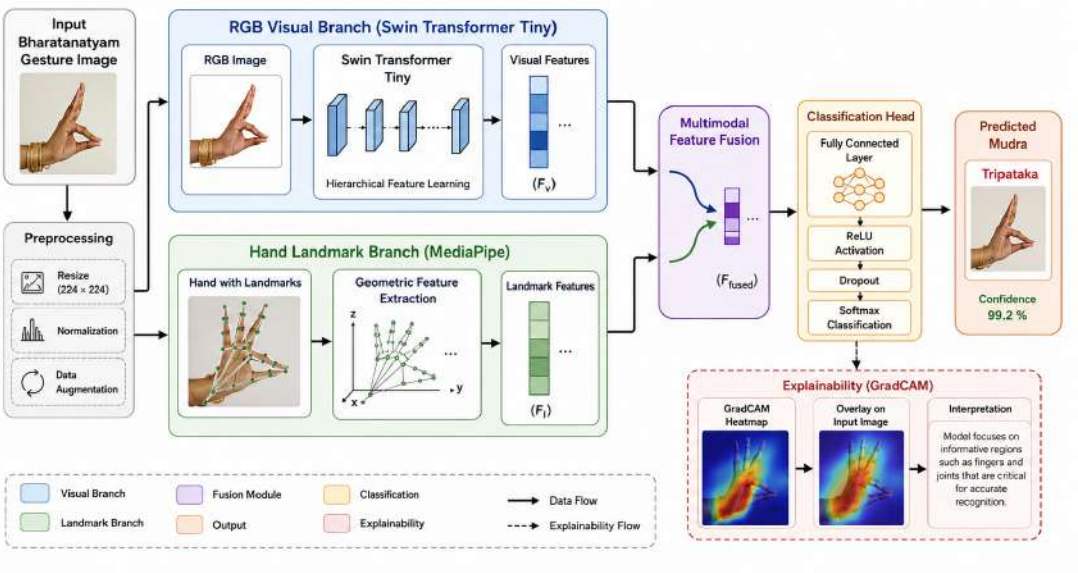



**Figure 1.3:** Overall architecture of the proposed multimodal dual-stage transfer learning framework

## 1.7 Organization of the Thesis

The remaining of this presented thesis is organized as given below:

- Chapter 2 presenting the literature review related to gesture recognition, transfer learning, transformer architectures, hand landmark analysis, and explainable artificial intelligence techniques,
- Chapter 3 describing the proposed methodology, including the tasks of dataset preparation, preprocessing, landmark extraction, multimodal fusion architecture, dual-stage transfer learning, and explainability framework.
- Chapter 4 discusses the experimental setup, implementation details, training configuration, evaluation metrics, and hardware specifications,
- Chapter 5 presents the experimental results, comparative analysis, explainability evaluation, robustness analysis, and cross-domain performance assessment, and
- Chapter 6 concluding the thesis by summarising the major findings, contributions, limitations, and possible future research directions.

# CHAPTER 2

# LITERATURE REVIEW

The literature surrounding automated Indian Classical Dance recognition and particularly Bharatanatyam mudra classification, reveals a significant evolution from the traditional approaches of machine learning towards advanced kind of deep learning and multimodal frameworks where existing research demonstrates continuous progress in visual representation learning, geometric hand analysis, transfer learning methodologies, and explainable artificial intelligence techniques for gesture understanding and despite these advancements, several important research gaps remain unresolved, particularly in terms of robustness, interpretability, and cross-domain generalization.

## 2.1 Evolution from using Hand-Crafted Features to techniques of Deep Learning

Early approaches to Bharatanatyam mudra recognition as shown in Fig 2.1 primarily relied on tradditional image processing and hancrafted feature extraction techniques in which researchers made use of the descriptors such as Histogram of Oriented Gradients (HOG)[7], Scale-Invariant Feature Transform (SIFT)[8], Sped-Up Robust Features (SURF)[9], Hu Moments, and normalized chain codes to represent the hand gestures mathematically and thus extracted features were subsequently classified with the employment of usually used machine learning algorithms including Support Vector Machines (SVM)[10], K-Nearest Neighbors (KNN)[11], Naive Bayes[12], and Random Forest[13] classifiers and although these approaches provided an initial understanding of hand orientation and shape representation, they suffered from several limitations because of the fact that handcrafted features were highly sensitive to illumination variations, background complexity, hand orientation changes, and performer-specific articulation differences and to add more these methods were lacking the ability to capture fine-grained hierarchical spatial relationships between finger joints and hand configurations[14].

In order to address these limitations, researchers hae gradually transitioned towardss deep learning methodologies, particularly Convolutional Neural Networks (CNNs) which automatically learn hierarchical visual representations directly from raw image data[15], eliminating the very need for a manually featureengineering where these Deep learning frameworks demonstrated substantially improved classification performance and robustness compared to traditional handcrafted approaches, partcularly in complex gesture recognition tasks involving subtle structural differences between the mudras[16].

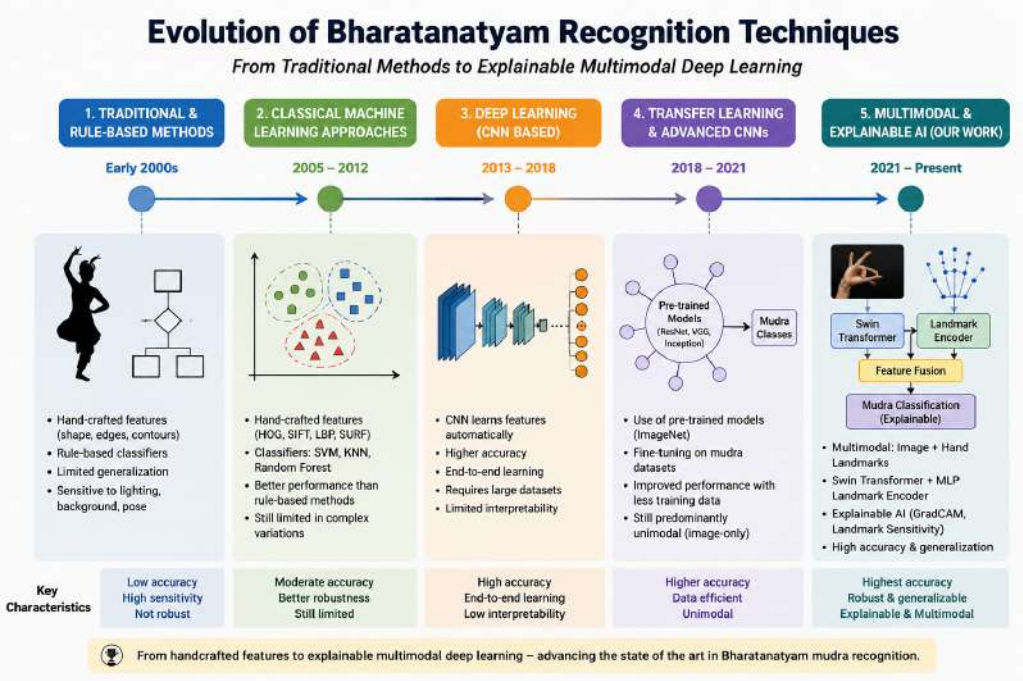



**Figure 2.1:** Evolution of gesture recognition approaches from traditional handcrafted features to deep learning frameworks

## 2.2 Dataset Challenges and Studio Bias

One among the most noticeable limitations identified all throughout the literature is that of the dearth in the case of large-scale, publicly available Bharatanatyam mudra datasets where most of the existing studies have shown the relying on small proprietary datasets collected in the highly controlled studio environments characterized by the presence of the uniform lighting, the plain backgrounds, and also the constrained movement of the dance performer.

Such controlled conditions introduce a phenomenon commonly referred to as *studio bias* and often these models which are trained exclusively on studio-based datasets are prone to fail in generalizing effectively when deployed in real-world stage environments containing cluttered backgrounds, complex costumes, dynamic illumination, overlapping performers, and varying camera viewpoints and so consequently, many existing systems demonstrate high accuracy under laboratory conditions but exhibit poor robustness in unconstrained practical scenarios and in order to mitigate these limitations, recent research stresses on the importance of the use of extensive data augmentation, domain randomization, and cross-domain evaluation strategies so that this simulation of realistic stage conditions through augmentation techniques can improve the generalization capability of the models by reducing overfitting to controlled datasets. These are summarized in Table 2.1.

**Table 2.1:** Common limitations of existing Bharatanatyam mudra datasets

| Limitation | Impact on Recognition Performance |
|---|---|
| Small dataset size | Reduced generalization capability |
| Controlled studio backgrounds | Poor robustness in real-world environments |
| Limited performer diversity | Increased subject-specific bias |
| Uniform lighting conditions | Sensitivity to illumination variations |
| Lack of cross-domain evaluation | Limited adaptability across datasets |

## 2.3 Geometric Landmarking and Spatial Features

To improve robustness against background complexity and visual occlusions, recent studies have explored geometric landmark-based representations for gesture recognition and frameworks such as Google MediaPipe[17] and OpenPose[18] enable real-time extraction of skeletal hand landmarks representing finger joints and palm structures as shown in Fig. 2.2 which provide a geometric abstraction of the hand independent of texture, background, and illumination conditions that helped the researchers to utilize spatial relationships between hand joints, including Euclidean distances, joint angles, and finger articulation patterns, to create structural representations suitable for gesture classification tasks.

Landmark-based approaches offer improved invariance to lighting and environmental variations compared to raw RGB image representations but then however, landmark-only frameworks may also end up losing on the important appearance-based contextual information such as texture, finger contours, and subtle visual semantics, so as a consequence the recent literature is increasingly advocating the integration of visual and geometric representations through multimodal learning frameworks.

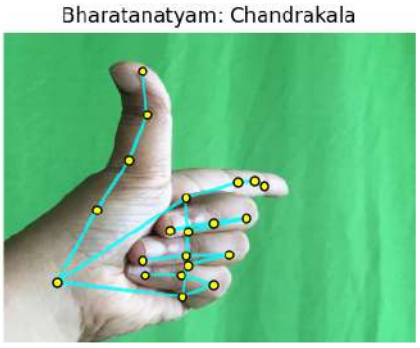



**Figure 2.2:** Example of skeletal hand landmark representation using MediaPipe

## 2.4 Transfer Learning and Double Transfer Learning

Pertaining to the limited size of Bharatanatyam mudra datasets, transfer learning has become a widely adopted strategy in gesture recognition research so that instead of training deep learning models from scratch, the researchers are in a position to utilize the architectures that are already pretrained usually on the datasets of large-scale such as ImageNet[19] and subsequently fine-tune them for gesture classification tasks and accordingly a good number of these pretrained architectures including VGG16, ResNet, EfficientNet, and MobileNet have demonstrated strong performance in hand gesture recognition applications and recently some researchers introduced specialized methodologies such as Double Transfer Learning (DTL) to further improve recognition accuracy and representation learning[6].

In Double Transfer Learning, a model pretrained on a large generic dataset undergoes an intermediate training stage using a generalized hand gesture or sign language dataset before final fine-tuning on Bharatanatyam mudras and so this progressive transfer learning strategy enables the network to learn both generalized visual patterns and domain-specific hand articulation structures supported by existing studies reporting that DTL significantly improves classification performance compared to conventional single-stage transfer learning approaches by enhancing feature generalization and reducing domain adaptation difficulty and so this methodology as depicted in Fig 2.3 is particularly effective for culturally specific gesture recognition tasks with not enough training data[6].

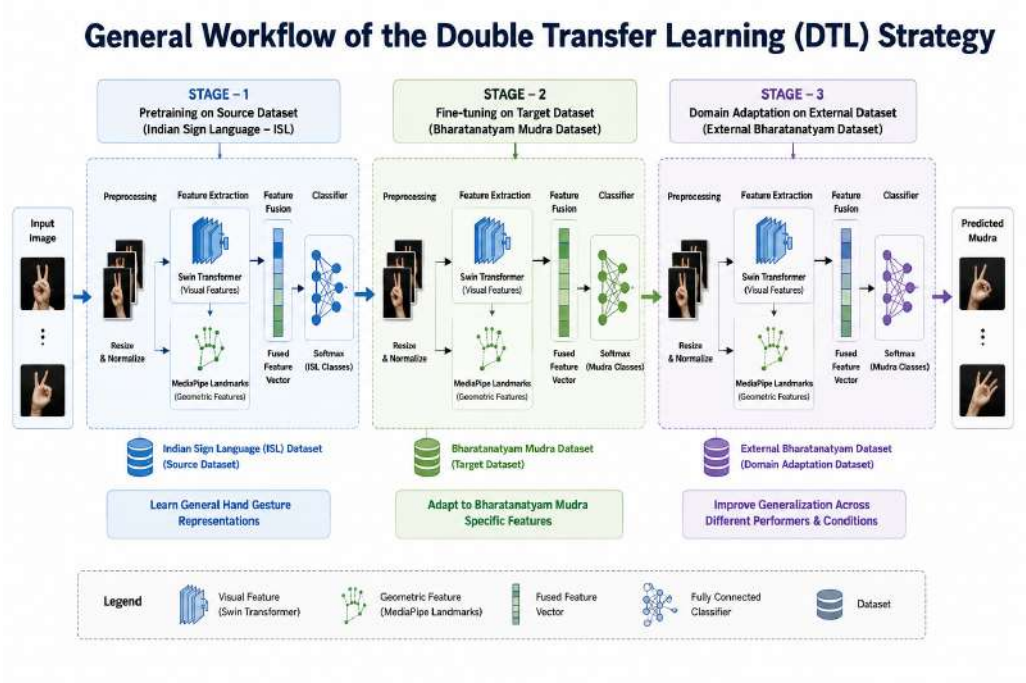



**Figure 2.3:** General workflow of the Double Transfer Learning (DTL) strategy

## 2.5 Explainable Artificial Intelligence in Gesture Recognition

Eevn though the deep learning models achieve remarkable classification performance, they often function as opaque blackbox systems that create difficulty to determine whether the model is learning meaningful gesture semantics or merely memorizing irrelevant background artifacts and this lack of transparency presents significant concerns in cultural heritage applications where interpretability and trustworthiness are essential which prompted reseacrh towards addressing these issues, whererecent literature increasingly incorporates Explainable Artificial Intelligence (XAI) techniques into gesture recognition frameworks and methods such as SHAP (SHapley Additive exPlanations) and Gradient-weighted Class Activation Mapping (GradCAM) as shown in Fig 2.4 generate visual explanations highlighting image regions that are significant in the contribution to that of the model predictions[20].

In Bharatanatyam mudra recognition, explainability techniques help identify critical finger articulations, hand contours, and anatomical regions influencing classification decisions. Such visual interpretations improve transparency and provide insight into the semantic reasoning process learned by deep learning models.

Recent studies suggest that integrating explainable visual representations with geometric landmark analysis and multimodal feature fusion leads to more robust and interpretable gesture recognition systems suitable for real-world deployment.


**Figure 2.4:** GradCAM-based explainability visualization highlighting discriminative gesture regions

**Table 2.2:** Literature Review on Indian Classical Dance Classification

| Author & Year | Methodology | Dataset | Strengths | Limitations |
|---|---|---|---|---|
| K. Thanikasalam et al. (2026)[14] | Ensemble of three EfficientNetV2S models using raw, skeleton, and embedded-landmark images | AHM-UoJ and Jisha Raj Bharatanatyam Mudra Dataset | Achieved high classification accuracy (98.28%) through multimodal feature integration | Computationally expensive due to ensemble learning and dependence on handcrafted feature extraction |
| C. Sarmah and P. Sarma (2024)[16] | CNN and ResNet-50 models using Watershed segmentation | Sattriya-08 Double Handed Mudra Dataset[16] | Effectively captures double-handed gestures with high training accuracy | Small dataset size and segmentation sensitivity under complex backgrounds |
| K. Adalarasu et al. (2025)[20] | Machine learning models using VGRF data with SHAP-based explainability | VGRF force platform dataset containing six dance postures | Provides interpretability and avoids optical occlusion problems | Limited to static poses and affected by class imbalance |
| S. Paul et al. (2025)[21] | YOLOv6-based hand detection with ResNet18 classification and flexion angle descriptors | Oxford Hand Dataset and Jisha Raj Bharatanatyam Dataset | Robust against rotation and scaling variations with real-time implementation capability | Depth estimation inaccuracies and dependency on performer precision |
| J. R. Challapalli et al. (2026)[22] | CNN architectures optimized using Flamingo Search metaheuristic algorithm | ICD dataset augmented using GANs along with CIFAR-10/100 | Efficient hyperparameter optimization and improved computational scalability | Limited generalization due to visually homogeneous datasets |
| S. S. and J. M.V. (2022)[23] | Deep Pose Estimator with GRU, 3D-CNN, and CNN-LSTM architectures | 300 HD dance video clips representing seven classes | Efficient handling of spatio-temporal dynamics for real-time inference | Requires high computational resources and GPU-intensive training |
| A. P. Parameshwaran et al. (2020)[5] | Double Transfer Learning using VGG16 and stacked ensemble models | Custom dataset containing 27 single-hand gestures | Effectively addresses data scarcity through progressive transfer learning | Limited robustness due to controlled environment dataset collection |

| Author & Year | Methodology | Dataset | Strengths | Limitations |
|---|---|---|---|---|
| S. Gupta and S. Singh (2024)[24] | Comparative survey of probabilistic, rule-based, geometric, and neural network models | Multiple gesture and dance datasets | Provides extensive comparative analysis of gesture recognition approaches | Limited experimental validation and dependence on existing literature |

## 2.6 Summary

This chapter is all about the reviewing of existing literature that is related to Bharatanatyam mudra recognition, gesture classification systems, deep learning architectures, transfer learning strategies, geometric landmark representations, and explainable artificial intelligence techniques. The review highlighted the transition from handcrafted feature extraction approaches that are traditional and conventional to modern multimodal deep learning frameworks while identifying major research limitations related to robustness, explainability, and cross-domain adaptability. Based on these observations, the next chapter presents the proposed explainable multimodal dual-stage transfer learning framework designed to address the identified research gaps and improve Bharatanatyam mudra recognition performance.

# CHAPTER 3

# PROPOSED WORK

## 3.1 Introduction

This chapter presents the proposed explainable multimodal dualstage transfer learning framework developed for robust Bharatanatyam mudra recognition. The proposed methodology integrates transformer-based visual learning with geometric hand landmark representations to improve classification performance, interpretability, and cross-dataset generalization capability. The framework combines Swin Transformer Tiny[25] for RGB image feature extraction and MediaPipe-based normalized hand landmark analysis[17] for capturing structural articulation patterns of Bharatanatyam mudras.

To address the limited availability of large-scale Bharatanatyam mudra datasets and improve transferability, a dual-stage transsfer learning strategy is employed. The proposed framework first learns generalized gesture representations from an Indian Sign Language dataset before fine-tuning on Bharatanatyam mudra data. Furthermore, cross-dataset evaluation and domain adaptation experiments are conducted using an external Bharatanatyam mudra dataset[14] containing the same mudras performed by different subjects under varying acquisition conditions.

Adding to the classification performance, the proposed framework as shown in Fig 3.1 incorporates explainable artificial intelligence techniques including GradCAM-based visual explanation and landmark sensitivity analysis for interpreting the decision-making behavior of the multimodal architecture[26].

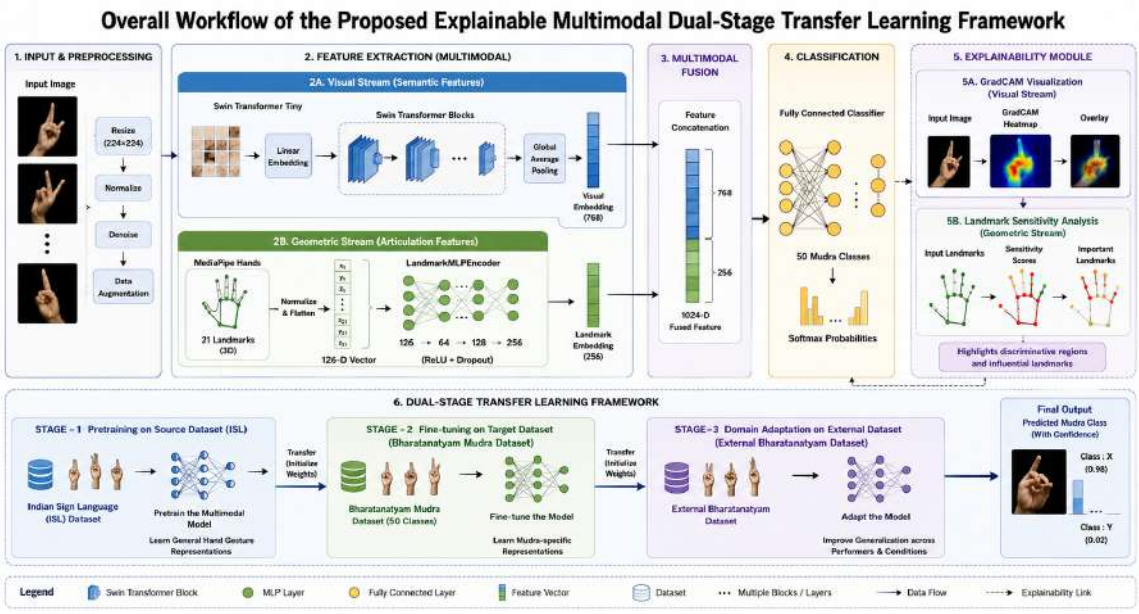



**Figure 3.1:** Overall workflow of the proposed explainable multimodal dual-stage transfer learning framework

## 3.2 Overall Framework

The proposed framework follows a multimodal architecture combining appearance-based visual features and geometric hand articulation features for Bharatanatyam mudra recognition. The full workflow is comprising of the following stages:

1. Dataset preparation
2. Image preprocessing and augmentation
3. Hand landmark extraction using MediaPipe
4. Visual feature extraction using Swin Transformer Tiny
5. Geometric feature encoding using LandmarkMLPEncoder
6. Multimodal feature fusion
7. Gesture classification
8. Explainability analysis

Initially, input Bharatanatyam gesture images are subjected to the preprocessing operations including resizing, normalization, and augmentation. The processed images are then simultaneously forwarded to two parallel branches:

- RGB image branch
- Landmark branch

The RGB branch extracts deep visual representations using Swin Transformer Tiny, while the landmark branch processes normalized three-dimensional hand keypoint coordinates extracted using MediaPipe Hands. The extracted visual and geometric features are fused into a unified multimodal representation, which is subsequently passed through classification layers for final mudra prediction.

The framework additionally incorporates explainability modules to visualize discriminative image regions and identify important hand joints contributing to classification decisions.

## 3.3 Dataset Description

Three datasets as shown in Table 3.1are utilized in the proposed study for transfer learning, Bharatanatyam mudra recognition, and cross-dataset evaluation.

### 3.3.1 Indian Sign Language Dataset

The intermediate transfer learning stage uses an Indian Sign Language (ISL) dataset[27] obtained from Kaggle which contains multiple hand gesture classes representing different ISL signs and is used for learning generalized gesture representations before we do the Bharatanatyam mudra specific fine-tuning using our Bharatantyam Mudra Dataset.

The ISL dataset helps the model to learn the following:

- generalized hand articulation patterns,

- finger configuration structures,
- gesture semantics, and
- visual hand representations.

### 3.3.2 Primary Bharatanatyam Mudra Dataset

The primary Bharatanatyam mudra dataset[28] is used for fine-tuning. This was also used for classification. This was collected as part of doctoral research conducted under the guidance of Dr. Sunil T.T., College of Engineering Attingal, Kerala, India[28]. The dataset contains multiple Bharatanatyam mudra classes represented using static hand gesture images captured under controlled conditions.

The dataset includes variations in:

- hand articulation,
- performer orientation,
- gesture appearance, and
- illumination conditions

### 3.3.3 External Bharatanatyam Mudra Dataset

To evaluate the robustness on cross-dataset and domain adaptability for the same cross-dataset, an external Bharatanatyam mudra dataset[14] is used. This was proposed by Kokul Thanikasalam et al. The external dataset contains the same mudra classes. These are but performed by different subjects under different acquisition conditions.

Unlike the generic gesture transfer evaluation, this cross-dataset analysis performed in our study investigates:

- performer variation,
- acquisition condition variation,
- dataset distribution shift, and
- cross-dataset generalization capability

The external dataset enables evaluation of the proposed framework under non-identical training and testing distributions.

**Table 3.1:** Datasets used in the proposed study

| Dataset | Data Type | Purpose | Role in Framework |
|---|---|---|---|
| Indian Sign Language Dataset | Hand gesture images | Intermediate transfer learning | Learning generalized gesture representations |
| Primary Bharatanatyam Mudra Dataset | Mudra images | Final fine-tuning and classification | Primary mudra recognition |
| External Bharatanatyam Mudra Dataset | Mudra images | Cross-dataset evaluation | Robustness and domain adaptation analysis |

## 3.4 Data Preprocessing

Data preprocessing is performed to improve training stability and enhance model generalization capability. The preprocessing pipeline consists of multiple operations applied uniformly across all datasets as shown in Fig 3.2.

The preprocessing operations include:

- Image resizing
- Pixel normalization
- Data augmentation
- Background standardization

All of the imput images are subjected to undergo the resizing to a fixed resolution that is compatible and works well with the Swin Transformer Tiny architecture. Pixel normalization is applied to standardize intensity distributions across samples.

To reduce overfitting and improve robustness against environmental variations, several augmentation techniques are employed, including:

- rotation,
- horizontal flipping,
- translation,
- zoom transformation, and
- brightness adjustment.

The augmentation process enables the model to learn invariant gesture representations under varying acquisition conditions.

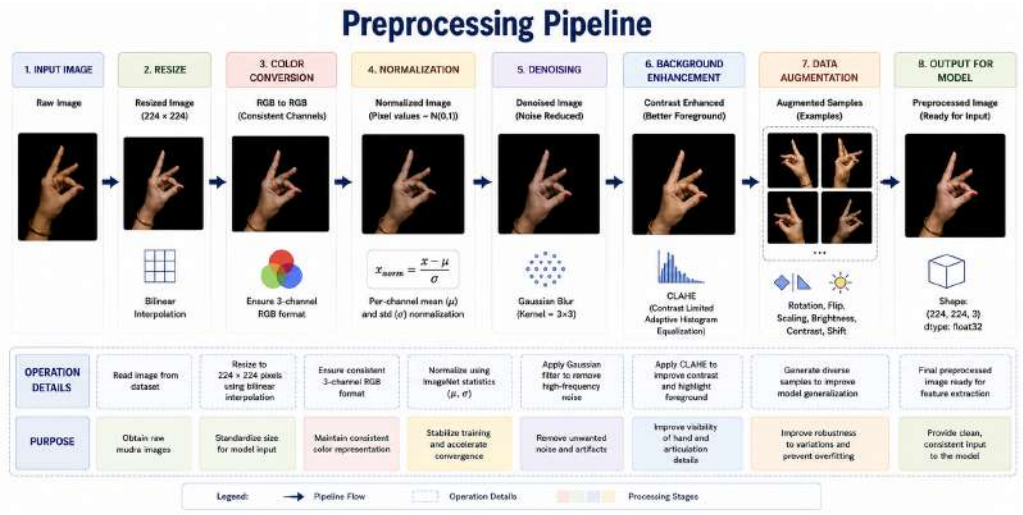



**Figure 3.2:** Image preprocessing and augmentation pipeline

## 3.5 Hand Landmark Extraction

In order to get the information features for geometric hand articulation, MediaPipe Hands is used for the extraction skeletal hand landmarks from input images.

MediaPipe provides 21 hand landmarks. These are then represented using normalized three-dimensional coordinates:

$$(x, y, z)$$

where:

- $x$ and $y$ represent normalized spatial coordinates, and
- $z$ represents relative depth information.

The landmark extraction process consists of:

1. Hand detection,
2. Landmark localization,
3. Coordinate normalization,
4. Feature vector generation.

Each detected hand produces:

$$21 \times 3 = 63$$

features corresponding to the 21 landmarks and their associated three-dimensional coordinates as shown in Fig 3.3.

For two-hand detection scenarios, the coordinates are concatenated to form a fixed-length feature vector of size:

126

For suppose only one hand is detected, in that case zero-padding is applied so that to preserve dimensional consistency across samples.

The resulting normalized coordinate vector is forwarded to the step/stage of the LandmarkMLPEncoder of our architecture to perform geometric representation learning.

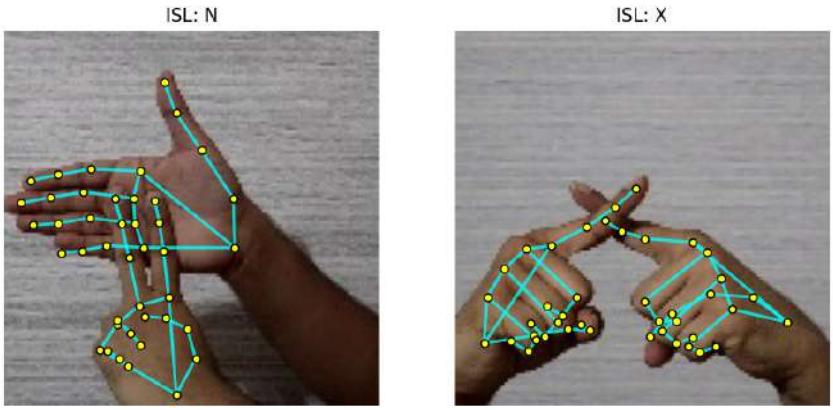



**Figure 3.3:** MediaPipe hand landmark extraction showing 21 skeletal keypoints

## 3.6 Visual Feature Extraction Using Swin Transformer Tiny

The RGB image branch utilizes Swin Transformer Tiny as the primary visual feature extraction backbone, and as discussed earlier this Swin Transformer is a vision transformer that has a hierachical architecture that employs self-attention mechanisms of shifted-window kind for an efficient local as well as that of the global representation learning. Compared to conventional convolutional neural networks, Swin Transformer as shown in Fig 3.4 provides improved capability for modeling long-range dependencies and subtle spatial relationships within gesture images and so this property is particularly important for Bharatanatyam mudra recognition, where fine-grained finger articulation differences significantly influence gesture semantics.

The Swin Transformer Tiny architecture performs:

- patch embedding,
- hierarchical feature extraction,
- window-based self-attention,
- shifted-window attention,and
- feature aggregation.

The extracted visual features capture:

- texture information,
- contour structures,

- finger configurations,and
- visual semantic patterns.

The generated visual embeddings are subsequently passed to the multimodal fusion stage.

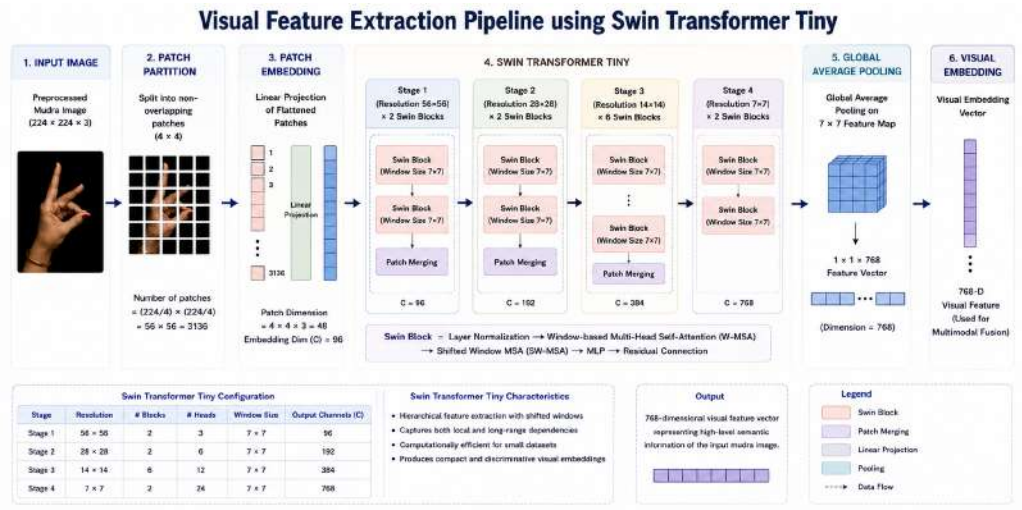



**Figure 3.4:** Visual feature extraction pipeline using Swin Transformer Tiny

## 3.7 LandmarkMLPEncoder

The landmark branch processes normalized geometric landmark coordinates using a multilayer perceptron-based LandmarkMLPEncoder.

The encoder receives the fixed-length 126-dimensional landmark vector generated from MediaPipe extraction and learns compact geometric representations corresponding to hand articulation structures as shown in Fig 3.5.

The LandmarkMLPEncoder enables the framework to learn:

- finger joint relationships,
- spatial articulation patterns,
- geometric gesture structures, and
- relative joint configurations.

The learned geometric embeddings complemented the visual representations extracted by the Swin Transformer branch.

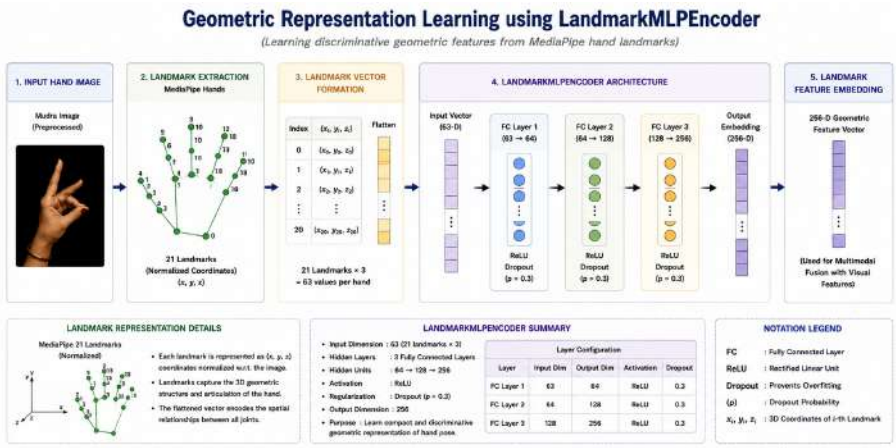



**Figure 3.5:** Geometric representation learning using LandmarkMLPEncoder

## 3.8 Dual-Stage Transfer Learning

To address limited Bharatanatyam dataset availability and improve representation learning, a dual-stage transfer learning strategy as in Fig 3.6 is employed.

Instead of directly fine-tuning an ImageNet-pretrained model on Bharatanatyam mudras, the proposed framework introduces an intermediate gesture adaptation stage using Indian Sign Language data.

The dual-stage transfer learning process consists of:

1. Initial pretraining on ImageNet,
2. Intermediate transfer learning on Indian Sign Language gestures,and
3. Final fine-tuning on Bharatanatyam mudras.

This progressive adaptation strategy enables the model to first learn generalized hand gesture representations before specializing in Bharatanatyam mudra recognition.

The DTL strategy improves:

- feature transferability,
- training stability,
- convergence capability,and
- cross-dataset adaptability.

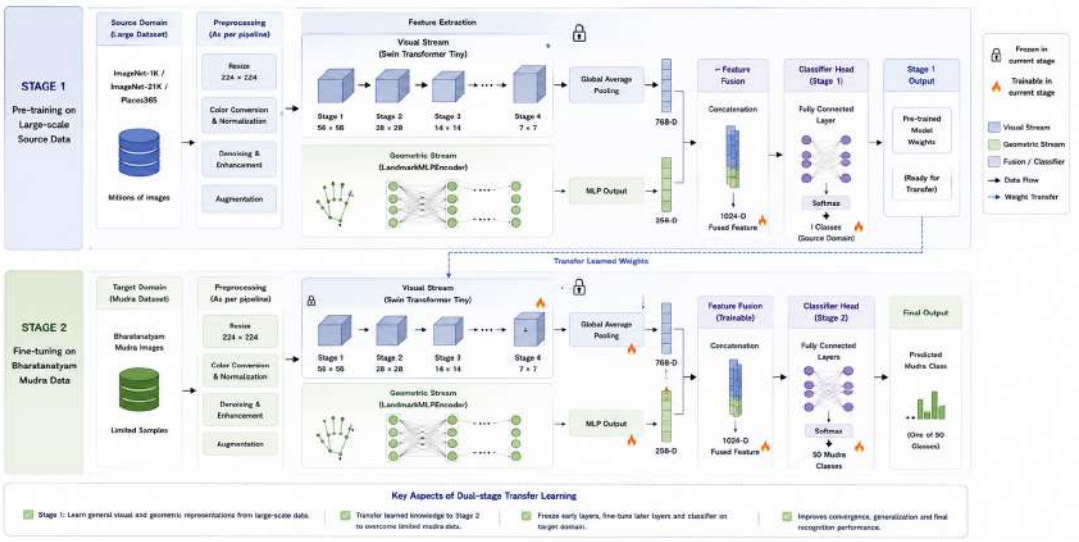



**Figure 3.6:** Dual-stage transfer learning workflow used in the proposed framework

## 3.9 Multimodal Feature Fusion

The proposed framework integrates visual embeddings extracted from Swin Transformer Tiny with geometric embeddings learned using LandmarkMLPEncoder as shown in the Fig 3.7.

This multimodal fusion strategy combines complementary information from:

- appearance-based visual representations
- geometric articulation representations

Visual features capture:

- texture
- contour
- visual semantics

while landmark embeddings capture:

- joint relationships
- spatial articulation
- geometric hand structure

The fused multimodal representation improves robustness against:

- illumination variation
- performer variation
- acquisition differences
- background complexity

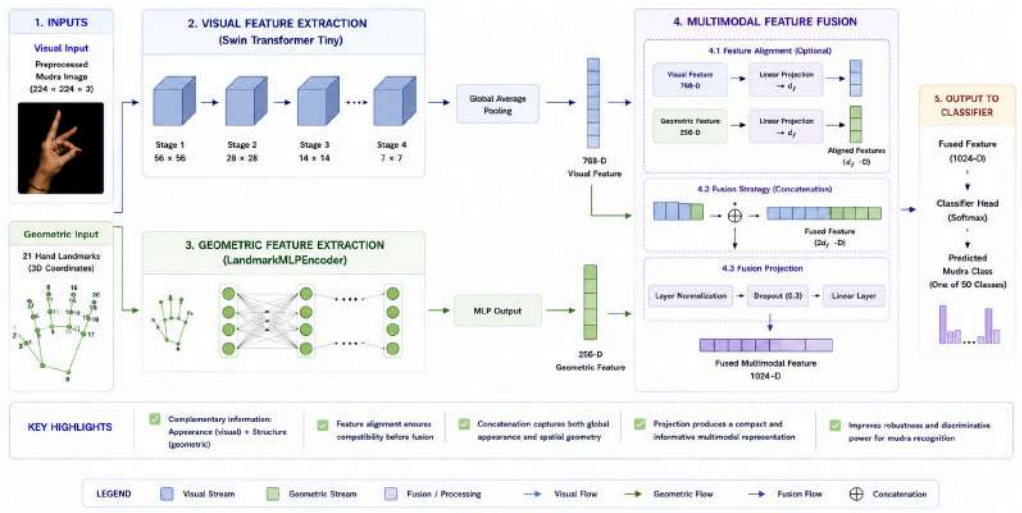



**Figure 3.7:** Multimodal feature fusion between visual and geometric representations

## 3.10 Gesture Classification

The fused multimodal features are processed through fullyconnected classification layers as in Fig 3.8 for final Bharatanatyam mudra prediction.

A softmax activation function is employed to generate probability distributions across all mudra classes.

The framework's optimization is done with categorical cross-entropy loss and gradient-based optimization techniques during training.

The classification stage predicts the most probable Bharatanatyam mudra corresponding to the input gesture image.

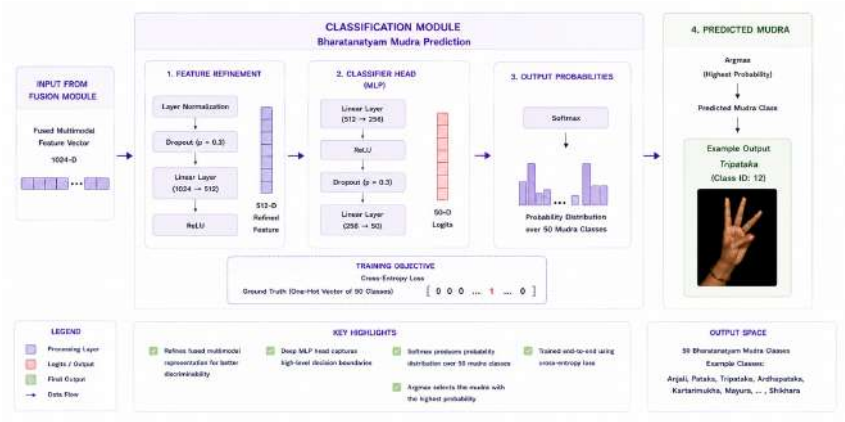



**Figure 3.8:** Classification module for Bharatanatyam mudra prediction

## 3.11 Cross-Dataset Evaluation and Domain Adaptation

To evaluate robustness and generalization capability, cross-dataset experiments are conducted using the external Bharatanatyam mudra dataset.

Two evaluation strategies are employed:

### 3.11.1 Zero-Shot Cross-Dataset Evaluation

The Bharatanatyam-trained model is directly evaluated on the external dataset without additional fine-tuning. This experiment evaluates the inherent cross-dataset generalization capability of the proposed framework under performer and acquisition variation.

### 3.11.2 Domain Adaptation Evaluation

The pretrained Bharatanatyam model is further fine-tuned using the training split of the external dataset and subsequently evaluated on the unseen testing split as in Fig 3.9.

This experiment evaluates the adaptability of the proposed framework under dataset distribution shift conditions.

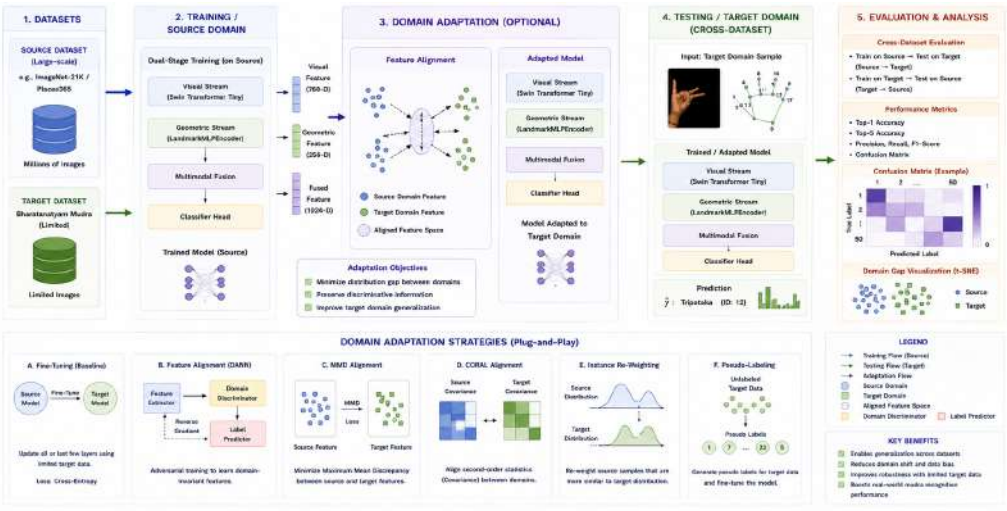



**Figure 3.9:** Cross-dataset evaluation and domain adaptation workflow

## 3.12 Explainability Analysis

To improve transparency and interpretability, the proposed framework incorporates dual-modal explainability analysis.

### 3.12.1 GradCAM-Based Visual Explainability

GradCAM is employed to visualize discriminative image regions contributing significantly to model predictions within the Swin Transformer branch as shown in Fig 3.10.

The generated attention heatmaps help identify:

- important finger regions
- discriminative contours
- gesture-specific visual semantics

### 3.12.2 Landmark Sensitivity Analysis

In addition to GradCAM, landmark sensitivity analysis as in Fig 3.11 is performed using gradient magnitude analysis on the landmark branch.

This analysis identifies:

- influential hand joints
- important articulation regions
- geometric structures affecting classification

The landmark importance patterns are further analyzed with reference to traditional Bharatanatyam Shastra-based gesture semantics.

**Figure 3.10:** GradCAM-based attention visualization highlighting discriminative gesture regions

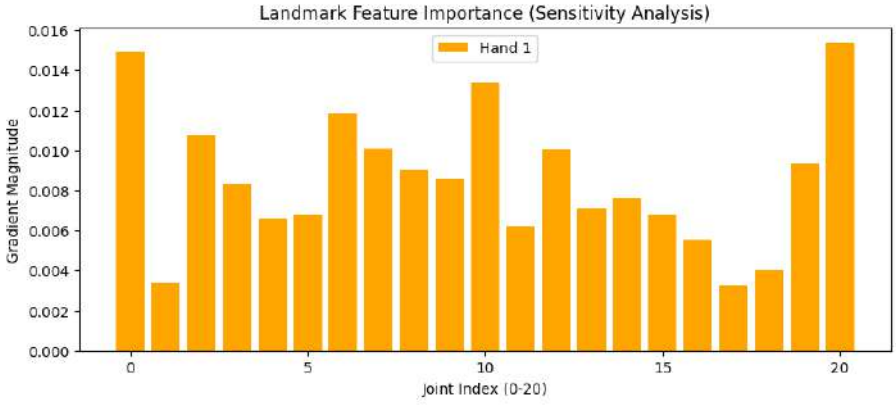



**Figure 3.11:** Landmark sensitivity analysis showing important hand joints influencing classification

## 3.13 Summary

This chapter presented the proposed explainable multimodal dual-stage transfer learning framework developed for Bharatanatyam mudra recognition. The methodology integrates transformer-based visual

learning, geometric landmark representation learning, progressive transfer learning, multimodal feature fusion, cross-dataset evaluation, and explainable artificial intelligence techniques to improve recognition robustness, adaptability, and interpretability.

The next chapter deals on the experimental setup, implementation details, training configuration, and evaluation metrics used for validating the proposed framework.

# CHAPTER 4

# EXPERIMENTAL SETUP

## 4.1 Introduction

This chapter details on the experimental setup and implementation details employed for evaluating the proposed explainable multimodal dual-stage transfer learning framework for Bharatanatyam mudra recognition. The chapter describes the computational environment, dataset preparation, preprocessing pipeline, model architecture, training protocols, evaluation methodologies, and explainability analysis techniques used throughout the study.

The experiments were designed to evaluate:

- Bharatanatyam mudra classification performance
- Cross-dataset generalization capability
- Domain adaptation effectiveness
- Multimodal representation learning
- Explainability and feature interpretability

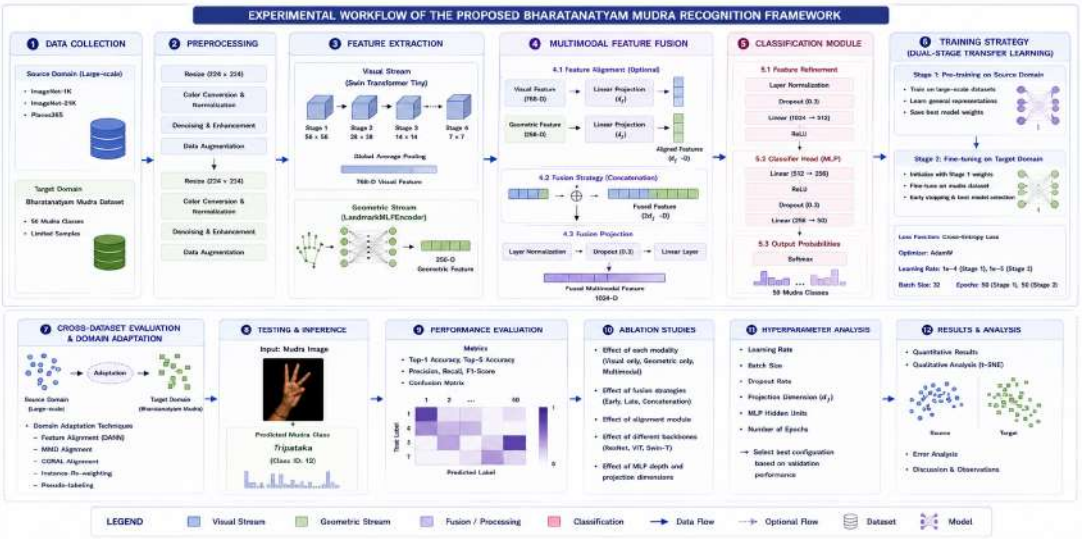



**Figure 4.1:** Experimental workflow of the proposed Bharatanatyam mudra recognition framework

## 4.2 Hardware and Software Environment

All of the experiments were run using the Google Colab environment with GPU acceleration support.

### 4.2.1 Hardware Configuration

The hardware specifications used for model training and evaluation are summarized in Table 4.1.

**Table 4.1:** Hardware configuration used for experimentation

| Component | Specification |
|---|---|
| Platform | Google Colab |
| GPU | NVIDIA T4 GPU |
| GPU Memory | 16 GB |
| Operating System | Linux-based cloud environment |

### 4.2.2 Software Environment

The proposed framework was implemented using Python and the PyTorch deep learning ecosystem. Several computer vision, explainability, and machine learning libraries were utilized throughout the experiments.

The software environment is summarized in Table 4.2.

**Table 4.2:** Software environment and libraries

| Software Component | Version / Library |
|---|---|
| Programming Language | Python 3.9 |
| Deep Learning Framework | PyTorch 2.0.1 + CUDA 11.8 |
| Vision Libraries | Torchvision, OpenCV |
| Transformer Libraries | Transformers, TIMM |
| Landmark Extraction | MediaPipe |
| Machine Learning Utilities | Scikit-learn |
| Visualization Libraries | Matplotlib, Plotly, Seaborn |
| Explainability Libraries | Grad-CAM |
| Feature Analysis Libraries | UMAP-learn, Scikit-image |

## 4.3 Datasets and Data Preparation

Three datasets as shown in Fig 4.3 were utilized in the proposed study for transfer learning, Bharatanatyam mudra classification, and cross-dataset evaluation.

### 4.3.1 Stage-1 Transfer Learning Dataset

The first stage of transfer learning utilized an Indian Sign Language (ISL) dataset obtained from Kaggle. The dataset contains 36 gesture classes representing different ISL hand signs.

The ISL dataset was used to learn generalized hand gesture representations before Bharatanatyam-specific fine-tuning.

The dataset split strategy included:

- 90% training split
- 10% validation split
- Separate held-out test set

The split was performed using `torch.utils.data.random_split()`.

### 4.3.2 Primary Bharatanatyam Mudra Dataset

The primary Bharatanatyam mudra dataset was obtained from the Hugging Face Hub. The dataset consists of 50 Bharatanatyam mudra classes represented using static hand gesture images.

The dataset split strategy was:

- 70% training set
- 15% validation set
- 15% testing set

The split was performed using `Dataset.train_test_split()` with:

```
seed = 42
```

The dataset splitting procedure preserved class distribution across splits through stratified partitioning.

### 4.3.3 External Bharatanatyam Mudra Dataset

To evaluate cross-dataset generalization and domain adaptation capability, an external Bharatanatyam mudra dataset[14] proposed by Kokul Thanikasalam et al. was utilized.

The external dataset contains:

- identical mudra classes
- different performers
- varying acquisition conditions
- different image distributions

This dataset was used for:

- zero-shot cross-dataset evaluation
- domain adaptation experiments

**Table 4.3:** Summary of datasets used in the proposed study

| Dataset | Classes | Dataset Size | Data Type | Purpose |
|---|---|---|---|---|
| Indian Sign Language Dataset | 36 | 42.7K | Gesture images | Stage-1 transfer learning |
| Bharatanatyam Mudra Dataset | 50 | 28,431 | Mudra images | Primary classification |
| External Bharatanatyam Mudra Dataset | 27 | 3,450 | Mudra images | Cross-dataset evaluation |

## 4.4 Image Preprocessing and Augmentation

All input images were resized to:

$$224 \times 224$$

pixels to match the input requirements of the Swin Transformer Tiny architecture.

### 4.4.1 Training-Time Augmentation

To improve robustness and reduce overfitting, multiple augmentation techniques were applied during training as shown in Fig 4.2:

- Random resizing and cropping
- Horizontal flipping
- Color jitter
- Image normalization

Normalization was performed using ImageNet mean and standard deviation values.

### 4.4.2 Validation and Testing Preprocessing

For validation and testing:

- Images were resized to 256 pixels
- Center cropped to 224 pixels
- Normalized using ImageNet statistics

**Figure 4.2:** Image preprocessing and augmentation examples

## 4.5 Hand Landmark Extraction

Hand landmarks were extracted using the MediaPipe HandLandmarker framework as shown in Fig 4.3.

Each detected hand produced:

$$21$$

skeletal landmarks represented using normalized three-dimensional coordinates:

$$(x, y, z)$$

The landmark coordinates were normalized relative to image dimensions.

For each hand:

$$21 \times 3 = 63$$

features were generated.

For two-hand detection:

$$63 \times 2 = 126$$

features were produced.

If only one hand was detected, zero-padding was applied to preserve fixed-dimensional input representation.

Approximately:

$$18\%$$

of Bharatanatyam training samples resulted in failed landmark detection. These cases were handled using zero-vector landmark representations.

The extracted landmark vectors were directly forwarded to the LandmarkMLPEncoder without additional z-score normalization or feature standardization.

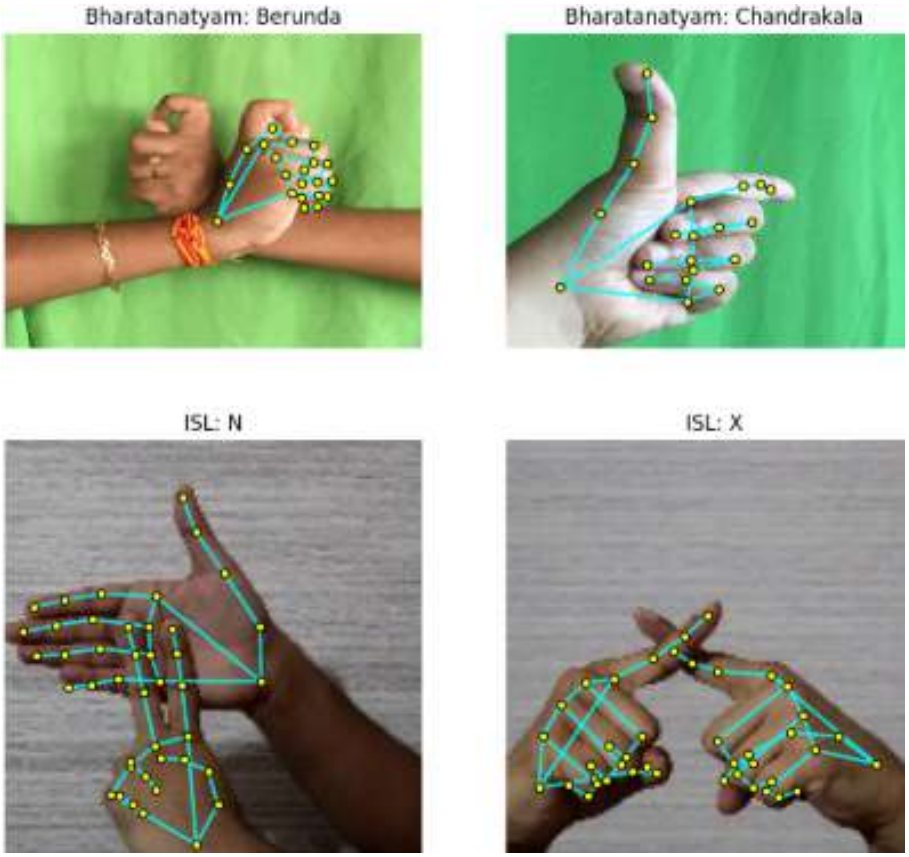



**Figure 4.3:** MediaPipe-based hand landmark extraction process

## 4.6 Model Architecture

The proposed multimodal framework integrates visual and geometric representation learning through two parallel branches as shown in Fig 4.4.

### 4.6.1 Visual Branch

The visual branch utilizes:

`swin_tiny_patch4_window7_224`

initialized using the pretrained Microsoft Swin Transformer[25] Tiny checkpoint.

The Swin Transformer backbone remained fully trainable during all training stages.

The backbone outputs:

768

dimensional visual embeddings obtained after global adaptive average pooling.

### 4.6.2 Landmark Branch

The geometric landmark branch utilizes a lightweight multilayer perceptron called `LandmarkMLPEncoder`.

The architecture follows:

$$126 \rightarrow 64 \rightarrow 128 \rightarrow 256$$

The encoder includes:

- Linear layers,
- ReLU activation, and
- Dropout with probability 0.2

The final output embedding dimension of the landmark branch is:

$$256$$

### 4.6.3 Feature Fusion

The visual and geometric embeddings are concatenated to form a fused multimodal representation:

$$768 + 256 = 1024$$

dimensional fused feature vector.

### 4.6.4 Classifier Head

The fused multimodal representation is directly mapped to the output classes using a single fully connected classification layer:

$$1024 \rightarrow \text{Number of Classes}$$

The final output dimension depends on the active training stage:

- 36 classes for ISL pretraining
- 50 classes for Bharatanatyam mudra recognition

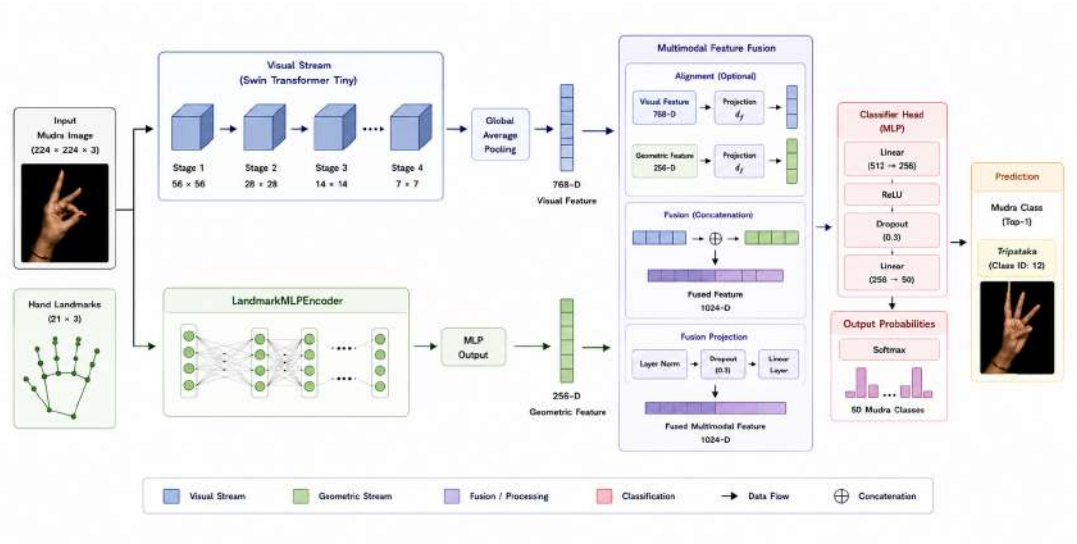



**Figure 4.4:** Architecture of the proposed multimodal Bharatanatyam mudra recognition framework

## 4.7 Training Protocols

### 4.7.1 General Training Configuration

The common training configuration used across all experiments is summarized in Table 4.4.

**Table 4.4:** General training configuration

| **Parameter** | **Value** |
|---|---|
| Batch Size | 32 |
| Optimizer | AdamW |
| Loss Function | CrossEntropyLoss |
| Learning Rate Scheduler | CosineAnnealingWarmRestarts |
| Mixed Precision Training | torch.cuda.amp |
| Model Selection Criterion | Highest validation accuracy |
| Gradient Clipping | Not applied |

### 4.7.2 Stage-1 ISL Pretraining

The first training stage focused on generalized gesture representation learning using the ISL dataset.

The configuration included:

- Epochs: 5
- Learning Rate: $1 \times 10^{-4}$
- Weight Decay: 0.01

### 4.7.3 Stage-2 Bharatanatyam Fine-Tuning

The second stage adapted the pretrained model to Bharatanatyam mudra recognition.

The configuration included:

- Epochs: 10
- Learning Rate: $5 \times 10^{-5}$
- Weight Decay: 0.01

Early stopping was employed using validation accuracy monitoring to reduce overfitting.

### 4.7.4 Stage-3 Domain Adaptation

The final stage performed domain adaptation using the external Bharatanatyam mudra dataset as shown in Fig 4.5.

The configuration included:

- Epochs: 5
- Learning Rate: $1 \times 10^{-5}$
- Weight Decay: 0.05

A conservative learning rate was used to preserve previously learned Bharatanatyam representations while adapting to external dataset variations.

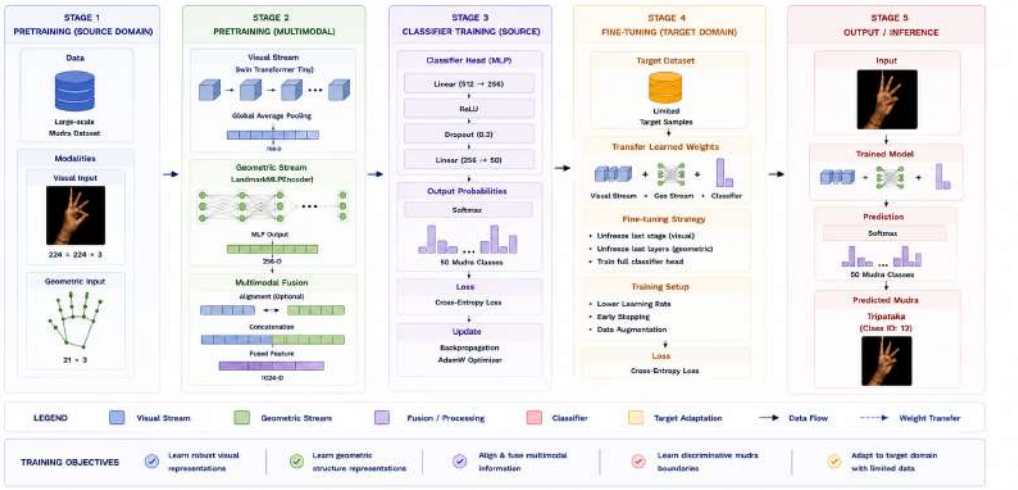



**Figure 4.5:** Multi-stage training pipeline of the proposed framework

## 4.8 Evaluation Metrics

Multiple quantitative metrics were used to evaluate classification performance and feature representation quality.

### 4.8.1 Classification Metrics

The primary evaluation metrics included:

- Accuracy,
- Precision,
- Recall,
- F1-score, and
- Cohen's Kappa Score

Confusion matrices were additionally generated to analyze inter-class misclassification patterns.

### 4.8.2 Cross-Dataset Evaluation

Cross-dataset robustness was evaluated using:

- Zero-shot evaluation
- Domain adaptation evaluation

The evaluation results included:

- Main Bharatanatyam classification accuracy:

  99.31%

- Zero-shot cross-dataset accuracy:

  66.90%

- Domain-adapted cross-dataset accuracy:

  86.11%

### 4.8.3 Feature Space Analysis

To evaluate latent feature separability, clustering metrics were computed on the fused feature embeddings:

- Silhouette Score[43],
- Calinski-Harabasz Index[44],and
- Davies-Bouldin Index[45].

Additionally, t-SNE visualization was employed to analyze feature-space clustering behavior as shown in Fig 4.6 [39].

### 4.8.4 Structural Similarity Analysis

Structural Similarity Index (SSIM)[46] was used to quantify inter-class visual similarity and analyze potential causes of classification confusion.

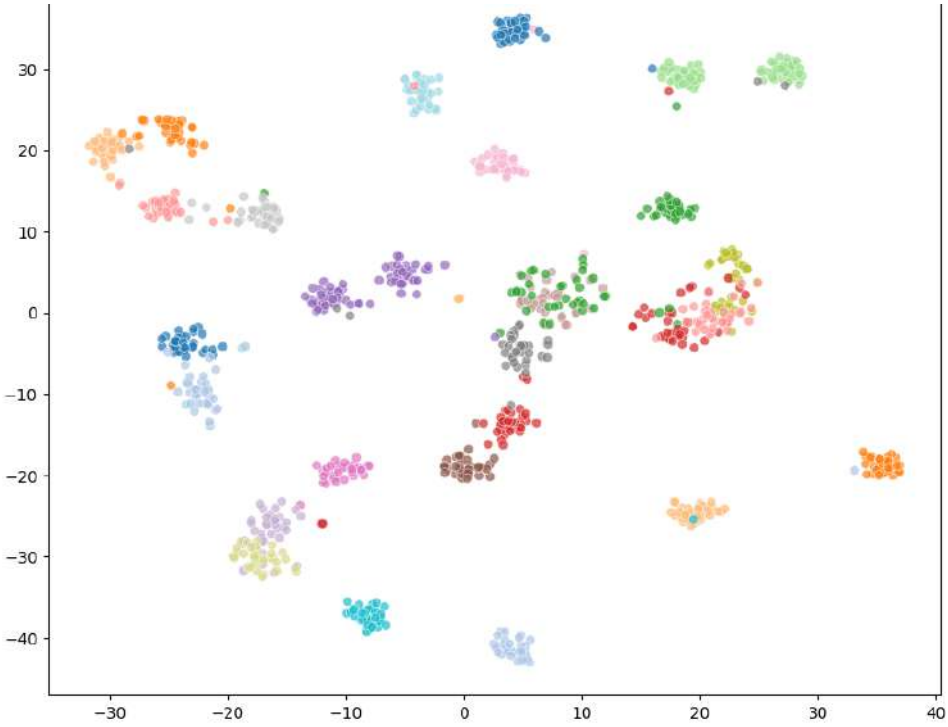



**Figure 4.6:** t-SNE visualization of fused multimodal feature representations

## 4.9 Explainability Analysis

To improve transparency and interpretability, two explainability mechanisms were employed.

### 4.9.1 GradCAM-Based Visual Explainability

GradCAM was applied to the Swin Transformer branch to visualize discriminative image regions contributing significantly to model predictions.

The generated heatmaps highlighted:

- important finger articulations,
- hand contours,and
- gesture-specific visual regions.

### 4.9.2 Landmark Sensitivity Analysis

Gradient-based landmark sensitivity analysis was performed on the landmark branch to identify influential hand joints affecting classification decisions, and the resulting importance patterns were interpreted using Bharatanatyam Shastra-based anatomical terminology for semantic alignment.

## 4.10 Summary

This chapter presented the complete experimental setup used for validating the proposed explainable multimodal dual-stage transfer learning framework for Bharatanatyam mudra recognition. The chapter described the computational environment, dataset preparation strategies, preprocessing pipeline, multimodal architecture, training configurations, evaluation methodologies, and explainability techniques employed throughout the study.

The next chapter presents the experimental results, comparative analysis, cross-dataset evaluation, feature-space analysis, and explainability outcomes obtained using the proposed framework.

# CHAPTER 5

# RESULTS AND DISCUSSION

## 5.1 Introduction

This chapter deals with the experimental results and as well as theperformance analysis of the proposed explainable multimodal dual-stage transfer learning framework for Bharatanatyam mudra recognition. The evaluation focuses on classification performance, cross-dataset generalization capability, domain adaptation effectiveness, latent feature representation quality, and explainability analysis.

The proposed framework was evaluated under multiple experimental settings including:

- Stage-1 gesture pretraining using Indian Sign Language data,
- Bharatanatyam mudra fine-tuning,
- Zero-shot cross-dataset evaluation,
- Domain adaptation evaluation,
- Feature-space clustering analysis,and
- Explainability analysis using GradCAM and landmark sensitivity.

The results demonstrate the effectiveness of multimodal representation learning, progressive transfer learning, and geometric feature integration for robust Bharatanatyam mudra recognition under varying performer and acquisition conditions.

## 5.2 Training Performance Analysis

### 5.2.1 Stage-1 ISL Pretraining Performance

The first stage of training is focused on learning generalized hand gesture representations using the Indian Sign Language dataset where the model demonstrated rapid convergence during the pretraining phase and achieved extremely high validation performance within a limited number of epochs and thus so the final Stage-1 performance metrics included:

- Validation Loss:

  0.0033

- Validation Accuracy:

  99.88%

The results indicate that the proposed multimodal architecture effectively learned generalized hand articulation and gesture representations prior to Bharatanatyam-specific adaptation.

### 5.2.2 Bharatanatyam Fine-Tuning Performance

During Stage-2 fine-tuning, the pretrained model was adapted to Bharatanatyam mudra recognition using the primary Bharatanatyam dataset and the validation performance improved progressively during the initial epochs, indicating successful transfer of generalized gesture knowledge into Bharatanatyam-specific semantic learning as also the highest validation accuracy was achieved during the:

Epoch 7

with:

99.34%

validation accuracy.

The training process demonstrated stable convergence with limited overfitting despite the highly fine-grained nature of Bharatanatyam mudra classification as shown in Table 5.1.

**Table 5.1:** Bharatanatyam fine-tuning performance across epochs

| Epoch | Validation Loss | Validation Accuracy |
|---|---|---|
| 1 | 0.1082 | 97.23% |
| 2 | 0.0669 | 97.93% |
| 3 | 0.0578 | 98.51% |
| 4 | 0.0263 | 99.30% |
| 5 | 0.0375 | 98.55% |
| 6 | 0.0538 | 98.37% |
| 7 | 0.0262 | 99.34% |
| 8 | 0.0721 | 98.11% |
| 9 | 0.0471 | 98.73% |
| 10 | 0.0308 | 99.03% |

The fluctuation observed in later epochs suggests that the model reached convergence relatively early and that extended training introduced mild overfitting behavior as shown in Fig 5.1.

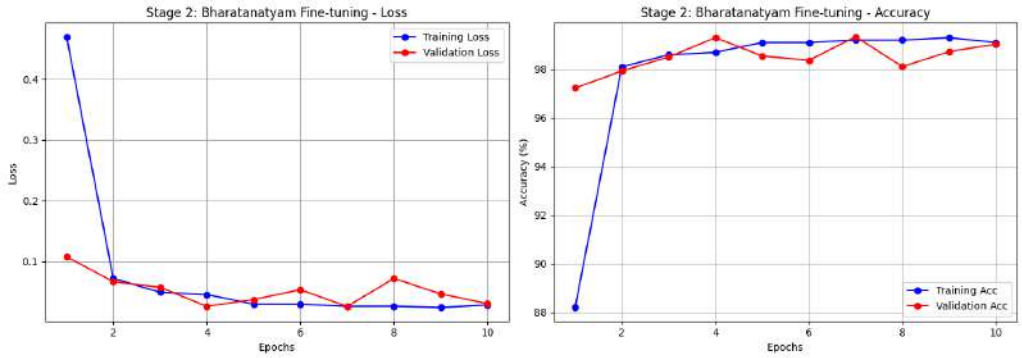



**Figure 5.1:** Training and validation performance during Bharatanatyam fine-tuning

## 5.3 Bharatanatyam Mudra Classification Results

The proposed multimodal framework achieved outstanding classification performance on the Bharatanatyam mudra test dataset.

The final test accuracy achieved was:

99.31%

across:

50

Bharatanatyam mudra classes.

The classification report demonstrated consistently high precision, recall, and F1-scores across most mudra categories as in Table 5.2.

**Table 5.2:** Overall Bharatanatyam mudra classification performance

| **Metric** | **Value** |
|---|---|
| Test Accuracy | 99.31% |
| Macro Precision | 0.99 |
| Macro Recall | 0.99 |
| Macro F1-Score | 0.99 |
| Weighted F1-Score | 0.99 |
| Number of Test Samples | 5687 |

Several mudra classes including:

- Anjali
- Garuda
- Kapotham
- Katakavardhana
- Mukulam
- Shanka
- Swastikam

achieved near-perfect classification performance.

The high accuracy as evident in Fig 5.2 demonstrates the effectiveness of combining transformer-based visual representation learning with geometric landmark encoding for fine-grained Bharatanatyam mudra classification.

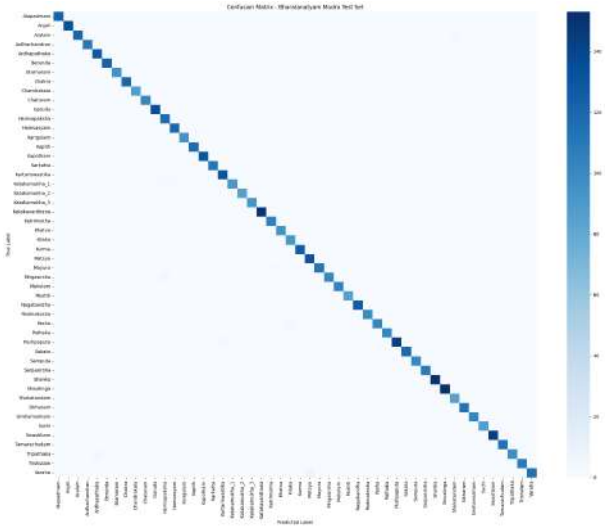


**Figure 5.2:** Confusion matrix for Bharatanatyam mudra classification

## 5.4 Cross-Dataset Generalization Analysis

### 5.4.1 Zero-Shot Cross-Dataset Evaluation

To evaluate robustness under performer and acquisition variation, the trained Bharatanatyam model was directly evaluated on the external Bharatanatyam mudra dataset without additional fine-tuning.

The zero-shot cross-dataset evaluation achieved:

66.90%

classification accuracy.

The significant reduction in performance compared to the in-domain Bharatanatyam test accuracy indicates the presence of:

- dataset distribution shift
- performer variation
- articulation variability
- acquisition condition differences

Despite the performance drop, several mudra classes maintained strong recognition performance, indicating partial transferability of the learned multimodal representations.

However, specific classes including:

- Aralam
- Bramaram
- Chaturam
- Katrimukha
- Trishulam

demonstrated substantial degradation under cross-dataset evaluation conditions.

**Table 5.3:** Zero-shot cross-dataset evaluation performance

| **Metric** | **Value** |
|---|---|
| Cross-Dataset Accuracy | 66.90% |
| Macro Precision | 0.75 |
| Macro Recall | 0.67 |
| Macro F1-Score | 0.67 |
| External Test Samples | 3335 |

The results as in Table 5.3 reveal that high in-domain accuracy alone does not guarantee robust generalization under performer and dataset variations.

### 5.4.2 Domain Adaptation Performance

To improve cross-dataset robustness, the pretrained Bharatanatyam model underwent additional domain adaptation using the training split of the external Bharatanatyam dataset.

Following adaptation which is depicted in training curves in Fig 5.3, the model achieved:

86.11%

accuracy on the unseen external testing set.

This represents a substantial improvement over the zero-shot evaluation setting as depicted in Table 5.4.

**Table 5.4:** Performance comparison between zero-shot and domain-adapted evaluation

| Evaluation Strategy | Accuracy |
|---|---|
| Zero-Shot Cross-Dataset Evaluation | 66.90% |
| Domain Adapted Evaluation | 86.11% |

The domain adaptation process enabled the framework to better accommodate:

- performer-specific articulation
- acquisition variability
- lighting differences
- dataset distribution shift

The results strongly validate the effectiveness of transfer learning and multimodal feature fusion for cross-dataset Bharatanatyam mudra recognition.

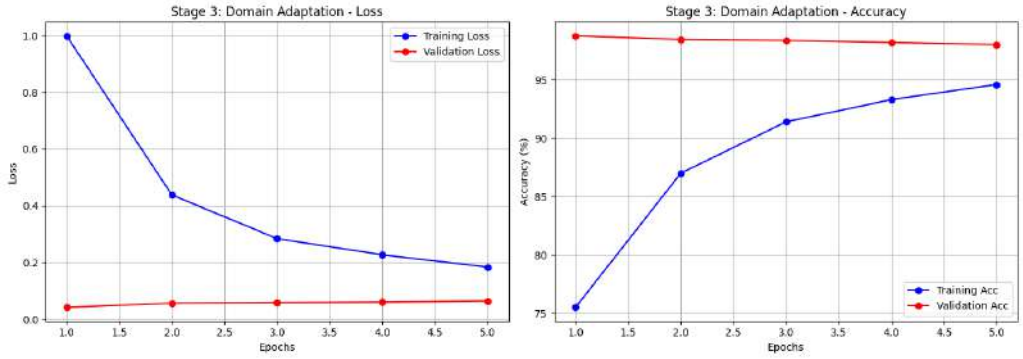



**Figure 5.3:** Training performance during domain adaptation

## 5.5 Feature Space Analysis

To analyze latent representation quality, clustering metrics were computed on the fused multimodal embeddings as shown in table 5.5.

**Table 5.5:** Feature-space clustering analysis

| Stage | Silhouette Score | Calinski-Harabasz Index | Davies-Bouldin Index |
|---|---|---|---|
| Stage-2 Fine-Tuned | 0.3611 | 49.0214 | 1.1793 |
| Stage-3 Domain Adapted | 0.1897 | 53.9176 | 1.9254 |

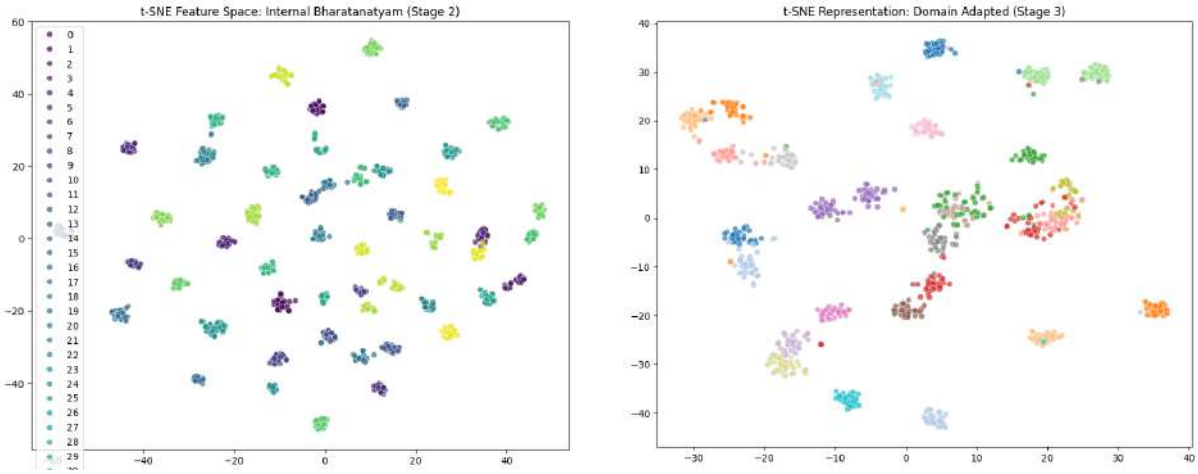
The Stage-2 fine-tuned model demonstrated stronger cluster compactness and separation, indicated by:

- higher silhouette score,and
- lower Davies-Bouldin index.

In contrast, the domain-adapted model exhibited more overlapping feature distributions, suggesting broader feature generalization at the expense of cluster compactness.

This behavior indicates a trade-off between:

- compact in-domain representation learning,and
- robust cross-dataset adaptability.




**(a)** For Primary Dataset **(b)** For Cross Dataset

**Figure 5.4:** t-SNE visualization of feature embeddings for different datasets used in the proposed framework

The t-SNE visualizations as shown in Fig 5.4 further support this observation, where the fine-tuned model exhibits relatively compact and well-separated mudra clusters, while the domain-adapted model demonstrates increased overlap among certain gesture classes.

## 5.6 Structural Similarity Analysis

Structural Similarity Index (SSIM) analysis was performed to investigate the visual similarities between Bharatanatyam mudra classes and the SSIM heatmaps as shown in Fig 5.5 revealed that several visually similar mudras shared overlapping structural characteristics, contributing to cross-class confusion during classification and also the analysis further indicated that subtle finger articulation difference plays a significant role in distinguishing the semantically similar Bharatanatyam mudras.

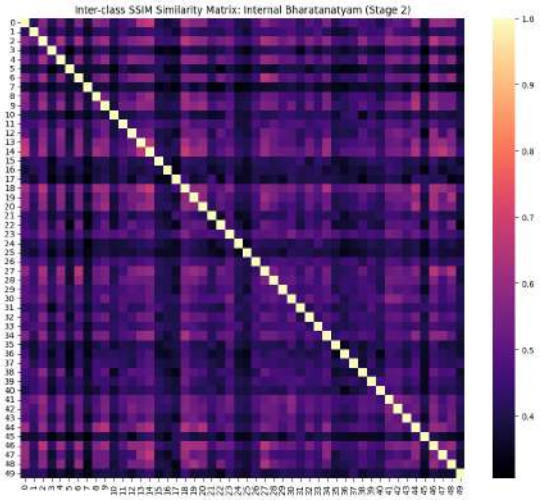



**Figure 5.5:** SSIM heatmap showing structural similarity among Bharatanatyam mudra classes

The SSIM analysis supports the necessity of combining:

- visual semantic learning,and
- geometric articulation encoding

for robust mudra recognition.

## 5.7 Explainability Analysis

### 5.7.1 GradCAM-Based Visual Explainability

GradCAM was employed to visualize the attention regions contributing significantly to the model's predictions.

The generated attention maps revealed that the Swin Transformer branch consistently focused on:

- finger articulation regions,
- fingertip structures,
- palm contours,and
- discriminative gesture boundaries

rather than irrelevant background regions.

This indicates that the proposed framework learned semantically meaningful gesture representations instead of memorizing dataset-specific artifacts.

### 5.7.2 Landmark Sensitivity Analysis

In addition to GradCAM visualization, landmark sensitivity analysis as shown in Fig 5.6 was conducted to identify influential hand joints contributing to classification decisions.

The analysis demonstrated that:

- fingertip landmarks,
- finger articulation joints,
- thumb positioning,and
- inter-finger spatial relationships

played significant roles in distinguishing Bharatanatyam mudras.

The landmark importance patterns aligned closely with traditional Bharatanatyam Shastra-based interpretations of mudra articulation.

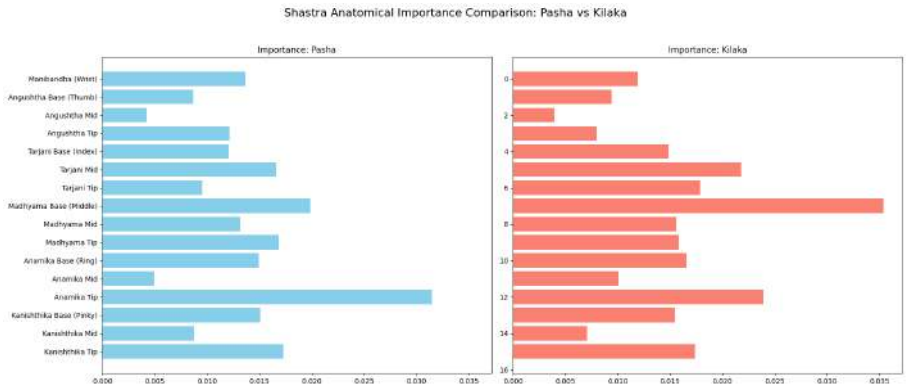



**Figure 5.6:** Landmark sensitivity analysis showing influential hand joints

The explainability analysis validates that the proposed framework learned anatomically and semantically meaningful Bharatanatyam gesture representations.

## 5.8 Discussion

The experimental results demonstrate that the proposed multimodal dual-stage transfer learning framework effectively combines visual and geometric representation learning for Bharatanatyam mudra recognition.

Several important observations can be made from the results:

- The Stage-1 ISL pretraining successfully enabled generalized gesture representation learning,
- Multimodal fusion between Swin Transformer embeddings and geometric landmark representations significantly improved fine-grained gesture discrimination,
- The proposed framework achieved outstanding in-domain Bharatanatyam classification accuracy of 99.31%,
- Zero-shot cross-dataset evaluation revealed substantial dataset distribution and performer variation challenges,
- Domain adaptation substantially improved external dataset generalization performance from 66.90% to 86.11%,
- Feature-space analysis demonstrated a trade-off between compact cluster separation and generalized adaptabilit, and

- Explainability analysis confirmed that the model focused on semantically meaningful Bharatanatyam articulation structures.

The results collectively demonstrate the effectiveness of combining:

- transformer-based visual learning,
- geometric landmark encoding,
- dual-stage transfer learning,and
- explainable artificial intelligence.

for robust Bharatanatyam mudra recognition under varying performer and acquisition conditions.

## 5.9 Summary

This chapter presented the experimental results and performance analysis of the proposed explainable multimodal dual-stage transfer learning framework for Bharatanatyam mudra recognition as the framework achieved strong classification performance, substantial cross-dataset adaptability, meaningful feature-space representations, and interpretable decision-making behavior through multimodal explainability analyssis.

The next chapter presents the conclusion, research contributions, limitations, and future research directions of the proposed study.

# CHAPTER 6

# CONCLUSION AND FUTURE WORK

## 6.1 Conclusion

This thesis presented an explainable multimodal dual-stage transfer learning framework for Bharatanatyam mudra recognition by integrating transformer-based visual representation learning with geometric hand landmark encoding. The proposed framework combined Swin Transformer Tiny for visual feature extraction and MediaPipe-based landmark representations for capturing fine-grained hand articulation semantics.

The study explored a progressive transfer learning strategy in which the model first learned generalized gesture representations using an Indian Sign Language dataset before being adapted for Bharatanatyam mudra recognition. The framework was further evaluated under cross-dataset conditions to analyze robustness against performer and acquisition variations. In addition to classification performance, explainability techniques such as GradCAM and landmark sensitivity analysis were incorporated to improve interpretability and semantic understanding of the learned gesture representations.

The overall findings indicate that combining visual semantic representations with geometric articulation features improves the capability of deep learning models to recognize complex Bharatanatyam mudras. The proposed framework also demonstrated the importance of multimodal representation learning and transfer learning for cultural gesture understanding tasks. Furthermore, the explainability analysis confirmed that the framework focused primarily on meaningful finger articulation and hand structure regions rather than irrelevant background features.

Although the proposed framework achieved strong performance, certain challenges remain. The current work primarily focuses on static mudra recognition and does not explicitly model temporal motion dynamics present in continuous dance sequences. In addition, variations in performer articulation and acquisition conditions continue to influence cross-dataset generalization capability. The framework also depends on the quality of landmark extraction, which may be affected under complex hand poses and partial occlusion conditions.

The proposed work contributes toward the development of robust and explainable artificial intelligence frameworks for Bharatanatyam mudra recognition and demonstrates the potential of multimodal deep learning for cultural heritage preservation and intelligent dance analysis applications.

## 6.2 Future Scope

Future research can extend the proposed framework toward video-based Bharatanatyam analysis incorporating temporal gesture modeling and spatio-temporal transformer architectures for capturing dynamic dance movements and gesture transitions. Additional improvements may be achieved through

advanced domain adaptation strategies, self-supervised representation learning, and lightweight deployment frameworks for real-time cultural AI applications. Furthermore, integrating deeper semantic knowledge from Bharatanatyam Shastra and choreography analysis may contribute toward more culturally grounded, interpretable, and semantically aware Bharatanatyam understanding systems.

# REFERENCES


[1] M. R. Reshma, B. Kannan, V. P. J. Raj, and S. Shailesh, "Cultural heritage preservation through dance digitization: A review," *Digital Applications in Archaeology and Cultural Heritage*, vol. 28, p. e00257, 2023.

[2] R. Amrutha and V. M. Ladwani, "Bharatanatyam hand gesture recognition using normalized chain codes and oriented distances," in *Proc. Int. Conf. Inventive Computation Technologies (ICICT)*, vol. 3, 2016, pp. 1–6.

[3] A. D. Naik and M. Supriya, "Classification of Indian Classical Dance Images using Convolution Neural Network," in *Proc. Int. Conf. Communication and Signal Processing (ICCSP)*, 2020, pp. 1245–1249.

[4] D. P. Akarsha, B. Monisha, K. L. Bhoomika, and D. R. V., "Bharatanatyam Mudra Classification using CNN," in *Proc. 1st Int. Conf. Software, Systems and Information Technology (SSITCON)*, 2024, pp. 1–6.

[5] A. P. Parameshwaran, H. P. Desai, R. Sunderraman, and M. Weeks, "Transfer Learning for Classifying Single Hand Gestures on Comprehensive Bharatanatyam Mudra Dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2019, pp. 1–3.

[6] A. P. Parameshwaran, H. P. Desai, M. Weeks, and R. Sunderraman, "Unravelling of Convolutional Neural Networks through Bharatanatyam Mudra Classification with Limited Data," in *Proc. 10th Annual Computing and Communication Workshop and Conference (CCWC)*, 2020, pp. 342–347.

[7] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2005, pp. 886–893.

[8] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[9] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.

[10] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[11] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[12] P. Langley, W. Iba, and K. Thompson, "An Analysis of Bayesian Classifiers," in *Proc. Nat. Conf. Artificial Intelligence (AAAI)*, 1992, pp. 223–228.

[13] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[14] K. Thanikasalam, A. Ramanan, and P. Kanmanirajah, "A comprehensive review and ensemble CNN approach for Bharatanatyam single-hand gesture classification," *Entertainment Computing*, vol. 56, p. 101069, 2026.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2012, pp. 1097–1105.

[16] C. Sarmah and P. Sarma, "A dataset of Sattriya dance: Classical dance of Assam," *Data in Brief*, vol. 52, p. 109878, 2024.

[17] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "MediaPipe Hands: On-device Real-time Hand Tracking," arXiv:2006.10214, 2020.

[18] Z. Cao, G. Hidalgo Martinez, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, 2021.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 248–255.

[20] K. Adalarasu, R. M. K. Chetty, K. G. Begum, S. Harini, and M. Janardhanan, "An Explainable Machine Learning (XAI) framework for classification of intricate dancing posture among Indian Bharatanatyam dancers," *Applied Soft Computing*, vol. 171, p. 112817, 2025.

[21] S. Paul, G. Sagar, P. P. Das, and K. S. Rao, "Two-stage pipeline based robust hand gesture recognition from Bharatanatyam dance images," *Multimedia Tools Appl.*, vol. 84, pp. 39667–39691, 2025.

[22] J. R. Challapalli, R. Durgam, B. L. Nandipati, and P. Malavath, "Intelligent fine-tuning of convolutional neural networks using flamingo search for traditional dance classification," *Systems and Soft Computing*, vol. 8, p. 200449, 2026.

[23] S. Shailesh and M. V. Judy, "Understanding dance semantics using spatio-temporal features coupled GRU networks," *Entertainment Computing*, vol. 42, p. 100484, 2022.

[24] S. Gupta and S. Singh, "Indian dance classification using machine learning techniques: A survey," *Entertainment Computing*, vol. 50, p. 100639, 2024.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10012–10022.

[26] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.

[27] S. Kumar, "Indian Sign Language Dataset," Kaggle, 2023. [Online]. Available: `https://www.kaggle.com/datasets/saurabh24999/indian-sign-language/data`. [Accessed: May 20, 2026].

[28] R. J. Raj, S. Dharan, and T. T. Sunil, "Optimal feature selection and classification of Indian classical dance hand gesture dataset," *The Visual Computer*, vol. 39, pp. 4049–4064, 2023.

[29] G. K. P. K. Nambiar, N. M. Kumar, V. G., and M. G. Thushara, "AI-Powered Bharatanatyam Mudra Identification and Description System: Preserving Heritage through Technology," in *Proc. 12th Int. Conf. Computing for Sustainable Global Development (INDIACom)*, 2025, pp. 1–7.

[30] S. U. Shetty, R. Puthran, and A. P. Shetty, "Asamyuta Hasta Mudras Recognition Using MediaPipe Keypoint Extraction and Random Forest Classification," in *Proc. Int. Conf. Artificial Intelligence and Data Engineering (AIDE)*, 2025, pp. 173–181.

[31] R. R. Subramanian et al., "Automated Real-Time Hand Gesture Detection for Bharatanatyam Dance," in *Proc. Int. Conf. Computational Robotics, Testing and Engineering Evaluation (ICCRTEE)*, 2025, pp. 1–6.

[32] P. Sadhana, N. Ravishankar, and S. Palaniswamy, "Bharatanatyam Mudra Recognition Using Deep Learning and Meta-Learning Techniques," in *Proc. Int. Conf. Communications and Computer Science (InCCCS)*, 2024, pp. 1–6.

[33] A. S. Nandeppanavar, S. S. Kallur, V. A. Sankannavar, and P. Thotad, "Bharatanatyam hasta mudra categorization using deep learning approaches," in *Proc. IEEE North Karnataka Subsection Flagship Int. Conf. (NKCon)*, 2023, pp. 1–6.

[34] D. Kilari and K. K. Singh, "Comparative study on the effect of HSV Segmentation and ORB Features on Transfer Learning models for recognition of Bharatanatyam Asamyukta Mudras," in *Proc. Int. Conf. Computational Intelligence, Communication Technology and Networking (CICTN)*, 2023, pp. 241–245.

[35] S. Haridas and V. R. Bai, "Detection and Classification of Indian Classical Bharathanatyam Mudras Using Enhanced Deep Learning Technique," in *Proc. Int. Conf. Innovations in Science and Technology for Sustainable Development (ICISTSD)*, 2022, pp. 18–23.

[36] P. Chavan, P. Choudhary, S. Upadhyay, S. Devarshi, S. Sureliya, and A. Kumar, "EfficientNetB0-based Feature Extraction for Single and Double-Hand Mudra Recognition," in *Proc. 5th Int. Conf. Sentiment Analysis and Deep Learning (ICSADL)*, 2026.

[37] V. Amrutha Raj and G. Malu, "EnGesto: An Ensemble Learning Approach for Classification of Hand Gestures," *IEEE Access*, vol. 12, pp. 85709–85723, 2024.

[38] S. Baskar, W. J. Hans, A. V. R., V. S. Solomif, and A. R., "MudraGyaan: A Novel Feature Extraction Algorithm for Machine Learning-Based Bharatanatyam Mudra Classification," in *Proc. Int. Conf. Advancement in Renewable Energy and Intelligent Systems (AREIS)*, 2024.

[39] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[40] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Int. Conf. Learn. Representations (ICLR)*, 2019.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[42] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Int. Conf. Learn. Representations (ICLR)*, 2021.

[43] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[44] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[45] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[46] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[47] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.