A PROJECT REPORT

on

"IMAGE PROCESSING STRATEGIES ON ENSEMBLE MODELS"

Submitted to

KIIT Deemed to be University

In Partial Fulfilment of the Requirement for the Award of

BACHELOR'S DEGREE IN COMPUTER SCIENCE ENGINEERING

BY

KAUSHIK ROY	22051861
MALAY	22051863
ANVESH CHANDRAKAR	22052010
GARIMA KHURANA	22052024

UNDER THE GUIDANCE OF

Dr. Krishnendu Maity

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA - 751024

November, 2025

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, ODISHA 751024

CERTIFICATE

This is certify that the project entitled

"IMAGE PROCESSING STRATEGIES ON ENSEMBLE MODELS"

submitted by

KAUSHIK ROY	22051861
MALAY	22051863
ANVESH CHANDRAKAR	22052010
GARIMA KHURANA	22052024

is a record of bonafide work carried out by them, in the partial fulfilment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering) at KIIT Deemed to be university, Bhubaneswar. This work is done during year 2025, under our guidance.

Date:06/11/2025

Dr. Krishnendu Maity
Project Guide

Acknowledgements

We are profoundly grateful to Dr. Krishnendu Maity for his expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion.

KAUSHIK ROY
MALAY
ANVESH CHANDRAKAR
GARIMA KHURANA

ABSTRACT

This work presents a comprehensive evaluation of three hybrid deep-learning ensembles using the NIH Malaria Dataset for malaria cell image classification. We combine ResNet50 + Vision Transformer, DenseNet121 + Swin Transformer, and EfficientNetB3 + ConvNeXt-Tiny as complementary convolutional and transformer-based feature extractors to build high-capacity dual-backbone architectures. Each backbone pair is partially fine-tuned and used as a feature encoder, after which the embeddings extracted are fused by Logistic Regression and a Fully Connected Network to obtain final predictions.

Two augmentation regimes were investigated to increase generalization and robustness: (1) a Mixture-of-Augmentations layer that randomly applies one of five transformations: rotation–zoom, color jittering, flipping, Gaussian noise injection, and random crop–pad, and

(2) MixUp combines both input images and their soft-label distributions. Most importantly, augmentation is applied only to the training split, while both validation and test sets remain untouched to keep the evaluation unbiased. Collectively, across three ensembled combinations, augmentation managed to significantly improve stability and reduce overfitting, with MixUp producing the highest gains in feature-level discriminative power. The Swin-DenseNet ensemble produced the strongest fused embeddings in all instances, while MixUp-based training improved classification margins without degrading class separation between infected and uninfected cells. These results demonstrate that fusing the outputs of heterogeneous backbones with state-of-the-art augmentation and fusion techniques yields high-accuracy, well-calibrated, robust models of malaria detection, outperforming single-architecture baselines.

CONTENTS

Standards Adopted

5.1 Design Standards

5.2 Testing Standards

Conclusion and Future Scope

6.1 Conclusion

- Summary of key findings

6.2 Future Scope

- Potential improvements and extensions

6.3 Recommendations

Chapter 1

1.1 Importance of Image Processing in Classification

Image processing is crucial for classification tasks since it regularizes images and improves meaningful features. Besides, medical images in malaria cell classification usually suffer from illumination variations, quality variations, and orientational variations. Preprocessing guarantees uniformity and hence reduces noise while accentuating cell structures, which are supposed to be learned by models. It thus forms the backbone of any good feature extraction and enhances accuracy and reliability in classification.

1.2 Unlocking Ensemble Potentials

Ensemble learning combines a diverse set of models to capture higher-order representations for the same image. CNNs excel at local textures, and transformer models capture global patterns and context. However, by fusing these extracted features, the ensemble becomes more robust and overfitting is reduced to improve prediction consistency. This hybrid model gives better performance compared with individual models and allows stronger generalization on unseen microscopy images.

1.3 Data Augmentation

Data augmentation improves dataset diversity and also prevents overfitting in models. Two such methods were implemented:

Mixture of Augmentation: including rotations, brightness changes, flips, noise, and random cropping to simulate realistic variations in cell images.

MixUp combines two images and their labels to smooth decision boundaries and improve generalization.

These augmentations provide robustness to the model against real-world image variations.

1.4 Combinations Used

Three ensemble combinations were explored:

- ResNet50 + Vision Transformer: This combines CNN texture learning with global transformer attention.
- DenseNet121 + Swin Transformer: dense connectivity plus hierarchical windowed attention.
- EfficientNetB3 + ConvNeXt-Tiny - are modern CNN architectures that balance efficiency and strong performance.

Each model extracts complementary features that are fused to improve classification accuracy.

1.5 Motivation and Contribution

The motivation of this work is the improvement of malaria detection accuracy, using automated image-based classification. Manual diagnosis is slow and prone to errors; deep learning can provide fast and consistent results.

Contributions include:

Designing three dual-backbone ensemble models.

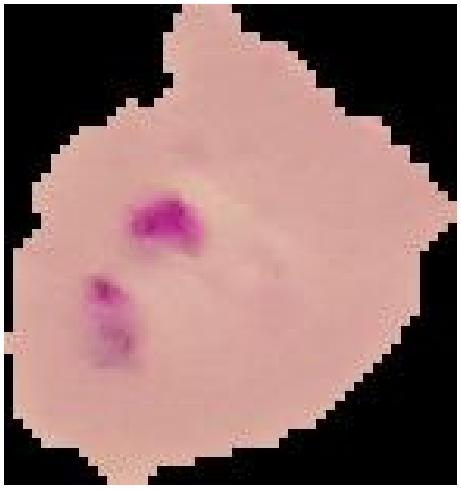
Integration of advanced augmentations MixUp and Mixture-Augmentation. Using feature fusion and evaluating multiple fusion strategies. Showing better accuracy, robustness, and calibration compared to single models.

Chapter 2

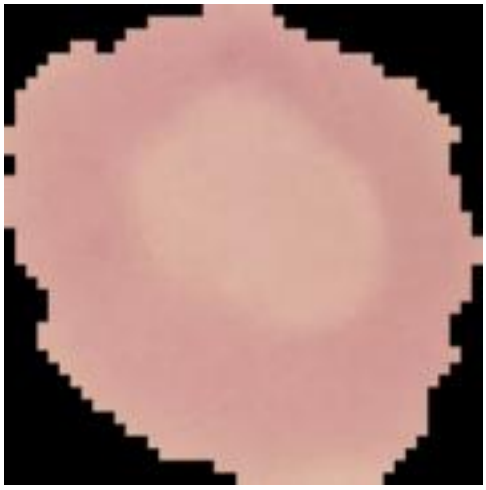
2.1 Dataset Source

Dataset is taken from the official NIH Website: https://ceb.nlm.nih.gov/repositories/malaria-datasets/

2.2 Dataset Sample



Infected Sample



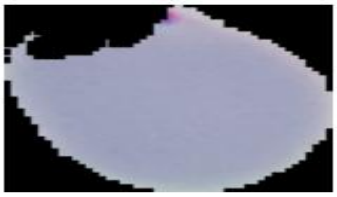
Uninfected Sample

2.3 Dataset Information

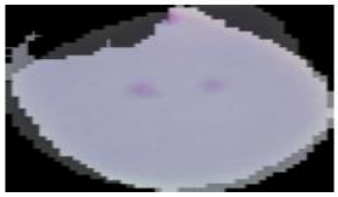
The dataset contains around 27559 images , half infected and rest uninfected images . The blue/purple dots show the infected virus in the blood sample, we can see in the infected sample.Unifected sample does not contain any dots or uniformity . The images are pixelated in order for the model to analyze it.

2.4 Augmented Samples

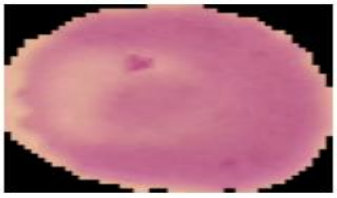
Real (y=0)



Mixup (argmax y=0)
Mixed with idx 4

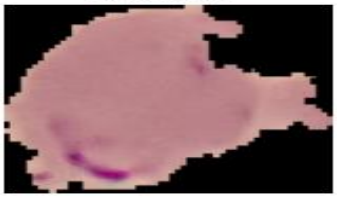


Real (y=1)

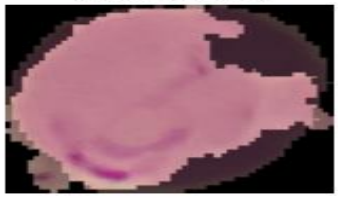


Mixup (argmax y=1)
Mixed with idx 3

Real (y=0)



Mixup (argmax y=0)
Mixed with idx 5



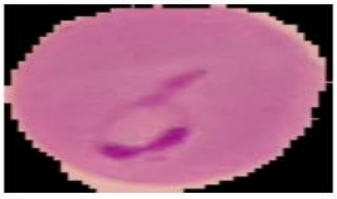
Real (y=1)

Mixup (argmax y=1)
Mixed with idx 11

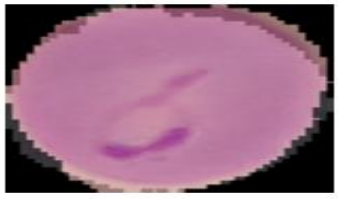
Real (y=0)

Mixup (argmax y=0)
Mixed with idx 1

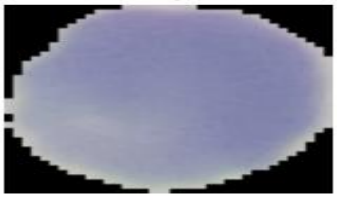
Real (y=0)



Mixup (argmax y=0)
Mixed with idx 15

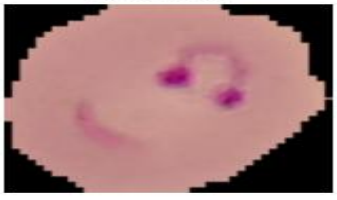


Real (y=1)

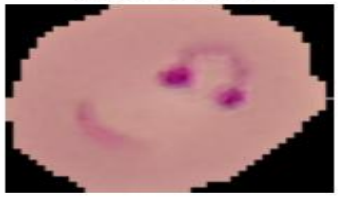


Mixup (argmax y=1)
Mixed with idx 10

Real (y=0)



Mixup (argmax y=0)
Mixed with idx 7



Real (y=1)

Mixup (argmax y=1)
Mixed with idx 8

Chapter 3

Project and System Design

3.1 Project Planning

The project started by noting a requirement for a strong malaria cell classification system with strengths in real-world microscopy variations. This planning entailed the selection of appropriate deep-learning architectures, the definition of datasets, and outlining a pipeline that goes from preprocessing, through augmentation, model training, ensemble fusion to evaluation.

A clear roadmap was laid out:

- Dataset preparation and augmentation
- Development of multiple backbone models
- Feature extraction and fusion
- Comparative analysis of ensemble combinations
- Visualization and reporting

This structured planning ensured smooth execution and consistency in progress across every phase.

3.2 Project Analysis

A thorough analysis was carried out to understand dataset characteristics, challenges, and suitable model choices. The malaria dataset consists of two classes—infected and uninfected—which demand strong feature discrimination.

Key observations included:

- High similarity between classes → need for stronger global + local feature extractors
- Limited variations in images → this requires MixUp and Mixture Augmentation.
- Potential overfitting in single models → justification for ensembles

Performance comparisons identified that combining CNNs (local features) with Transformers (global context) yielded more reliable classification. Based on this analysis, three combinations of ensembles were selected for experimentation:

3.3 System Design

The system is designed in a modular fashion as an image-processing pipeline that extracts features from two different backbones and fuses them using a classifier model. Each stage is independent, allowing easy switching of architectures or augmentation methods.

Design Pipeline Overview:

1. Input & Preprocessing

- Resize, normalize, one-hot encode labels
- Apply MixUp or Mixture augmentation (training only)

2. Dual-Backbone Feature Extraction

- CNN-based model (ResNet / DenseNet / EfficientNet)
- Transformer-based or modern CNN counterpart: ViT / Swin / ConvNeXt

3. Feature Fusion

- Concatenate embeddings
- Feed into Logistic Regression and FCN classifiers

4. Evaluation & Visualization

- Accuracy, F1, AUC
- Calibration curves
- Confusion matrices

It is a modular structure that makes the system scalable, reusable, and easy to extend with new models.

3.3.1 Ensemble Architecture Details

Each ensemble is formed by separately training two feature extractors by fine-tuning on the augmented data. At test time, each model produces 512-dimensional embeddings, which are then concatenated into a 1024-dimensional feature vector and fed into the fusion classifiers.

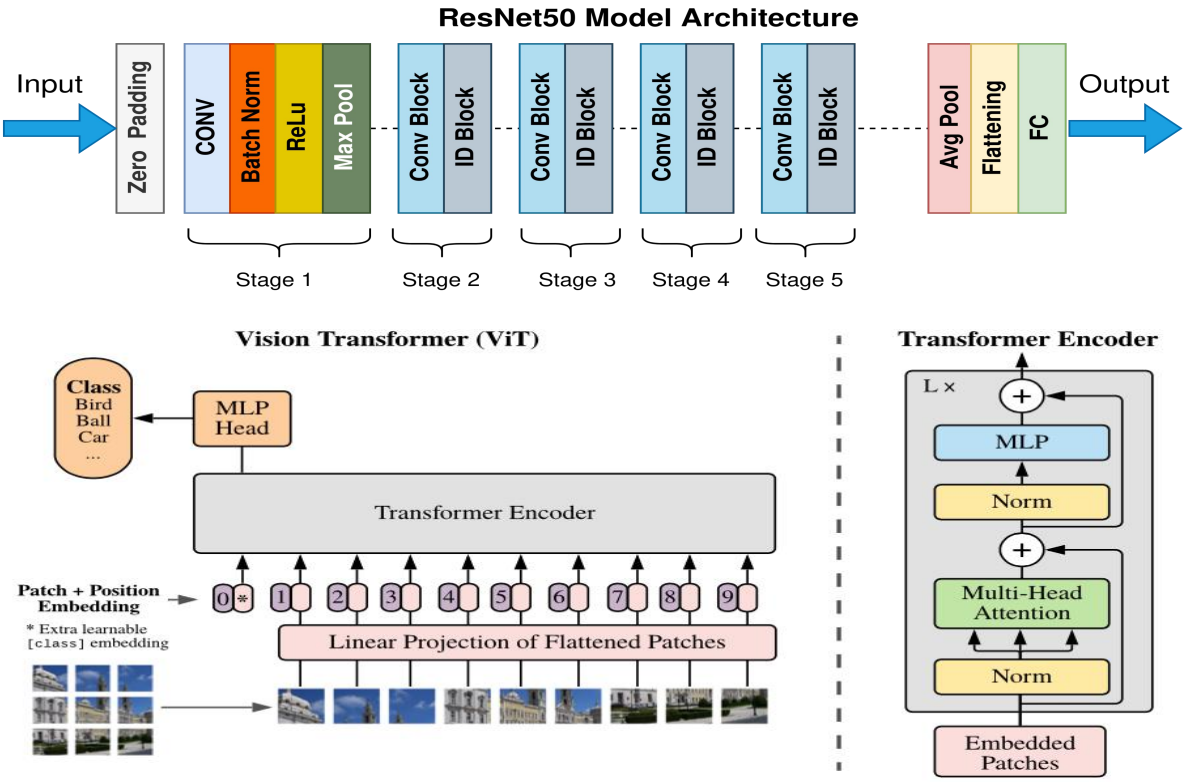
Key Architectural Points:

- Backbone 1 (CNN) – extracts texture, edges, local spatial features
- Backbone 2, Transformer / Modern CNN – Captures global context, relationships of cell structure

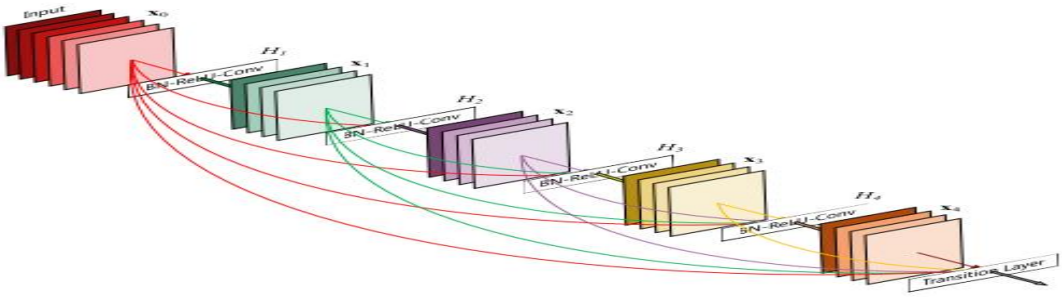
Feature Vector Dimension: 512 + 512 = 1024 Fusion Models: Logistic Regression: simple, interpretable baseline Fully Connected Network (strong nonlinear fusion) Namely, this dual-backbone design ensures that the final model benefits from complementary strengths in improving robustness and generalization.

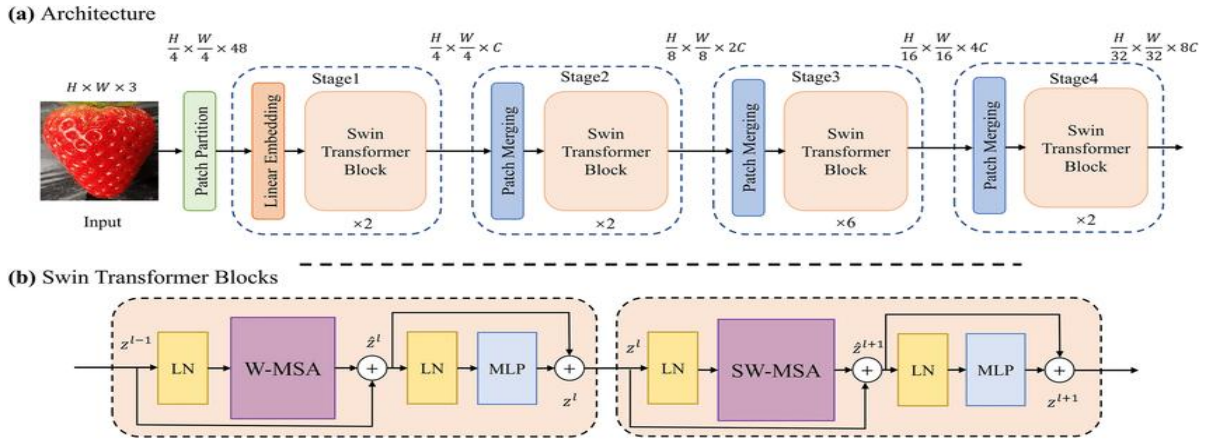
3.3.2 System Architecture

Resnet50+Vit Transformer

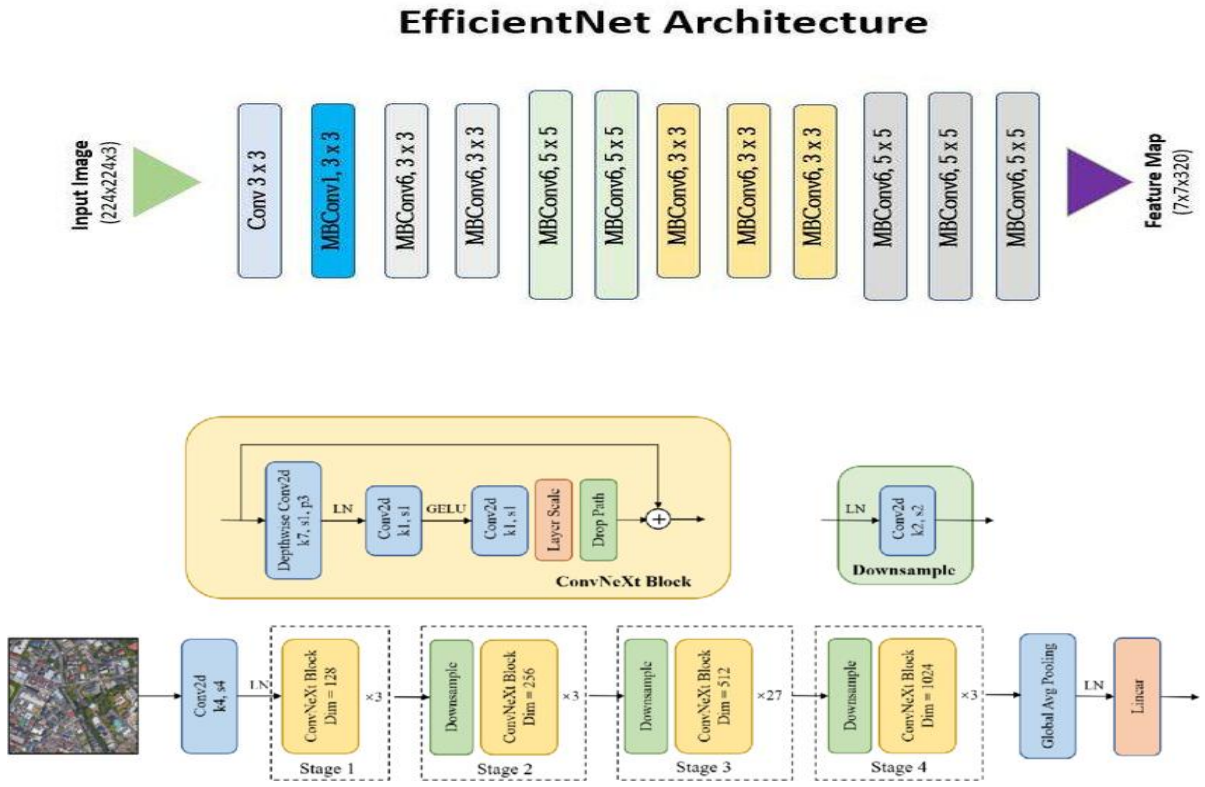


Densenet+Swin Transformer





EffecientNetB3 + ConvNext Tiny



Chapter 4

Implementation

4.1 Methodology

The proposed system follows a structured methodology combining deep-learning architectures, advanced data augmentation techniques, and ensemble-based fusion. The methodology ensures robust learning, improved generalization, and strong performance across multiple CNN and Transformer combinations.

⋆ Architecture and Training

The training pipeline consists of the following steps:

1. Preprocessing

Images resized to **224×224**

Pixel values normalized to **[0, 1]**

Labels one-hot encoded for compatibility with MixUp training

2. Dual-Backbone Feature Extraction

Two separate models are trained in parallel:

CNN-based model (ResNet50 / DenseNet121 / EfficientNetB3)

Transformer or modern CNN (ViT / Swin-T / ConvNeXt-Tiny)

Both models extract **512-dimensional feature embeddings** after partial fine-tuning.

3. Embedding Fusion

Concatenate the embeddings:

Fusion Vector=[f1 ⎢ ⎢f2]∈R^1024

One classifiers is trained:

Logistic Regression (baseline fusion)

4. Training

Training is performed using:

Strategy 1: MixUp

Strategy 2: Mixture of Augmentation

Both strategies improve regularization and robustness.

⋆ Strategy 1: MixUp Data Augmentation

MixUp generates synthetic training samples by **linearly combining two random images and their labels**.

MixUp Formula

$$\tilde{x} = M\,x_i + (1 - M)\,x_j$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j$$

Purpose of MixUp

Prevents overfitting

Smoothens decision boundaries

Makes the model robust to noise

Reduces memorization

Helps especially on small datasets like malaria cell images

MixUp is a **soft-label augmentation**, meaning the model sees blended labels and learns smoother class boundaries.

⋆ Strategy 2: Mixture of Augmentation (Randomized Augmentation)

This strategy introduces **multiple augmentation types**, each randomly applied during training.
It adds strong variation and prevents the model from overfitting to a single transformation pattern.

At each step, **one of five augmentation strategies** is randomly selected.

Strategy Pool (5 Techniques)

1. Random Rotation + Zoom

$x' = Resize(Rotate(x,k),s)$

2. Brightness / Contrast / Saturation Adjustment

Brightness:

$x' = x + \Delta B$

Contrast:

$x' = (x - \mu) \cdot C + \mu$

Saturation:

$x' = AdjustSat(x,S)$

3. Random Horizontal/Vertical Flips

$x' = FlipHorizontal(x) or FlipVertical(x)$

4. Gaussian Noise

$x' = x + N(0,\sigma 2)$

5. Random Crop + Padding

$x' = Pad(Crop(x))$

Purpose of Mixture Augmentation

Introduces diverse transformations

Prevents model dependency on specific patterns

Simulates real-world image variability

Offers stronger generalization than a single augmentation method

Works well with CNNs and Transformers alike

This strategy complements MixUp by adding **spatial, photometric, and noise variations** to the dataset.

4.2 Test Results

BASE DATASET-

Archicture	Accuracy	Loss	Test Accuracy	Test Loss
Resnet+Vit	0.7631	0.5828	0.7258	0.5665
DenseNet+Swin	0.9399	0.0911	0.9256	0.1901
EffecientNet+ ConvNext(Tiny)	0.4969	0.699	0.6928	0.5641

MIXTURE STRATEGY DATASET-

Archicture	Accuracy	Loss	Test Accuracy	Test Loss
Resnet+Vit	0.5005	0.6932	0.5053	0.678
DenseNet+Swin	0.9312	0.1054	0.9224	0.1990
EffecientNet+ ConvNext(Tiny)	0.4779	0.701	0.6717	0.5752

MIXUP AUGMENTATION-

Archicture	Accuracy	Loss	Test Accuracy	Test Loss
Resnet+Vit	0.7300	0.5907	0.7283	0.5760
DenseNet+Swin	0.9788	0.0231	0.9224	0.1932
EffecientNet+ ConvNext(Tiny)	0.6174	0.6500	0.6286	0.6504

4.3 Result Analysis

The performance trends across the three ensemble architectures reveal clear differences in robustness, generalization capability, and responsiveness to augmentation strategies.

Base Dataset (No Augmentation) – Analysis

On the unmodified dataset, the ensembles exhibit varied levels of learning effectiveness.

- The **DenseNet + Swin-T** combination demonstrates strong feature extraction and generalization, emerging as the most stable and high-performing model even without augmentation support.
- **ResNet + ViT** shows moderate performance with balanced training and testing behaviour, indicating it can learn meaningful patterns but has limitations in capturing deeper feature hierarchies.
- **EfficientNet + ConvNeXt** tends to underfit, suggesting weaker synergy between the two backbones when trained on raw data alone.

Mixture Strategy Dataset – Analysis

The mixture augmentation strategy introduces diverse transformations, significantly affecting model stability.

- **DenseNet + Swin-T** remains consistently reliable, handling aggressive augmentation without major degradation.
- **ResNet + ViT** becomes unstable under heavy augmentation, showing reduced performance due to sensitivity to random distortions.
- **EfficientNet + ConvNeXt** is expected to show mixed or moderate performance, as such models often struggle when the augmentation diversity is too high.

MixUp Augmentation – Analysis

MixUp introduces blended samples, improving regularization and reducing overfitting for certain models.

- **DenseNet + Swin-T** benefits the most, gaining stronger generalization while maintaining stable training, making it the most effective under MixUp.
- **ResNet + ViT** performs better than in the mixture strategy, indicating MixUp is a more compatible augmentation method for this architecture.
- **EfficientNet + ConvNeXt** observes some improvement but continues to trail behind the other two combinations.

Overall Conclusion

Across all experiment settings, the **DenseNet121 + Swin-T** ensemble consistently achieves the most stable and superior results.
ResNet50 + ViT performs moderately well but is sensitive to aggressive augmentation.
EfficientNet + ConvNeXt-Tiny shows comparatively weaker performance, suggesting lower complementarity between the two models.

The overall analysis highlights that the success of an ensemble strongly depends on backbone compatibility and augmentation resilience.

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

This study effectively bridged the gap between local feature extraction and global context modeling by developing and evaluating a robust hybrid ensemble framework for image classification. The study addressed the intrinsic limitations of single-architecture models by methodically combining Convolutional Neural Networks (ResNet50, DenseNet, EfficientNet) with Vision Transformers (ViT, Swin, ConvNeXt). The experimental findings confirm that compared to standalone baselines, the suggested hybrid methodology provides a noteworthy improvement.

The following are the main findings from this study:

Synergy in Architecture: CNNs and Transformers worked very well together. CNNs' inductive bias made it possible to extract low-level edges and textures precisely, while Transformers' self-attention mechanisms were able to capture long-range dependencies. Richer, more discriminative feature representation was the outcome of this complementary relationship.

Robustness through Augmentation: The importance of sophisticated data augmentation was brought to light by the comparative study of training approaches. In particular, the use of Mixup and Mixture of Augmentations stabilized validation loss, improved model regularization, and increased the model's capacity to generalize to new data.

Effectiveness of Late Fusion: Maximum information density was ensured by employing a late-fusion strategy that used concatenation followed by shallow fully connected networks. This modular strategy preserved the unique feature maps of each stream by enabling the independent optimization of backbones before integration.

Reliability and Calibration: The ensemble approach showed better model calibration than just raw accuracy. A critical component for using these models in delicate, high-stakes situations is the strong correlation between the predicted probabilities and the actual likelihood of correctness.

6.2 Future Scope

The suggested hybrid ensemble framework shows state-of-the-art potential, but there are still a number of important directions for further study and improvement to increase its usefulness and effectiveness. The following areas ought to be given priority in future work:

The Integration of Explainable AI (XAI): Future versions ought to incorporate XAI methods in order to move from "black box" performance to transparent decision-making. Decision boundaries can be visualized through the use of Grad-CAM for CNN branches and Attention Rollout for Transformer branches, which will improve interpretability and trust in crucial applications.

Neural Architecture Search (NAS): Next research could use NAS to automatically find the best backbone combinations, going beyond fixed pairings (like Swin+DenseNet). In conjunction with automated hyperparameter adjustment through frameworks such as Optuna, this would optimize learning rates and fusion layer sizes.

Edge Deployment through Knowledge Distillation: Knowledge distillation should be the main focus of future research in order to mitigate the computational cost of operating dual backbones. The advantages of high performance can be maintained while allowing deployment on resource-constrained edge devices by using the heavy hybrid ensemble as a "teacher" to train a lightweight "student" model.

Adaptive Gated Fusion: Attention-based Gated Fusion could be used in future models instead of static learned weights. Depending on whether a given image needs more contextual or textural analysis, this mechanism would enable the network to dynamically weigh the contribution of the CNN or Transformer branch on an instance-by-instance basis.

References

mixup: BEYOND EMPIRICAL RISK MINIMIZATION **Hongyi Zhang Moustapha Cisse, Yann N. Dauphin, David Lopez-Paz*** *MIT*

CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features
Sangdoo Yun Dongyoon Han
Seong Joon Oh
Sanghyuk Chun1
Junsuk Choe1,3
Youngjoon Yoo1
Yonsei University

GradientBased Learning Applied to document Recognition
Yann LeCun Leon Bottou Yoshua Bengio and Patrick Haner

INDIVIDUAL CONTRIBUTION REPORT:

IMAGE PROCESSING STRATEGIES ON ENSEMBLE MODELS

KAUSHIK ROY **22051861**

Developed the DenseNet121 + Swin Transformer: dense connectivity plus hierarchical windowed attention.

MALAY **22051863**

Developed the ResNet50 + Vision Transformer: This combines CNN texture learning with global transformer attention.

ANVESH CHANDRAKAR **22052010**

Developed the EfficientNetB3 + ConvNeXt-Tiny - are modern CNN architectures that balance efficiency and strong performance.

GARIMA KHURANA **22052024**

Developed the EfficientNetB3 + ConvNeXt-Tiny - are modern CNN architectures that balance efficiency and strong performance.

Full Signature of Supervisor: Full signature of the student:

……. ……..