**LEVERAGING MuRIL FOR PHISHING DETECTION IN THE INDIAN CONTEXT**

| **Muskan**  **New Delhi, India** | **Ranjan**  **New Delhi, India** | **Utkarsh Gupta**  **New Delhi, India** |
| --- | --- | --- |
| **Netaji Subhas University of Technology** | **Netaji Subhas University of Technology** | **Netaji Subhas University of Technology** |
| **muskan.ug21@nsut.ac.in** | **ranjan.ug21@nsut.ac.in** | **utkarsh.gupta.ug21@nsut.ac.in** |

***Abstract*— With the proliferation of mobile internet services in India, phishing via SMS (smishing) has become an urgent cybersecurity concern, especially for users of Indian regional languages. In this study, we leverage the MuRIL transformer model. Our results highlight the potential of MuRIL in building scalable smishing detection systems for low-resource languages.**

**Keywords- Cyber-security, Few-shot learning, Indic languages, Low-resource languages, MuRIL, Smishing detection**

## **I. INTRODUCTION**

Phishing via SMS (smishing) poses a significant cybersecurity threat in multilingual India, where Hindi internet users now outnumber English users. Despite this linguistic shift, most detection systems remain optimized solely for English content, leaving speakers of Indian languages

vulnerable to localized attacks.  
**A. Background and Literature Review**

Traditional phishing detection techniques rely on rule-based systems and classical

machine learning that struggle with semantic complexities of non-English languages . Key limitations include focus on English-language content, scarcity of annotated datasets in Indian languages, and

Inability to adapt to culturally-contextualized persuasion techniques. [1][2].

Recent transformer-based models like Google's MuRIL [3] offer promising solutions, being pre trained on 17 Indian languages in both native and transliterated scripts (Kakwani et al., 2020). Prior approaches include Hybrid BERT-CNN models for Bangla smishing (Ahmed et al., 2025) and text normalization-enhanced Naïve Bayes for Hindi (Goel et al., 2024), though both struggled with either computational demands or semantically sophisticated attacks.

Few-Shot Learning (FSL) has emerged as a valuable approach for low-resource languages, enabling models to generalize from minimal examples (Yin et al., 2020; Snell et al., 2017).

**B. Research Objectives**

This study aims to develop an effective smishing detection model for low-resource Indian languages (primarily Hindi and Bangla) using Google's Multilingual Representations for Indian Languages (MuRIL) transformer model. We propose supervised fine-tuning and few-shot learning approaches, addressing three critical gaps in existing research: inefficient handling of code-mixed texts, dependency on large labeled datasets, and prohibitive computational costs for real-time deployment. We also aim to examinine cross-lingual transfer capabilities while comparing MuRIL-based classification against traditional machine learning methods.

Our work addresses a critical security gap for millions of non-English-speaking internet users, contributing to more inclusive cybersecurity solutions for India's linguistically diverse population.

## **II. MATERIALS AND METHODS**

This study implemented a systematic approach for detecting SMS phishing (smishing) in low-resource Indian languages, specifically Hindi and Bangla. The methodology encompassed dataset preparation, model selection, supervised learning, and few-shot learning techniques.

**A. Dataset Description**Two primary datasets were utilized:

1. Multilingual Spam Data (Patel, 2022) with 5,572 entries (87% ham, 13% spam) containing English messages with Hindi translations.
2. Bangla SMS Dataset for Smishing Detection (Das, 2023) with 2,602 entries in Bangla script, labeled as spam or ham.

**B. Model Selection**MuRIL was selected as the primary transformer model due to its advantages over alternatives like IndicBERT, as shown in Table I.

**Table I.** Comparative Analysis of MuRIL and IndicBERT

| **Feature** | **MuRIL** | **IndicBERT** |
| --- | --- | --- |
| Language Coverage | Supports 17 Indian languages | Supports 12 Indian languages (subset of MuRIL) |
| Corpus Size & Diversity | Trained on Wikipedia, news, and supervised translation pairs in native and Roman scripts | Trained on Wikipedia and OSCAR corpus in native script only |
| Handling of Transliterated Text | Explicitly trained on both native-script and transliterated (Romanized) Indian languages | Not trained on transliterated text; performs poorly on Romanized inputs |
| Multilingual Robustness | Better generalization to code-switched, noisy, and multilingual inputs, short, informal texts | Limited generalization; works best on clean native-script monolingual text |

**C. Implementation**Two distinct approaches were implemented:  
**a. Supervised Fine-Tuning Pipeline:**  
The supervised approach involved:

* Data preprocessing with label encoding: (spam=1, ham=0)
* 80/20 train-test split with stratification
* MuRIL tokenization with maximum sequence length
* Fine-tuning the MuRIL model on the complete training data

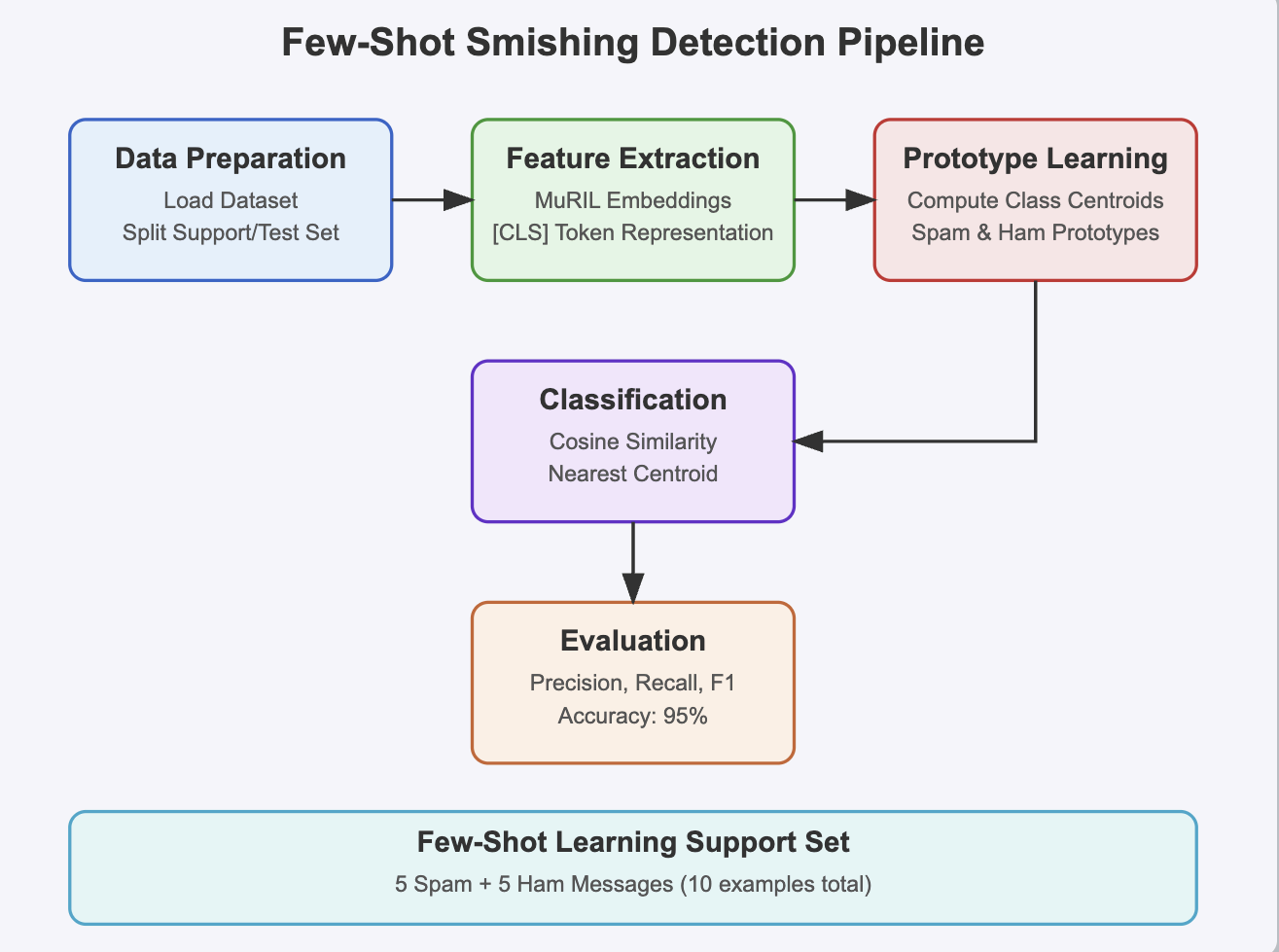
**b. Few-Shot Learning Pipeline:**(Y. Wang et al.,2020) discusses various Few Shot Learning approaches, as shown in Table II :

**Table II**. Feasible Few Shot Learning Approaches with MuRIL

| **Approach** | **Supported by MuRIL** | **Summary** |
| --- | --- | --- |
| Small-scale fine tuning | Yes | Fine-tune MuRIL on small labeled Hindi dataset |
| Parameter-Efficient Fine Tuning (PEFT)  eg. LoRA, Prefix Tuning | Yes | Only fine-tune a small number of parameters (lighter, faster) |
| Embedding-Based (Prototypical networks) | Yes | Use MuRIL to get embeddings → classify new samples based on distance to class centroids |
| Prompt based (inference only) | No | MuRIL isn’t trained to generate/comprehend prompts like T5/GPT model |

For low-resource settings, an embedding-based few-shot learning approach was implemented, selected for its lightweight nature and data efficiency.The pipeline (as shown in Figure 1) for Few Shot Learning involves:

* Support/test split with balanced class representation (5 examples per class)
* MuRIL-based sentence embedding generation
* Classification using cosine similarity with support set centroids

**Figure 1. Pipeline for Smishing Detection Code using Few Shot Learning**

**D. Tools and Libraries**Implementation was carried out in Python using:

* Data handling: Pandas, NumPy
* NLP: Hugging Face Transformers (for MuRIL), Datasets
* Machine Learning: Scikit-learn (for evaluation metrics and similarity measures)
* Deep Learning: PyTorch (backend framework)
* Development: Kaggle Notebooks (providing GPU resources)

Both approaches were evaluated using standard classification metrics including accuracy, precision, recall, and F1-score, with particular emphasis on balanced performance across languages and minimizing false negatives in spam detection. **III. RESULTS**This section evaluates the performance of multilingual spam classification using supervised fine-tuning, and few-shot learning. Experiments were conducted across English, Bangla, and Hindi datasets, using standard metrics: accuracy, precision, recall, and F1-score, results as shown in Table III, and Figure 2.

**A. Supervised Learning (Fine-Tuned MuRIL)**

* **English:** Achieved 99% accuracy with excellent balance across classes. Misclassifications were rare, mostly involving promotional messages lacking clear spam signals.
* **Bangla:** Strong spam precision (0.99). False negatives emerged in informal or dialect-rich spam messages, slightly lowering F1-score compared to English.
* **Hindi:** Spam recall dropped to 0.92. Errors were linked to code-mixed inputs and transliterated content. Some legitimate service messages were flagged as spam due to formatting similarities.

**B. Few-Shot Learning**

Few-shot learning used MuRIL in an in-context setup without additional training. Performance varied by language.

* **English:** Strong performance with F1-scores of 0.97 (Ham) and 0.81 (Spam). Errors mostly involved mistaking OTP or service messages as spam.
* **Bangla:** F1-scores were balanced (Ham: 0.79, Spam: 0.82). Errors arose from lack of linguistic grounding in low-resource contexts.
* **Hindi:** Lower precision for Spam (0.65). Ambiguous sentence structure and implicit language cues reduced accuracy. Several transactional messages were misclassified.

#### **C. Cross-Language and Method Comparison**

### **Table** III: Language wise comparison across methods (Supervised vs. Few Shot vs Traditional ML)

| **Lang.** | **Method** | **Accuracy** | **Spam Precision** | **Spam Recall** | **Remarks** |
| --- | --- | --- | --- | --- | --- |
| **English** | Supervised (MuRIL) | 0.99 | 0.96 | 0.97 | Excellent results; nearly perfect classification |
| Few-Shot (LLM) | 0.95 | 0.80 | 0.83 | Slightly weaker Spam detection |
| Traditional ML[9-10] | ~0.91 | ~0.90 | ~0.90 | Reasonable performance |
| **Hindi** | Supervised (MuRIL) | 0.975 | 0.92 | 0.92 | Balanced, slight drop in Ham F1 |
| Few-Shot (LLM) | 0.91 | 0.65 | 0.78 | High Ham accuracy, but weak Spam detection |
| Traditional ML [9-10] | ~0.85 | ~0.82 | ~0.84 | Performs moderately, outperformed by LLMs |
| **Bangla** | Supervised (MuRIL) | 0.9725 | 0.99 | 0.95 | Very strong across both classes |
| Few-Shot (LLM) | 0.81 | 0.81 | 0.84 | Acceptable for zero training, but lower than supervised |
| Traditional ML [9-10] | ~0.83 | ~0.80 | ~0.81 | Reasonable baseline, but significantly outperformed |

English outperformed other languages across all methods. Bangla showed relatively stable results, while Hindi’s spam classification lagged, especially in few-shot settings. Supervised MuRIL consistently outperformed traditional and few-shot approaches, particularly for Spam detection.

#### **D. Error Analysis and Observations** Supervised MuRIL showed minimal errors across all datasets, while Few-shot models struggled with nuanced or culturally embedded spam cues.

#### 

#### **Figure 2: Accuracy Comparison Across Languages and Methods**

**E. Key Takeaways**

* **Supervised fine-tuning on MuRIL** yielded the best results, especially for spam identification in high-stakes applications.
* **Few-shot learning** is valuable for low-resource settings, though it requires careful prompt design and improved language grounding.
* **Traditional models**, while fast and simple, lack the semantic depth needed for reliable multilingual spam detection.

**IV. DISCUSSION**  
The results demonstrate that supervised fine-tuning of MuRIL delivers consistently high classification accuracy across all languages, with English showing near-perfect detection. Importantly, this indicates that contextualized embeddings from domain-relevant pretrained language models can effectively address nuanced classification tasks such as spam detection, even in the presence of code-mixing and informal syntax. It highlights the superior semantic generalization capacity of MuRIL

A notable contribution of this work is the application of few-shot learning using prototypical embeddings from MuRIL, which allowed the system to generalize from only five labeled samples per class. Despite the significant reduction in training data, the model exhibited competitive performance in English and Bangla. While Hindi lagged slightly, the performance still surpassed that of traditional machine learning baselines. This finding is significant for real-world deployments where labeled data is often scarce, particularly in underrepresented Indian languages.   
  
The study also reveals certain limitations. Code-mixing, transliteration variability, and imbalance in dataset sizes for low-resource languages like Hindi and Bangla reduced classification robustness, especially for spam messages with subtle or culturally embedded cues. Few-shot models, although efficient, were more sensitive to the selection of support examples and lacked deeper cultural and contextual grounding.

The broader implication of this research is the practical viability of few-shot and embedding-based classification approaches in multilingual cybersecurity applications. The reduced reliance on large annotated datasets and the portability of pretrained embeddings pave the way for lightweight, deployable, and scalable solutions, particularly on mobile platforms.  
  
The contributions of this research are threefold:

1.⁠ ⁠Development of a multilingual smishing detection system leveraging both data-rich (fine-tuned MuRIL) and data-scarce (few-shot) strategies.

2.⁠ ⁠Empirical demonstration of the effectiveness of LLMs in Indian regional language security contexts.

3.⁠ ⁠Provision of a benchmark dataset and reproducible notebooks, facilitating future work and industrial adaptation.

In conclusion, this study validates the potential of multilingual transformers like MuRIL in combating phishing in linguistically diverse settings. It lays the groundwork for integrating explainable AI, extending language support (e.g., to Dravidian languages), and developing real-time, resource-efficient spam detection systems for deployment in India’s complex digital landscape.

### **V. CONCLUSION**

This study makes a meaningful contribution to the growing intersection of multilingual NLP and cybersecurity. By investigating smishing detection in English, Hindi, and Bangla using MuRIL, it highlights the potential of both supervised and few-shot learning strategies in low-resource environments. Few-shot learning, in particular, demonstrated strong generalization despite limited data, pointing toward scalable solutions for underserved languages.

However, several challenges were encountered, including dataset scarcity for non-English languages, class imbalance in spam/ham examples, and constrained computational resources limiting model training epochs. Additionally, the small support set in few-shot setups affected performance for linguistically diverse and context-dependent spam patterns.

Future work could focus on incorporating Explainable AI techniques for model transparency, expanding coverage to Dravidian languages through curated datasets, and exploring advanced few-shot or zero-shot generalization approaches. Cross-lingual transfer learning and real-time, lightweight deployment—especially on mobile devices—remain promising directions.

Overall, this research lays a foundational framework for building inclusive, efficient, and practical smishing detection systems tailored for India’s multilingual population.

### **VI. REFERENCES**

[1] J. Zhang et al., "When LLMs Meet Cybersecurity: A Systematic Literature Review," arXiv preprint arXiv:2405.03644, 2024.

[2] D. Gada, "Indian Phishing Landscape: A Machine Learning and Deep Learning Approach for Detecting Malicious URLs and Curating an Indigenous Dataset," Int. J. Intell. Syst. Appl. Eng., vol. 12, no. 4, pp. 1670–1679, Jun. 2024.

[3] S. Khanuja et al., "MuRIL: Multilingual Representations for Indian Languages," arXiv preprint arXiv:2103.10730, 2021.

[4] D. Kakwani et al., "IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages," in Findings of the Association for Computational Linguistics: EMNLP 2020, 2020.

[5] D. Goel, H. Ahmad, A. K. Jain, and N. K. Goel, "Machine Learning Driven Smishing Detection Framework for Mobile Security," arXiv preprint arXiv:2412.09641, 2024.

[6] W. Yin, J. Hay, and D. Roth, "Meta-learning for few-shot natural language processing," in Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 927–937, 2020.

[7] J. Snell, K. Swersky, and R. Zemel, "Prototypical Networks for Few-Shot Learning," in Advances in Neural Information Processing Systems (NeurIPS), 2017.

[8] Y. Wang et al., "Generalizing from a Few Examples: A Survey on Few-Shot Learning," ACM Computing Surveys, 2020.

[9] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text Classification Algorithms: A Survey," Information, vol. 10, no. 4, p. 150, 2019.

[10] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, "Contributions to the Study of SMS Spam Filtering: New Collection and Results," in Proc. 11th ACM Symposium on Document Engineering, pp. 259–262, 2011.