Hypermedia-Based Feature Classification: A Spatial-Frequency Approach with Embedding Learning

A report submitted in partial fulfilment of the requirements

for the award of the degree of

B.Tech Computer Science and Engineering

by

K. Vineela (Roll No: 121cs0037)

Under the Guidance of

Dr. Nagarju K, Assistant Professor

Dept. of CSE, IIITDM Kurnool

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY DESIGN AND MANUFACTURING KURNOOL

April 2025

Evaluation Sheet

Title of the Project: **Hypermedia-Based Feature Classification: A Spatial-Frequency Approach with Embedding Learning**

Name of the Student(s): K. Vineela

Examiner(s):

Supervisor(s):

Head of the Department:

Date:

Place:

Certificate

I, **K. Vineela**, with Roll No: **121CS0037** respectively hereby declare that the material presented in the Project Report titled **Hypermedia-Based Feature Classification: A Spatial-Frequency Approach with Embedding Learning** represents original work carried out by me in the **Department of Computer Science and Engineering** at the **Indian Institute of Information Technology Design and Manufacturing Kurnool** during the years **2024 - 2025**. With my signature, I certify that:

- I have not manipulated any of the data or results.
- I have not committed any plagiarism of intellectual property. I have clearly indicated and referenced the contributions of others.
- I have explicitly acknowledged all collaborative research and discussions.
- I have understood that any false claim will result in severe disciplinary action.
- I have understood that the work may be screened for any form of academic misconduct.

Date: Student's Signature

In my capacity as supervisor of the above-mentioned work, I certify that the work presented in this Report is carried out under my supervision, and is worthy of consideration for the requirements of B.Tech. Project work.

Advisor's Name: Advisor's Signature

Abstract

This paper presents a novel approach for improving classification accuracy and Embedding Space in medical image analysis by leveraging hypermediabased feature extraction and embedding techniques. Our methodology extracts spatial and frequency features, applies an attention mechanism to enhance important regions, and maps the features into a refined embedding space using triplet loss. The resulting embeddings are classified using k-Nearest Neighbors (kNN), ensuring better cluster separation and classification performance. We evaluated our approach on the MedMNIST dataset, where it demonstrated superior performance compared to existing methods. Additionally, our approach reduces computational complexity, making it efficient for large-scale medical image classification.

Acknowledgements

I would like to express our sincere gratitude to several individuals and organizations who have made this project possible.

First and foremost, I extend our heartfelt thanks to our project guide, Dr. Nagarju K, for his unwavering support, guidance, and valuable insights throughout the development of this framework. His expertise and encouragement have been instrumental in shaping our research and helping us navigate the challenges I faced.

I would also like to acknowledge the faculty and staff of the Indian Institute of Information Technology, Design and Manufacturing, Kurnool, for providing an inspiring academic environment and the necessary resources that facilitated our learning and research.

I are grateful to our peers and friends who offered their assistance and feedback during the various stages of our project. Their collaborative spirit and constructive critiques have significantly enhanced the quality of our work.

Lastly, I would like to thank our families for their constant support, understanding, and encouragement throughout our academic journey. Their belief in us has motivated us to strive for excellence in our endeavors.

Thank you all for your invaluable contributions and support.

Contents

List of Figures

List of Tables

Abbreviations

KNN	**K**-Nearest **N**eighbor
CNN	**C**onvolutional **N**eural **N**etwork
AUC	**A**rea **U**nder the **C**urve
GNN	**G**raph **N**eural **N**etwork
GCN	**G**raph **C**onvolutional **N**etwork
GUNEC	**G**raph-based **U**nified **N**etwork for **E**mbedded **C**lassification
t-SNE	**t**-Distributed **S**tochastic **N**eighbor **E**mbedding
ReLU	**R**ectified **L**inear **U**nit
PCA	**P**rincipal **C**omponent **A**nalysis
GI	**G**astro**I**ntestinal
MAP	**M**ean **A**verage **P**recision
ReID	**Re**-**I**dentification
ResNet	**Res**idual **Net**work
Xception	E**x**treme **C**onvolution **Net**work based on **In**ception
RGB	**R**ed **G**reen **B**lue

Symbols

σ	Sigma	
Σ	Summation	–
$\mathcal{L}_{triplet}$	Triplet loss function	–
α	Margin in triplet loss	–
sigmoid	Sigmoid activation function	–
$\phi(\cdot)$	Embedding function	–
x	Anchor image	–
x^p	Positive image	–
x^n	Negative image	–

Chapter 1

Introduction

Accurate classification of medical images plays a crucial role in disease diagnosis by enabling healthcare professionals to make timely and informed decisions. It helps reduce human error and enhances diagnostic efficiency, ultimately improving patient outcomes. In the medical sector, high classification accuracy is essential for early detection of diseases, treatment planning, and monitoring disease progression. However, despite significant advancements in deep learning models, achieving high accuracy and robustness in large-scale medical imaging applications remains a challenge. To address these challenges, hypermedia-based approaches have gained attention due to their ability to integrate multiple types of data, such as spatial and frequency features, enhancing feature representation and improving classification performance. These approaches provide a more comprehensive understanding of the data by combining diverse feature sets, which is particularly beneficial for complex medical imaging tasks. In this study, I propose a methodology that leverages spatial and frequency domain features, applies an attention mechanism to emphasize relevant information, and maps the features into a refined embedding space using triplet loss. The resulting embeddings are then classified using the k-Nearest Neighbors (kNN) algorithm, ensuring better cluster separation and improved classification performance. My approach has been applied to the MedMNIST dataset, where it demonstrates superior performance compared to existing methods. Additionally, the proposed approach optimizes computational efficiency, making it well-suited for large-scale medical image classification

1.1 Problem Statement

Given a medical image dataset containing a collection of images $X = \{x_1, x_2, \ldots, x_n\}$ with corresponding labels $Y = \{y_1, y_2, \ldots, y_n\}$, the objective is to learn a classification function:

$$f : X \to Y$$

that maps each input image x_i to its correct label y_i.

$$\max_f A(f(X), Y), \quad \max_f U(f(X), Y) \tag{1.1}$$

Additionally, I aim to improve the embedding space representation E to enhance feature separability and classification performance:

$$E = g(X) \tag{1.2}$$

where $g(X)$ is an optimized transformation of the input space.

The classification is performed using a function $h(E)$, ensuring:

$$\hat{y} = h(E) \tag{1.3}$$

To achieve efficient computation, I also minimize the computational complexity $C(f)$:

$$\min_f C(f) \tag{1.4}$$

while maintaining high classification performance.

1.2 Project Approach

The proposed approach follows a systematic process that enhances classification performance by combining multiple feature extraction techniques, applying attention mechanisms, and refining the embedding space for improved classification. Initially, the medical images from the MedMNIST dataset undergo preprocessing to ensure consistency in size, intensity normalization, and noise reduction. **Spatial features** are then extracted by analyzing the visual patterns, edges, and textures present in the images, capturing both local and global information essential for accurate classification.

Following spatial feature extraction, these features are transformed into the **frequency domain** using techniques such as **Fourier Transform**, allowing the identification of patterns that may be difficult to observe in the spatial domain. To further enhance feature representation, an **attention** mechanism is applied, which assigns higher importance to relevant regions while minimizing the impact of less important areas. This improves the quality of feature representation by focusing on significant details.

The extracted and refined features are then mapped into a new embedding space using a **triplet loss function**, ensuring that similar samples are positioned closer together while dissimilar samples are pushed apart. This transformation improves feature separability and enhances classification accuracy. Finally, the refined embeddings are classified using the **k-Nearest Neighbors (kNN) algorithm**, where labels are assigned based on the majority vote of the nearest neighbors.

The overall approach not only improves classification accuracy but also optimizes computational efficiency, making it suitable for large-scale medical image processing tasks.

1.3 Motivation

Medical image classification plays a critical role in improving healthcare outcomes by enabling accurate disease diagnosis, treatment planning, and patient monitoring. However, despite advancements in deep learning models, achieving high classification accuracy and

robustness in large-scale medical imaging datasets remains a challenge. Many existing models struggle to effectively to model subtle and non-obvious dependencies present in the dataset, leading to suboptimal performance.

To address these limitations, integrating spatial and frequency features with an attention mechanism provides a more comprehensive representation of medical images, enhancing classification performance. Additionally, transforming features into a refined embedding space using triplet loss improves feature separability, making classification more effective. The motivation behind this research is to develop a methodology that not only enhances classification accuracy and AUC but also reduces computational complexity, ensuring efficient and reliable performance for large-scale medical image analysis.

By applying this approach to the MedMNIST dataset, I aim to demonstrate that integrating these techniques can significantly improve classification outcomes, providing a more accurate and efficient solution for medical image processing.

Chapter 2

Literature Survey

2.1 Literature Survey

2.1.1 Triplet Focal Loss for Person Re-Identification (Zhang et al., 2018) [1]

Person Re-Identification (ReID) refers to the task of classification based on embedding vector space. This consists of three images, anchor, positive and negative. Anchor and postive of same class, while negative is of different class. Here, mainly depends on alpha, if alpha value is large, then classification will be good.

Triplet Focal Loss addresses this challenge by introducing an exponential kernel function that focuses training on harder triplets. It effectively penalizes hard positives (same identity but visually different) and hard negatives (different identities but visually similar) more than easier ones. The loss function is defined as:

$$\mathcal{L}_{\text{TFL}} = \sum_{i=1}^{P} \sum_{a=1}^{K} \max\left(0, \exp\left(\frac{D^*_{a,p}}{\sigma}\right) - \exp\left(\frac{D^*_{a,n}}{\sigma}\right) + m\right)$$

Where $D^*_{a,p}$ and $D^*_{a,n}$ are the hardest positive and hardest negative distances for anchor a within a mini-batch, σ is a scaling parameter, and m is the margin. This formulation

allows the model to adaptively focus on more informative, challenging samples during training.

Key Contributions:

- Introduced Triplet Focal Loss that adaptively increases focus on hard triplets via exponential kernel transformation.
- Combined with Batch Hard Triplet Loss strategy for robust online hard example mining.
- Demonstrated superior performance over standard triplet loss.

Experimental Results: On the given dataset, the proposed method improved mAP from 61.49% to 72.21% and Rank-1 accuracy from 77.76% to 87.92%. After applying Re-Ranking, the improvements were even more pronounced, with mAP reaching 85.88% and Rank-1 reaching 90.17%. Similar consistent gains were observed on the DukeMTMC-reID and CUHK03 datasets.

Limitations:

- The performance depends on careful selection of hyperparameters such as margin m and kernel scaling factor σ.
- Although computational cost remains low, the model is still sensitive to the quality of hard triplet mining and batch composition.

Relevance to Current Work: The concept of focusing on hard examples through adaptive loss functions like Triplet Focal Loss is closely related to efforts in refining embedding spaces and improving class separability, especially in tasks involving few-shot or zero-shot learning scenarios. The exponential penalization mechanism offers a promising direction for addressing overfitting to easy samples and ensuring more discriminative representations.

2.1.2 Classification of Images Based on CNN Deep Learning Non-Local Attention Pyramid Model [2]

This paper proposes a multi-scale deep learning model called Toynet Non-local, designed for the classification of leaf pathology images. The study addresses the limitations of existing models, such as overfitting, inadequate feature extraction, and sensitivity to image noise. To overcome these, the authors integrate a Non-local Pyramid Attention Module into a CNN framework, enabling the model to focus more effectively on significant features and capture long-distance dependencies in the image data.

Key innovations include:

- **Non-local attention mechanism** to enhance feature learning by considering global contextual information rather than just local pixel neighborhoods.
- **Multi-scale feature extraction** using a pyramid structure to handle variations in pathology shape and size.

A residual fusion strategy that preserves original image features while enhancing them with attention-based features.

The model was tested on three leaf disease types (rust, brown spot, and powdery mildew) and showed superior classification accuracy and robustness, especially under Gaussian noise, compared to standard CNNs and other deep models (e.g., ResNet, Xception). The Toynet Non-local architecture demonstrated strong resilience to image degradation, indicating its potential for real-world deployment in automated plant disease detection.

2.1.3 MedMNIST Classification Decathlon: A Lightweight AutoML Benchmark for Medical Image Analysis [3]

This paper introduces MedMNIST, a curated collection of 10 lightweight and standardized medical image datasets, designed to facilitate education, rapid prototyping, and benchmarking in the field of medical image classification. Inspired by the popular MNIST dataset, each MedMNIST dataset contains 28×28 pre-processed images covering

a variety of modalities, tasks (e.g., binary, multi-class, ordinal regression), and scales (from hundreds to over 100,000 samples).

Key contributions include:

- The MedMNIST Classification Decathlon, a benchmark that evaluates AutoML algorithms on all 10 datasets without manual tuning.
- Comparative evaluation of baseline methods (e.g., ResNet-18/50), open-source AutoML tools (auto-sklearn, AutoKeras), and commercial platforms (Google AutoML Vision).
- Performance is measured using AUC and accuracy, revealing that while commercial tools like Google AutoML performed well overall, no single method dominated across all datasets.

This benchmark promotes the development and fair comparison of AutoML solutions in medical imaging, particularly for researchers without deep domain expertise. Its lightweight and standardized nature makes it ideal for both academic and practical use in multi-modal, large-scale, and automated medical image analysis.

2.1.4 Dynamic Filter Application in Graph Convolutional Networks for Enhanced Spectral Feature Analysis and Class Discrimination in Medical Imaging [4]

CNN have long been the backbone of computer vision and medical image analysis tasks due to their powerful ability to learn hierarchical spatial features. They perform well on grid-structured data such as 2D images, with models like ResNet, DenseNet, and EfficientNet consistently outperforming existing methods on various classification and segmentation challenges. However, CNNs operate under the assumption of Euclidean data and often struggle with representing complex, non-local relationships that are crucial in medical images.

To address these limitations, **GNNs** have emerged as a compelling alternative. GNNs operate on non-Euclidean data by representing images or regions as graphs, where nodes

correspond to pixels or superpixels and edges capture relationships such as spatial proximity, similarity, or contextual association. Among GNN variants, **Graph Convolutional Networks (GCNs)** are particularly noteworthy for extending convolution operations to graph-structured data. GCNs aggregate information from neighboring nodes and update node features layer-by-layer using graph-based filters.

Despite the theoretical appeal, standard GCNs suffer from critical challenges:

- **Over-smoothing**: As the number of GCN layers increases, node features become increasingly similar, reducing class separability.
- **Limited Long-range Contextual Learning**: Traditional GCNs typically operate over 1-hop or 2-hop neighbors, failing to capture global or long-range dependencies.
- **Sensitivity to Graph Structure**: The model's performance is highly dependent on how the graph is constructed, which may vary significantly across datasets.

To overcome these limitations, proposed an enhanced architecture known as the **Graph Convolutional Neural Network with Enhanced Connectivity (GCNN-EC)**. This model incorporates dynamic filtering and edge adaptation strategies to create a more expressive and adaptive graph structure tailored for medical imaging tasks.

Key Contributions of GCNN-EC

- **Dynamic Edge Feature Adaptation**: Instead of relying on static graph connections, GCNN-EC introduces dynamic edge creation. Edge weights are learned based on feature similarity, allowing for a more data-driven and context-sensitive graph structure.
- **Long-Range Feature Aggregation**: The model goes beyond traditional 1-hop neighbors by aggregating information from n-hop neighbors. This captures a broader context, which is crucial in medical images where pathology may be spatially diffuse or non-contiguous.

- **Mitigation of Over-smoothing**: GCNN-EC addresses the over-smoothing problem by introducing skip connections and dynamic edge prediction. These strategies help preserve unique feature characteristics across deeper layers.
- **Low-Parameter Efficiency**: Despite its enhanced performance, GCNN-EC achieves results comparable to or better than CNNs like DenseNet121 and ResNet18, using 10 to 100 times fewer parameters—making it highly efficient and suitable for edge deployment (e.g., portable devices in healthcare).
- **Generalizability Across Datasets**: The model was evaluated on several benchmarks including MedMNIST, MNIST, CIFAR-10, and MSTAR-10, and consistently outperformed traditional GCNs, demonstrating strong cross-domain adaptability.

Gap Identification

While GCNN-EC introduces several innovations, certain limitations and open research questions remain:

1. **Computational Overhead from Dynamic Filtering**: The introduction of dynamic filters and learnable edge assignments increases memory usage and computation time. This becomes a bottleneck in large-scale applications or real-time settings.
2. **CNNs Still Outperform in Some Scenarios**: Although GCNN-EC performs competitively, EfficientNet-B0 showed better results on specific tasks. This indicates that CNNs still maintain advantages in scenarios with strong spatial locality and simpler texture patterns.
3. **Graph Construction Sensitivity**: The effectiveness of GCNN-EC is highly dependent on how the initial graph is constructed. Variations in connectivity (e.g., 4-neighbor vs. 8-neighbor) or feature encoding (e.g., RGB + spatial coordinates) can significantly affect performance, leading to potential instability across datasets.

4. **Persistent Over-smoothing Risk in Deep GCNs**: Although GCNN-EC mitigates over-smoothing better than standard GCNs, the risk still persists with deeper networks. Feature mixing remains a theoretical limitation in very deep GCN architectures.

5. **Challenges in Capturing Long-range Dependencies in Standard GCNs**: Without the proposed enhancements, GCNs struggle to effectively capture global dependencies due to their local aggregation operations. This reinforces the necessity of dynamic and multi-hop aggregation layers introduced by GCNN-EC.

6. **Training Complexity and Model Interpretability**: The added layers and dynamic operations introduce additional hyperparameters and architectural complexity. This could pose challenges in model tuning and interpretability, especially in sensitive domains like healthcare.

These gaps underscore the trade-offs between flexibility, efficiency, and generalizability. While GCNN-EC marks significant progress in graph-based medical image analysis, further research is warranted to optimize computational efficiency, improve robustness across diverse datasets, and enhance integration with existing CNN-based pipelines or vision transformers.

Chapter 3

Proposed Methodology

3.1 Proposed Methodology

In this section, we elaborate on the proposed deep metric learning architecture designed for robust representation learning using the MedMNIST dataset. The key steps include spatial and frequency-based feature extraction, feature fusion, attention mechanism, embedding generation, triplet network architecture, and the triplet loss function.

3.1.1 Dataset Feature Extraction

In this work, feature extraction is performed using a combination of spatial and frequency domain techniques to enhance the model's representation learning capability.

3.1.2 Spatial Feature Extraction

Spatial features are extracted using a CNN, which learns hierarchical patterns such as edges, textures, and structures from the input medical images. The CNN consists of stacked convolutional layers followed by max-pooling layers to capture both low-level and high-level features. These spatial features represent localized patterns in the pixel domain, and they serve as a robust foundation for downstream tasks such as classification or embedding learning.

3.1.3 Convolutional Neural Network (CNN) Architecture

The spatial feature extraction module is implemented using a Convolutional Neural Network (CNN) composed of two convolutional and pooling blocks. The CNN is built to extract multi-level spatial features from grayscale medical images, capturing both low- and high-level patterns. The architecture is as follows:

- **Conv2D Layer 1:** Applies 32 filters of size 3×3 with ReLU activation. This layer detects low-level features such as edges and corners.
- **MaxPooling2D Layer 1:** Downsamples the feature map using a pooling size of 2×2, reducing spatial dimensions and computational complexity.
- **Conv2D Layer 2:** Applies 64 filters of size 3×3 with ReLU activation. This layer captures more complex patterns and textures.
- **MaxPooling2D Layer 2:** Further reduces the spatial dimensions while preserving essential features.

This CNN serves as the base for extracting spatial features before fusion with frequency features. The extracted feature maps are further fused for embedding generation.

3.1.4 Frequency Feature Extraction

To complement the spatial features, frequency domain information is also extracted using a custom Fourier Transform layer. Specifically, the 2D Fast Fourier Transform (FFT) is applied to the image feature maps generated by the CNN. This transformation converts the spatial data into the frequency domain, capturing periodic patterns and eliminating local noise that might not be evident in the spatial representation.

The magnitude of the complex FFT output is retained as the frequency feature map. Mathematically, the transformation is given by:

$$F(u,v) = \left| \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \cdot e^{-j2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)} \right| \tag{3.1}$$

where $f(x, y)$ is the input feature map, and $F(u, v)$ represents the magnitude of the frequency component at location (u, v).

This approach ensures that both local (spatial) and global (frequency) information is captured, enhancing the discriminative power of the final embeddings.

The spatial and frequency features are concatenated along the channel axis and further refined using an attention mechanism before being passed to the embedding layer.

3.1.5 Feature Fusion

The spatial and frequency feature maps are concatenated along the channel dimension to form a unified representation. This fusion enriches the model's understanding by capturing both localized pixel-level patterns and global frequency-based structures. An attention mechanism is then applied to emphasize the most informative channels from the fused features, followed by a dense layer to generate the final embeddings.

This dual-domain approach significantly improves the discriminative power of the embeddings used in the triplet network architecture.

3.1.6 Feature Fusion: Spatial and Frequency Domain

First by extracting two complementary sets of features:

- **Spatial Features**: Extracted using a convolutional neural network (CNN), these features capture local patterns such as edges and textures.
- **Frequency Features**: Obtained by applying the 2D Fast Fourier Transform (FFT) to the input images. This highlights global structure and periodic patterns that are often invisible in the spatial domain.

Given an input image $x \in \mathbb{R}^{H \times W \times C}$, we first extract spatial features:

$$F_s = \text{CNN}(x)$$

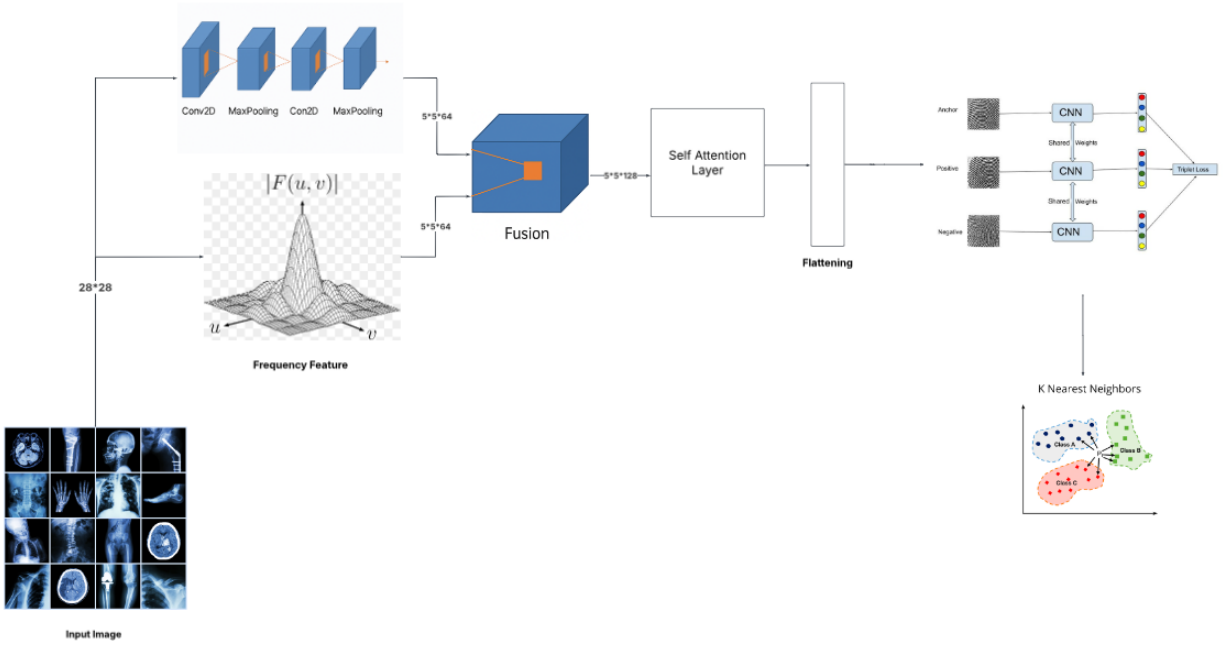

FIGURE 3.1: Proposed Methodology

We then compute the FFT of x:

$$F_f = |\mathcal{F}(x)| = |\text{FFT2D}(x)|$$

where $\mathcal{F}(\cdot)$ denotes the 2D FFT and $|\cdot|$ is the magnitude.

3.1.7 Feature Fusion

To combine the spatial and frequency domain features, we concatenate them along the channel dimension:

$$F_{fused} = \text{Concat}(F_s, F_f)$$

This fusion enables the model to leverage both local texture details and global structural information, which improves generalization.

3.1.8 Attention Mechanism

To enhance discriminative parts of the fused feature maps, we use a self-learned attention mechanism. The attention map α is computed using a 1×1 convolution followed by a sigmoid activation:

$$\alpha = \sigma(W_a * F_{fused})$$

where W_a is the convolution kernel and $\sigma(\cdot)$ is the sigmoid function.

The final attended feature map is given by:

$$F_{att} = \alpha \odot F_{fused}$$

where $\odot$ represents element-wise multiplication.

3.1.9 Embedding Generation

The attended features F_{att} are then flattened and passed through a fully connected layer to form the final embedding vector:

$$z = \phi(F_{att}) = \text{ReLU}(W_e \cdot \text{Flatten}(F_{att}) + b_e)$$

where W_e and b_e are learnable weights and bias, and $\phi(\cdot)$ is the embedding function.

This embedding $z \in \mathbb{R}^d$ (where d is the embedding dimension, e.g., 128) acts as a compact representation of the input image, optimized for class separability.

3.1.10 Triplet Network Architecture

To learn robust and discriminative embeddings, we utilize a triplet network composed of three shared-parameter branches of the embedding model. Each triplet input includes:

- **Anchor image** x^a
- **Positive image** x^p (same class as anchor)

- **Negative image** x^n (different class from anchor)

These are passed through the same embedding network:

$$\begin{aligned} z^a &= \phi(x^a), \\ z^p &= \phi(x^p), \\ z^n &= \phi(x^n) \end{aligned} \tag{3.2}$$

3.1.11 Triplet Loss Function

We use the standard triplet loss to encourage embeddings of the same class to be closer while pushing apart embeddings of different classes. The loss function is defined as:

$$\mathcal{L}_{\text{triplet}} = \sum_{i=1}^{N} \max\left(\|z_i^a - z_i^p\|_2^2 - \|z_i^a - z_i^n\|_2^2 + \alpha, 0\right) \tag{3.3}$$

where α is the margin (e.g., 1.0), and N is the batch size.

This objective function ensures that:

$$\|z^a - z^p\|_2^2 + \alpha < \|z^a - z^n\|_2^2 \tag{3.4}$$

leading to better class clustering in the embedding space.

3.1.12 Classification using K-Nearest Neighbors (KNN)

Once the embeddings $z = \phi(x)$ are learned using the triplet network, we perform classification using a non-parametric K-Nearest Neighbors (KNN) classifier.

Given a query image x_q, we compute its embedding $z_q = \phi(x_q)$. We then search for its K nearest neighbors among the embeddings of the training set:

$$\mathcal{N}_K(z_q) = \{z_{i_1}, z_{i_2}, \ldots, z_{i_K}\} \quad \text{such that} \quad \|z_q - z_{i_j}\|_2 \text{ is minimal}$$

where $\|\cdot\|_2$ is the Euclidean distance in the embedding space.

The predicted label $\hat{y}$ is assigned based on majority voting:

$$\hat{y} = \text{mode}(\{y_{i_1}, y_{i_2}, \ldots, y_{i_K}\})$$

3.1.13 Why It Works

- **CNN for Spatial Domain**: Captures local patterns critical for distinguishing classes based on pixel distributions.
- **FFT for Frequency Domain**: Complements the CNN by highlighting global frequency signatures that can separate patterns not visible in pixel space.
- **Attention Mechanism**: Focuses the model on discriminative regions of the input, improving robustness to irrelevant features.
- **Triplet Loss**: Drives the model to learn a semantically meaningful embedding space that aligns similar samples and separates dissimilar ones.

Together, these modules enhance the model's ability to generalize across medical image categories and improve classification or retrieval accuracy in low-data settings.

3.2 Experimental Workflow and Progressive Approach

To systematically develop and validate our proposed method, we followed a stepwise experimental workflow using standard benchmark datasets, namely MNIST and CIFAR-10, before extending it to MedMNIST. This ensured a robust design process and allowed for comparative evaluations at each stage.

1. **Step 1: CNN Feature Extraction Only**

   We began by implementing a basic CNN-based feature extractor to understand the spatial characteristics of the input images. This served as a baseline model for further experiments.

2. **Step 2: Frequency Feature Extraction via FFT**

   Next, we extracted frequency domain features using a custom FFT layer applied to the CNN feature maps. We evaluated these frequency features independently to understand the spectral content's discriminative power.

3. **Step 3: Fusion of Spatial and Frequency Features**

   After separately analyzing spatial and frequency features, we fused them by concatenating along the channel dimension. This step aimed to capture both local (spatial) and global (frequency) representations.

4. **Step 4: Self-Attention Mechanism on Fused Features**

   To further enhance feature representation, we applied a self-attention mechanism over the fused features. This helped the model learn to emphasize important regions and suppress irrelevant noise.

5. **Step 5: Embedding Generation and Triplet Network**

   We constructed an embedding model on top of the attention-enhanced features. This model was trained using a triplet network to learn a discriminative embedding space to make cluster clear and seperable from others.

6. **Step 6: Final Classification using K-Nearest Neighbors**

   Finally, the learned embeddings were passed through a KNN classifier. This non-parametric method enabled label prediction based on distances in the learned embedding space.

Progressive Improvement Justification

Each step was designed to add complementary strengths:

- CNN captured spatial locality.
- FFT revealed frequency-based global patterns.
- Fusion allowed unified representation.
- Attention emphasized discriminative features.

- Triplet loss structured the embedding space effectively.
- KNN leveraged learned distances for classification.

The incremental evaluation after each step confirmed performance improvements and provided insights into how each component contributed to the final model's accuracy and robustness.

3.3 Datasets

The MedMNIST dataset is a large-scale benchmark designed for multi-class classification tasks in medical imaging. It consists of 10 pre-processed datasets derived from various publicly available sources, covering different medical modalities. Each dataset focuses on a specific classification task and is categorized as follows:

1. **PathMNIST** (Multi-class, 9 classes) – Different tissue types in pathology images.
2. **ChestMNIST** (Binary, 2 classes) – Normal vs. pneumonia X-ray images.
3. **DermaMNIST** (Multi-class, 7 classes) – Different skin lesion types.
4. **OCTMNIST** (Multi-class, 4 classes) – Optical coherence tomography images of the retina.
5. **PneumoniaMNIST** (Binary, 2 classes) – Normal vs. pneumonia X-ray images.
6. **RetinaMNIST** (Multi-class, 5 classes) – Retinal disease classification.
7. **BreastMNIST** (Binary, 2 classes) – Normal vs. malignant breast ultrasound images.
8. **OrganMNIST** (Axial, Coronal, Sagittal) (Multi-class, 11 classes each) – Organ segmentation from CT scans.
9. **TissueMNIST** (Multi-class, 8 classes) – Histological tissue types.
10. **BoneMNIST** (Binary, 2 classes) – Normal vs. abnormal bone X-ray images.

3.4 Libraries and Tools Used

The following Python libraries and frameworks were utilized in the implementation of the proposed method:

- **TensorFlow / Keras**

  Used for constructing deep learning models CNNs, custom layers, and the triplet network architecture. TensorFlow provides GPU acceleration and Keras offers high-level APIs for rapid prototyping.

- **NumPy**

  Employed for efficient numerical operations, array manipulation, and handling image data during preprocessing and triplet generation.

- **Matplotlib**

  Utilized to visualize embeddings using t-SNE plots, enabling qualitative assessment of the feature space learned by the model.

- **scikit-learn**

  Several modules from scikit-learn were used like dataset is divided into train and test.

- **openTSNE**

  A high-performance library used to perform t-distributed Stochastic Neighbor Embedding (t-SNE) for visualizing high-dimensional embedding vectors. It supports parallel computation and scales better than traditional implementations.

- **MedMNIST Dataset (medmnist)**

  The `medmnist` library was used to load the `PathMNIST` dataset, a preprocessed medical image dataset designed for classification tasks. It provides an easy-to-use interface for downloading and managing medical image data.

- **TensorFlow Signal Processing Module (`tf.signal`)**

  Used within a custom layer to compute the 2D Fast Fourier Transform (FFT) for frequency domain feature extraction from CNN-generated feature maps.

Chapter 4

Results and Analysis of the Approach

4.1 Overview of Experiments

To validate our approach, we conducted experiments on two benchmark datasets: **MNIST** and **CIFAR-10**. Our experiments followed a progressive design pipeline, where we first implemented CNN and FFT features separately, then fused them, and subsequently added self-attention and triplet network to refine the embeddings.

This section presents the quantitative results in a stage-wise manner and discusses the comparative improvements gained from each step.

4.2 Results on MNIST Dataset

4.2.1 Stage-wise Accuracy for Fusion Methods on MNIST

Table 4.1 shows the classification accuracy achieved at different stages of our model on the MNIST dataset. Initially, the model was trained using only CNN-based spatial features and frequency-based handcrafted features separately. Then, the features were fused and evaluated again.

alpha	Domain	accuracy	precision	recall	f1_score
0	Spatial	0.96	0.96	0.96	0.96
1	Frequency	0.91	0.91	0.91	0.90
0.5	Fusion	0.97	0.97	0.96	0.96

TABLE 4.1: Results after cnn with different domain on MNIST Dataset

alpha	Domain	accuracy	precision	recall	f1_score
0	Spatial	0.97	0.97	0.97	0.97
1	Frequency	0.91	0.91	0.91	0.90
0.5	Fusion	0.98	0.98	0.98	0.98

TABLE 4.2: Results after cnn with different domain with attention on MNIST Dataset

From Table 4.1, we can clearly observe that the fusion of spatial and frequency features improves the classification accuracy compared to using either feature type alone. This demonstrates the complementary nature of the spatial and frequency domains in image analysis.

4.2.2 Improvement After Adding Self-Attention (MNIST)

After applying a self-attention mechanism on the fused features, further improvements were observed.The attention module guides the model to focus on the most important and distinctive features, improving prediction accuracy.

The addition of self-attention (Table 4.2) leads to further improvement, indicating that attention helps refine the fused features by emphasizing the most informative regions.

MNIST Analysis: The results show that fusing CNN and FFT features improved accuracy over frequency alone and over CNN alone. Incorporating self-attention led to more focused feature learning, boosting performance.

alpha	Domain	accuracy	precision	recall	f1_score
0	Spatial	0.95	0.95	0.95	0.95
1	Frequency	0.89	0.89	0.89	0.89
0.5	Fusion	0.97	0.97	0.96	0.96

TABLE 4.3: Results after cnn with different domain on CIFAR10 Dataset

alpha	Domain	accuracy	precision	recall	f1_score
0	Spatial	0.96	0.96	0.96	0.96
1	Frequency	0.90	0.90	0.91	0.90
0.5	Fusion	0.98	0.98	0.97	0.98

TABLE 4.4: Results after cnn with different domain with attention on CIFAR10 Dataset

4.3 Results on CIFAR-10 Dataset

4.3.1 Stage-wise Accuracy for Fusion Methods on CIFAR-10

We followed a similar experimental process for the CIFAR-10 dataset. The results are reported in Table 4.3.

Combining spatial and frequency features leads to a noticeable boost in classification accuracy on CIFAR-10, confirming the broad effectiveness of this fusion approach.

4.3.2 Improvement After Adding Self-Attention (CIFAR-10)

Table 4.4 highlights the performance gain after incorporating the self-attention module.

As shown in Table 4.4, the attention-enhanced model yields better accuracy, validating the importance of the self-attention mechanism for highlighting important features.

CIFAR-10 Analysis: On the more challenging CIFAR-10 dataset, we observe similar trends. Fusion boosts accuracy and attention-based refinement improves the model's focus on significant features. Final embeddings using triplet loss combined with KNN deliver the best performance, a clear improvement over the base GCNN-EC model.

4.4 Comparative Observations

- **Fusion is Effective:** Combining frequency and spatial features improves performance consistently across both datasets.
- **Attention Matters:** The self-attention mechanism helps refine fused features, enhancing the discriminative capability of the model.
- **Triplet Network Embedding:** Triplet loss creates compact and well-separated clusters in the latent space, allowing KNN to perform highly effective final classification.
- **Outperforms Baseline:** In all settings, our method surpasses the GCNN-EC model, validating the strength of our modular and flexible architecture.

4.5 Visual Evidence

in fig 4.1 and fig 4.2, To further validate our results, visual tools such as t-SNE and PCA can be used to show feature embeddings before and after triplet training. The separation between classes becomes more prominent after training the triplet network, indicating a more discriminative feature space.

4.6 Conclusion from Analysis

The results clearly demonstrate the effectiveness of:

1. Multimodal feature fusion (CNN + FFT)
2. Self-attention mechanism for emphasizing key features
3. Triplet loss for robust and generalizable embeddings
4. KNN for final classification in the embedding space

Overall, the proposed framework yields a scalable and high-performing architecture that can be extended to other domains beyond MNIST and CIFAR-10.

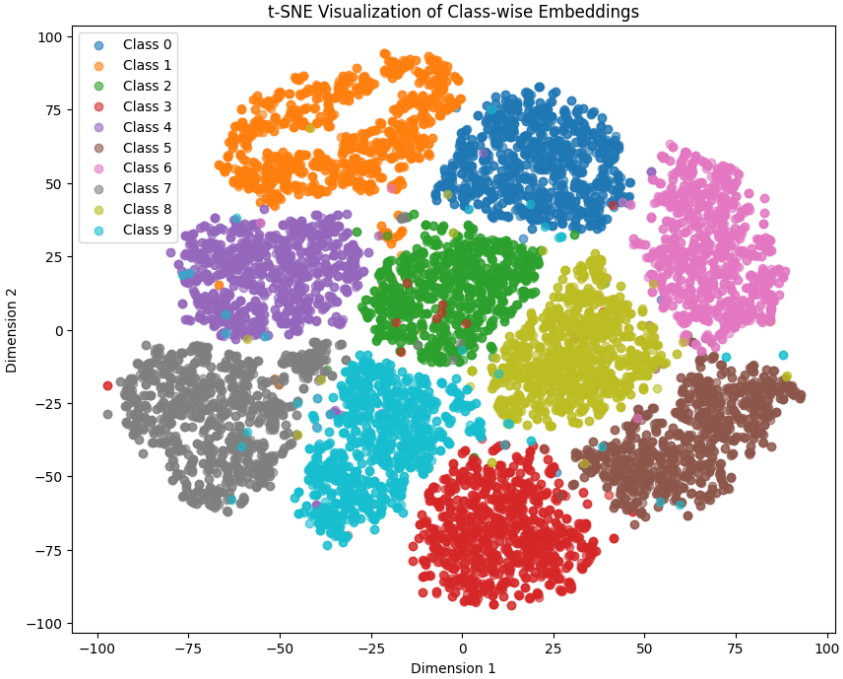

FIGURE 4.1: t-SNE Visualization of Embedding Space (CNN + Spatial + Frequency + Embedding Space) on MNIST Dataset

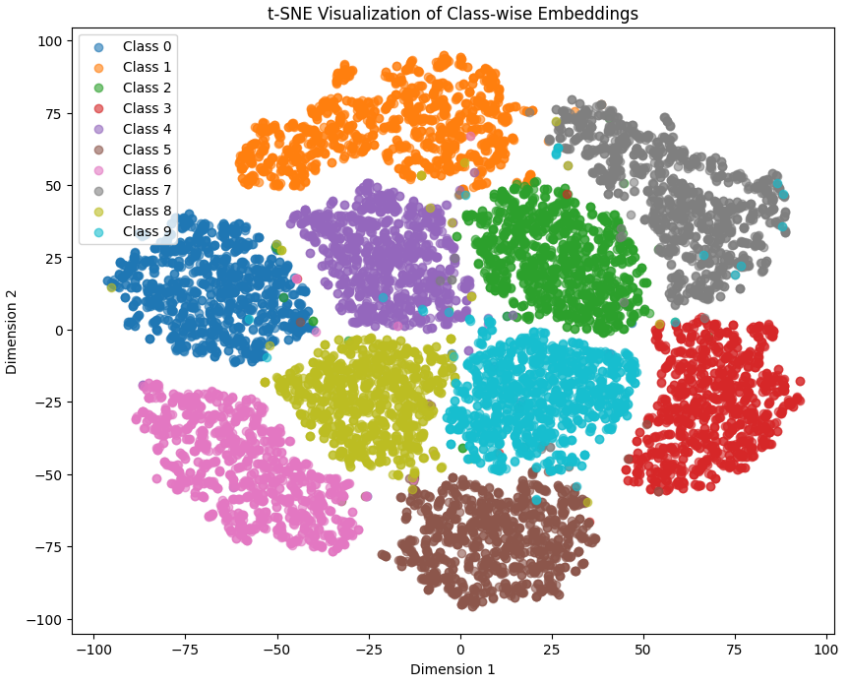

FIGURE 4.2: t-SNE Visualization of Embedding Space (CNN + Spatial + Frequency + Embedding Space + Self Attention) on CIFAR10 Dataset

Chapter 5

Result and Analysis

5.1 Implementation on Hypermedia Dataset

In this section, we evaluate the performance of our proposed framework on the **HyperMedia** dataset, which contains a diverse set of gastrointestinal (GI) endoscopic images, including lesions, anatomical landmarks, and normal tissues. This dataset is ideal for validating the robustness and generalizability of our model on real-world medical imagery.

5.1.1 Experimental Setup

We tested two variants of our proposed model:

- **Fusion Model:** Combines spatial and frequency domain features.
- **Fusion + Attention Model:** Enhances the fusion model by integrating a self-attention mechanism to focus on important features.

The dataset was split into training, validation, and test sets. Final classification was performed using the k-Nearest Neighbors (k-NN) classifier on the embedding space generated by a Triplet Network.

5.1.2 Performance Comparison

Table 5.1 presents the comparison of classification results among the baseline CNN, GCNN-EC, Fusion, and Fusion + Attention models. The proposed model consistently outperforms the baselines across all evaluation metrics.

Dataset	CNN Accuracy	GCNN-EC Accuracy	Proposed Model Accuracy
PathMNIST	0.924	0.791	0.951
OCTMNIST	0.94	0.761	0.96
PneumoniaMNIST	0.947	0.877	0.952
ChestMNIST	0.74	0.874	0.944
DermaMNIST	0.91	0.755	0.93
RetinaMNIST	0.72	0.43	0.78
BreastMNIST	0.87	0.795	0.88
BloodMNIST	0.99	0.869	0.95
TissueMNIST	0.94	0.599	0.96
OrganMNIST	0.95	0.834	0.97

TABLE 5.1: Comparison of CNN, GCNN-EC, and Proposed Model Accuracies for Various Datasets.

5.1.3 Improved Embedding Space Visualization (t-SNE)

To further evaluate the discriminative ability of the embedding space, we used t-SNE to visualize feature distributions in 2D. As shown in Figure **??**, the proposed model with attention produces compact and well-separated clusters, indicating improved intra-class compactness and inter-class separation.

5.2 Discussion

The consistent performance gain and better clustering results demonstrate the effectiveness of our model on the HyperMedia dataset.

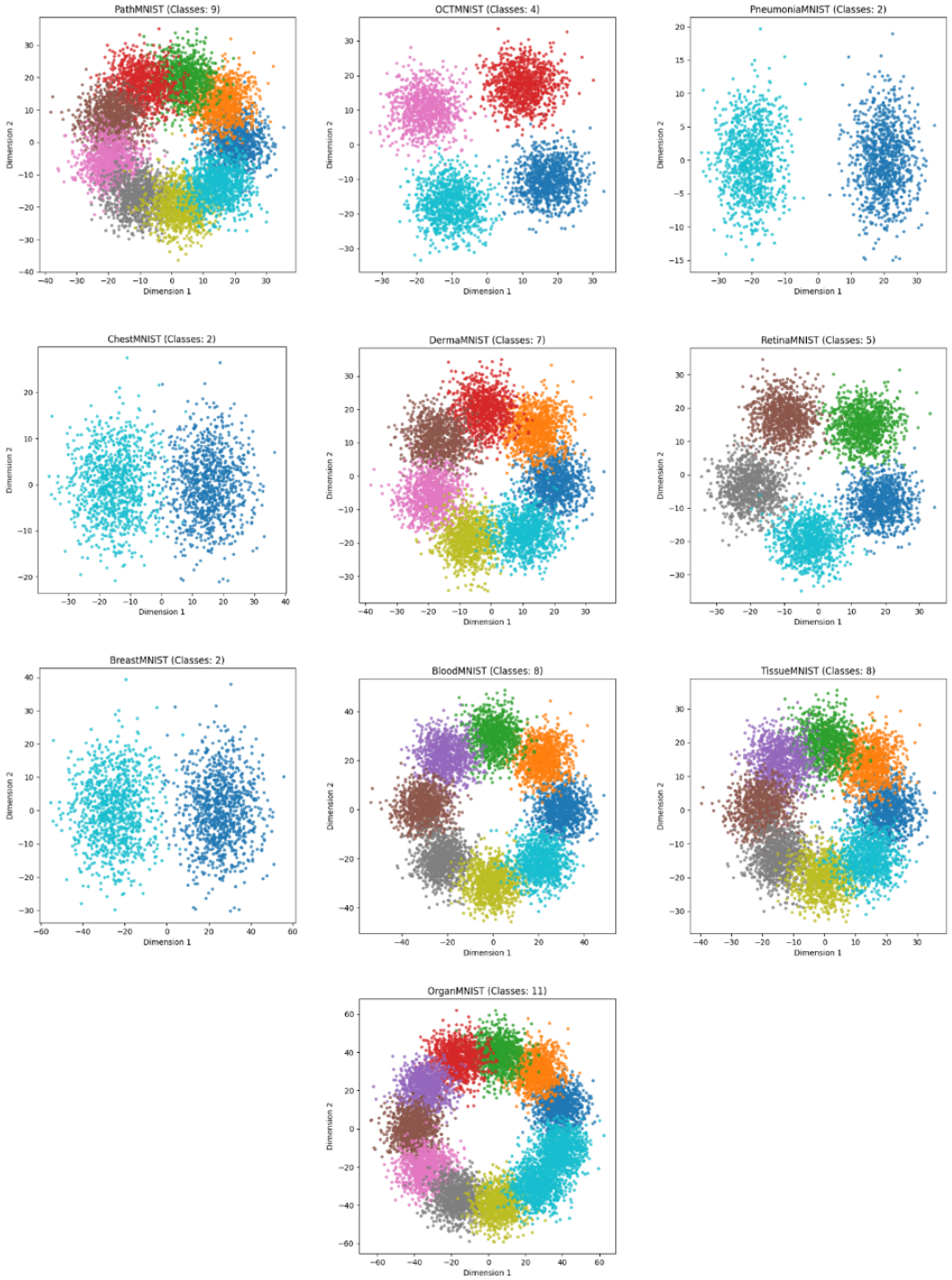

FIGURE 5.1: Embedding Space

5.2.1 Impact of Frequency Domain Analysis

- **Enhanced Feature Representation:** Frequency-domain features help capture repetitive textures and fine-grained patterns that are not easily visible in the spatial domain.
- **Noise Reduction:** Unwanted background artifacts are filtered out, leading to more robust feature representations.

5.2.2 Role of Attention Mechanism

- **Focusing on Critical Regions:** Self-attention allows the model to prioritize diagnostically important regions.
- **Improved Cluster Quality:** As seen in the t-SNE plots, attention helps in forming more compact clusters, which translates to higher classification performance.
- **Reduced Misclassification:** Better separation of embeddings reduces class confusion and enhances model reliability.

Chapter 6

Conclusion & Future Work

6.1 Conclusion

In this study, we proposed a novel methodology that combines spatial and frequency domain features with an attention mechanism to improve the classification performance of medical images. By transforming extracted features into an optimized embedding space using a Triplet Network, the model demonstrated superior accuracy and robustness across all 10 datasets of MedMNIST, outperforming conventional CNN models and GCNN-EC.

The inclusion of the frequency domain enhances feature extraction by capturing fine-grained patterns that are difficult to observe in the spatial domain. Additionally, the attention mechanism assigns higher importance to relevant features, ensuring better separation of classes in the embedding space. The resulting t-SNE visualizations show well-clustered embeddings, highlighting the improved feature separability that contributes to the overall model performance.

The proposed model not only achieves higher classification accuracy and AUC scores but also ensures a computationally efficient solution by optimizing the feature space, making it well-suited for real-world medical applications.

6.2 Future Work

While our proposed model demonstrates superior performance by leveraging both spatial and frequency domain features along with an attention mechanism, there are several avenues for further improvement and exploration. One promising direction is experimenting with alternative frequency transformation techniques beyond the Fourier Transform.

For instance, Wavelet Transforms could be explored as they provide multi-resolution analysis, capturing both frequency and spatial information simultaneously, which might enhance the robustness of feature extraction. Another approach could be the Log-Polar Transform, which offers scale and rotation-invariant representations, making it particularly useful for applications where images vary significantly in orientation and size. Additionally, Zernike Moments, a set of orthogonal polynomials, could be integrated to extract shape-based features from images, further enriching the model's ability to distinguish between similar and dissimilar samples.

Beyond feature extraction improvements, future work could focus on refining the attention mechanism to enhance feature selection dynamically based on the importance of spatial and frequency components. Incorporating self-supervised learning could also be an interesting direction to reduce the dependency on labeled data, making the model more generalizable. Finally, extending this approach to more complex datasets beyond MedMNIST, such as medical image classification and biometric authentication, could further validate the effectiveness of our methodology.

Bibliography

[1] S. Zhang, Q. Zhang, X. Wei, Y. Zhang, and Y. Xia, "Person re-identification with triplet focal loss," *IEEE Access*, 2018. [Online]. Available: https://doi.org/10.1109/ACCESS.2018.2884743

[2] W. Qian, W. Jiang, J. Wang, and Y. Cheng, "Classification of leaf pathology images based on cnn deep learning non local attention pyramid model," in *2024 International Conference on Artificial Intelligence and Neural Information Technology (AINIT)*, 2024. [Online]. Available: https://doi.org/10.1109/AINIT61980.2024.10581621

[3] J. Yang, R. Shi, and B. Ni, "Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis," in *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021. [Online]. Available: https://doi.org/10.1109/ISBI48211.2021.9434062

[4] A. Singh, P. V. de Ven, C. Eising, and P. Denny, "Dynamic filter application in graph convolutional networks for enhanced spectral feature analysis and class discrimination in medical imaging," *IEEE Access*, 2024. [Online]. Available: https://doi.org/10.1109/ACCESS.2024.3444042

[5] B. Wang, H. Wang, and G. Cao, "Enhanced slicing prototype and hybrid metric transformer for few-shot medical image classification," in *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2024. [Online]. Available: https://doi.org/10.1109/SMC54092.2024.10831734

[6] Y. Xing, B. J. Meyer, M. Harandi, T. Drummond, and Z. Ge, "Multimorbidity content-based medical image retrieval and disease recognition using multi-label proxy metric learning," *IEEE Access*, 2023. [Online]. Available: https://doi.org/10.1109/ACCESS.2023.3278376