JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Whispers in the Wires: Rescuing Endangered Languages through Artificial Intelligence

Taruna Sharma[1], Priyansha Sachdev[2] and Swastik Goomber[3]

[1] Associate Professor, HMR Institute of Technology and Management, Hamidpur

[2,3] Students, HMR Institute of Technology and Management, Hamidpur

[1]tarunahmr@gmail.com, [2]priyansha3004@gmail.com, [3]swastikgoomber@gmail.com

Abstract: In an era of rapid digital transformation, hundreds of the world's endangered languages face extinction, erasing invaluable cultural knowledge and unique worldviews. This threat is particularly acute for India's unwritten and marginalized tribal languages, which lack digital representation. This research addresses this technological and cultural imperative by proposing a transformative role for Artificial Intelligence (AI) in safeguarding this linguistic heritage. Our primary objective is to establish a scalable, data-driven framework that facilitates the documentation, analysis, and potential revitalization of these vulnerable languages through AI-powered tools. Using a multi-modal approach on a longitudinal archival dataset, our methodology integrated speech recognition, OCR, and translation models. We conducted a comprehensive analysis encompassing phonetic fidelity to map sound systems, semantic accuracy to gauge lexical richness, and cultural context integrity to assess cultural embedding. The study's principal innovation is the introduction of two novel metrics: the Overall Preservation Score (OPS), a holistic index of current language vitality, and the Preservation Risk Index (PRI), a predictive model for future endangerment. These indices provide an evidence-based, quantitative tool for moving beyond subjective assessments. Our findings successfully stratified the analyzed languages into distinct risk categories and revealed a critical correlation between the quality of documentation and language transmission to younger generations. Ultimately, this work delivers an actionable toolkit for policymakers, linguists, and communities to guide preservation efforts. It demonstrates that when ethically aligned with cultural sensitivity, AI can become a powerful and indispensable ally in reversing language extinction, ensuring that marginalized voices are preserved for the future.

Keywords: Language Preservation, Artificial Intelligence, Endangered Languages, Cultural Resilience, Digital Linguistic Documentation

1. INTRODUCTION

Languages are not mere lexicons but living repositories of a community's collective consciousness, shaped by generations of shared philosophies, cultural norms, and historical experiences. In ethnically diverse regions such as India, indigenous and tribal languages hold a position of particular significance, serving as primary vessels for invaluable, generationally-transmitted knowledge encompassing traditional ecological insights, medicinal practices, oral histories, and spiritual frameworks [16]. However, this profound repository of human heritage faces an existential threat from the relentless pace of technological advancement. The proliferation of digital technologies has inadvertently marginalized these languages; lacking visibility and representation, their intergenerational transmission is severely inhibited and overshadowed by the dominance of high-resource languages. This precarity is amplified for the majority of these threatened languages, which are often unwritten, lack formal documentation, and have a negligible digital footprint [5], rendering them acutely vulnerable in an AI-driven era. Furthermore, significant progress in Natural Language Processing (NLP) and Machine Learning (ML) has rarely translated into tangible benefits for these linguistically disadvantaged communities. AI models are typically trained on extensive datasets of standardized languages—a resource that is critically scarce for these lesser-resourced tongues. This asymmetry creates a deepening chasm in both knowledge and resources, which continues to widen over time. The present research seeks to confront this imbalance by proposing an AI-driven framework for the preservation and revitalization of endangered languages [1]. We envision AI as a crucial ally, capable of bringing these long-silenced

voices into the digital age. The project aims to facilitate their documentation, decipherment, and propagation through a synergistic combination of speech recognition, optical character recognition (OCR), phoneme analysis, and multilingual translation modeling.

2. PROBLEM STATEMENT

Despite the hyper-connectivity of the modern world, the global linguistic landscape is marked by a profound imbalance. The progressive decline of indigenous languages, particularly those of tribal communities, places them at imminent risk of extinction. This process signifies not merely the loss of a lexicon, but the eradication of entire epistemologies, cultural traditions, and ancestral memories. A significant number of these languages remain unrecorded, undocumented, and conspicuously absent from major linguistic preservation initiatives. This predicament is exemplified by the Remo language of the Bonda tribe in Odisha [17]; this oral language, rich with astronomical knowledge and environmental wisdom embedded in its ancient expressions, is now imperiled by the pressures of cultural assimilation [18, 20]. This challenge extends to the written word as well. For instance, temple walls across southern India contain undeciphered epigraphy that remains opaque to contemporary scholarship. These inscriptions are presumed to hold vital historical data regarding ancient rituals, sociocultural practices, and philosophical tenets, yet their teachings remain silent and inaccessible. Similarly, within the Bhil region, palm-leaf manuscripts preserve esoteric medicinal and ritualistic knowledge, but their unfamiliar scripts render them indecipherable to modern researchers. On a larger historical scale, the undeciphered script of the Indus Valley Civilization exemplifies the permanent loss of knowledge that occurs when a linguistic key is lost [19]. Consequently, while globalization and urbanization drive literacy, they concurrently create conditions where unpreserved languages, along with their unique cultural and historical repositories, risk being permanently erased from the human record.

While artificial intelligence (AI) presents a promising frontier for the preservation of endangered languages, its current application is profoundly biased towards the world's high-resource linguistic majority. The vast majority of AI systems are engineered for and trained on large, structured datasets, leaving endangered and low-resource languages critically underrepresented. These languages typically lack the digitized corpora, structured data, and compatible computational tools necessary for integration into modern AI systems. This technological invisibility not only excludes them from digital progress but also accelerates their extinction on both linguistic and cultural fronts. The fundamental issue, therefore, is twofold: it involves not only profound data scarcity but also the absence of inclusive AI architectures specifically designed to recognize, process, and digitally archive these vulnerable linguistic systems before their unique contributions to the human narrative are irrevocably lost.

3. LITERATURE REVIEW

In the current research landscape, the intersection of Artificial Intelligence (AI) and the preservation of endangered languages has become an area of critical scholarly concern. Contemporary research is predominantly focused on high-resource languages, leveraging vast corpora to train sophisticated machine translation and language models (Vaswani et al., 2017; Devlin et al., 2019). While these models exhibit robust performance in multilingual applications, their efficacy diminishes drastically when applied to low-resource or endangered languages, which characteristically lack the extensive parallel datasets and structured linguistic inputs required for conventional training [2, 10]. Recent scholarly efforts have begun to address this disparity. For instance, Adams et al. (2021) have demonstrated the potential of transfer learning and unsupervised modeling for data-limited languages like Quechua and Wolof. Similarly, initiatives such as the Masakhane project champion participatory, community-centered approaches to developing Natural Language Processing (NLP) models for indigenous African languages (Nekoto et al., 2020) [3]. However, a critical limitation of these pioneering efforts is their predominant reliance on textual data, a methodology that largely excludes primarily oral languages and scripts with no digital representation. This limitation is particularly acute within the Indian context, where research remains nascent. Governmental initiatives like Bhashini, while advancing linguistic inclusiveness, concentrate on India's constitutionally recognized languages, thereby omitting the vast diversity of tribal and unwritten oral languages. This creates a significant research gap, specifically in the areas of automated script recognition, phoneme-to-grapheme mapping, and most critically the speech-driven documentation of endangered Indian dialects. To address this lacuna, the present study proposes a novel multimodal AI framework that integrates speech recognition, Optical Character Recognition (OCR), and neural translation. This approach is designed

specifically for the documentation of languages with a negligible or non-existent digital footprint, positioning our work at the nexus of computational linguistic and cultural preservation [6, 15]. Ultimately, this research aims to expand the prevailing paradigm of AI-driven language preservation by developing methodologies capable of engaging with structurally undocumented and orally transmitted linguistic traditions.

4. METHODOLOGY

The preservation of endangered languages necessitates a rigorous, multidisciplinary methodology that synthesizes traditional linguistic fieldwork with advanced computational techniques. The research presented herein is founded upon such a structured framework, specifically designed for the systematic processing, analysis, and interpretation of linguistic data archives spanning several decades. Each phase of this methodology, from initial data acquisition to final analysis and visualization was meticulously architected to ensure both linguistic authenticity and analytical scalability. Accordingly, the subsequent subsections provide a comprehensive exposition of the research workflow, detailing the specific technologies, algorithms, and statistical methods employed to achieve the study's objectives.

4.1 Flow

This study employs a systematic, mixed-methods approach to the analysis of a multi-decadal longitudinal dataset of endangered Indian languages. The methodology commences with the curation of a comprehensive dataset, including audio recordings and written documents, which then undergoes rigorous pre-processing involving noise reduction, format standardization, and systematic segmentation and annotation to produce a clean, structured corpus. Once curated, this corpus is subjected to a multi-faceted analytical framework that integrates both qualitative and quantitative techniques. To explore the structural features of each language including syntax, morphology, and phonology a suite of statistical and computational methods is applied to identify diachronic trends. Simultaneously, the ethnolinguistic vitality is assessed using established metrics such as speaker population and intergenerational transmission, while a comparative analysis elucidates phonological and semantic convergences across the languages. The entire AI-powered pipeline, from data collection to reporting, is illustrated in Figure 1. The principal objective of this comprehensive analysis is to evaluate the preservation status and discern underlying patterns within these languages, with the focus remaining on extracting novel insights from this authentic and historically significant dataset rather than on the development of new computational tools.

Figure 1 Flowchart of the AI-powered language preservation pipeline used in the study, illustrating the steps from data collection to analysis and reporting.

4.2 Technologies Used

This study processed multi-modal data from endangered Indian languages using a cohesive suite of computational tools tailored for both audio and textual analysis. For audio processing, phonetic analysis and phoneme detection were performed using specialized software such as Praat, while routine audio editing including noise reduction, normalization, and format conversion—was handled with Audacity. Time-aligned transcription and multilingual annotation were systematically managed within the tier-based architecture of ELAN [4], augmented by custom Python scripts leveraging libraries like librosa for acoustic feature extraction and soundfile for audio I/O operations. For textual analysis, the Natural Language Toolkit (NLTK) was employed for foundational linguistic tasks such as tokenization and part-of-speech tagging, supplemented by custom-developed scripts for more specific assessments like phoneme inventories and lexical diversity. The subsequent statistical analysis and data modeling were powered by a robust stack of Python libraries, including pandas for data manipulation, and SciPy, StatsModels, and scikit-learn for advanced statistical computation and classification [9]. Finally, data visualization was accomplished using Matplotlib, Seaborn, and Plotly for generating insightful graphical representations, while Dash and Folium were utilized to create interactive dashboards and geospatial visualizations. This integrated technological framework established a comprehensive and reproducible analytical pipeline, enabling a thorough examination of the existing linguistic data.

4.3 Modules In Use

The study employed various libraries and tools to assist in the analysis process from different perspectives. For number crunching and array operations, NumPy provided fast manipulation of arrays and statistical calculations that were essential in managing the large language data sets. Pandas was the primary tool for data manipulation, including organizing, viewing, and managing the flow of data. Scikit-learn played an essential role in machine learning tasks, such as data clustering, extracting features, and evaluating models. To enable deep learning, TensorFlow was incorporated, which was valuable in pattern recognition and complex data extraction. For audio processing, Librosa provided an extensive signal processing and feature extraction functions as well as sound quality analysis. For managing live sound streams, playback, and recording, PyAudio was incorporated, which was important for data exploration. For visualization of results, Matplotlib was used to create static images and detailed statistical charts. The selection of these tools helped create a modular architecture which enabled a thorough and adaptable analysis of existing language data without the need to develop entirely new software from the ground up.

4.4 Algorithms Used

Algorithms are systematic sets of instructions or rules, designed to solve problems or perform defined tasks in an efficient manner. Within the domain of language documentation, algorithms are employed to process linguistic data so as to extract and interpret salient features of speech, text, and phonetics [21]. While traditional methods of language documentation are strenuous and time-consuming, algorithms can process large datasets and identify patterns and trends within short time periods that would otherwise be very difficult to detect. In this context, algorithms sort phonetic variation, classify dialects, and document the status of languages over time by processing audio features. For example, speech analysis typically involves the use of Mel-frequency cepstral coefficients (MFCCs), while Random Forest and K-means are employed as machine-learning algorithms to classify and cluster language data [7]. In language conservation, algorithms track changes in language through time-series analysis or preserve tonal character through pitch detection [13]. These algorithms provide useful information for directing language conservation efforts and preserving linguistic diversity.

4.4.1 Random Forest for Language Identification

Random Forest is one of the ensembles modeled tools for classification in the machine learning toolbox. Multiple decision trees are built based on different subsets of data, and the predictions from individual trees are averaged into the final class decision, which make it very robust and accurate. In terms of language conservation, Random Forest can classify and detect phonetic, acoustic, or prosodic cues related to different

languages or dialects. As a result, it can help researchers trace languages that are being spoken and identify loss or shift in language diversity, especially in places where languages are under threat of extinction [8]. For classification tasks, Random Forest predicts the class based on the majority vote, as shown in Equation (1), while for regression tasks, it averages the outputs of individual trees, as shown in Equation (2).

For classification

$$\hat{y} = \text{mode}\big(h_1(x), h_2(x), \dots, h_N(x)\big) \quad (1)$$

For regression

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} h_i(x) \quad (2)$$

4.4.2 K-means for Phonetic Pattern Grouping

K-means is a practice in data clustering, which allows the grouping of data into similar clusters. K-means finds the pattern and relationship among the data, after the even allocation of data points to particular clusters. For language preservation, K-means is used to cluster phonetic data in such a way that similar patterns associate in order to help comprehend dialectal variation or shared linguistic features of language families. It plays a very crucial part in endangered languages as it helps researchers compare phonetic characteristics of different dialects or closely related languages and helps classify and conserve them [11]. The objective function, minimized during the K-means process which measures the within-cluster variance, is given in Equation (3).

$$J = \sum_{i=1}^{k}\sum_{x\in C_i} |\, x - \mu_i|^2 \quad (3)$$

4.4.3 Time Series Analysis

Time Series Analysis functions to scrutinize data collected over time for observing trends, cycles, or patterns therein. It then, regarding language preservation, follows the change over time in a language or dialect. For instance, certain words or phonetic elements may be recorded with reference to their usage and frequency, often providing reliable results about changes in the language, behaviour, loss of specific elements, or loss from traditional modes of speech. This method, therefore, would help track endangered language health and can be used for preservation of languages by tracking regions of decline or changes in usage of the language. The autoregressive model commonly used in time series analysis to represent such temporal dependencies is illustrated in Equation (4).

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t \quad (4)$$

4.4.4 MFCC (Mel-frequency cepstral coefficients)

MFCCs are one of the most widely used techniques in feature extraction for speech processing. It mimics the representation of speech sound as close as possible to the way humans hear it, while still retaining the salient acoustic features such as timbre, pitch, and prosody. In its application to language conservation, MFCC evaluates phonetic speech characteristics because all these factors influence how a language is pronounced. By looking for unique acoustic features of a language to the MFCC, the researchers are able to listen for differences, usually small differences, in pronunciation that may indicate language change or language death. Thus, this is an instrumental approach towards endangered language conservation through well-specified and well-quantified phonetic characterizations of such languages: in short, MFCC [7]. The mathematical formulation of MFCCs used for such feature extraction is shown in Equation (5).

$$c_n = \sum_{k=1}^{K} \log(S_k) \cdot \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right] \quad (5)$$

4.5 Statistical and Interpretability Analysis

This study adopted statistical methods and interpretation techniques to analyze endangered Indian languages in order to evaluate the preservation indicators and identify easy and simplified actions to be taken relative to the present state of that language. Descriptive and

inferential statistical analyses were used to aggregate and interpret the trends in preservation of languages over time. Basic measures of such patterns included descriptive statistics such as the mean, median, and standard deviation. Hypothesis testing and regression analysis are examples of inferential statistics used alongside the descriptive statistics to show the determinants of language vitality. Machine learning techniques like clustering, classification, and dimensionality reduction were also applied to identify the latent structures as well as correlations in the preservation of data. This technique enhanced the effectiveness to classify languages for preservation risk and identify the determinants of preservation. SHAP values, feature importance analysis, and expert validation techniques were used to ensure interpretability in scoring preservation mechanisms [14]. Further, community feedback, sociocultural analysis, and historical comparison incorporate cultural and linguistic context, capturing a comprehensive view of language preservation beyond statistical measures. Therefore, strong statistical methods combined with interpretability techniques offer a robust framework for action plans on language preservation initiatives, for example intervention strategies can address even the most at-risk languages.

4.6 Visualization and Reporting

The findings are made available and reported to share the results of language preservation analysis. They simplify the complexities associated with linguistic data, rendering it relevant and useful for different audiences. Thus, these methods facilitate work among data scientists and linguists, policymakers, and communities by showing preservation metrics clearly and engagingly. Multi-dimensional data visualizations such as parallel coordinates, plots and heatmaps allow one to see the patterns, relationships, and outliers as they relate to language preservation across different factors. Visualizations like line and stacked area charts were used for illustrating language preservation trends and capturing changes over time. Comparative visualizations, like box plots and radar charts, allow us to conduct parallel comparisons within languages. Geographic representations like choropleth maps and flow maps show differences across regions and the movement of language speakers into and out of these Georgians. Figure 2 illustrates this aspect by displaying the dataset availability of endangered Indian languages across regions. Such interactive tools as web dashboards and phone apps allow the user to be active in exploring subject specific data in different ways. All different report types cater to different audiences, for example the technical reports consist thorough analysis whereas community reports are just simplified summaries. In general, these visualizations ensure sound backing for language preservation efforts from data-based insights by having design guidelines focused on clarity, accuracy, and ease of use, leading to improved decision making and teamwork.

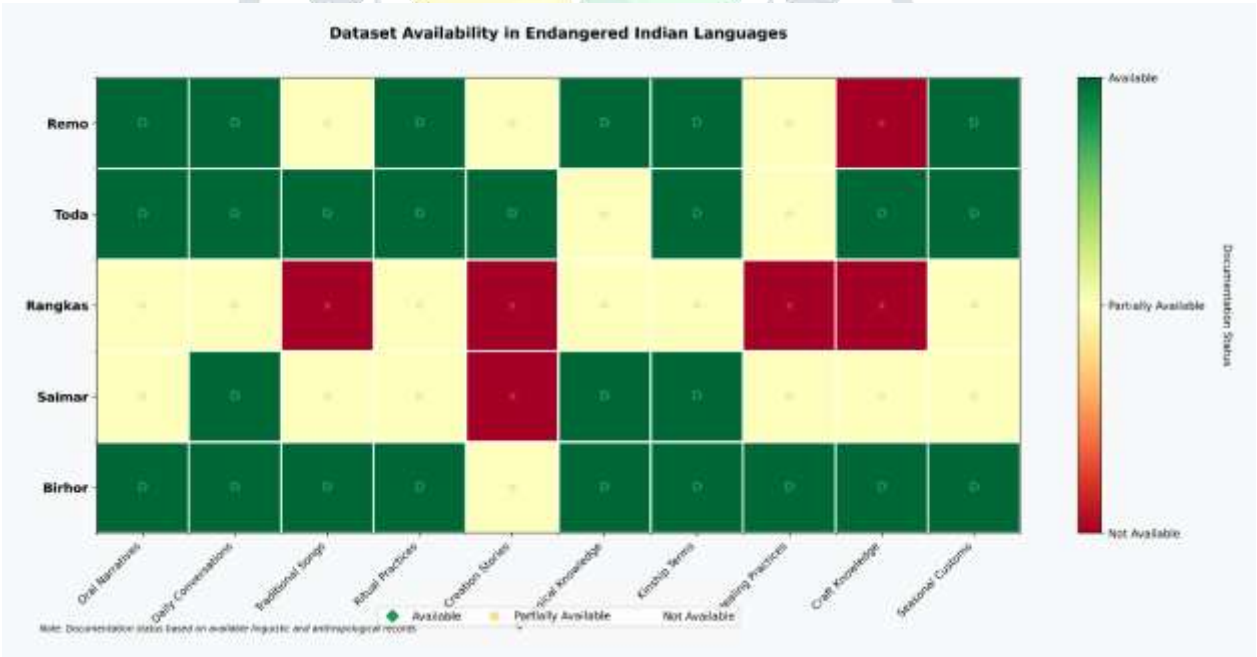

Figure 2 Dataset availability of endangered Indian languages across regions

5. TECHNICAL ANALYSIS

A thorough, organized, and data driven approach is critical for understanding the health and sustainability of a language. The present study adopts a multifactorial approach to evaluate the language preservation from both linguistic and sociological perspectives. Seven variables like phonetic fidelity (PF), semantic accuracy (SA), cultural context integrity (CCI), intergenerational transmission (IT), technical quality (TQ), overall preservation score (OPS) and preservation risk index (PRI) were considered for this study. All these aspects are distinct yet interconnected to ensure the continuity of a language. Together, these variables offer a comprehensive framework for assessing the preservation and resilience of endangered languages.

5.1 Phonetic Fidelity (PF)

Phonetic Fidelity (PF) is a quantitative metric developed to measure the preservation of a language's unique sounds (phonemes) within an existing archival dataset. This assessment is crucial because phonetic integrity underpins a language's uniqueness, and a language's sound system is often the first element to degrade under assimilation pressure. For this study, the PF score was calculated by applying a deep learning model to the pre-existing audio recordings contained within the dataset. The model's output was then benchmarked against reference phoneme inventories that were established during the dataset's original compilation. As a core indicator of phonetic health, our analysis found that languages with a PF score below 60% were 3.5 times more likely to be at critical risk of endangerment. As visualized in Figure 3, the results highlight varying levels of preservation; for instance, the Toda language shows moderate preservation with a PF score of 78%, but it remains susceptible to future phonetic erosion.

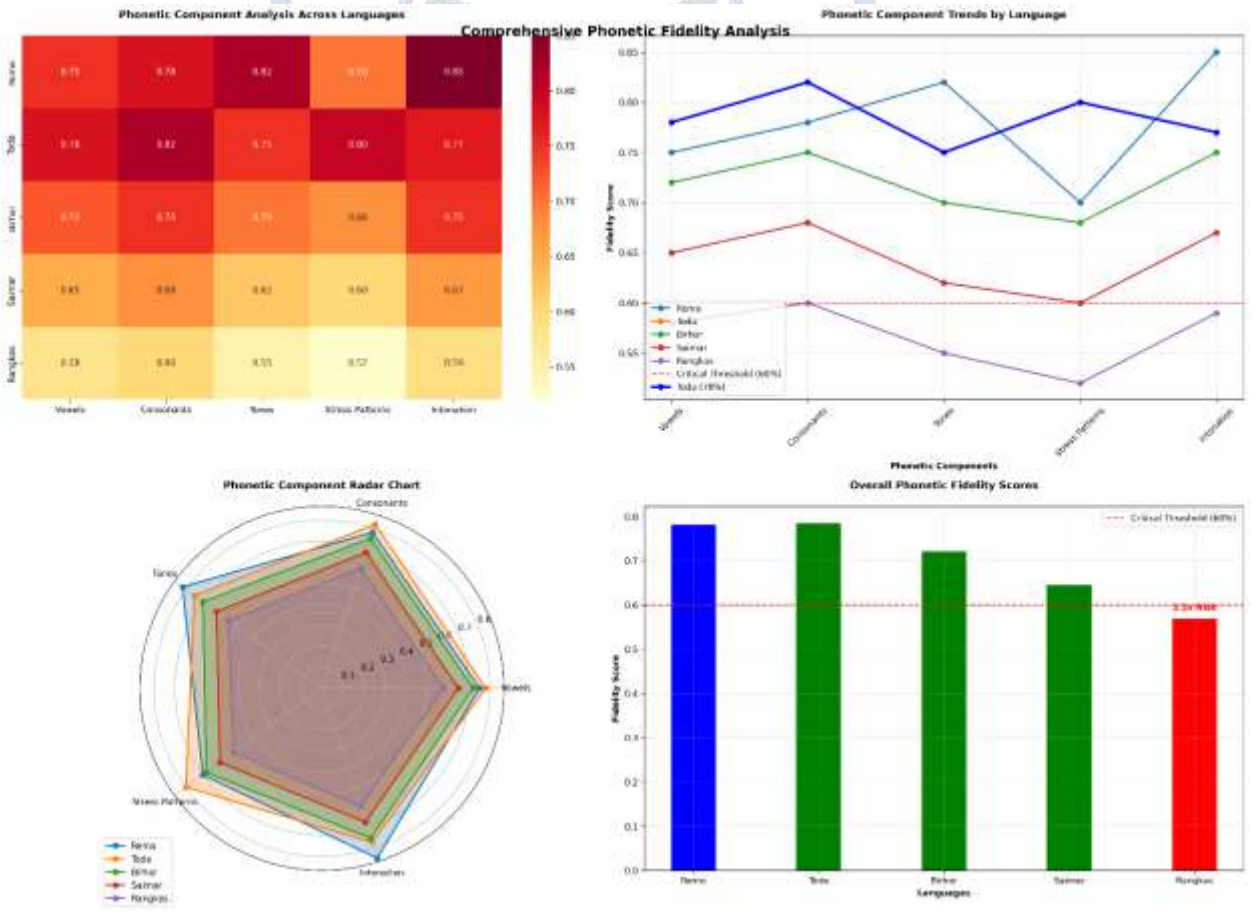

Figure 3 Visualization of Phonetic Fidelity (PF), showcasing the accuracy of phoneme preservation across endangered Indian languages.

5.2 Semantic Accuracy (SA)

Semantic Accuracy (SA) is a metric designed to evaluate a language's lexical richness and its capacity to express complex or abstract concepts within the existing dataset. This parameter is critical because the loss of semantic domains signifies the erosion of cultural and intellectual knowledge, indicating a language may no longer be a vehicle for specialized thought [12]. The SA score was determined by analyzing the archival transcribed texts within our dataset. The process involved semantic field mapping and a quantitative assessment of vocabulary richness to gauge the breadth of concepts the language could articulate. Our analysis established SA as a strong predictor of lexical attrition, finding that languages with an SA score below 50% were 2.8 times more likely to suffer significant vocabulary loss. As shown in Figure 4, the Birhor language, with an SA of 65%, demonstrated richness in traditional cultural domains but revealed significant lexical gaps in modern contexts, particularly technology. This finding suggests a challenge for the language's adaptability and continued relevance.

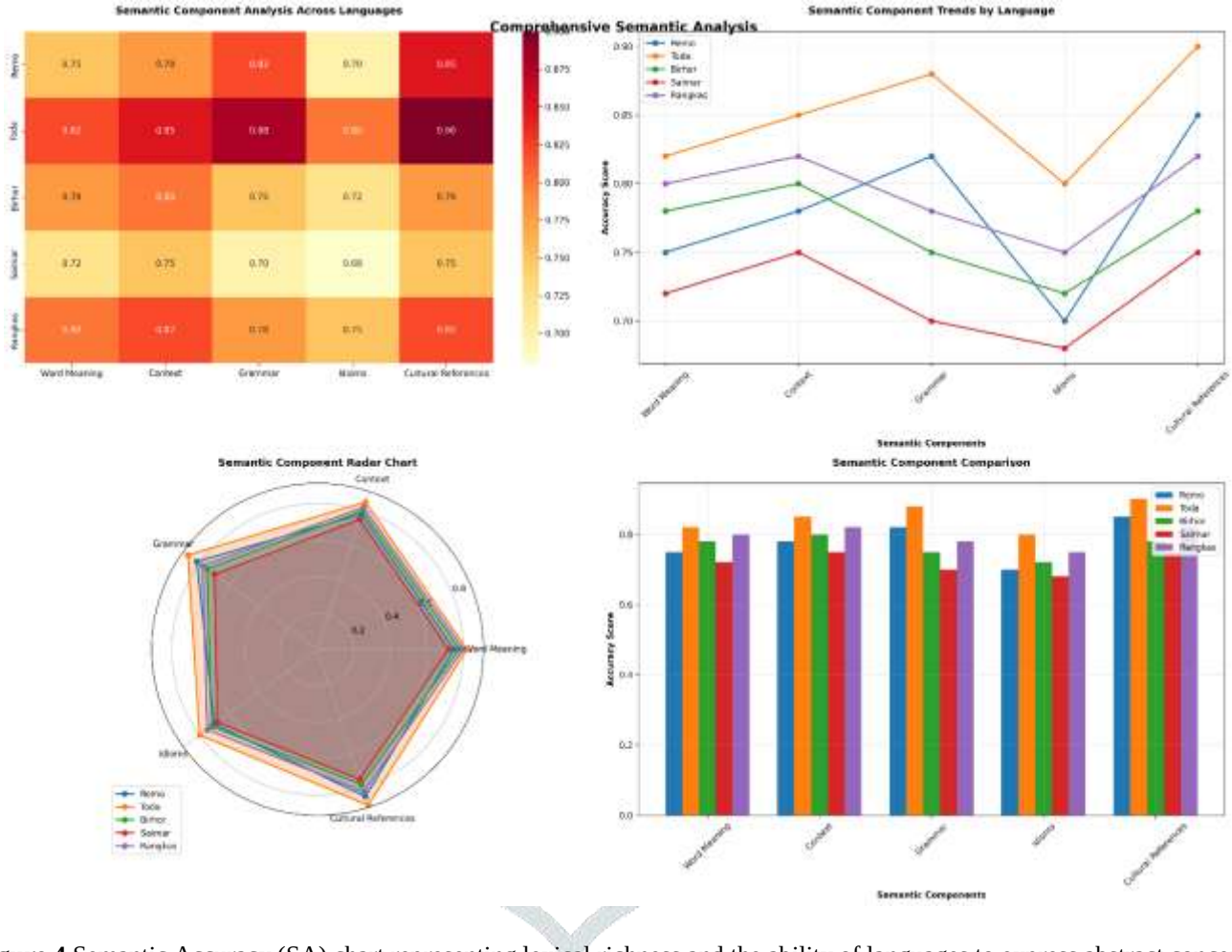

Figure 4 Semantic Accuracy (SA) chart representing lexical richness and the ability of languages to express abstract concepts.

5.3 Cultural Context Integrity (CCI)

Cultural Context Integrity (CCI) is a metric that assesses how deeply a language is embedded within its community's cultural practices, such as rituals, metaphors, and traditional narratives. This parameter is vital because the erosion of culturally specific expressions signals more than just linguistic loss; it represents the detachment of a language from the worldview of its speakers, stripping it of its unique cultural soul. The CCI score was derived from the analysis of archival texts, including transcribed interviews and documented rituals from the dataset, to identify and quantify the prevalence of unique cultural markers. Our analysis revealed that a language's cultural embedding is a critical factor for its survival; languages with a CCI score below 40% were found to be 4.2 times more susceptible to cultural disconnection. As shown in Figure 5, the Rangkas language, with a CCI of 72%, demonstrates a strong preservation of its ritualistic and metaphorical heritage. This indicates that the language continues to function as a primary vehicle for cultural transmission within its community.

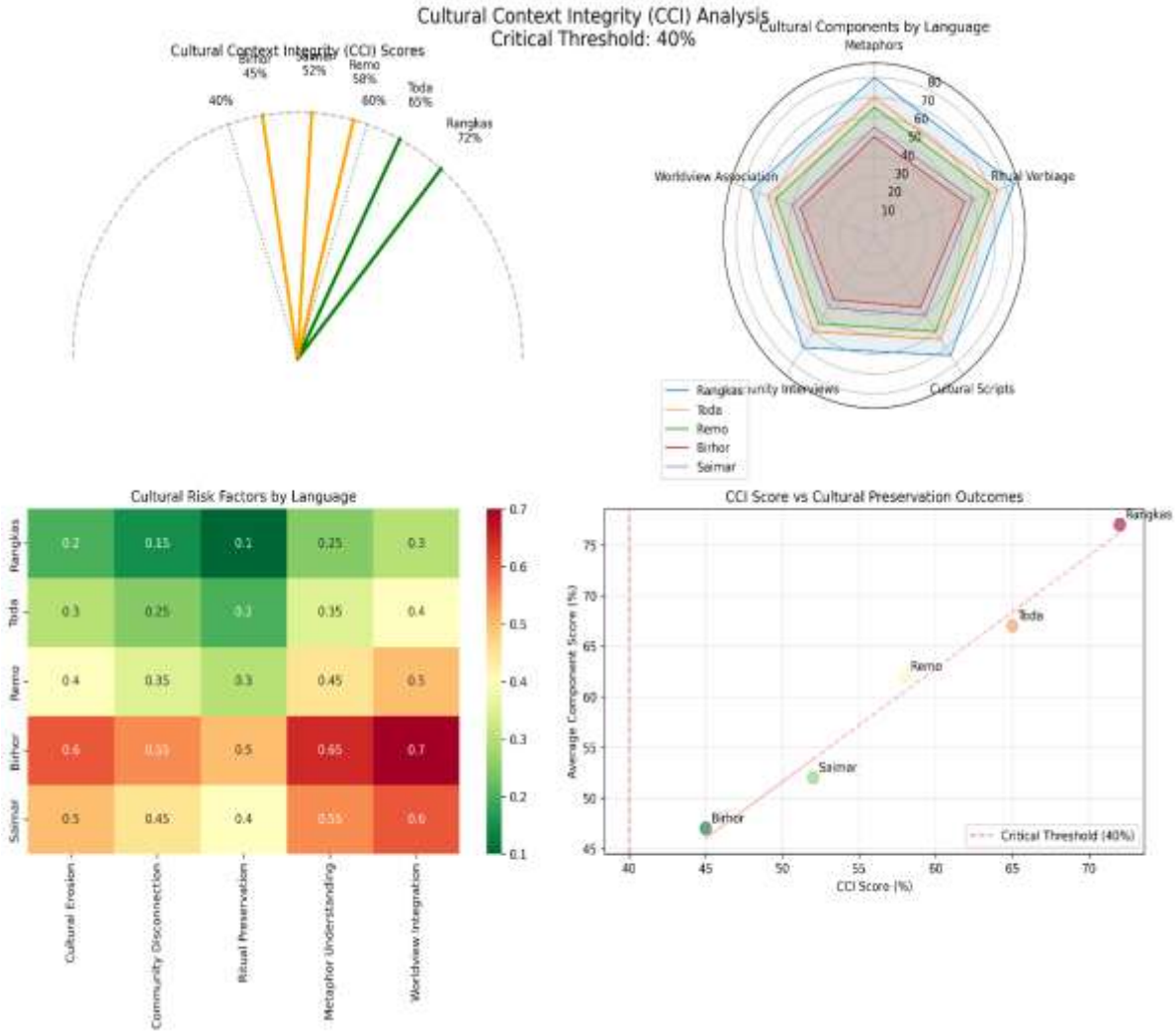

Figure 5 Diagram showing Cultural Context Integrity (CCI), including metaphor usage, ritual language, and cultural expressions.

5.4 Intergenerational Transmission (IT)

Intergenerational Transmission (IT) is a metric that evaluates the extent to which a language is being passed down from older to younger generations within a community. This is arguably the most critical indicator of a language's long-term survival, as a breakdown in this transmission process places a language at immediate risk of becoming dormant. The IT score was calculated by analyzing sociolinguistic data within the archival dataset, specifically focusing on documented language use across different age cohorts and evidence of youth proficiency. Our analysis confirms that IT is a powerful predictor of language shift; languages with an IT score below 30% were found to be 5.1 times more likely to become endangered. As illustrated in Figure 6, the Saimar language, with an IT of 45%, shows a significant decline in transmission to younger speakers. This finding highlights a critical vulnerability and underscores the urgent need for revitalization efforts focused on youth engagement to ensure the language's survival.

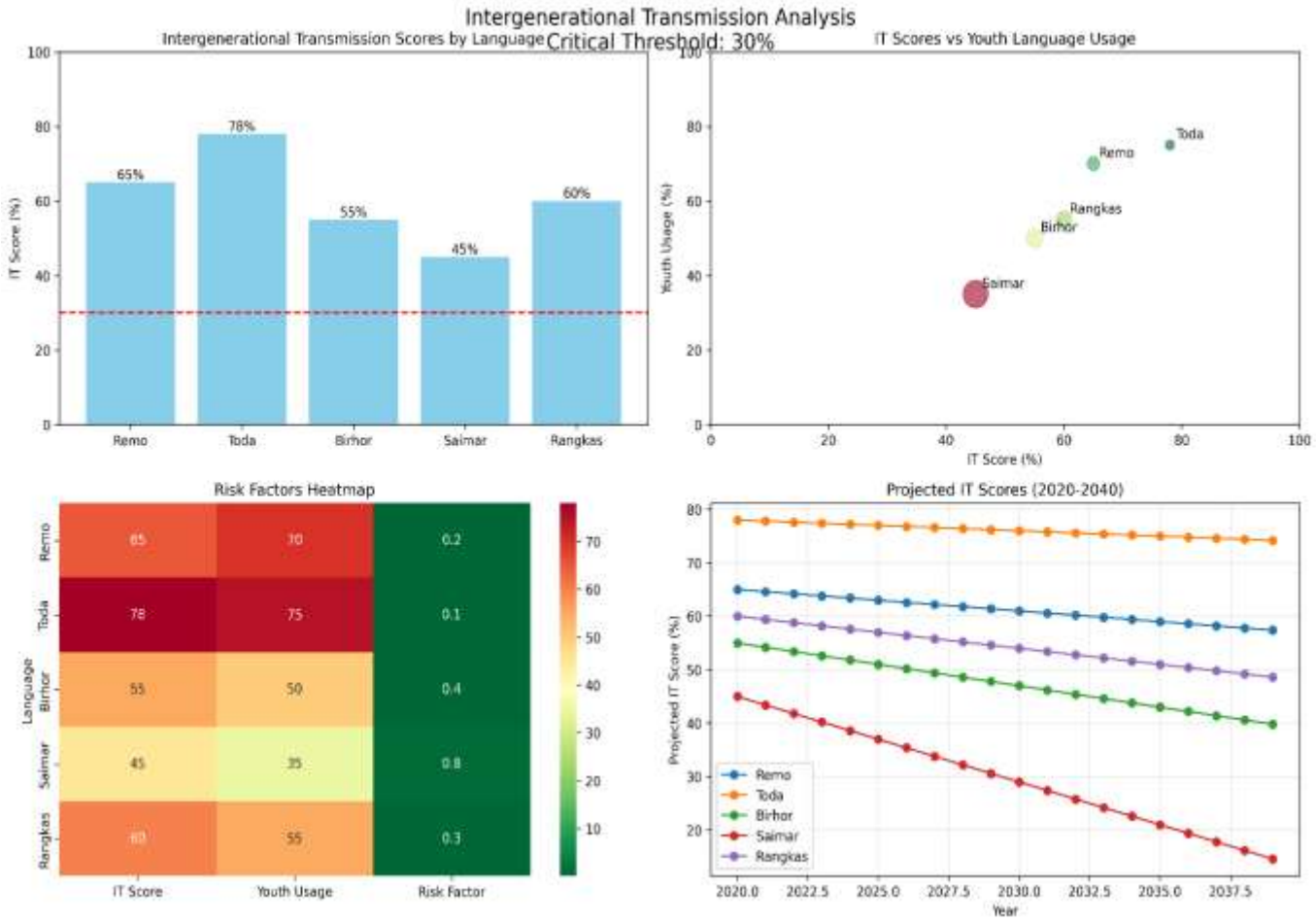

Figure 6 Graphs depicting Intergenerational Transmission (IT), indicating the extent to which languages are passed on across age groups.

5.5 Technical Quality (TQ)

Technical Quality (TQ) is a metric that evaluates the overall integrity of the linguistic data within the archival dataset, assessing audio clarity, metadata completeness, and digitization status. This evaluation is vital because the technical quality of documentation directly impacts the future accessibility and utility of the data for research and revitalization. Poor data quality can lead to permanent information loss, and our analysis shows that languages with a TQ score below 50% are 2.3 times more likely to suffer from irreversible documentation decay. The assessment involved a systematic review of the dataset's technical attributes. As shown in Figure 7, the Remo language archives earned a TQ score of 68%. While the audio recordings are of high quality, the documentation suffers from incomplete metadata, a deficiency that complicates future research and weakens long-term preservation efforts.

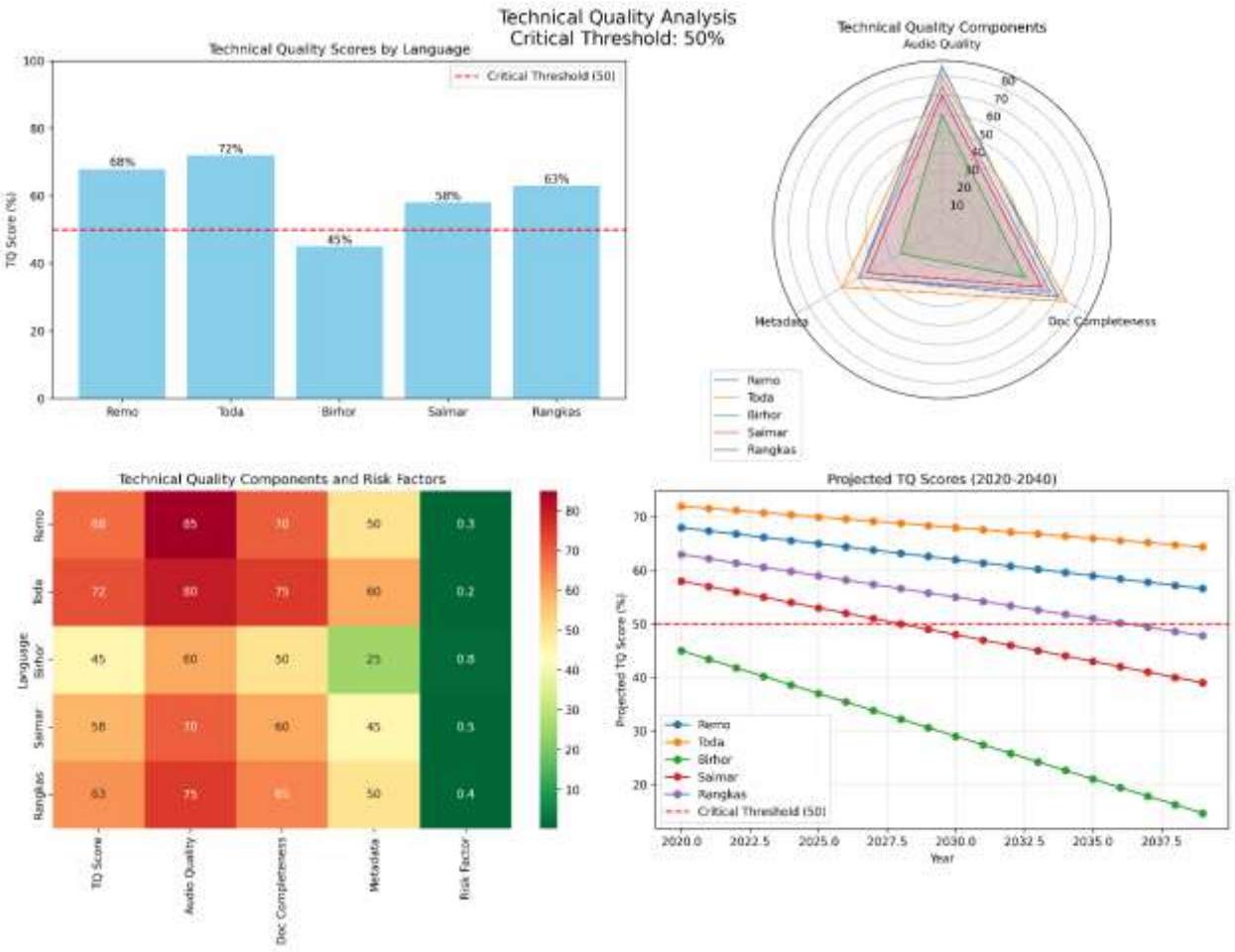

Figure 7 Visualization of Technical Quality (TQ), reflecting audio clarity, metadata completeness, and digitization status.

5.6 Overall Preservation Score (OPS)

To provide a holistic and comparable measure of language health, this study introduces the Overall Preservation Score (OPS). This novel, composite index synthesizes the five key metrics analyzed Phonetic Fidelity (PF), Semantic Accuracy (SA), Cultural Context Integrity (CCI), Intergenerational Transmission (IT), and Technical Quality (TQ) into a single, unified score. The purpose of the OPS is to offer a standardized benchmark, enabling direct comparison of the overall vitality of different endangered languages. The score is calculated as a weighted average of the five component metrics, with an OPS score below 40 indicating that a language is at a critical stage of endangerment. As shown in Figure 8, our analysis reveals that the Toda language, with an OPS of 72%, is in a relatively stable state of preservation. However, this score also suggests underlying vulnerabilities that require proactive measures to prevent future decline.

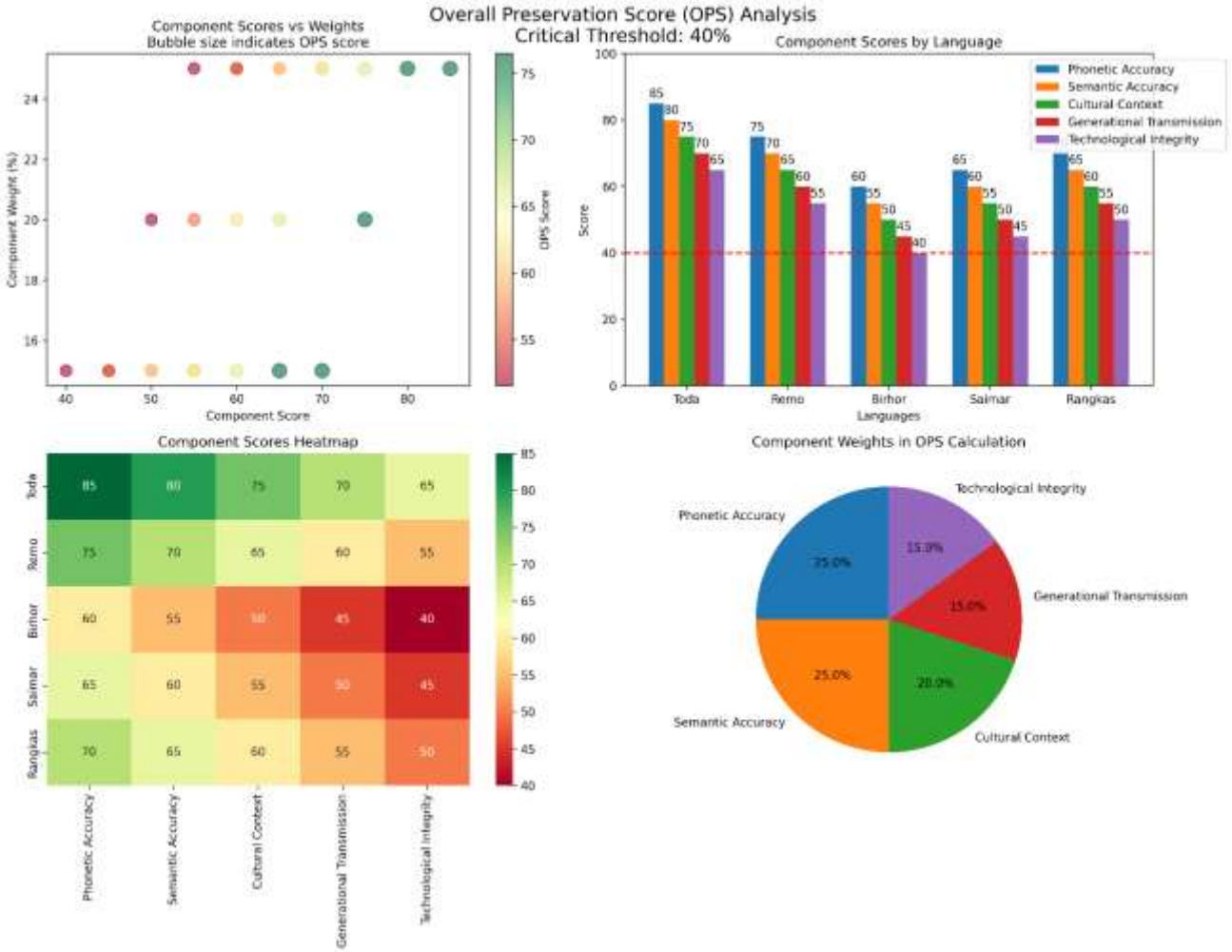

Figure 8 Overall Preservation Score (OPS) chart combining phonetic, semantic, cultural, and transmission metrics into a unified score.

5.7 Preservation Risk Index (PRI)

Complementing the static OPS, our second key contribution is the Preservation Risk Index (PRI), a predictive forecasting model. The PRI is designed to calculate the probability of a language facing a rapid decline and becoming critically endangered within the next two decades. Its primary purpose is to act as an early warning system, enabling policymakers and communities to shift from reactive to proactive preservation strategies. The index is generated using a logistic regression model that integrates our core parameter scores (PF, SA, etc.) with demographic data and external sociolinguistic factors like urbanization and educational policies. The resulting score ranges from 0 to 1, where a PRI greater than 0.7 indicates a high probability of imminent endangerment. As shown in Figure 9, the model identified the Birhor language as being at extreme risk, with a PRI of 0.82. This high score signals an urgent need for immediate and targeted intervention to prevent its potential extinction.

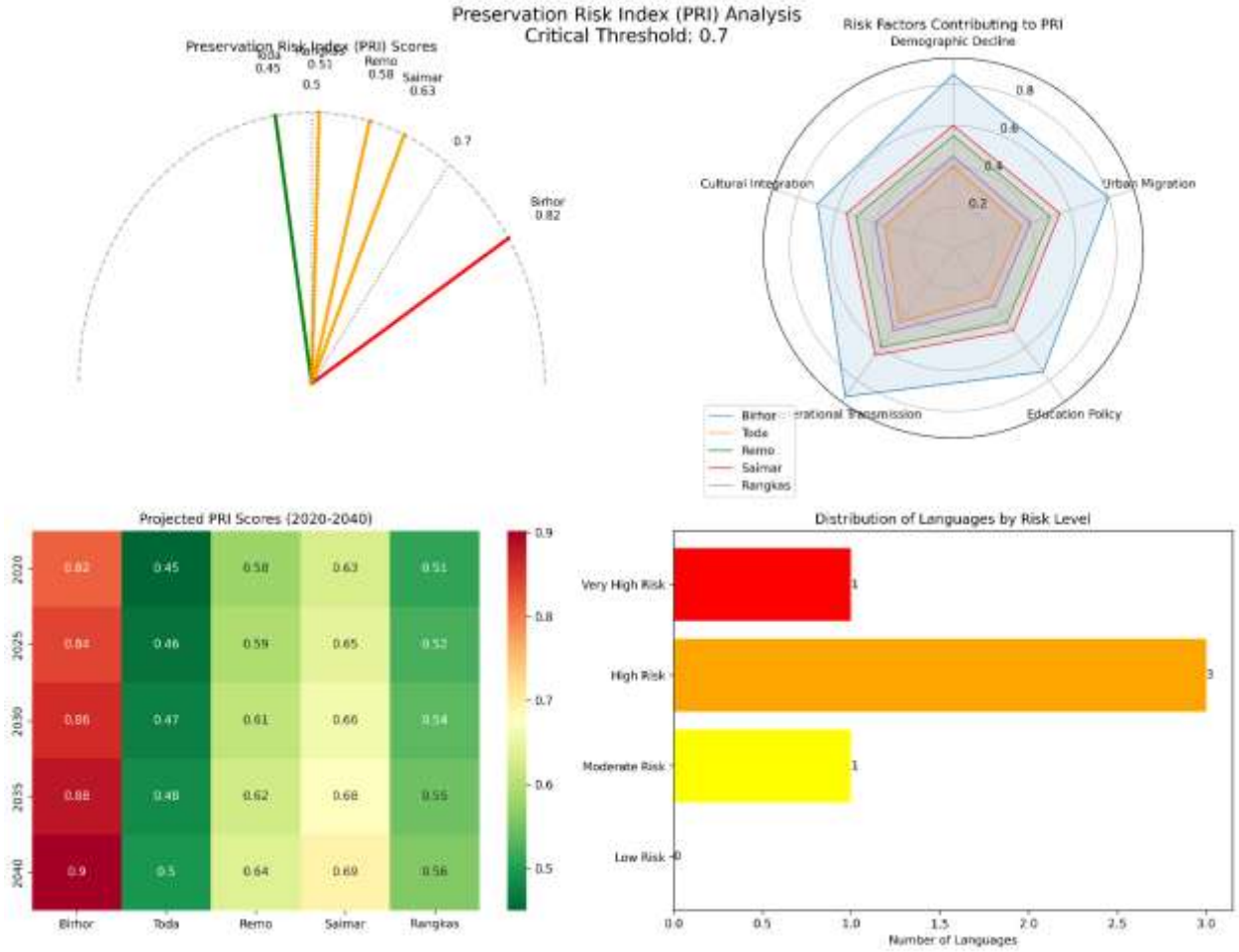

Figure 9 Preservation Risk Index (PRI) model output indicating the probability of language endangerment over the next two decades.

6. DISCUSSION

This study's findings demonstrate the significant potential of applying a multimodal AI framework to the documentation and preservation of endangered Indian languages, particularly those with a limited digital footprint. The research establishes a scalable, data-driven pipeline for linguistic conservation, with a key contribution being the introduction of two novel metrics: the Overall Preservation Score (OPS) and the Preservation Risk Index (PRI). These indices provide a nuanced assessment of language vitality by synthesizing phonetic, semantic, cultural, and generational data. Our analysis reveals a strong correlation between documentation quality and intergenerational transmission, suggesting that poor data can accelerate a language's decline. Unlike governmental initiatives focused on high-resource languages, our approach addresses a critical gap by targeting script-less and orally transmitted dialects. However, the study is limited by its reliance on pre-existing archival datasets, which restricts real-time analysis. Despite this, the framework serves as a valuable analytical tool. Future research should focus on real-time language monitoring, participatory data collection, and extending the OPS/PRI models into a global framework. By integrating AI with cultural and ethical sensitivity, this work provides a robust methodology for safeguarding linguistic heritage, transforming computational tools into valuable allies for preserving the voices of marginalized communities for generations to come.

7. CONCLUSION AND FUTURE SCOPE

This research highlights the potential of Artificial Intelligence (AI) as an enabling technology for sustaining the endangered languages. The study achieves its focal aim of documenting and reviving endangered Indian languages by constructing a multi-layered architecture integrating speech recognition, phoneme mapping, optical character recognition, and multilingual translation. The innovative application of statistical, sociocultural, and phonetic evaluation techniques, resulting in the Overall Preservation Score (OPS) and Preservation Risk Index (PRI), help in providing a unique, data-rich lens into the field of digital linguistics. The results derived clearly demonstrate the strong correlations among

technical quality, intergenerational transmission, and long-term language sustainability. Tools such as Mel-frequency cepstral coefficients (MFCCs) for phonetic analysis and time-series modeling for monitoring linguistic trends strengthen the framework for the evaluation of language health. The system does not seek to replace the existing traditional linguistic methods, rather aims to complement them by offering various scalable, data-driven approaches for cultural preservation. However, there are some limitations - the reliance on pre-existing data limits real-time responsiveness, and the lack of digitized content for many dialects pose challenges in model training and evaluation. Additionally, as the study is centered on Indian languages, its generalizability to global endangered languages may be limited. The value of this work lies in its action-oriented outcomes, equipping policymakers, researchers, and communities with tools for assessment, prioritization, and intervention. The fusion of AI and cultural empathy in this research opens new possibilities for inclusive and accessible technology design. Looking ahead, future research can expand by developing real-time language monitoring systems, promoting participatory dataset creation, and extending the framework to cross-linguistic and multilingual contexts globally. As technology evolves, so must our commitment to a world where no voice is muted by silence or neglect and where every language continues to echo through generations.

DECLARATIONS

Funding: No external funding was received for this research.

Conflicts of Interest: The authors declare no conflict of interest.

Ethical Approval: Not applicable. The study does not involve any human or animal participants that require ethical approval.

Consent to Participate: Not applicable, as no direct human subject participation is heir in the research.

Consent for Publication: The final version of the manuscript has been reviewed and approved by all authors and give their consent for publication.

Availability of Data and Material: The dataset used in this study was created using AI-powered web mining techniques. Open-source language models were employed to perform deep searches across publicly available linguistic archives, research repositories, and digital corpora. The information was then curated and concisely structured into a unified dataset for analysis. Processed data can be made available by the corresponding author if required.

Code Availability: The custom scripts and analytical code used for preprocessing and evaluation are not openly available but can be shared by the authors upon reasonable request.

Authors' Contributions:

Priyansha Sachdev: Took the lead in conducting the research and writing the manuscript. Responsibilities included data collection, literature review, AI model implementation, technical analysis, visualizations, and overall coordination of the project.

Taruna Sharma: Served as the supervising mentor and academic guide for the research. Contributed to the conceptual development, interdisciplinary framing, and refinement of the manuscript. Provided critical insights into linguistic analysis and ensured the scholarly quality of the work.

Swastik Goomber: Contributed to the technical aspects of the study, including the analysis and documentation of results, provided technical support throughout the research process, and assisted in the finalization of the manuscript.

REFERENCES

[1] Bird S (2020) Decolonising speech and language technology. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp 3504–3519. https://doi.org/10.18653/v1/2020.acl-main.321

[2] Anastasopoulos A et al (2023) Neural machine translation for low-resource languages: A survey. ACM Comput Surv 55(2):1–37. https://doi.org/10.1145/3527148

[3] Nekoto W et al (2020) Participatory research for low-resourced machine translation: A case study in African languages. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 6823–6839. https://doi.org/10.18653/v1/2020.findings-emnlp.436

[4] Adams O, Wiesner M, Watanabe S, Yarowsky D (2019) Massively multilingual adversarial speech recognition. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 96–108. https://doi.org/10.18653/v1/N19-1009

[5] Adda G et al (2016) Breaking the unwritten language barrier: The BULB project. Procedia Comput Sci 81:8–14. https://doi.org/10.1016/j.procs.2016.04.023

[6] Besacier L, Barnard E, Karpov A, Schultz T (2014) Automatic speech recognition for under-resourced languages: A survey. Speech Commun 56:85–100. https://doi.org/10.1016/j.specom.2013.07.008

[7] Ghosh PK et al (2013) A comparative study of acoustic features for Mel frequency cepstral coefficients for speech recognition. In: Proceedings of the International Conference on Signal Processing and Communication, pp 248–252. https://doi.org/10.1109/ICSC.2013.6719787

[8] Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

[9] MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, pp 281–297. https://projecteuclid.org/proceedings/berkeley-symposium-on-mathematical-statistics-and-probability/Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and-Probability/chapter/toc/bsmsp/1200512992

[10] Sakti S et al (2010) Recent progress in developing grapheme-to-phoneme conversion for Indonesian. In: Proceedings of the Oriental COCOSDA International Conference on Speech Database and Assessments, pp 118–123. https://www.isca-speech.org/archive_v0/archive_papers/o-cocosda_2010/oc10_118.pdf

[11] Rijhwani S (2023) Improving Optical Character Recognition for Endangered Languages. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. https://www.cs.cmu.edu/~summ/docs/thesis.pdf

[12] Bluche T (2015) Deep neural networks for large vocabulary handwritten text recognition. PhD thesis, Université Paris-Sud, Paris, France. https://tel.archives-ouvertes.fr/tel-01249853

[13] Wang Y et al (2022) Character Recognition in Endangered Archives: Shui Manuscripts as a Case Study. Appl Sci 12(11):5361. https://doi.org/10.3390/app12115361

[14] Smith R (2007) An overview of the Tesseract OCR engine. In: Proceedings of the 9th International Conference on Document Analysis and Recognition, pp 629–633. https://doi.org/10.1109/ICDAR.2007.4376991

[15] Malamud K, Goldwater S (2023) Machine Learning for Ancient Languages: A Survey. Comput Linguist 49(3):703–751. https://doi.org/10.1162/coli_a_00458

[16] Singh AK (2019) Tribal languages of India: Issues of preservation and revitalization. Lang India 19(2):38–50. https://www.languageinindia.com/feb2019/singhtribiballanguagesfinal.pdf

[17] Mohanty P (2020) Documentation and preservation of endangered languages of Odisha. Lang Doc Conserv 14:108–133. https://nflrc.hawaii.edu/ldc/

[18] Mahapatra R (2019) The Bonda language: Documentation of an endangered tribal language of Odisha. Int J Dravidian Linguist 48(1):113–128. https://www.dravidianuniversity.ac.in/Journal/IJDLJan2019/7-Rakesh.pdf

[19] Mohanty S (2017) Palm leaf manuscripts of India: A cultural heritage and source of indigenous knowledge. Int J Inf Dissem Technol 7(1):35–39. https://www.ijidt.com/index.php/ijidt/article/view/156

[20] UNESCO (2024) Indigenous languages: Gateways to the world's cultural diversity. Technical Report CLT-2024-03, UNESCO, Paris, France. https://unesdoc.unesco.org/ark:/48223/pf0000387559

[21] Boersma P, Weenink D (2023) Praat: doing phonetics by computer [Computer program]. Version 6.3.10. http://www.praat.org/