International Conference on Signal Processing, Information, Communication and Systems 2025 (SPICSCON 2025)

Date: 21-22 November, 2025

BanglaMM-Disaster: A Multimodal Transformer-Based Deep Learning Framework for Multiclass Disaster Classification in Bangla

Authors: Ariful Islam[1], Md Rifat Hossen[1], Md. Mahmudul Arif[1], Abdullah Al Noman[2], Md Arifur Rahman[3]

Affiliations: [1]Chittagong University of Engineering and Technology, Bangladesh [2]Wilmington University, USA [3]Trine University, USA

Paper ID. 122

Overview

- Problem Statement & Motivation
- Dataset & Disaster Categories
- Multimodal Methodology
- Experimental Results
- Conclusions & Future Work

Problem Statement & Motivation

The Challenge

- Bangladesh faces frequent natural disasters.
- Massive multilingual social media data during disasters.
- Bangla lacks robust disaster classification tools.
- Need for rapid, accurate disaster identification.

Why Multimodal?

- Text alone misses visual context.
- Images alone lack situational details.
- Combining both improves accuracy.

Research Gap: Limited multimodal classification for low-resource Bangla.

Research Objectives

1

Create Bangla Multimodal Dataset

5,037 text-image pairs across 9 disaster categories.

2

Develop Multimodal Framework

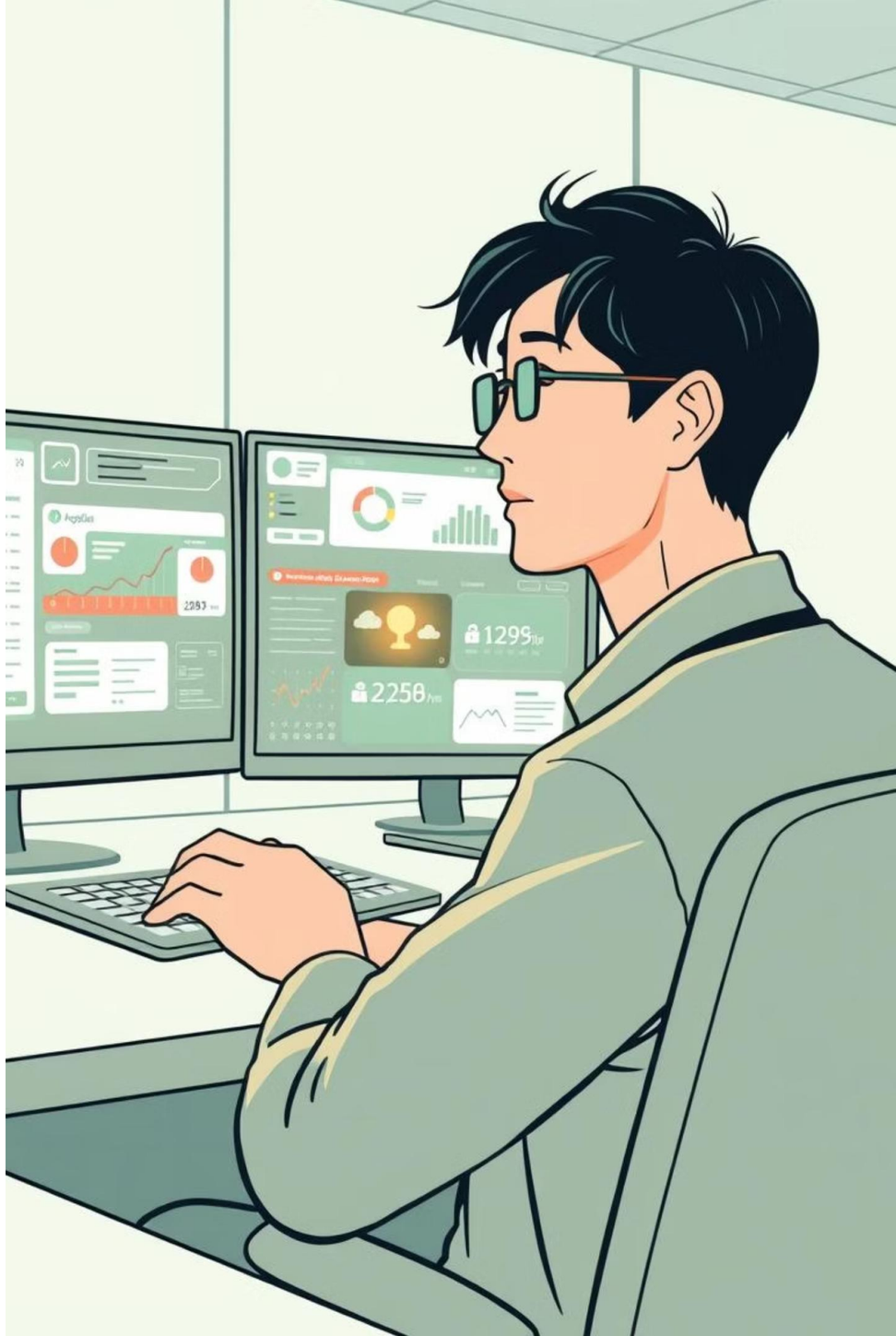
Transformer-based text, CNN-based image encoders, early fusion.

3

Evaluate Performance

Compare model combinations and state-of-the-art approaches.

Target: Improve disaster response through automated social media monitoring.



Dataset Overview: BanglaMM-Disaster

9 Disaster Categories

- Earthquake (13.87%)
- Flood (12.35%)
- Fire (11.30%)
- Landslide (10.66%)
- Drought (10.49%)
- Cyclone (10.03%)
- Pandemic (9.27%)
- Humanitarian Crisis (11.14%)
- Irrelevant (10.89%)

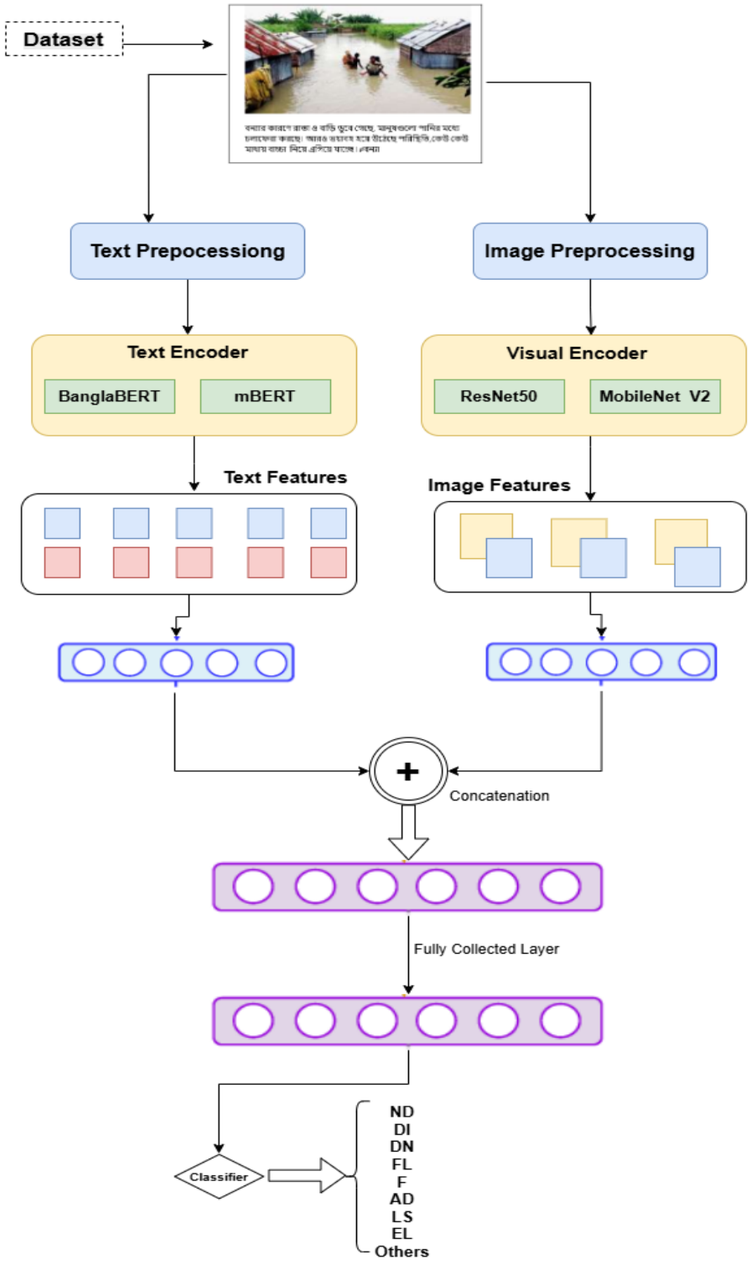
Dataset Statistics

- Total samples: 5,037 text-image pairs.
- Source: Social media posts (Facebook, Twitter).
- Annotation: Manual labeling by native Bangla speakers.
- Split: 70% training, 10% validation, 20% testing.

Key Features

- Balanced class distribution.
- Real-world social media content.
- Authentic Bangla language usage.

Proposed Methodology: System Architecture



1

Text Processing

Bangla text input, preprocessing, then BanglaBERT, mBERT, XLM-RoBERTa for feature embeddings.

2

Image Processing

Disaster images input, preprocessing, then ResNet50, DenseNet169, MobileNetV2 for feature embeddings.

3

Multimodal Fusion

Early fusion (concatenate text + image features), fully connected layers, Softmax classification (9 classes).

Model Configurations

Text Encoders

- **BanglaBERT:** Pre-trained on Bangla corpus, 110M parameters.
- **mBERT:** Multilingual BERT, 172M parameters.
- **XLM-RoBERTa:** Cross-lingual model, 270M parameters.

Image Encoders

- **ResNet50:** Residual networks, 25.6M parameters.
- **DenseNet169:** Dense connections, 14.1M parameters.
- **MobileNetV2:** Lightweight architecture, 3.5M parameters.

Training Configuration

- **Optimizer:** Adam (LR: 1e-5 text, 1e-4 images).
- **Batch size:** 32.
- **Epochs:** 20 with early stopping.
- **Loss function:** Categorical cross-entropy.

Experimental Results: Unimodal Performance

Image-Only Models

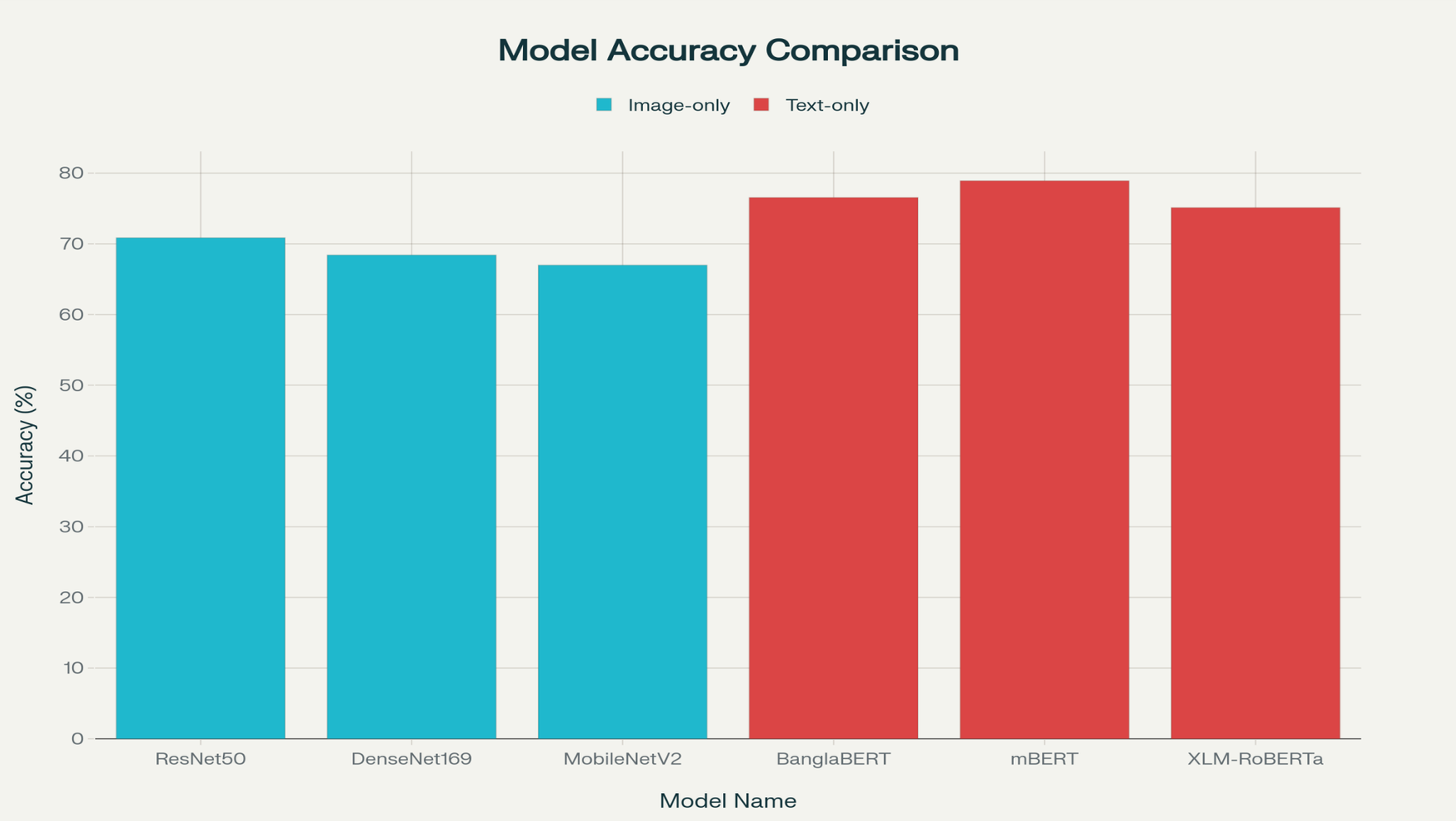
- **ResNet50:** 66.85% accuracy
- **DenseNet169:** 65.47% accuracy
- **MobileNetV2:** 62.41% accuracy

Text-Only Models

- **BanglaBERT:** 76.73% accuracy
- **mBERT:** 78.15% accuracy
- **XLM-RoBERTa:** 79.92% accuracy

Key Observation

Text models generally outperform image-only models. mBERT shows the strongest text performance. Visual features alone are insufficient for complex disaster classification.



Multimodal Fusion Results

mBERT + ResNet50: 83.76%

Precision: 84.12%, Recall: 83.76%, F1-Score: 83.89%

mBERT + DenseNet169: 82.45%

XLM-RoBERTa + ResNet50: 81.33%

Performance Gains

- +3.84% over best text-only model.
- +16.91% over best image-only model.

Why This Works

Complementary information from text and images. mBERT captures linguistic context, ResNet50 extracts robust visual features.

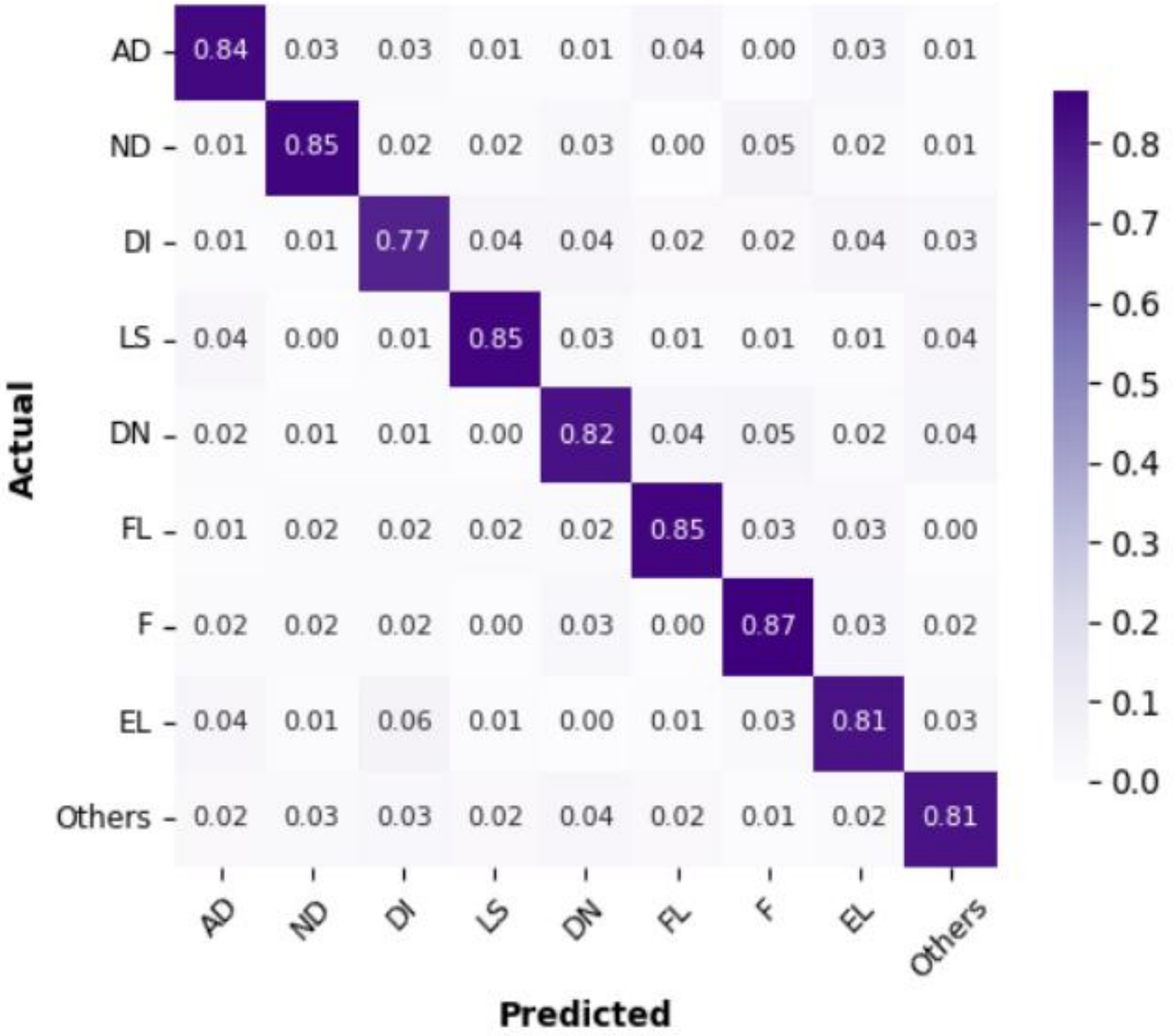
Confusion Matrix Analysis

Key Insights

- **Strong Performance:** Earthquake, Flood, Fire (>85% precision).
- **Challenging Classes:** Humanitarian Crisis vs Pandemic, Drought vs Irrelevant.
- **Class-wise Accuracy:** Best: Earthquake (91.2%), Most challenging: Humanitarian Crisis (76.8%).

Error Analysis

Multimodal fusion reduces misclassification by 23% compared to unimodal approaches.



Comparison with State-of-the-Art

Traditional ML (SVM)	62.4%
CNN-only	68.7%
LSTM-only	71.3%
BERT + CNN (Related work)	79.2%
Our BanglaMM-Disaster	83.76%

Significant Improvement

+4.56% over previous best methods.

Tailored for Bangla

Optimized for the unique characteristics of the Bangla language.

Robust Data

Utilizes a larger, more diverse dataset for training.

Performance Metrics

Computational Efficiency

Model Complexity

197.6M parameters
(mBERT + ResNet50), 1.8 GB size.

Inference Performance

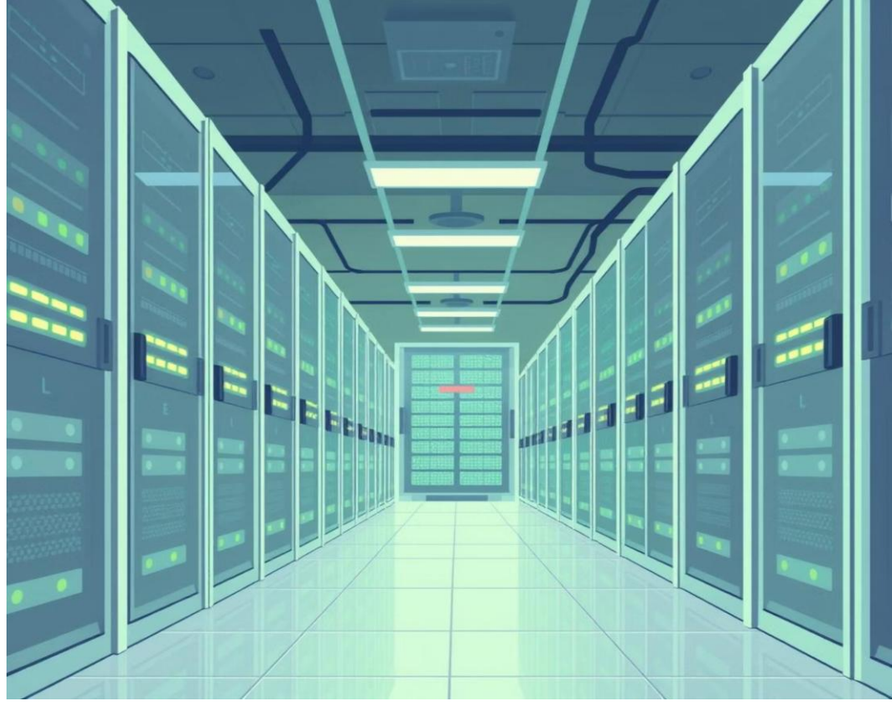
0.45 seconds per sample, suitable for real-time.

Resource Requirements

Training on NVIDIA RTX 3090
(~6 hours).

Scalability

Processes 2,000+ posts/minute for disaster monitoring.



Qualitative Analysis

Success Cases

→ Clear imagery + explicit disaster mentions yield 95%+ accuracy.

→ Consistent terminology across text and image.

Failure Cases

→ Sarcastic/metaphorical language.

→ Poor quality images or ambiguous contexts.

Model Strengths

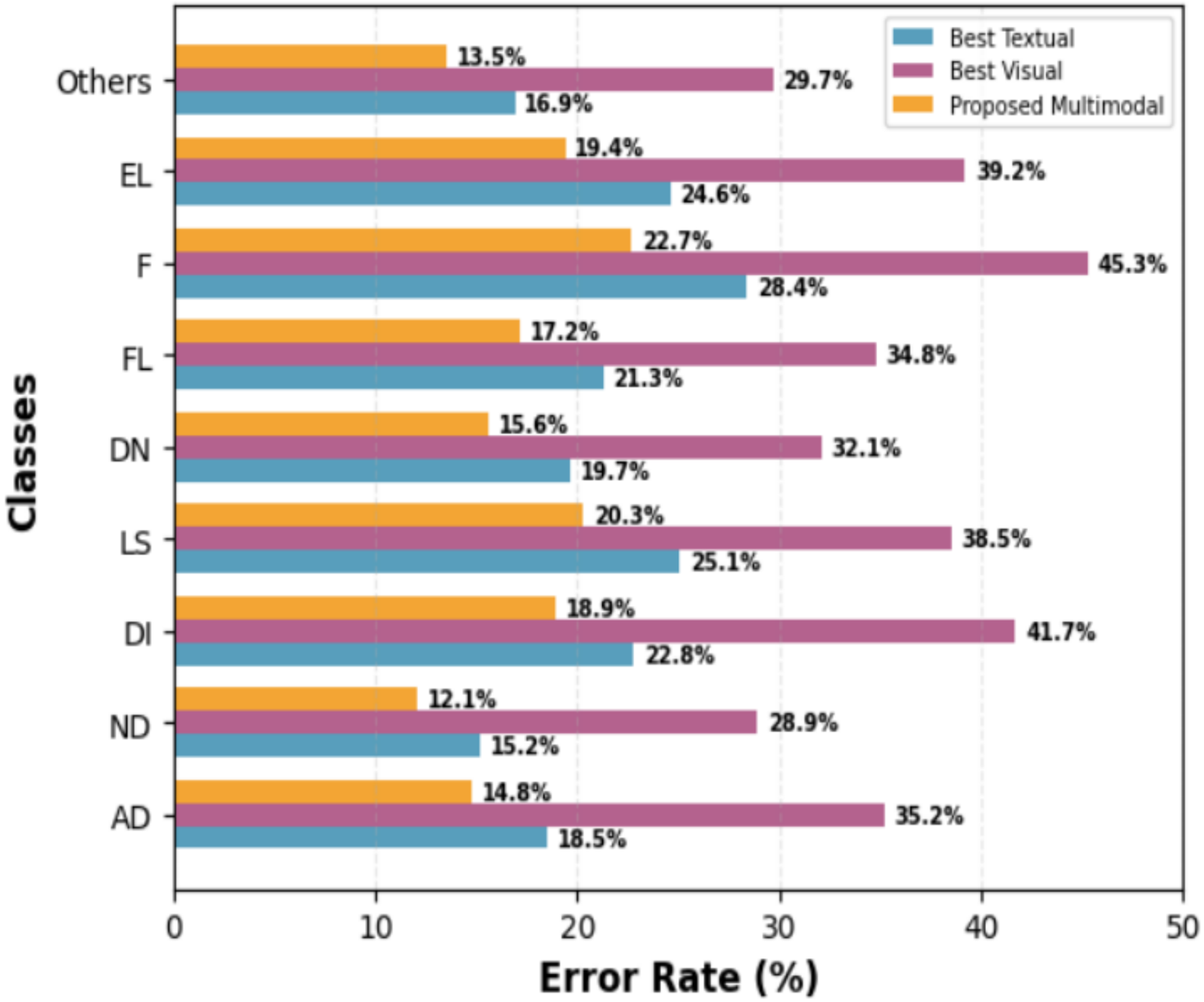
→ Handles code-mixing (Bangla-English).

→ Robust to social media noise and varied image quality.

Limitations

→ Struggles with rare disaster types.

→ Needs more training data for minority classes.



Key Contributions

Novel Dataset

First large-scale Bangla multimodal disaster dataset: 5,037 annotated samples, 9 categories, publicly available.

Multimodal Framework

Transformer + CNN architecture with optimized early fusion, achieving 83.76% accuracy (SOTA)

Comprehensive Evaluation

Tested 9 model combinations, detailed performance analysis, and validated computational efficiency.

Real-world Applicability

Fast inference (0.45s), scalable architecture, and deployment-ready for disaster management.

Conclusions & Future Work

Conclusions

- Developed effective multimodal framework for Bangla disaster classification.
- Multimodal approach significantly outperforms unimodal methods (+3.84%).
- mBERT + ResNet50 achieves best performance (83.76%).
- Framework suitable for real-time disaster monitoring.

Future Research Directions

01

Dataset Expansion

Include more disaster types and increase samples for minority classes.

02

Model Enhancement

Explore late/hybrid fusion, attention mechanisms, and domain adaptation.

03

Real-world Deployment

Integrate with disaster management systems, mobile apps, and multi-language support.

04

Advanced Features

Add temporal analysis, geolocation, and severity classification.

Thank You!

Questions & Discussion

Contact: Md Rifat Hossen, Department of Computer Science and Engineering, Chittagong University of Engineering and Technology

Email: rifat8851@gmail.com

Acknowledgments:

CUET Department of CSE, Dataset annotation team, IEEE SPICSCON 2025 organizing committee.