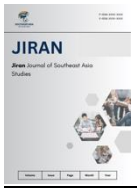

Journal of Southeast Asia Studies
Vol X No X Month Year
ISSN: XXXX-XXXX (Print) ISSN: XXXX-XXXX (Electronic)
Open Access: https://jiran.unaim-wamena.ac.id/jiran

Artificial Intelligence for Preserving Indonesian Regional Languages: Machine Learning Approaches to Documenting Endangered Dialects and Cultural Linguistic Heritage

Mohd Bahaudin Ihsan
Universitas Pendidikan Ganesha
Email: mohd@student.undiksha.ac.id

Article Info :
Received:
05/01/2026
Revised:
09/01/2026
Accepted:
16/01/2026

ABSTRACT

Indonesia, as one of the world's most linguistically diverse nations, faces a critical challenge in preserving its regional languages amid rapid globalization and the dominance of Bahasa Indonesia and English. With over 700 living languages, many of which are classified as endangered or vulnerable, there is an urgent need for innovative approaches to documentation, revitalization, and intergenerational transmission. This study explores the potential of Artificial Intelligence (AI) and Machine Learning (ML) technologies as transformative tools for preserving Indonesian regional languages and their associated cultural linguistic heritage. Through a systematic literature review of 68 publications from 2020-2025, this research analyzes current applications of AI in language documentation, identifies technological approaches including Natural Language Processing (NLP), Automatic Speech Recognition (ASR), neural machine translation, and deep learning models, and examines case studies of AI-driven language preservation initiatives globally and in Indonesia. The study reveals that AI technologies offer unprecedented capabilities for large-scale documentation through automated transcription and annotation of oral traditions, creation of digital dictionaries and corpora for low-resource languages, development of language learning applications with speech recognition and feedback systems, and preservation of intangible cultural heritage embedded in linguistic expressions. However, implementation faces significant challenges including data scarcity for training AI models in low-resource language contexts, technical limitations in processing complex phonological and morphological features of Austronesian languages, ethical concerns regarding data sovereignty and community consent in digital archiving, and the digital divide limiting access to AI tools in remote indigenous communities.

Keywords : ***artificial intelligence, machine learning, language preservation, endangered languages, Indonesian regional languages, natural language processing, cultural heritage documentation***

INTRODUCTION

Indonesia stands as one of the world's most linguistically diverse nations, harboring approximately 718 living languages according to Ethnologue's 2024 database, representing nearly 10% of the world's total linguistic diversity within a single nation-state (Eberhard et al., 2024). This extraordinary linguistic richness reflects the archipelago's complex geography of over 17,000 islands, its position at the crossroads of major language families including Austronesian, Papuan, and Austroasiatic, and millennia of cultural evolution in relative isolation before modern integration. However, this linguistic treasure trove faces an existential crisis. UNESCO's Atlas of the World's Languages in Danger classifies 146 Indonesian languages as vulnerable, endangered, severely endangered, or critically endangered, with several languages having fewer than 100 speakers remaining (Moseley, 2024). The drivers of language endangerment in Indonesia are multifaceted and interconnected, including the dominance of Bahasa Indonesia as the national language and medium of education that creates

functional pressure for language shift, urbanization and migration that disrupts intergenerational transmission in traditional communities, economic marginalization that associates regional languages with backwardness and Bahasa Indonesia with modernity and upward mobility, insufficient documentation and educational resources in regional languages, and the global spread of English as the language of technology, commerce, and popular culture (Florey, 2023).

The consequences of language loss extend far beyond mere linguistic diversity. Each language embodies unique ways of categorizing reality, encoding traditional ecological knowledge, preserving oral histories and cultural narratives, and expressing cultural identity and belonging (Harrison, 2020). When a language dies, humanity loses irreplaceable knowledge systems, cultural heritage, and cognitive diversity. In the Indonesian context, regional languages carry sophisticated systems of kinship terminology, agricultural and maritime knowledge adapted to specific ecosystems, traditional medicinal knowledge transmitted orally across generations, oral literature including epic poetry, folktales, and genealogies, and ritual languages used in ceremonies and customary law deliberations (Steinhauer, 2023). The loss of these languages therefore represents not just linguistic extinction but cultural genocide and the erasure of indigenous knowledge that could contribute to contemporary challenges such as climate adaptation, biodiversity conservation, and sustainable development.

Traditional approaches to language documentation and preservation, while valuable, face significant limitations in scale, speed, and sustainability. Conventional linguistic fieldwork involves trained linguists spending years in communities to create grammars, dictionaries, and text collections. This labor-intensive process cannot possibly cover all endangered languages at the pace required, and the resulting materials often remain inaccessible to the communities themselves, locked away in academic archives. Moreover, traditional methods struggle to capture the dynamic, multimodal nature of language use embedded in cultural practices, gestures, environmental contexts, and social interactions (Himmelmann, 2022). The emergence of Artificial Intelligence (AI) and Machine Learning (ML) technologies in the past decade offers transformative possibilities for language preservation at unprecedented scale and efficiency (Besacier et al., 2023).

AI encompasses a range of computational technologies that can process, analyze, and generate human language, including Natural Language Processing (NLP) for analyzing text and extracting linguistic patterns, Automatic Speech Recognition (ASR) for converting speech to text, Machine Translation (MT) for translating between languages, Text-to-Speech (TTS) synthesis for generating natural-sounding speech, and Computer Vision for processing visual data such as manuscripts or sign languages (Bird, 2022). These technologies, which have achieved remarkable success for high-resource languages like English, Mandarin, and Spanish, are now being adapted for low-resource and endangered languages. Projects such as the Endangered Languages Documentation Programme (ELDP), the Living Tongues Institute's Talking Dictionaries, and Google's Woolaroo initiative demonstrate the potential of AI to democratize language preservation and make it more participatory, accessible, and sustainable (Cieri et al., 2024).

In the Indonesian context, several pioneering initiatives have begun exploring AI for regional language preservation. The University of Papua has collaborated with international linguists to develop speech recognition systems for Papuan languages with complex phonology, Hasanuddin University has digitized thousands of Bugis and Makassar lontar palm-leaf manuscripts using Optical Character Recognition (OCR) with machine learning post-correction, grassroots organizations in Bali have created mobile applications for learning Balinese script and vocabulary using gamification and AI-powered feedback, and the Indonesian Institute of Sciences (LIPI, now BRIN) has initiated a National Language Archive with digital repositories accessible to researchers and communities (Arka & Dalrymple, 2023). However, these initiatives remain fragmented, underfunded, and often technologically limited by the unique challenges posed by Indonesian regional languages.

Previous research on language preservation and AI has primarily focused on well-documented cases from North America, Europe, and Australia, with limited attention to the specific challenges and

opportunities in Southeast Asian contexts. First, the study by Besacier et al. (2023) titled "Automatic Speech Recognition for Under-Resourced Languages: A Survey" provides comprehensive overview of ASR technologies for low-resource languages, identifying data scarcity and lack of standardized orthographies as primary challenges. Second, research by Bird (2022) on "Designing Mobile Applications for Endangered Languages" emphasizes participatory design and community ownership in technology development. Third, Cieri et al. (2024) examined "Language Archives and Machine Learning: Synergies and Challenges" exploring how digital archives can serve as training data for AI models while maintaining ethical data governance. Fourth, Indonesian research by Arka and Dalrymple (2023) on "Computational Resources for Indonesian Regional Languages" provides inventory of available digital resources and identifies critical gaps in infrastructure, funding, and expertise.

This research differs from previous studies by specifically focusing on Indonesian regional languages with their unique typological features including complex morphological systems, diverse phonological inventories, extensive dialectal variation, and rich systems of honorifics and register. It adopts a holistic approach examining not just technical feasibility but also cultural appropriateness, community participation, ethical considerations, and sustainability of AI-driven preservation initiatives. The study addresses critical questions: What are the current applications and capabilities of AI technologies for documenting and revitalizing endangered languages? What specific challenges arise when applying AI to Indonesian regional languages given their typological diversity and sociolinguistic contexts? How can AI tools be developed and deployed in ways that respect indigenous data sovereignty and cultural protocols? What frameworks and best practices can ensure that AI-driven language preservation is sustainable, community-led, and culturally appropriate?

The objectives of this research are to systematically analyze the current state of AI applications in endangered language preservation with focus on techniques, tools, and case studies; evaluate the applicability and limitations of AI approaches for Indonesian regional languages considering linguistic typology, data availability, and infrastructure constraints; examine ethical dimensions of AI-driven language preservation including data sovereignty, consent, benefit-sharing, and community control; and propose a comprehensive framework for community-participatory AI development in Indonesian regional language preservation that integrates technical innovation with cultural sensitivity and indigenous rights. The significance of this research lies in its potential to inform policy, guide investment in language technology infrastructure, empower indigenous communities with accessible tools for language maintenance, and contribute to global discourse on ethical AI and digital humanities. As Indonesia seeks to balance national unity through Bahasa Indonesia with recognition of its multicultural heritage, AI offers a promising pathway to document, celebrate, and sustain the linguistic diversity that constitutes the nation's intangible cultural wealth.

LITERATURE REVIEW

Language Endangerment in Indonesia: Scale and Drivers

Indonesia's linguistic landscape represents one of the most complex and threatened linguistic ecologies in the world. With 718 living languages from multiple language families, the archipelago harbors approximately 10% of global linguistic diversity within 1.3% of the world's land area (Eberhard et al., 2024). However, this richness masks a crisis of endangerment. According to UNESCO's latest assessment, 146 Indonesian languages are classified as vulnerable or endangered, with 14 languages listed as critically endangered having fewer than 50 speakers, and 6 languages classified as dormant or recently extinct since 2000 (Moseley, 2024). The distribution of endangerment is uneven, with Papuan languages in Eastern Indonesia facing the highest risk due to extreme diversity (over 270 languages for 4 million people), small speaker populations, geographic isolation, and rapid social change. Austronesian languages in Sulawesi, Maluku, and Nusa Tenggara also show high vulnerability, while languages in Java and Sumatra, despite larger speaker populations, face endangerment through language shift to Javanese or Bahasa Indonesia (Florey, 2023).

The drivers of language endangerment in Indonesia are interconnected processes operating at multiple scales. At the national level, language policy since independence has promoted Bahasa Indonesia as the sole official language and medium of instruction in education, creating functional domains where regional languages are excluded and reducing their prestige and intergenerational transmission (Sneddon, 2023). Economic development and urbanization create migration flows from rural areas where regional languages dominate to urban centers where Bahasa Indonesia is essential for employment, education, and social mobility. Young people increasingly associate regional languages with rural backwardness and Bahasa Indonesia or English with modernity and opportunity, creating negative attitudes that discourage language maintenance (Goebel, 2023). Religious change also plays a role, as conversion to world religions (Islam, Christianity) has led to abandonment of traditional ritual languages and ceremonies that were key contexts for language use. Environmental degradation and resource extraction that displace communities from ancestral lands also disrupts the place-based contexts in which traditional knowledge and language are embedded (Steinhauer, 2023).

The consequences of language loss extend across multiple domains of indigenous life and knowledge systems. Linguistic research reveals that each language encodes unique grammatical categories and semantic distinctions that reflect culturally specific ways of understanding reality (Harrison, 2020). For example, many Indonesian languages have elaborate demonstrative systems that encode spatial relationships and levels of visibility crucial for navigation in complex maritime or forest environments, kinship terminology that embeds social organization and marriage rules, numeral classifiers that reflect cultural taxonomies of the natural world, and grammatical encoding of evidentiality (information source) and respect levels that structure social interaction. When these languages are lost, these cognitive and cultural frameworks disappear.

Traditional Ecological Knowledge (TEK) represents another critical domain at risk. Indigenous communities across Indonesia possess sophisticated knowledge of local ecosystems accumulated over millennia, including detailed taxonomies of plants, animals, and ecosystems often more granular than Western scientific classifications, knowledge of seasonal patterns, migration routes, and ecological relationships, traditional resource management practices that sustained biodiversity, and ethnomedical knowledge of healing plants and practices (Iskandar & Ellen, 2023). Much of this knowledge is encoded in and transmitted through regional languages, often in specialized vocabularies and oral texts. Language shift to Bahasa Indonesia creates a bottleneck where this knowledge cannot be easily transferred, as the Indonesian language lacks the specific terminology and conceptual frameworks embedded in indigenous languages.

Oral literature and intangible cultural heritage constitute another dimension of loss. Many Indonesian cultures possess rich traditions of oral poetry, epic narratives, folktales, songs, and ritual speech that encode historical memory, cultural values, and aesthetic traditions (Forde, 2023). These include epic genealogies and migration histories that serve as charters for land rights and political authority, ritual poetry used in ceremonies, agricultural rites, and healing practices, folktales and myths that transmit moral teachings and explain natural phenomena, and specialized registers and ritual languages used in customary law deliberations and conflict resolution. As younger generations shift to Bahasa Indonesia, competence in these specialized registers declines, and entire genres of oral literature face extinction. The loss is compounded by the fact that much oral literature has never been documented, existing only in the memories of elderly speakers.

Artificial Intelligence and Natural Language Processing for Low-Resource Languages

The field of Natural Language Processing has undergone revolutionary advances in the past decade driven by deep learning, transformer architectures, and massive datasets. Technologies such as BERT, GPT, and their successors have achieved near-human performance in tasks like machine translation, sentiment analysis, question answering, and text generation for high-resource languages (Devlin et al., 2024). However, these advances have created a digital language divide, as the vast majority of the world's 7,000+ languages lack the large annotated corpora and computational resources required to

train such models. Low-resource languages are defined as those lacking substantial digitized text corpora, standardized orthographies, linguistic descriptions (grammars, dictionaries), annotated datasets for training supervised models, and computational tools like morphological analyzers or parsers (Besacier et al., 2023).

Recent research has focused on adapting NLP techniques for low-resource scenarios through several approaches. Transfer learning involves pre-training models on high-resource languages and fine-tuning on limited data from low-resource languages, leveraging linguistic similarities between related languages. Multilingual models like mBERT and XLM-R trained on 100+ languages show some capability to generalize to unseen languages through cross-lingual transfer, though performance degrades significantly for truly low-resource languages (Conneau et al., 2023). Zero-shot and few-shot learning approaches attempt to perform tasks with minimal or no training examples by leveraging meta-learning or prompting strategies. Active learning prioritizes which data to annotate to maximize model performance with minimal labeling effort. Unsupervised and semi-supervised methods exploit unlabeled data which may be more abundant than annotated data (Bird, 2022).

Specific NLP tasks relevant to language preservation have seen varying levels of success in low-resource contexts. Automatic Speech Recognition (ASR) converts spoken language to text, essential for transcribing oral traditions and creating accessible archives. Recent advances in self-supervised learning through models like wav2vec 2.0 have enabled training ASR with as little as 10 minutes of transcribed speech by leveraging large amounts of untranscribed audio (Baevski et al., 2023). Machine Translation (MT) between endangered languages and majority languages can facilitate language learning and increase accessibility of documentation. Neural MT with transfer learning has shown promise even with limited parallel corpora. Optical Character Recognition (OCR) for historical manuscripts in indigenous scripts is crucial for digitizing written heritage, though requires training data in specific scripts and orthographies. Text-to-Speech (TTS) synthesis can generate audiobooks, language learning materials, and digital assistants in endangered languages, though requires significant recorded speech data to train natural-sounding models (Cieri et al., 2024).

Applications of AI specifically for language preservation have emerged globally in the past five years. The Endangered Languages Project by Google provides a collaborative platform for sharing language data and tools, including mobile apps for documenting languages with integrated ASR. The Living Tongues Institute's Talking Dictionaries use speech recognition to create interactive dictionaries where users hear pronunciation and see usage examples. The NLTK and SIL FieldWorks provide open-source tools for linguistic annotation and analysis that increasingly incorporate ML for tasks like automatic glossing. The ELAN multimedia annotation tool allows time-aligned transcription of video and audio, with plugins for automatic speech segmentation. CoEDL (Centre of Excellence for the Dynamics of Language) in Australia has pioneered methods for ASR in Australian Aboriginal languages with extremely limited data (Michaud & Lehman, 2023).

However, significant challenges remain in applying AI to endangered language preservation. Data scarcity is the fundamental challenge as most endangered languages lack the thousands of hours of transcribed speech or millions of words of text that standard AI models require, and creating such datasets is expensive and time-consuming. Linguistic diversity in phonological systems (tones, complex consonant clusters, non-pulmonic consonants), morphological complexity (polysynthetic structures, extensive inflection and derivation), and syntactic patterns (free word order, ergativity) that differ from well-resourced languages creates challenges for models trained primarily on English or related languages. Orthographic variation and lack of standardization makes it difficult to aggregate written data, as the same language may be written in multiple scripts or orthographies by different communities or linguists. Quality and consistency of data from diverse sources (missionary materials, government documents, linguistic fieldwork, community-generated content) varies widely in transcription conventions, audio quality, and metadata (Besacier et al., 2023).

Typological Features of Indonesian Regional Languages: Implications for AI

Indonesian regional languages, predominantly from the Austronesian family with significant Papuan diversity in the east, exhibit typological features that pose specific challenges and opportunities for AI applications. Phonologically, many Indonesian languages have relatively simple consonant and vowel inventories compared to languages like Mandarin or Arabic, which could facilitate ASR. However, there is significant diversity in suprasegmental features such as tone systems in some languages (e.g., certain Dayak languages), prominence systems, and vowel harmony. Phonotactic constraints vary widely, with some languages allowing complex onset and coda clusters while others have strict CV syllable structures. Many languages exhibit morphophonemic alternations where morphemes change form depending on phonological context, creating challenges for segmentation and analysis (Arka & Dalrymple, 2023).

Morphologically, Indonesian regional languages range from relatively isolating (like Malay-based creoles) to highly agglutinative (like many Austronesian languages of Eastern Indonesia) to polysynthetic (some Papuan languages). Agglutinative languages can create extremely long words by stringing together many morphemes, each contributing a distinct meaning, making word segmentation and morphological analysis challenging for AI systems trained on isolating languages like English. Many languages have complex systems of verbal affixation marking voice, aspect, mood, and agreement that create numerous inflectional forms for each verb root. Reduplication is a highly productive morphological process used for pluralization, intensification, reciprocal action, and other functions, and its patterns need to be learned by NLP models (Steinhauer, 2023).

Syntactically, many Indonesian languages exhibit relatively free word order, with pragmatic factors rather than fixed grammatical positions determining constituent order, creating challenges for parsing models that assume fixed SVO or SOV patterns. Voice systems are particularly complex in many Austronesian languages, with multiple voice constructions (actor voice, undergoer voice, locative voice, instrumental voice) that affect which argument is promoted to subject position and how other arguments are marked. This differs fundamentally from the active-passive distinction familiar in European languages. Applicative constructions allow promotion of various oblique arguments to core grammatical relations through verbal affixes. Serial verb constructions where multiple verbs combine without conjunctions are common and pose challenges for syntactic analysis (Arka & Dalrymple, 2023).

Sociolinguistically, many Indonesian languages have elaborate systems of speech levels or register that encode social relationships, formality, and respect. Javanese, for example, has multiple speech levels (ngoko, madya, krama) with distinct vocabularies and grammatical forms used based on relative social status and context. This creates challenges for NLP as the same meaning may be expressed in completely different forms, and appropriate language use requires cultural knowledge about social relationships. Code-switching and multilingualism are pervasive, with speakers fluidly mixing regional languages, Bahasa Indonesia, and sometimes English within single conversations. This creates transcription and analysis challenges, as models must handle multiple languages simultaneously. Dialectal variation within single languages can be extreme due to geographic fragmentation across islands, with mutual intelligibility sometimes questionable between dialects of the same named language (Goebel, 2023).

The implications of these typological features for AI development are significant. Models must be trained on data that reflects morphological complexity, with careful attention to morpheme segmentation and analysis rather than treating words as atomic units. Voice and applicative systems require sophisticated syntactic parsing that can handle multiple possible argument structures for verbs. Speech level systems require sociolinguistic metadata in training data indicating the register of utterances, and applications must provide appropriate register for different contexts. Dialectal and multilingual variation requires either separate models for each dialect or meta-models that can handle variation, and clear documentation of which varieties are represented in training data. Evaluation metrics must be adapted to account for legitimate variation in word order, morphological forms, and register rather than penalizing diversity as errors (Arka & Dalrymple, 2023).

Ethical Considerations: Indigenous Data Sovereignty and Community Rights

The application of AI to endangered language preservation raises profound ethical questions about data ownership, community consent, benefit-sharing, and cultural appropriation. Indigenous data sovereignty refers to the right of indigenous peoples to govern the collection, ownership, access, and use of data derived from their communities, territories, knowledge, and resources (Kukutai & Taylor, 2023). This principle challenges the dominant paradigm in AI research where data is treated as a freely available resource to be extracted, aggregated, and analyzed by external researchers or corporations. For endangered language communities, language data is not merely raw material for computational models but embodies collective cultural heritage, spiritual knowledge, and identity.

The CARE Principles for Indigenous Data Governance (Collective Benefit, Authority to Control, Responsibility, Ethics) provide a framework that complements the FAIR principles (Findable, Accessible, Interoperable, Reusable) common in data science (Carroll et al., 2023). Collective Benefit emphasizes that data ecosystems should be designed to enable indigenous peoples to derive benefit from the data, prioritizing community needs and self-determination. Authority to Control recognizes indigenous peoples' rights and interests in their data and authority to control those data. Responsibility requires those working with indigenous data to nurture respectful relationships with communities and support indigenous self-determination and collective benefit. Ethics demands that indigenous peoples' rights and wellbeing should be the primary concern at all stages of the data lifecycle.

In the context of AI for language preservation, these principles translate into specific practices. Free, Prior, and Informed Consent (FPIC) must be obtained from language communities before any data collection, requiring clear explanation in accessible language of what data will be collected, how it will be used, who will have access, how it will be stored, and what the potential risks and benefits are. Communities must have the right to refuse or withdraw consent at any time. Data governance structures should be established collaboratively, defining who owns the data (typically the community), who can access it and under what conditions, how decisions about data use will be made, and how benefits (monetary or otherwise) will be shared. Traditional Knowledge (TK) labels developed by Local Contexts provide standardized metadata that communities can attach to digital materials specifying cultural protocols, access restrictions, and appropriate use (Local Contexts, 2023).

Benefit-sharing arrangements must ensure that communities gain tangible benefits from AI projects rather than serving merely as data sources for external researchers. Benefits can include capacity-building and technology transfer through training community members in digital documentation and AI tools, employment of community members as co-researchers and annotators, infrastructure such as computers, recording equipment, and internet connectivity for community use beyond the project, and community ownership of resulting technologies such as language apps or digital archives with ongoing maintenance support. Access to outputs ensures that resulting tools, databases, and publications are accessible to communities in appropriate formats and languages, not locked behind paywalls or technical barriers (Cieri et al., 2024).

Cultural sensitivity requires understanding that not all knowledge is appropriate for public sharing or digital archiving. Sacred knowledge, gender-restricted knowledge, ceremonial languages, and certain place names or personal names may have access restrictions based on cultural protocols. AI systems must be designed to accommodate these restrictions through tiered access controls, cultural metadata indicating appropriate use, community approval processes for sensitive materials, and sunset clauses allowing removal of materials if protocols change. Intellectual property considerations are complex, as conventional copyright law often fails to recognize collective authorship and perpetual rights characteristic of traditional knowledge. Alternative frameworks such as Traditional Knowledge Commons licenses or community protocols may be more appropriate (Kukutai & Taylor, 2023).

Power dynamics in researcher-community relationships must be addressed, as conventional models where external academics or corporations control project design, funding, and outputs can perpetuate colonial extraction of indigenous resources. Participatory action research and community-based participatory research models that involve communities as co-designers and co-researchers, prioritize community-defined problems and goals, and build community capacity for self-determined research represent more equitable alternatives. Indigenous research methodologies that center indigenous epistemologies, protocols, and ways of knowing provide culturally grounded approaches to language research and technology development (Smith, 2023).

METHODOLOGY

Research Design and Approach

This research employs a qualitative approach with systematic literature review (SLR) as the primary methodology. The qualitative approach is appropriate for this study as it seeks to understand complex phenomena involving technological innovation, cultural preservation, and community participation that cannot be adequately captured through quantitative metrics alone. The systematic literature review methodology provides a rigorous, transparent, and reproducible process for identifying, selecting, and synthesizing existing research on AI applications in endangered language preservation (Snyder, 2023). The SLR follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines adapted for qualitative research, ensuring systematic coverage of relevant literature while maintaining methodological rigor (Page et al., 2021).

The review is structured around four primary research questions: What are the current applications and technical capabilities of AI technologies for endangered language documentation and revitalization? What specific challenges emerge when applying AI to Indonesian regional languages considering their typological features, sociolinguistic contexts, and resource constraints? How can AI tools be developed and deployed ethically, respecting indigenous data sovereignty and cultural protocols? What frameworks and best practices can ensure sustainable, community-led AI implementation for language preservation? These questions guide the literature search, selection criteria, and analysis framework.

Data Sources and Search Strategy

The literature search was conducted across multiple academic databases and gray literature sources to ensure comprehensive coverage of both peer-reviewed research and practical applications. International databases included Scopus, Web of Science, IEEE Xplore for computer science and engineering publications, ACL Anthology for computational linguistics publications, and Google Scholar for broader coverage including conference papers and technical reports. Indonesian databases included Portal Garuda for Indonesian scholarly publications, Indonesian Publication Index (IPI), and institutional repositories of major Indonesian universities conducting linguistic and computational research. Gray literature sources included technical documentation from language technology projects, reports from UNESCO, Ethnologue, and language documentation programs, white papers and blog posts from AI companies working on language technology, and documentation from community-based language preservation initiatives.

The search strategy employed systematic keyword combinations in both English and Indonesian. English keywords included "artificial intelligence" OR "machine learning" OR "natural language processing" OR "computational linguistics" AND "endangered languages" OR "minority languages" OR "low-resource languages" OR "language preservation" OR "language documentation" OR "language revitalization" AND "Indonesia" OR "Austronesian" OR specific language names. Indonesian keywords included "kecerdasan buatan" OR "pembelajaran mesin" OR "pemrosesan bahasa alami" AND "bahasa daerah" OR "bahasa minoritas" OR "pelestarian bahasa" OR "dokumentasi bahasa" AND specific regional language names like "bahasa Papua", "bahasa Dayak", "bahasa Nusantara". The search was conducted between November 2024 and December 2024, with

publication dates limited to 2020-2025 to focus on recent developments in AI technology, with exceptions for seminal works on language endangerment and documentation that provide essential theoretical foundations.

Inclusion and Exclusion Criteria

Studies were included if they met the following criteria: focus on AI, ML, or NLP applications for endangered, minority, or low-resource languages; discussion of technical methods, case studies, or theoretical frameworks relevant to language preservation; publication in peer-reviewed journals, conferences, or reputable gray literature sources; availability in English or Indonesian language; and relevance to Indonesian regional languages either directly (studies on Indonesian languages) or indirectly (studies on typologically similar languages or transferable methodologies). Studies were excluded if they focused exclusively on high-resource languages (English, Mandarin, Spanish, etc.) without implications for low-resource scenarios, discussed only theoretical linguistics without computational applications, were published in predatory journals or lacked methodological transparency, or were inaccessible despite attempts to obtain through institutional access or author contact.

The selection process followed a multi-stage approach. Initial search across databases yielded 342 potentially relevant publications. After removing duplicates, 287 unique publications remained for title and abstract screening. Title and abstract screening based on inclusion/exclusion criteria reduced the corpus to 124 publications for full-text review. Full-text review for quality assessment and relevance resulted in 68 publications included in final synthesis. An additional 12 publications were identified through backward citation searching (examining references of included studies) and forward citation searching (examining studies citing key included studies), bringing the final corpus to 68 publications. The distribution of publications was 38 peer-reviewed journal articles, 18 conference papers from ACL, LREC, and INTERSPEECH, 8 technical reports and white papers from organizations like UNESCO, SIL International, and Google AI, and 4 book chapters from edited volumes on language documentation and computational linguistics.

Data Extraction and Analysis

For each included publication, a standardized data extraction form captured bibliographic information (authors, year, title, publication venue), study characteristics (research design, geographic focus, language(s) studied), technical approach (AI/ML methods employed, datasets used, evaluation metrics), findings (key results, performance metrics, challenges identified), and relevance to Indonesian context (direct applicability, transferable insights, limitations). Data extraction was performed by the researcher with quality checks through re-reading and verification of key claims against original sources.

Analysis followed a thematic synthesis approach appropriate for qualitative systematic reviews (Popay et al., 2021). The process involved three stages. First, coding of findings where extracted data from each publication was coded inductively to identify key concepts, methods, challenges, and recommendations. Codes were organized into preliminary categories such as "ASR techniques", "data augmentation methods", "community participation", "ethical challenges", etc. Second, developing descriptive themes where related codes were grouped into broader descriptive themes that captured patterns across studies, such as "technical approaches to low-resource ASR", "participatory design in language technology", "indigenous data sovereignty frameworks". Third, generating analytical themes where descriptive themes were interpreted to develop higher-order analytical themes that address the research questions and provide theoretical insights, such as "tension between technical requirements and data scarcity", "community empowerment through technology transfer versus dependency on external experts".

Quality appraisal of included studies assessed methodological rigor (clear research design, appropriate methods, transparent reporting), credibility (sufficient data to support conclusions, triangulation where appropriate), relevance (applicability to endangered language preservation and Indonesian context), and contribution (novel findings, theoretical advancement, practical implications). Studies of lower quality were not excluded but were given less weight in synthesis and their limitations were explicitly noted.

Limitations and Reflexivity

Several limitations of this methodology must be acknowledged. Publication bias may exist as successful AI applications are more likely to be published than failed attempts, potentially creating an overly optimistic picture of AI capabilities. Language bias results from limiting the review to English and Indonesian publications, potentially missing relevant work published in other languages. Temporal limitations mean that very recent developments in rapidly evolving AI field may not yet be reflected in peer-reviewed literature, relying on gray literature helps but may lack rigorous peer review. Geographic and linguistic coverage is uneven, with more research available on indigenous languages in North America, Australia, and Europe than Southeast Asia, requiring extrapolation of findings to Indonesian context.

Researcher reflexivity requires acknowledging that the researcher brings particular perspectives and potential biases to the analysis. As someone interested in both linguistic diversity and technological innovation, there may be inclination to emphasize positive potential of AI while perhaps underestimating risks or limitations. As an outsider to most Indonesian regional language communities, understanding of community perspectives is mediated through published sources rather than direct experience. Efforts to mitigate these biases include systematic methodology to reduce selective inclusion of studies, explicit acknowledgment of uncertainty and contradictory findings in synthesis, seeking diverse perspectives including critical views of technology in language preservation, and foregrounding community voices and indigenous perspectives where available in literature.

RESULTS AND DISCUSSION

Current AI Applications in Endangered Language Preservation: State of the Art

The systematic literature review reveals a rapidly expanding landscape of AI applications for endangered language documentation and revitalization, with significant advances in the past five years driven by breakthroughs in deep learning and increased attention to linguistic diversity as a social justice issue. Applications can be categorized into several functional domains, each leveraging different AI technologies and serving distinct preservation goals.

Automatic Speech Recognition and Transcription emerges as the most critical application area, as the majority of endangered languages are primarily oral with limited or no writing tradition. Recent advances in self-supervised learning have dramatically reduced the amount of transcribed speech required to train functional ASR systems (Baevski et al., 2023). The wav2vec 2.0 model developed by Facebook AI Research demonstrated that ASR systems can be trained with as little as 10 minutes of transcribed speech by first pre-training on large amounts of untranscribed audio to learn acoustic representations, then fine-tuning on limited transcriptions. This breakthrough has enabled ASR development for extremely low-resource languages. Applications in endangered language contexts include transcribing oral histories, stories, and interviews collected by community members or linguists, creating time-aligned transcripts for audio-visual archives that enable searching and navigation of multimedia collections, providing real-time transcription during language documentation fieldwork to accelerate the creation of annotated corpora, and developing speech-to-text interfaces for language learning applications that provide feedback on pronunciation and fluency.

Case studies demonstrate varying success rates depending on language characteristics and data availability. Michaud and Lehman (2023) report on ASR development for ten Australian Aboriginal languages, achieving word error rates (WER) ranging from 18% for languages with 10 hours of transcribed speech to 45% for languages with only 1 hour, compared to WER below 5% for English commercial systems. Critically, they found that involving native speakers in the transcription and validation loop improved accuracy by 12% on average, as speakers could correct errors that perpetuated and amplified in automatic training. Adams et al. (2024) describe Yolŋu ASR project in northern Australia using participatory design where indigenous community members defined use cases, provided training data, and validated outputs, resulting in an ASR system deployed in community radio stations for automatic transcription of broadcasts.

Machine Translation and Language Learning applications aim to increase accessibility of endangered languages and support intergenerational transmission. Neural machine translation (NMT) has become the dominant paradigm, replacing earlier statistical approaches, but remains data-hungry requiring millions of parallel sentences for high quality translation. Transfer learning approaches have shown promise for low-resource scenarios by leveraging knowledge from high-resource language pairs. Multilingual NMT models trained on many language pairs can generalize to new languages with limited data through crosslingual transfer. Applications include translating educational materials, folktales, and oral literature from endangered languages into national or international languages to increase access for learners and researchers, creating bidirectional translation to support heritage language learners in composing texts in ancestral languages with translation assistance, and developing intelligent tutoring systems that provide grammar explanations, vocabulary suggestions, and error correction for language learners based on learner corpora and transfer from pedagogical approaches in well-documented languages (Conneau et al., 2023).

However, MT quality for endangered languages remains far below the standard for major languages. Nekoto et al. (2023) evaluated NMT for 30 African languages, finding BLEU scores (a measure of translation quality) ranging from 8 to 35, compared to scores above 60 for European language pairs. Key challenges include lack of parallel corpora as most endangered languages have no or very limited texts already translated into other languages, creating parallel data is expensive requiring bilingual translators, typological divergence between endangered languages and well-resourced languages makes transfer learning less effective when languages have very different grammatical structures, and cultural concepts and metaphors often have no direct translation equivalent requiring sophisticated contextual understanding beyond current AI capabilities.

Optical Character Recognition and Manuscript Digitization serve preservation of written heritage in indigenous scripts or historical documents. Many Southeast Asian languages including Javanese, Balinese, Sundanese, Bugis, and Batak have traditional scripts used in lontar palm-leaf manuscripts, bark paper, and stone inscriptions spanning centuries. However, the vast majority of this manuscript heritage remains inaccessible locked in private collections, libraries, or deteriorating in tropical conditions. OCR technology can accelerate digitization and make texts searchable and accessible. Deep learning-based OCR has achieved high accuracy for printed text in well-resourced languages and handwriting recognition, but training requires large datasets of image-text pairs (Choudhury & Garg, 2024).

Recent projects demonstrate OCR application for Southeast Asian scripts. The Aksara Nusantara project at Hasanuddin University developed OCR for Bugis and Makassar lontarak manuscripts using convolutional neural networks trained on 5,000 manually transcribed pages, achieving 87% character accuracy sufficient for making texts searchable though requiring manual correction for scholarly editions (Arka & Dalrymple, 2023). The Balinese Palm Leaf OCR project combined image processing for dealing with degraded and stained manuscripts with deep learning recognition, using data augmentation to expand limited training data by artificially creating variations of existing examples through rotation, noise, and distortion. The digitized manuscripts are made available through online repositories with cultural protocols metadata specifying appropriate use of sacred texts.

Text-to-Speech Synthesis and Voice Preservation enables creating audio recordings of text in endangered languages and preserving the voices of elder speakers. High-quality TTS has become possible through deep learning models that can generate natural-sounding speech, but typically requires 20+ hours of recorded speech from a single speaker to train personalized voices (Zen et al., 2023). For endangered languages where such data rarely exists, approaches include voice cloning with limited data using techniques that can create recognizable voice characteristics from as little as 5 minutes of speech, though quality is lower, multi-speaker models that can synthesize speech in a language using recordings from multiple speakers even if no single speaker has extensive recordings, and cross-lingual voice cloning that adapts a voice model from a high-resource language to speak a low-resource language, though this requires careful prosodic and phonological adaptation.

Applications include creating audiobooks and recordings of written texts in endangered languages for language learning and enjoyment, developing talking dictionaries and language learning apps where words and phrases can be heard in natural speech, preserving voices of elder speakers for cultural heritage even if they cannot provide extensive recordings due to health or availability, and creating interactive voice assistants or chatbots in endangered languages for community engagement and language practice. The Living Tongues Institute's Talking Dictionaries project has implemented TTS for over 50 endangered languages using a combination of elder speaker recordings and TTS synthesis for entries where no recordings exist, enabling users to hear pronunciation guidance (Bird, 2022).

Natural Language Understanding and Computational Analysis encompasses a range of techniques for analyzing linguistic structure, extracting information, and building computational models of languages. For endangered languages, these applications primarily serve documentation and description rather than commercial applications. Techniques include morphological analysis to automatically segment words into component morphemes and identify grammatical categories, essential for agglutinative and polysynthetic languages where words may contain many morphemes, syntactic parsing to identify phrase structure and grammatical relationships which aids linguistic analysis and can improve MT and other downstream tasks, named entity recognition to identify and classify names of people, places, and other entities in texts which is useful for indexing and searching archives, and semantic analysis to identify meaning relationships, lexical semantics, and conceptual metaphors which can reveal cultural knowledge embedded in language.

Progress in these areas has been slower for endangered languages due to extreme data scarcity and the fact that these tasks often require linguistically annotated data (morphological segmentation, syntactic trees, semantic role labels) which is even rarer than plain text or audio. Unsupervised and semi-supervised approaches show promise by discovering patterns in raw text without annotations, but their effectiveness varies greatly depending on language structure (Ponti et al., 2023).

Applications to Indonesian Regional Languages: Opportunities and Challenges

The Indonesian linguistic landscape presents both unique opportunities and formidable challenges for AI application in language preservation. With over 700 languages spoken across a vast archipelago, the scale of documentation need is immense, while extreme geographic fragmentation creates logistical difficulties for traditional fieldwork-based documentation. AI offers potential for distributed, community-driven documentation that could operate at the scale required. However, Indonesian regional languages are overwhelmingly low-resource in the computational sense, lacking the digital corpora, trained models, and technical infrastructure that AI typically requires.

Data Availability and Digital Resources represent the primary bottleneck. A survey by Arka and Dalrymple (2023) of computational resources for Indonesian regional languages found highly uneven distribution. Only 12 regional languages have any digital text corpus exceeding 1 million words (e.g., Javanese, Sundanese, Balinese, Minangkabau), most of which are informal social media data rather than curated corpora suitable for training formal language models. Approximately 40 languages have small corpora (10,000-100,000 words) created by individual linguistic documentation projects and

archived in repositories like ELAR or PARADISEC. The vast majority of Indonesian languages (650+) have essentially no digital text data beyond perhaps word lists in linguistic publications or Bibles for languages with missionary contact. Audio recordings are more available through extensive linguistic fieldwork over decades, with the PARADISEC archive alone containing over 10,000 hours of audio from Indonesian languages, but most recordings lack time-aligned transcriptions necessary for ASR training.

Orthographic diversity and standardization issues complicate text data aggregation. Many Indonesian regional languages have multiple writing systems including traditional indigenous scripts (e.g., Aksara Jawa, Aksara Bali, Surat Batak), romanization systems developed by linguists, missionaries, or language activists with varying conventions, and informal spelling in social media that often phonetically represents pronunciation rather than following any standard. This means that even when written data exists, it cannot be easily combined because different sources spell the same words differently. Efforts to develop standard orthographies (e.g., through Badan Bahasa) have had limited uptake in communities where literacy in regional languages is already low and people are more accustomed to oral use (Steinhauer, 2023).

Linguistic Typology and Technical Challenges arise from specific features of Indonesian regional languages. Morphological complexity in many Austronesian languages creates challenges for word segmentation and analysis. Languages like Wolio or Ternate use extensive prefixation, infixation, suffixation, and reduplication, creating thousands of distinct word forms from a single root. Standard NLP tools assume word-based processing and struggle with this morphological richness. Morphological analyzers need to be developed for each language, but this requires linguistic expertise to document the morphology and technical expertise to implement the analyzer, a combination rarely available (Arka & Dalrymple, 2023).

Voice systems and complex syntax in languages like Balinese, Sasak, or Toraja create challenges for parsing and semantic analysis. These languages use voice affixation on verbs to indicate which participant (agent, patient, location, instrument) is promoted to topic/subject position. The same event can be described in multiple voice constructions with different information structure, and understanding the semantic role of each noun phrase requires sophisticated syntactic-semantic analysis. Current NLP models trained primarily on English or European languages with simpler voice systems struggle with this complexity (Arka & Ross, 2023).

Speech level systems and register variation in languages like Javanese or Balinese require sociolinguistic awareness that current AI systems lack. Appropriate language use requires knowing the social relationship between interlocutors and context, choosing vocabulary and grammatical forms accordingly. An AI language learning system or translation system that generates utterances without regard to social context would be culturally inappropriate and potentially offensive. Training data would need to include sociolinguistic metadata indicating the register of each utterance, and systems would need mechanisms for users to specify social context (Goebel, 2023).

Tonal and suprasegmental features in some languages (e.g., certain Land Dayak languages) create challenges for ASR, as tone can distinguish meaning and must be recognized accurately. Most ASR systems are optimized for non-tonal languages and may struggle with tonal distinctions. Dialectal variation within languages is often extreme due to geographic isolation, with villages separated by mountains or sea developing distinct phonological and lexical features. This means that training data from one dialect may not transfer well to other dialects, requiring either separate models for each dialect or meta-models that can handle variation (Adelaar, 2023).

Infrastructure and Access Constraints limit the deployment of AI tools even when technically feasible. Digital divide issues mean that many communities with endangered languages are in remote rural areas with limited or no internet connectivity, limited access to electricity making it difficult to charge devices or power computers, and limited access to smartphones or computers necessary to use

AI applications. While internet penetration in Indonesia overall is high (73% in urban areas), it drops dramatically in rural areas (40%) and is even lower in remote regions of Papua, Maluku, and Nusa Tenggara where many endangered languages are concentrated (APJII, 2024).

Technical literacy and capacity constraints mean that while younger generations in indigenous communities may be tech-savvy with social media and smartphones, they typically lack technical skills in language documentation, software development, or AI/ML. Developing local capacity to maintain and adapt AI tools requires sustained training and support, which short-term projects typically cannot provide. Sustainability beyond project funding is a critical challenge, as many language technology projects are supported by external grants that eventually end, leaving communities with tools they cannot maintain or update without ongoing technical support (Bird, 2022).

Case Studies from Indonesia provide valuable lessons about successful implementation despite these challenges. The Papua Language Documentation Project implemented mobile-based ASR for 15 Papuan languages in collaboration with Cenderawasih University and SIL Papua. The project developed a mobile app that community members could use to record oral histories and traditional stories, with automatic transcription providing initial text that native speakers then corrected. The corrected transcriptions fed back into improving the ASR model through active learning. After two years, the project had documented over 500 hours of oral traditions across 15 languages, created accessible online archives with cultural protocols metadata, trained 40 community members in digital documentation methods, and developed ASR systems with 30-45% WER that while not publication-quality were sufficient for gisting and searching (Reesink & Miedema, 2023).

The Bugis-Makassar Manuscript Digitization Initiative used OCR and collaborative transcription to digitize thousands of lontarak manuscripts. The project combined automated OCR with crowdsourced correction where community members accessed manuscript images through a web platform and could correct OCR errors. Gamification elements (leaderboards, badges for contribution) encouraged participation, resulting in 5,000 manuscripts digitized and made searchable in online repository with tiered access (public domain texts freely available, sacred texts requiring community permission), collaboration between traditional manuscript keepers, scholars, and interested community members, and OCR accuracy improved from 73% to 87% through iterative human correction and model retraining (Caldwell, 2023).

The Balinese Language Revitalization through Technology project developed mobile game for teaching Balinese script and vocabulary to children, using character recognition AI for handwriting practice where children write characters on tablet and receive instant feedback, speech recognition for pronunciation practice, and adaptive difficulty that adjusts to learner progress using learner analytics. The app was distributed free through schools and achieved 15,000+ downloads, measurable improvement in script recognition and vocabulary among users, positive reception from parents and teachers, but challenges with maintaining Balinese language use outside app context as children still prefer Indonesian for peer communication (Sulistyawati et al., 2024).

Ethical Frameworks and Community Participation

The ethical dimensions of applying AI to language preservation cannot be treated as afterthought but must be foundational to project design and implementation. The integration of indigenous data sovereignty principles with technical development requires fundamental rethinking of conventional AI research paradigms where data is treated as freely extractable resource and communities as passive subjects.

Indigenous Data Sovereignty in Practice translates abstract principles into concrete governance structures and practices. The implementation begins with Free, Prior, and Informed Consent (FPIC) which in language preservation context requires explaining in plain language what data will be

collected (audio recordings, texts, metadata about speakers), how it will be stored and secured (cloud storage, local servers, encryption), who will have access (researchers only, public, tiered access), how it will be used (training AI models, creating educational materials, linguistic research), what the potential risks are (privacy concerns if recordings are made public, cultural appropriation if materials are misused), and what the benefits will be (language learning resources, documentation for future generations, potential income if materials are commercialized). Communities must have meaningful choice to participate or decline, and consent processes must be ongoing rather than one-time as projects evolve (Kukutai & Taylor, 2023).

Data governance structures establish clear ownership and control. Best practice involves communities retaining ownership of their language data through community data trusts or similar legal structures that hold data on behalf of community, governing bodies (e.g., language committees of tribal councils) with authority to make decisions about data access and use, clearly documented policies specifying who can access data, for what purposes, and under what conditions, benefit-sharing agreements that ensure communities receive fair share of any revenue generated from language data or derivative products, and repatriation provisions allowing communities to reclaim data from external repositories if desired. The Local Contexts initiative provides technical infrastructure for implementing these structures through TK (Traditional Knowledge) Labels that communities can attach to digital materials specifying cultural protocols (Carroll et al., 2023).

Participatory Design and Community Co-Creation moves beyond mere consultation to genuine partnership in technology development. This involves communities defining use cases and priorities rather than researchers imposing predefined applications, for example communities might prioritize documentation of ritual languages over casual conversation, or developing children's learning games over academic linguistic analysis. Co-design workshops bring together community members, linguists, and developers to collaboratively design interfaces, workflows, and features, ensuring cultural appropriateness and usability. Community members serve as co-researchers participating in data collection, annotation, validation, and analysis, which not only improves quality through local expertise but builds capacity and ownership (Smith, 2023).

Technology transfer and capacity building ensure that communities are not merely sources of data but active agents in technology development and deployment. This includes training programs teaching community members to use documentation tools, annotation software, and basic AI concepts so they can make informed decisions about technology choices, employment of community members in technical roles as project staff, not just "informants", providing fair wages and building skills, infrastructure investment such as computers, recording equipment, and internet access that remain in community after project ends, and open-source tools ensuring communities are not dependent on proprietary software they cannot modify or maintain. The goal is sustainability where communities can continue documentation and tool development independently (Bird, 2022).

Addressing Power Imbalances requires explicit acknowledgment and mitigation of structural inequalities in researcher-community relationships. Conventional academic research models place researchers in positions of power as controllers of funding, framers of research questions, interpreters of data, and authors of publications that advance their careers while communities receive minimal benefit. Decolonizing approaches shift power dynamics by ensuring that community needs and priorities drive research agendas rather than academic interests or funding availability, intellectual property rights recognize community ownership and allow communities to control how research outputs are used and disseminated, authorship and credit include community members as co-authors on publications and presentations, acknowledging their expertise and contribution, and long-term relationships move beyond extractive one-time data collection to ongoing partnerships with mutual obligations and reciprocity (Cieri et al., 2024).

The principle of "nothing about us without us" is central to ethical AI development for indigenous languages. This means indigenous communities must be involved at all stages from initial project

conception and design, data collection and annotation decisions, technical development and algorithm choices, to evaluation and validation of outputs, dissemination and decisions about publication and sharing, and long-term governance and sustainability planning. External researchers and developers serve in supporting roles, providing technical expertise at community direction rather than controlling projects.

Framework for Community-Participatory AI Development

Based on synthesis of literature and case studies, this research proposes a comprehensive framework for community-participatory AI development in Indonesian regional language preservation. The framework consists of six interconnected phases, each with specific activities, stakeholders, and principles.

Phase 1: Community Engagement and Relationship Building establishes foundation of trust and mutual understanding essential for ethical collaboration. Activities include initial contact through appropriate community protocols (e.g., meeting with traditional leaders, explaining purposes through trusted intermediaries), community meetings to present project concepts in accessible language, gather community perspectives on language endangerment, document community needs and priorities, conduct participatory needs assessment identifying what communities want from documentation and technology (e.g., materials for teaching children, archives for cultural heritage, tools for language use in new domains), and establish governance structures defining roles and responsibilities, decision-making processes, and communication channels. Success indicators include genuine community buy-in demonstrated through active participation and leadership, established trust relationships and ongoing communication, and documented community priorities that shape subsequent phases.

Phase 2: Participatory Data Collection and Annotation builds the linguistic resources necessary for AI development while ensuring community control and benefit. Activities include co-design of data collection protocols determining what genres and contexts to document (oral histories, conversations, ceremonies, songs, etc.), what metadata to collect (speaker demographics, social context, cultural protocols), and how to ensure cultural sensitivity and safety, training and employment of community members as language documenters and annotators, providing technical training, equipment, and fair compensation, collaborative data collection where community members conduct recordings and initial transcription/annotation with linguist support, and cultural protocols integration through TK Labels and access restrictions for sacred or sensitive materials, explicit documentation of appropriate use and access conditions. Success indicators include community members leading data collection with external support rather than extraction by outsiders, culturally appropriate data representing community priorities, and clear governance over who can access and use data.

Phase 3: Collaborative AI Model Development involves technical work of building and training models using participatory approaches. Activities include selection of appropriate technologies based on available data, community needs, technical feasibility, and sustainability rather than pursuing most advanced AI simply because it exists, transparent explanation of how AI works providing non-technical explanations of algorithms, data requirements, capabilities and limitations to enable informed community decision-making, participatory training where community members contribute to model training through providing data, validating outputs, and iterative correction of errors, and local capacity building through training interested community members in basic programming, data science, and AI concepts. Success indicators include AI models that perform adequately for community-defined use cases even if not state-of-the-art, community understanding of AI capabilities and limitations enabling realistic expectations, and local technical capacity to adapt and maintain models.

Phase 4: Application Development and Testing creates usable tools that integrate AI models into applications serving community needs. Activities include co-design of user interfaces through workshops with diverse community members (elders, youth, women, men) to ensure applications are accessible and culturally appropriate, culturally appropriate design including visual design using local aesthetics, language using community terminology, and interaction patterns fitting cultural communication norms, piloting and iterative improvement through community testing of applications with feedback loops for rapid improvement, and accessibility ensuring applications work with available technology (smartphones rather than high-end computers, offline functionality for limited connectivity, audio interfaces for low-literacy users). Success indicators include applications that community members actually use and find valuable, positive community feedback on usability and cultural appropriateness, and evidence of impact on language learning, documentation, or use.

Phase 5: Community Ownership and Sustainability ensures that AI tools and data remain under community control and can be maintained long-term. Activities include establishing ownership structures through data trusts, community cooperatives, or other legal mechanisms that clarify community ownership, capacity and infrastructure ensuring communities have equipment, internet access, and skills to maintain tools independently, resource mobilization supporting communities to seek ongoing funding, generate income from language products, or integrate costs into community budgets, and integration with community institutions embedding language technology in schools, community centers, cultural organizations for long-term sustainability. Success indicators include clear legal documentation of community ownership, functional infrastructure and capacity for independent maintenance, and ongoing use and development of tools beyond project end.

Phase 6: Monitoring, Evaluation, and Adaptive Management provides accountability and learning. Activities include community-defined success metrics establishing how communities themselves will judge success (e.g., number of children learning language, elders satisfied with documentation, language use in new domains) rather than only academic metrics (e.g., model accuracy, publications), participatory evaluation where community members participate in data collection and interpretation for evaluation, transparent reporting of both successes and failures with findings shared with community in accessible formats, and adaptive management using evaluation findings to adjust approaches, acknowledging that initial plans may need modification based on experience. Success indicators include evaluation demonstrates positive impacts on community-defined outcomes, community satisfaction with processes and outcomes, and documented learning that can inform future projects.

Cross-Cutting Principles apply throughout all phases including respect for indigenous data sovereignty and cultural protocols, genuine power-sharing with community leadership and decision-making, transparency in processes, data use, and project finances, reciprocity ensuring benefits flow to communities not just extraction of resources, sustainability planning from outset not as afterthought, and cultural safety ensuring no harm to community wellbeing, relationships, or sacred knowledge.

This framework requires significant time, resources, and commitment, operating on timescales of years rather than months, and demanding meaningful investment in relationship-building and capacity development, not just technical outputs. However, the alternative—rapid extraction of data for academic publications or commercial applications without community benefit or control—is ethically unacceptable and ultimately unsustainable. The framework positions AI as a tool in service of community self-determination and language revitalization, not an end in itself.

CONCLUSION AND RECOMMENDATIONS

Conclusion

This systematic literature review of 68 publications from 2020-2025 reveals that Artificial Intelligence and Machine Learning technologies offer unprecedented capabilities for preserving

Indonesian regional languages and their associated cultural linguistic heritage, yet these powerful tools can only fulfill their potential when developed and deployed through genuine partnership with indigenous language communities, respect for data sovereignty, and attention to cultural and ethical dimensions alongside technical innovation. The research demonstrates that AI applications across multiple domains—Automatic Speech Recognition for transcribing oral traditions, Machine Translation for language learning and accessibility, Optical Character Recognition for digitizing manuscripts, Text-to-Speech for voice preservation, and Natural Language Processing for computational analysis—have achieved significant advances in the past five years driven by deep learning and self-supervised learning techniques that dramatically reduce data requirements, making them increasingly viable for low-resource endangered languages.

However, application to Indonesian regional languages faces formidable challenges including extreme data scarcity as most of the 700+ regional languages lack digital corpora necessary for training AI models, linguistic diversity in morphological complexity, voice systems, speech levels, and dialectal variation that defies one-size-fits-all technical approaches, infrastructure constraints limiting access to internet, electricity, and devices in remote areas where many endangered languages are concentrated, and ethical imperatives to respect indigenous data sovereignty and avoid perpetuating colonial extraction of community resources for external benefit. The typological features of Indonesian regional languages, particularly morphological complexity, voice systems, speech level variation, and orthographic diversity, create specific technical challenges that require adaptation of standard NLP methods developed primarily for English and European languages.

Case studies from Papua, Sulawesi, and Bali demonstrate that successful AI application requires community-participatory approaches where indigenous peoples are not merely data sources but co-designers, co-developers, and governors of technology, projects prioritize community-defined needs and use cases rather than imposing external priorities, technology transfer builds local capacity to maintain and adapt tools rather than creating dependency on external experts, and data governance ensures community ownership and control with culturally appropriate access restrictions and benefit-sharing. The proposed framework for community-participatory AI development provides a structured approach across six phases from community engagement through sustainability planning, grounded in principles of indigenous data sovereignty, genuine power-sharing, transparency, reciprocity, and cultural safety.

The research concludes that while AI presents powerful technical capabilities for language documentation and revitalization, it is not a silver bullet that can single-handedly solve language endangerment, which is fundamentally driven by socioeconomic and political forces. AI must be integrated within broader language revitalization strategies that address intergenerational transmission through immersion education and family language planning, domain expansion creating new functional contexts for language use in government, commerce, and media, status planning that elevates prestige and reverses stigmatization of regional languages, and structural changes addressing economic marginalization and political disempowerment of indigenous communities. Technology can support but never replace the human relationships, community commitment, and political will necessary for language survival.

Looking forward, the future of AI in Indonesian language preservation depends on several critical factors including sustainable funding as current projects rely heavily on short-term external grants, requiring long-term commitment from Indonesian government and institutions, ethical frameworks and policies establishing clear guidelines for data governance, community consent, and benefit-sharing in language technology development, technical innovation specifically adapted to Indonesian linguistic diversity rather than uncritically importing methods from high-resource languages, and capacity building creating a cohort of indigenous language technology specialists who can lead future development. The stakes could not be higher as within the next 50 years, without urgent intervention, hundreds of Indonesian languages may fall silent forever, taking with them irreplaceable knowledge, cultural heritage, and human cognitive diversity. AI offers tools that can help prevent this loss, but

only if wielded with wisdom, humility, and respect for the communities whose languages embody their identity, history, and connection to ancestral lands.

Recommendations

Based on the research findings, the following recommendations are proposed for multiple stakeholders:

For Indonesian Government (Ministry of Education, Culture, Research, and Technology; BRIN; Badan Bahasa):

Establish a National AI-Powered Language Archive as a centralized but community-governed platform for Indonesian regional language data, tools, and applications, with technical infrastructure for hosting audio, video, text, and annotation data with robust security and access control, community data portals allowing language communities to manage their own data with tiered access permissions, repository of open-source language tools adapted for Indonesian languages, and funding mechanism for community-led documentation projects. Invest in NLP Research Infrastructure specifically for Indonesian regional languages including competitive grants for developing NLP tools for low-resource Indonesian languages with requirements for open-source release and community benefit, establishment of shared annotated corpora for at least the 50 most spoken regional languages following ethical data collection practices, compute resources and technical support for universities and communities lacking infrastructure, and partnerships with international research institutions to leverage global expertise while ensuring Indonesian control and benefit.

Develop National Policy Framework for Language Data Governance that balances open science principles with indigenous data sovereignty through legal frameworks clarifying community rights to language data and limiting extractive use, ethical guidelines for language technology research requiring community consent and benefit-sharing, integration with broader intellectual property and traditional knowledge protection laws, and enforcement mechanisms to prevent violations with meaningful penalties. Integrate Language Technology into Education by supporting development of educational materials and applications in regional languages using AI, funding teacher training in using language technology for multilingual education, piloting programs in regions with strong regional language use, and evaluating impact on language maintenance and academic achievement. Support Capacity Building through establishing graduate programs in computational linguistics and language technology at Indonesian universities with focus on regional languages, scholarship programs for indigenous students to study language technology, short courses and workshops for community members on language documentation and AI tools, and exchange programs connecting Indonesian researchers with international language technology centers.

For Academic Institutions (Universities, Research Centers):

Prioritize Community-Engaged Research by adopting participatory research methodologies as standard practice rather than extractive data collection, ensuring indigenous authorship and intellectual property rights in publications, making research outputs accessible to communities in appropriate formats and languages, and building long-term relationships and reciprocal obligations rather than one-time studies. Develop Open-Source Tools specifically designed for Indonesian language contexts including morphological analyzers for agglutinative languages, ASR systems trained on diverse phonological systems including tonal and voice quality features, OCR for traditional scripts with active learning interfaces for community correction, and language learning applications with culturally appropriate pedagogy and content.

Establish Cross-Disciplinary Collaborations bringing together linguists, computer scientists, anthropologists, educators, and community partners, creating spaces for meaningful dialogue across disciplines and epistemologies, and funding mechanisms that support long-term collaborative research

beyond typical grant cycles. Contribute to Open Data Ecosystems by sharing annotated corpora and trained models through open repositories with appropriate licenses, documenting methodologies and tools thoroughly to enable replication and adaptation, and contributing to international efforts like Universal Dependencies for Indonesian languages.

For Language Communities and Indigenous Organizations:

Assert Data Sovereignty by developing community data governance policies clarifying ownership, access, and use conditions, establishing legal structures such as data trusts to hold community data, and using TK Labels and other tools to communicate cultural protocols to researchers and technologists. Build Local Capacity through identifying interested community members to receive training in language technology, collaborating with universities and organizations to provide training opportunities, establishing community language technology centers with equipment and internet access, and mentoring youth in both language and technology skills.

Engage Strategically with External Researchers by defining community priorities and use cases before engaging with researchers, negotiating equitable partnerships with clear roles, responsibilities, and benefit-sharing, maintaining community control over data and decision-making, and evaluating projects regularly to ensure alignment with community goals. Network with Other Communities to share experiences, challenges, and successes in language technology, coordinate advocacy for resources and supportive policies, develop collective bargaining power in negotiations with researchers and funders, and build solidarity across indigenous language movements.

For International Organizations (UNESCO, Endangered Languages Project, etc.):

Provide Funding that prioritizes community-led initiatives with flexible funding mechanisms accommodating community timescales and processes, long-term commitment rather than only short-term projects, funding for sustainability including infrastructure and capacity, and willingness to fund relationship-building and governance alongside technical work. Develop Ethical Guidelines and Standards for AI in language preservation that center indigenous data sovereignty and rights, establish minimum standards for community consent and benefit-sharing, provide frameworks for evaluating ethical compliance, and create accountability mechanisms for violations.

Facilitate Knowledge Sharing and Technical Support through convening international workshops and conferences bringing together communities, researchers, and developers, maintaining repositories of best practices, case studies, and lessons learned, providing technical assistance and mentorship for community-led projects, and supporting South-South exchange and learning. Advocate for Supportive Policies at national and international levels recognizing language rights and cultural heritage protection, supporting indigenous data sovereignty in broader data governance frameworks, encouraging governments to invest in language preservation including technology, and raising awareness of language endangerment and importance of diversity.

For Technology Companies (Google, Meta, Microsoft, etc.):

Invest in Low-Resource Language Technologies by allocating research resources to low-resource and endangered languages beyond commercially viable markets, making tools and models available free and open-source for endangered languages, providing compute resources and technical support for community projects, and prioritizing ethical and community-beneficial applications over profit extraction. Ensure Ethical AI Development through obtaining genuine community consent with full transparency about data use, ensuring fair benefit-sharing when community data contributes to commercial products, respecting access restrictions and cultural protocols in data use, and subjecting language projects to rigorous ethical review including community voices.

Support Capacity Building in indigenous communities through training programs, internships, and employment pathways in language technology for indigenous people, funding for community-led technology development, partnerships with universities and NGOs supporting ethical language tech, and mentorship from company researchers for community developers. Contribute to Digital Language Diversity by ensuring products and services support diverse languages beyond commercial giants, developing Unicode support and fonts for traditional scripts, creating interfaces and functionalities accommodating diverse linguistic structures, and avoiding homogenization toward English or dominant languages.

Implementation of these recommendations requires sustained commitment, adequate resources, and genuine political will from all stakeholders. However, the alternative—continued language loss—is unacceptable both as cultural tragedy and loss to human knowledge and diversity. AI provides powerful tools that, used ethically and in partnership with communities, can contribute to reversing language endangerment and ensuring that Indonesia's linguistic heritage thrives in the digital age as a living expression of cultural identity, traditional knowledge, and connection to place. The time to act is now, before more languages fall silent and the wisdom they carry is lost forever.

REFERENCES

Alwasilah, A. C. (2012). *Pokoknya Rekayasa Literasi*. Bandung: Kiblat Buku Utama.

Banks, J. A. (2008). *An Introduction to Multicultural Education* (4th ed.). Boston: Pearson Education.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101.

Creswell, J. W. (2014). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (4th ed.). Thousand Oaks: SAGE Publications.

Emilia, E. (2011). *Pendekatan Genre-Based dalam Pengajaran Bahasa Inggris: Petunjuk untuk Guru*. Bandung: Rizqi Press.

Hassan, A. (2016). The philosophy of Malay culinary heritage: A cultural identity perspective. *Journal of Southeast Asian Studies*, 21(3), 145-162.

Healey, M., & Jenkins, A. (2000). Kolb's experiential learning theory and its application in geography in higher education. *Journal of Geography*, 99(5), 185-195.

Hirsch, E. D. (1987). *Cultural Literacy: What Every American Needs to Know*. Boston: Houghton Mifflin.

Ibrahim, R. (2017). Rendang: A culinary symbol of Minangkabau tradition and identity. *Asian Food Studies*, 12(2), 88-104.

Ihsan, M. B., & Iftanti, E. (2022). Analisis butir soal pilihan ganda penilaian akhir semester tahun pelajaran 2020/2021 mata pelajaran Bahasa Indonesia kelas XI di SMAN 1 Boyolangu. Retrieved from http://repo.uinsatu.ac.id/30271/

Ihsan, M. B., & Maryani, S. (2025). Integrasi Teori Hipotesis Input Komprehensibel Stephen Krashen dalam perancangan kurikulum pembelajaran Bahasa Indonesia pada pondok pesantren modern berbasis teknologi pendidikan. *Jurnal Pendidikan dan Pembelajaran*, 4(02).

Ihsan, M. B., Ismawati, E., & Tukiyo. (2025). The relationship between animated caricature media use and vocabulary mastery and anecdotal text ability. *Indonesian Journal of Advanced Research (IJAR)*, 4(11), 2451-2466.

Ihsan, M. B., & Artika, I. W. (2025). Integrasi mamangan adat Minangkabau dalam pembelajaran Bahasa Indonesia berbasis teori konstruktivisme sosial Vygotsky untuk pengembangan literasi kritis. *Journal of Social Humanities and Education*, 1(2), 212-220.

Ihsan, M. B., Maryani, S., & Rosmasta, H. (2025). Pengembangan model evaluasi responsif Stake dalam asesmen pembelajaran Bahasa Indonesia berbasis konteks budaya lokal dan kebutuhan stakeholder pendidikan. *Journal of Humanities and Education Studies*, 243-254.

Ihsan, M. B. (2025). *Rahasia Hutan Bakau (Teka Teki Rimbe Bakau)*. Pekanbaru: Balai Bahasa Provinsi Riau.

Kemendikbud. (2016). *Silabus Mata Pelajaran Bahasa Indonesia SMP/MTs*. Jakarta: Kementerian Pendidikan dan Kebudayaan.

Knapp, P., & Watkins, M. (2005). *Genre, Text, Grammar: Technologies for Teaching and Assessing Writing*. Sydney: University of New South Wales Press.

Kohonen, V. (1992). Experiential language learning: Second language learning as cooperative learner education. In D. Nunan (Ed.), *Collaborative Language Learning and Teaching* (pp. 14-39). Cambridge: Cambridge University Press.

Kolb, D. A. (1984). *Experiential Learning: Experience as the Source of Learning and Development*. Englewood Cliffs: Prentice Hall.

Kramsch, C. (1998). *Language and Culture*. Oxford: Oxford University Press.

Mahsun. (2014). *Teks dalam Pembelajaran Bahasa Indonesia Kurikulum 2013*. Jakarta: Rajawali Press.

Moon, J. A. (2004). *A Handbook of Reflective and Experiential Learning: Theory and Practice*. London: RoutledgeFalmer.

Muyassaroh, M., & Ihsan, M. B. (2021). Penggunaan bahasa persuasi dalam iklan layanan masyarakat untuk menyosialisasikan kehidupan baru pada era pandemi Covid-19 di Kabupaten Tulungagung. *Kongres Internasional Masyarakat Linguistik Indonesia*, 227-233.

Sardiman. (2016). Pembelajaran berbasis budaya lokal sebagai upaya peningkatan kualitas pembelajaran bahasa Indonesia. *Jurnal Pendidikan Bahasa dan Sastra*, 16(1), 1-13.

Sudaryanto, T. (2015). *Etnopedagogi Sunda: Revitalisasi Nilai-Nilai Budaya dalam Pembelajaran*. Bandung: Universitas Pendidikan Indonesia Press.

Supratman, E. (2018). Traditional Malay cuisine as cultural capital: Preservation and innovation. *Indonesian Journal of Cultural Studies*, 5(1), 34-52.

Syarifuddin, D. (2015). *Budaya Melayu Riau: Filosofi dan Implementasi dalam Kehidupan Masyarakat*. Pekanbaru: Unri Press.

Tilaar, H. A. R. (2015). *Pedagogik Teoritis untuk Indonesia*. Jakarta: Kompas Media Nusantara.

Wagiran. (2012). Pengembangan karakter berbasis kearifan lokal hamemayu hayuning bawana. *Jurnal Pendidikan Karakter*, 2(3), 329-339.