Lightweight Multi-modal Emergency Detection and Translation for Extremely Low-Resource Contexts

Anonymous ACL submission

Abstract

Emergency reporting in remote regions is often delayed by infrastructural challenges and language barriers. While multimodal AI offers a solution, its deployment is hindered by scarce localized data and computational constraints. This paper addresses extreme data scarcity by proposing and evaluating a lightweight, Vision-to-Telugu emergency classification pipeline. We use a novel, imbalanced 70-image dataset (6 categories, e.g., fire, snake bite) to simulate this data constraint. We benchmark 15 vision encoders, pairing the classifier with a zero-overhead dictionary for 100% accurate Telugu translation. To validate our small-set results, we conduct a Bootstrap-Wilcoxon statistical analysis. Our findings show DINOv2-Base (82.45% mean accuracy) statistically significantly outperforms a CLIP-ViT-B32 baseline (53.91%) with a large effect size ($p < 0.001$, $\delta = +0.820$). This work provides a blueprint and robust validation methodology for effective multi-modal systems in severe data-constrained, social-impact settings.

1 Introduction

In many remote and underserved communities, access to rapid emergency services is severely limited. This "digital divide" is compounded by infrastructural gaps, low network connectivity, and significant language barriers, which can prevent critical alerts from reaching authorities (Chakravarthi et al., 2022). While AI-driven systems, including multimodal approaches for categorizing crisis events (Abavisani et al., 2020), present a promising solution, they face two practical hurdles: (1) the computational constraints of deploying models on low-power, "edge" devices, and (2) the extreme scarcity of localized, labeled training data.

This second challenge defines a "few-shot learning" problem, where a model must generalize from a tiny number of examples. For a practical solution, a model must also be highly efficient, adhering to design principles established by lightweight architectures (Sandler et al., 2019).

Given these constraints, the choice of a pre-trained vision "backbone" is the most critical factor for success. Modern-day models are dominated by new paradigms, such as vision-language supervision (Radford et al., 2021), powerful self-supervised methods (Oquab et al., 2024), and advanced transformer architectures (Liu et al., 2021). However, it is unclear which of these SOTA pre-training strategies performs best in an extremely low-resource, social-impact context.

This paper tackles this question directly. We present a feasibility study and robust benchmark for a lightweight, multimodal (Vision-to-Telugu) emergency alert system. We introduce a novel, imbalanced 70-image dataset to simulate a realistic, data-scarce environment. Our primary contribution is a rigorous, statistical comparison of modern vision backbones. We find that the self-supervised features from DINOv2(Oquab et al., 2024) are dramatically more effective for this few-shot task than the features learned by the vision-language model CLIP (Radford et al., 2021), providing a clear direction for building practical, low-resource intervention tools.

2 Related Work

Our work is positioned at the intersection of few-shot learning, efficient model deployment, and low-resource multi-modal contexts.

Few-Shot & Low-Resource Learning: Standard deep learning assumes large, balanced datasets. In contrast, our 70-image dataset represents an "extreme" or "few-shot" learning problem, requiring models to learn effectively from minimal data. In such scenarios, the quality of a model's pre-trained features is paramount. Our work does not propose a new few-shot algorithm, but rather evaluates the practical utility of existing SOTA models in this critical, low-resource setting.

Efficient Vision Models: A practical solution for remote areas cannot rely on cloud connectivity. This necessitates models designed for on-device efficiency. The principles of lightweight architectures, such as those pioneered by MobileNetV2 (Sandler et al., 2019), set the standard for low-power mobile vision. We build on this by evaluating modern architectures for their suitability in a high-accuracy, low-resource, low-compute benchmark.

Modern Pre-training Strategies: Our core contribution is the comparison of dominant pre-training paradigms. We compare:

- **Vision-Language Supervision:** Models like CLIP (Radford et al., 2021) that learn features by mapping images to text from web-scale datasets.
- **Hierarchical Vision Transformers:** Architectures like the Swin Transformer (Liu et al., 2021), which have become a powerful, SOTA backbone for many vision tasks.
- **Self-Supervised Learning:** Models like DINOv2 (Oquab et al., 2024) that learn rich, robust features from images alone, without any human-provided labels or text.

Our work directly compares the efficacy of these strategies in a practical, data-scarce social impact problem.

Low-Resource Language Contexts: For the multimodal translation component, we target Telugu, a language underrepresented in many large-scale NLP datasets (Chakravarthi et al., 2022). This work acknowledges that a full, power-hungry translation model is often brittle and impractical in this context. We therefore opt for a zero-overhead, deterministic dictionary, which is 100% accurate for our defined classes and represents the most robust, lightweight solution for this application.

3 Methodology

3.1 Proposed System Workflow

The system is specifically designed for practical deployment in extremely low-infrastructure environments, such as remote and tribal regions, where users typically rely on basic, low-resource smartphones and experience intermittent network connectivity. As illustrated in Figure 1, the workflow begins when a tribal or remote user captures an image of a perceived emergency. This image is immediately processed by a lightweight, on-device vision model, which efficiently classifies the situation into one of six pre-defined categories. Crucially, this processing happens locally on the device, minimizing the need for immediate network access.

The resulting classification (e.g., ’fire’) is paired with the device’s location coordinates and instantly translated into the local language (Telugu) using a zero-overhead, deterministic dictionary (e.g., ’agni pramadam’ in telugu scrpt). Finally, this minimal, text-based alert (consisting of the Telugu label and GPS coordinates – only a few bytes of data) is transmitted to local authorities. The critical advantage of this "classify-then-translate" design is its extreme resilience to poor connectivity: by sending only a minimal text string instead of a large image file, the system can reliably deliver urgent information even over very low-bandwidth or intermittent mobile signals, ensuring timely intervention in areas previously disconnected from rapid response services.

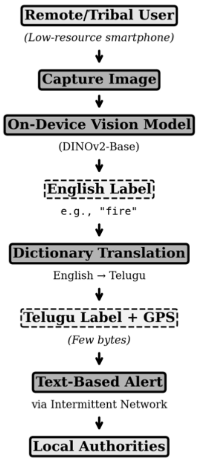


Figure 1: System Workflow

3.2 Emergency-Vision Dataset

To simulate a realistic, "found data" scenario for a low-resource deployment, we manually curated and labeled a novel dataset. This dataset consists of 70 total images sourced from public web domains.

Each image was manually labeled into one of six categories: fire, flood, building collapse, road accident, snake bite, or no condition.

Crucially, the dataset is intentionally small and highly imbalanced, reflecting the practical difficulty of acquiring labeled data for rare or sensitive events. The class distribution is highly skewed, with no condition (24 images, 34.3%) being the largest class and snake bite (4 images, 5.7%) being the smallest.

Furthermore, to ensure the benchmark reflects real-world challenges, the dataset includes images captured under diverse lighting conditions, containing 38 "day" images and 28 "night" images. This challenging, low-data, and imbalanced dataset forms the basis of our few-shot learning evaluation.

4 Experimental Setup

Our experiment is designed to identify the most robust and lightweight vision model for our proposed workflow, given the extreme data constraints.

Training Pipeline: We split our 70-image dataset into a 70% training (49 images), 15% validation (10 images), and 15% test set (11 images). We employed a standard transfer learning approach: for each model, the pre-trained backbone was frozen, and only a simple classifier head—consisting of a 50% dropout layer and a final linear layer—was trained (to account for the low data).

All models were trained for a maximum of 30 epochs using the AdamW optimizer (learning rate $1e-4$, 0.01 L2 weight decay) and a CrossEntropyLoss function. We used a ReduceLROnPlateau scheduler and EarlyStopping with a patience of 10 epochs. To compensate for the small dataset, we applied heavy data augmentation during training, including random cropping, horizontal flipping, rotation, color jitter, and random erasing.

Model Benchmarking: We conducted a comprehensive benchmark of 15 models, including standard CNNs (e.g., MobileNetV2, EfficientNet-B0), Vision Transformers (ViT, BEiT), and state-of-the-art self-supervised (DINO, DINOv2) and vision-language (CLIP) models. Due to the 4-page limit, we present results for the top 3 performing models and the CLIP-ViT-B32 model, which serves as our robust baseline.

Translation Module: The translation step is implemented as a zero-overhead, $O(1)$ deterministic dictionary lookup. This component maps the 6 English class labels from the classifier to their validated Telugu translations (e.g., 'fire' → 'agni pramaadam' in telugu script). This approach was chosen over an ML translation model as it is 100% accurate for our domain, instantaneous, and adds zero computational or memory load to the on-device application, making it ideal for the low-resource context.

4.1 Statistical Validation

Given our extremely small test set (11 images), a single test accuracy score is highly sensitive to the specific data split and is not a reliable measure of performance. A model achieving 90% vs. 80% accuracy could be a difference of one single image.

To rigorously validate our findings, we employed a non-parametric statistical comparison. We used Bootstrap Resampling to simulate 1,000 different test sets. For each of the $N = 1,000$ iterations, we created a new test set by sampling with replacement from our original 11-image test set.

On each bootstrapped sample, we calculated the accuracy for our baseline (CLIP-ViT-B32) and for each of the top-performing models. This process generated 1,000 paired accuracy scores for each comparison. We then used the Wilcoxon signed-rank test—a robust test for non-normally distributed data—to determine if the observed performance difference between a model and the baseline was statistically significant ($p < 0.05$). We also computed Cliff's Delta (δ) to measure the effect size (i.e., the magnitude of the performance gain).

5 Results and Discussion

The results of our statistical validation are presented in Table 1.

Table 1: Statistical comparison of top models against the CLIP-ViT-B32 baseline, based on $N = 100$ bootstrap samples. All comparisons show a statistically significant difference ($p < 0.001$).

Our key finding is that DINOv2-Base is the clear winner, achieving a mean accuracy of 82.45% across the bootstrap samples—a massive +28.55% absolute gain over the CLIP-ViT-B32 baseline (53.91%). This improvement is statistically significant ($p < 0.001$) with a "large" effect size ($\delta = 0.82$). The other models, DINOv2-Large ($\delta = 0.64$) and Swin-Tiny ($\delta = 0.29$), also significantly outperformed the baseline.The superior performance of DINOv2-Base over DINOv2-Large (74.91%) suggests the larger model overfit our small 49-image training set, while the "Base" model offered better generalization. This strongly supports using DINOv2's self-supervised features for few-shot tasks over CLIP's language-supervised ones.

We then analyzed per-class F1 scores of the models and confusion matrix (DINOv2-Base) from the original 11-image test set.

Model	Mean Acc.	Baseline Acc.	Diff.	p-value	Cliff's δ
DINOv2-Base	**82.45%**	53.91%	+28.55%	< 0.001	**0.82**
DINOv2-Large	74.91%	53.91%	+21.00%	< 0.001	0.64
Swin-Tiny	63.55%	53.91%	+9.64%	< 0.001	0.29

Table 1: Statistical comparison against the 'CLIP-ViT-B32' baseline, based on $N = 100$ bootstrap samples. All comparisons show a statistically significant difference ($p < 0.001$).

Figure 2 compares the per-class F1 scores of our top two models against the baseline. This chart reinforces DINOv2-Base's dominance.

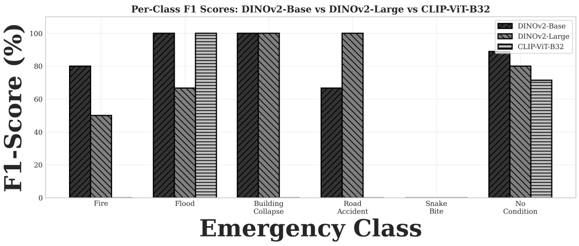


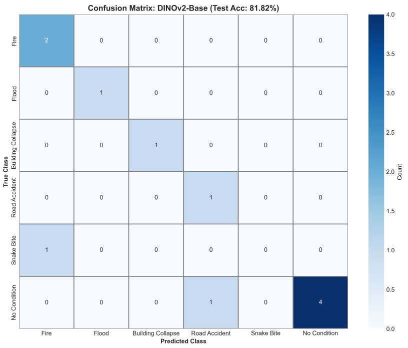
Figure 2: Per-class F1 scores on the test set.

The confusion matrix for DINOv2-Base (Figure 3) provides a clear diagnosis of the 81.82% accuracy (9/11 correct). We see two misclassifications: one 'No Condition' image was mistaken for a 'Road Accident', and, critically, the single 'Snake Bite' test image was misclassified as 'Fire'.

This 'Snake Bite' error is the reason for the 0% F1 score for that class in Figure 2. Notably, this was a universal failure across all benchmarked models, as seen in Figure 2. This is an expected and direct consequence of the extreme class imbalance in our 70-image dataset, which contained only four 'snake bite' images (5.7% of the data). With such a minimal training signal, no model was able to learn distinguishing features for this rare class. This highlights a clear limitation and an important area for future work, which would involve targeted data augmentation or sourcing for this critical, under-represented category.

6 Conclusion

We demonstrated the feasibility of a lightweight, multi-modal (Vision-to-Telugu) emergency alert system for extremely low-resource settings. Using a 70-image few-shot dataset, our statistical analysis proved that DINOv2-Base provides a +28.55% absolute accuracy gain over the CLIP-ViT-B32 baseline, validating the use of self-supervised models for data-scarce social good tasks. The primary limitation was the failure of all models to classify 'snake bite' due to insufficient training data.



	Fire	Flood	Building Collapse	Road Accident	Snake Bite	No Condition
Fire	2	0	0	0	0	0
Flood	0	1	0	0	0	0
Building Collapse	0	0	1	0	0	0
Road Accident	0	0	0	1	0	0
Snake Bite	1	0	0	0	0	0
No Condition	0	0	0	1	0	4

Figure 3: Confusion matrix for DINOv2-Base on the 11-image test set (81.82% accuracy). The 'Snake Bite' class was misclassified as 'Fire'.

7 Limitations and Ongoing Work

The primary limitation of our study is the 70-image dataset, which is insufficient for production-level deployment. This extreme data scarcity, particularly for the 'snake bite' class (only 4 training images), led to a universal classification failure for that category.

Our ongoing work focuses on three areas. First, we are actively collecting a larger and more diverse dataset across all categories to improve model robustness and address these data-scarce limitations. Second, we are packaging the DINOv2-Base pipeline into a lightweight PWA (Progressive Web App) for field testing. Finally, to scale beyond our initial 6 classes, we plan to replace the static dictionary with a lightweight, on-device translation model that can handle more nuanced alert messages.

References

Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. 2020. Multimodal categorization of crisis events in social media. *Preprint*, arXiv:2004.04917.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Subalalitha Cn, Sangeetha S, Malliga Subramanian, Kogilavani Shanmugavadivel, Parameswari Krishnamurthy, Adeep Hande, Siddhanth U Hegde, Roshan Nayak, and Swetha Valli. 2022. Findings of the shared task on multi-task learning in Dravidian languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 286–291, Dublin, Ireland. Association for Computational Linguistics.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *Preprint*, arXiv:2103.14030.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, and 7 others. 2024. Dinov2: Learning robust visual features without supervision. *Preprint*, arXiv:2304.07193.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2019. Mobilenetv2: Inverted residuals and linear bottlenecks. *Preprint*, arXiv:1801.04381.