Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15507
e-ISSN : 2541-2019
p-ISSN : 2541-044X

IndoBERT-Based Pediatric Disease Classification and Symptom-Based Traditional Medicine Recommendation from Lontar Usada Rare

I Putu Erick Prawira Winata[1)], I Gede Iwan Sudipa[2)*], Ni Putu Suci Meinarni [3)], Dewa Ayu Putri Wulandari[4)], Christina Purnama Yanti[5)]
[1)2)*3)4)5)]Fakultas Teknologi dan Informatika, Program Studi Teknik Informatika, Institut Bisnis dan Teknologi Indonesia, Bali, Indonesia
[1)]prawirawinata123@gmail.com, [2)*] iwansudipa@instiki.ac.id, [3]sucimeinarni@instiki.ac.id, [4)]putri.wulandari@instiki.ac.id, [5)]christinapy@instiki.ac.id

Submitted :Oct 29, 2025 | **Accepted** : Dec 19, 2025 | **Published** : Jan 04, 2026

Abstract: This study aims to develop a Balinese traditional text-based pediatric disease classification model using a fine-tuned IndoBERT model on the Lontar Usada Rare dataset. The dataset used consists of 422 entries containing disease symptoms, disease types, medicinal ingredients, and treatment procedures obtained from transliteration of lontar manuscripts and interviews with traditional medicine experts. Pre-processing was done through case folding, cleansing, and normalization, followed by label encoding on 35 disease classes. The IndoBERT model was fine-tuned using the AdamW optimizer with a learning rate of 5e-5, batch size 8, and 15 epochs. Evaluation results showed the model was able to achieve 90.59% accuracy, 94.71% precision, 90.59% recall, and 90.99% F1-score, indicating excellent performance in understanding the linguistic context of traditional medical text. The developed recommendation system integrates model prediction with TF-IDF-based cosine similarity method to provide the most relevant treatment recommendations based on user symptom input. This research makes an important contribution to the digitization and preservation of Balinese traditional medical knowledge through the development of a structured and widely accessible digital knowledge base.

Keywords: IndoBERT Method, Lontar Usada Rare, pediatric disease classification, traditional Balinese medicine

INTRODUCTION

Bali is renowned for its rich cultural heritage, including local knowledge transmitted through generations via ancient palm leaf manuscripts known as lontar. Lontar Usada Rare contains detailed descriptions of pediatric diseases, herbal formulations, and healing methods grounded in local wisdom and ethnomedicinal practices. Various medicinal plants mentioned in the text have been validated to possess biological properties that ameliorate inflammation and enhance immune function (Widhiantara et al., 2024). Despite its cultural and medical significance, the preservation of lontar manuscripts faces substantial challenges, including physical deterioration and limited digitalization efforts. Recent research demonstrates that the majority of lontar manuscripts are in damaged or severely deteriorated conditions, with many characters lost or rendering automated digitization highly challenging (Siahaan et al., 2022). Government commitment has been formalized through Law No. 5 of 2017, Health Ministry Regulation No. 37 of 2017, and Balinese Provincial Regulation No. 55 of 2019, Lontar Usada Rare remains accessible only at limited institutions such as Gedong Kirtya Singaraja. Community-based digitalization projects such as WikiLontar 2021 have rescued 606 lontar manuscripts from Balinese community collections from more severe damages, raising awareness of the importance of this fragile cultural heritage to Balinese literary and cultural life (Ayu et al., 2024).

As information technology advances, the digitalization of cultural heritage presents unprecedented opportunities for education, research, and healthcare applications. Transformer-based natural language processing models, particularly IndoBERT (Indonesian Bidirectional Encoder Representations from Transformers), have been specifically designed for Indonesian language with bidirectional architecture enabling contextual understanding from both directions of sentences (Koto & Baldwin, 2020). Previous research has demonstrated the effectiveness of such models across diverse Indonesian NLP tasks, achieving state-of-the-art performance in sentiment analysis, emotion classification, and language understanding benchmarks with F1-scores exceeding 90%

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

through hybrid architectural approaches (Ahmadian et al., 2024). Furthermore, NLP applications in modern medical texts have proven effective for information extraction and automated disease classification (Wang et al., 2020) (Liu et al., 2020).

While transformer-based models have demonstrated efficacy across various Indonesian NLP tasks and modern medical text processing, their application to traditional medical texts remains severely limited. Low-resource languages and cultural heritage documents present unique computational challenges that differ fundamentally from high-resource scenarios (Pakray, 2025). According to recent research, low-resource languages often lack the extensive and standardized datasets required for effective training, with critical NLP tools such as part-of-speech taggers and annotated datasets being frequently non-standardized or entirely absent (N. Wang, 2025). Furthermore, ancient texts exhibit archaic grammatical structures and vocabulary incompatible with modern linguistic paradigms, requiring specialized models for processing (Wei, 2025). The digitization of cultural heritage through transformer-based models faces significant hurdles including diverse handwriting styles, deteriorated manuscript conditions, multilingual content, and archaic or non-standard orthography (Romein & Ströbel, 2025). Previous studies have concentrated on contemporary text types such as social media, product reviews, and electronic health records (EMRs), yet none have explored cultural heritage documents such as Lontar Usada Rare. Ancient medical texts present unique and underexplored challenges: grammatical and linguistic expression differences between classical and modern language varieties, specialized and culturally-specific medical terminology, and overlapping symptom descriptions (Hou et al., 2025). Research on traditional medical texts from other Asian cultures demonstrates that semi-structured or unstructured corpus formats introduce additional complexity in digitizing local medical knowledge. Consequently, a significant research gap exists in developing automated systems capable of interpreting traditional Balinese medical knowledge and rendering it digitally accessible for educational and research purposes.

This study addresses this gap by providing the first transformer-based computational model for pediatric disease classification derived from Lontar Usada Rare, filling the absence of NLP models for Indonesian traditional medical texts. Specific contributions include: (1) the first application of fine-tuned IndoBERT for automated pediatric disease classification from Lontar Usada Rare; (2) construction of a structured and annotated knowledge base containing 422 entries extracted from the lontar manuscript; (3) development of a hybrid approach combining TF-IDF and IndoBERT embeddings for robust symptom feature extraction; and (4) implementation of a classification system that matches complex symptom presentations to diseases using cosine similarity based on semantic representations. Through this approach, this study provides the first computational solution for interpreting Indonesian traditional medical texts, supporting digital preservation, knowledge dissemination, and practical application of Balinese local wisdom in the era of healthcare digitalization (Liu et al., 2020) (Hou et al., 2025).

The specific objective is to develop a fine-tuned IndoBERT model capable of classifying 35 pediatric disease categories from Lontar Usada Rare with high accuracy, and to validate the model's performance in recommending traditional treatments based on input symptom presentations.

LITERATURE REVIEW

IndoBERT is a pre-trained model based on BERT architecture specifically developed for Indonesian language using the Indo4B corpus containing more than 4 billion words (Koto & Baldwin, 2020). The bidirectional architecture of IndoBERT enables understanding of word context from both directions in a sentence, which is a fundamental advantage in processing medical texts where a single word can have different meanings depending on context (Rasmy, 2021). Research by Koto et al. (2020) demonstrated that IndoBERT outperformed multilingual models such as mBERT and MalayBERT with state-of-the-art performance across multiple Indonesian natural language understanding tasks spanning morpho-syntax, semantics, and discourse.

The effectiveness of IndoBERT has been proven across multiple domains. Koto et al. (2020) demonstrated that IndoBERT achieves 96.52% accuracy for POS tagging tasks, outperforming multilingual models on morpho-syntactic tasks. The bidirectional contextual understanding inherent in transformer-based architectures has proven especially valuable for medical text processing, where semantic discrimination depends heavily on surrounding context (Rasmy, 2021). These models have been successfully adapted across diverse Indonesian NLP applications including sentiment analysis, named entity recognition, and document classification tasks through domain-specific fine-tuning strategies.

Although IndoBERT has proven effective on modern texts, a significant research gap exists in its application to traditional texts and cultural heritage documents (Ginzel & Girsang, 2026; Kątek et al., 2025). Recent advances in transformer-based models for biomedical applications demonstrate their potential for specialized domain adaptation. (Li et al., 2020) showed that transformer architectures, particularly BERT-based models, excel in capturing contextual relationships in medical texts through self-attention mechanisms that process information bidirectionally.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15507
e-ISSN : 2541-2019
p-ISSN : 2541-044X

However, the application of transformer models to ancient medical manuscripts presents distinct challenges. As documented in recent heritage digitization studies, historical documents require sophisticated algorithms that combine expertise from digital humanities, history, linguistics, and computer science (Romein & Ströbel, 2025). Ancient Tamil inscription recognition research demonstrates that hybrid approaches combining transformer architectures with preprocessing techniques (such as GAN-based augmentation and perspective correction) significantly improve recognition accuracy in degraded, low-resource historical texts (Murugan & Visalakshi, 2025).

Ancient medical texts introduce additional complexity that remains unresolved: grammatical and linguistic expression differences between classical and modern language varieties, archaic and culturally-specific medical terminology, and semantically overlapping symptom descriptions (Li et al., 2020).

Research on processing traditional medical texts in other Asian contexts reveals that corpus with semi-structured or unstructured formats introduces additional complexity in digitizing local medical knowledge (Hou et al., 2025). Specifically, Indonesian-language medical NLP research remains severely limited, with most NLP applications focused on high-resource modern texts rather than cultural heritage documents (Pakray, 2025). To date, lontar digitalization efforts have concentrated on physical preservation and transliteration, yet have not integrated AI technologies for automated extraction and classification of traditional medical knowledge.

To address the challenges of low-resource domains in traditional medical texts, recent literature identifies several effective strategies. First, transfer learning through fine-tuning pre-trained models has proven to be the most effective approach for low-resource domains, enabling models to adapt to the unique linguistic characteristics of traditional texts without requiring large datasets (Pakray, 2025). Second, integration of traditional features such as TF-IDF with deep learning embeddings creates a hybrid system that combines the power of deep learning in capturing complex patterns with the transparency and interpretability of traditional similarity methods (Li et al., 2020) (Sogandi, 2024).

Third, class imbalance mitigation strategies through stratified sampling and weighted loss functions have proven effective in maintaining proportional distribution across classes in small datasets, preventing model bias toward majority classes (Kim & Id, 2022). Fourth, consultation with domain experts (in this case, Balinese traditional medicine practitioners) for manual validation and cultural interpretation ensures that disease clustering and category labels align with actual ethnomedicinal practice, rather than purely algorithmic grouping (Ginzel & Girsang, 2026).

Research on traditional medical texts is not merely a technical challenge but also a critical effort in cultural heritage preservation during the digitalization era (Pakray, 2025)**.** By applying fine-tuned IndoBERT to Lontar Usada Rare texts for pediatric disease classification and traditional treatment recommendations, this research addresses a significant gap in developing automated systems capable of interpreting Balinese traditional medical knowledge and rendering it digitally accessible for educational and ethnomedicinal research purposes.

METHOD

This study was organized systematically to produce a classification model of pediatric diseases based on Lontar Usada Rare. The data came from interviews with herbal medicine experts and texts from Gedung Kirtya Singaraja that have been translated into Indonesian, resulting in 422 data covering symptoms of diseases, types of diseases, herbal medicines, and treatment procedures.

Methodological design of this research addresses specific challenges of traditional medical text processing: (1) linguistic domain shift from archaic Balinese language in lontar to modern Indonesian used for model pre-training, requiring specialized preprocessing and careful fine-tuning strategy; (2) class imbalance inherent in small datasets where rare diseases have limited documentation, requiring stratified sampling and weighted loss functions; (3) transliteration bias from converting ancient manuscript to digital text, addressed through expert validation; and (4) preservation of cultural context where manual expert consultation ensures disease grouping aligns with traditional ethnomedicinal practice, not purely algorithmic clustering (Ginzel & Girsang, 2026)(Park et al., 2025). These challenges necessitate a hybrid approach combining deep learning model adaptation with transparent evaluation and cultural domain expertise.

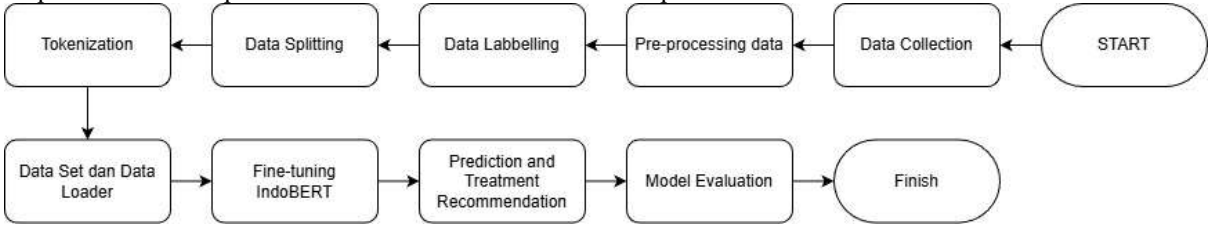

Fig. 1 Research flow

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15507
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Based on the research pipeline above, the methodology systematically addresses each stage with explicit consideration of threats to validity. The research integrates data quality assurance (expert validation, inter-rater reliability), domain adaptation strategies (custom preprocessing, low learning rate fine-tuning), statistical rigor (stratified sampling, weighted loss functions), and transparent evaluation (per-class metrics, confusion matrix analysis). This comprehensive approach ensures that the model not only achieves high accuracy but also maintains validity and reliability when processing culturally-specific ancient medical texts with inherent class imbalance and linguistic domain shift challenges.

Data Collection

Primary data was collected through interviews with Dr. Nyoman Sridana, S.Kes.H., M.Si., owner of PT Vision Bali Usada Taru Pramana Herbal Production. Secondary data in the form of transliterated and translated Lontar Usada Rare was collected from Gedong Kirtya Singaraja.

Table 1. Lontar Usada Rare dataset

Lontar Usada Rare data
Jika ada bayi yang sakit, jika terlihat pada putih matanya biru, tangan dan kakinya dingin, tangisnya agak serak dan merdupkan mata, sakit bayi ini, penyakitnya pada tulang ubun-ubun renggang disebut belahan. bahan obatnya sembung kedis, ambil akarnya, beras merah, lempuyang, mesui, Semua lumatkan. tempelkan pada ubun-ubunnya
Jika seperti agak putih matanya si bayi, serta uratnya seperti berdarah, tenaganya tidak stabil, bibirnya keropos, panas badan si bayi itu, penyakitnya disebut guaman. Sarana : biji mentimun, buah paspasan, subatah enau, wong papah idung, semua dikuskus, lumatkan pada mulutnya.
Jika seperti ada gerakan pada tubuh menyebar, sikap kaki dan tangan ditelungkupkan , mata kemerahan , disebut tiwang penyu. Sarana obatnya : babakan ceremai, pala, menyan, lungid, sinrong, gegambiran, arak, lulurkan

Meta data

The total dataset consisted of 422 entries representing traditional treatment cases for pediatric illnesses. Each entry includes a full description of the symptoms of the disease, the name of the disease according to lontar terminology, the herbal medicine ingredients used, and the treatment application procedure. The data was then organized into a structured metadata structure with five main columns to facilitate the classification process: (1) Disease Symptoms; (2) Disease Type; (3) Disease Type Class; (4) Medicinal Ingredients; and (5) Treatment Procedure. This metadata structure is designed to facilitate the information extraction and automatic classification process. Table 2 shows an example of the metadata structure used in the study.

Table 2. Metadata of the Lontar Usada Rare Dataset

No	**Disease Symptom**	**Disease Type**	**Class of Disease Type**	**Medicine Materials**	**Treatment Procedure**
1.	Putih matanya biru, tangan dan kakinya dingin, tangisnya agak serak dan merdupkan mata	Penyakitnya pada tulang ubun-ubun renggang disebut belahan.	Belahan	Bahan obatnya sembung kedis, ambil akarnya, beras merah, lempuyang, mesui	Semua lumatkan. tempelkan pada ubun-ubunnya
2.	Agak putih matanya si bayi, serta uratnya seperti berdarah, tenaganya tidak stabil, bibirnya keropos, panas badan si bayi itu	Penyakitnya disebut guaman	Guaman	Biji mentimun, buah paspasan, subatah enau, wong papah idung	Semua dikuskus, lumatkan pada mulutnya
3.	Agak kemerahan tubuh sang bayi	Disebut tiwang brahma	Tiwang	Daun dapdap yang muda, empol ending merah, triketuka	Lulurkan
4.	Hidung bayi tersumbat dan berair	Pilek	Gangguan pernafas	Padang lepas, daging kemiri, pulasai, lengkuas 3 iris	Sembar pada taneng, tengah-tengah dada, ujung tulang rusuk
.....					
100.	Muncul bercak merah pada tubuh si bayi	Gatal	Gatal-gatal	Daun dapdap wong yang sedang, kapur tohor	Dilulurkan

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15507
e-ISSN : 2541-2019
p-ISSN : 2541-044X

101.	Telinga bayi terlihat seperti mengeluarkan air	Curek	Curek	Sirih lanang, minyak wijen	Dipipis lalu ditiupkan pada telinganya
102.	Bayi batuk-batuk dan suhu badan panas dingin	Demam	Panas dingin	Daun kemuning, sulasih harum, kencur, temu tis, kelapa bakar	Disembar
.....					
422.	Jika matanya terlihat biru dan putih, tangan serta kaki terasa dingin, menangis dengan suara agak serak, serta tubuh meredup dan melemas, dan terdapat belahan di ubun-ubun.	Belahan pada ubun ubun	Belahan	Sembung kedis lulurnya, beras merah, lempuyang, mesui, semua lumatkan.	Di tempel pada ubun ubunnya.

Data pre-processing

Pre-processing is an important stage in Natural Language Processing to prepare raw text data to be ready for processing by machine learning models (Y. H. Id et al., 2020). In this research, pre-processing is performed on the Disease Symptoms column to clean and standardize the text before modeling. This process includes three main stages that are performed sequentially.

The first stage is Case Folding, which converts all characters into lowercase letters to avoid duplication of words with different capitalization, such as "White", "WHITE", and "white" which are then considered the same (Y. H. Id et al., 2020)**.** The second stage is Cleansing, which removes irrelevant elements such as punctuation marks, numbers, and symbols so that the text focuses on meaningful words, for example the text "White has blue eyes, cold hands and feet!" becomes "White has blue eyes, cold hands and feet" (Y. H. Id et al., 2020). The third stage is Normalization, which homogenizes non-standard words into standard forms, for example "serek" becomes "serak" or "kropos" becomes "keropos" (Y. H. Id et al., 2020)**.**

This process uses a custom dictionary tailored to the lontar text and traditional Balinese medical terminology so that words with the same meaning are recognized as identical entities by the model. Figure 2 shows the results of each pre-processing stage.

Data Labelling

The clustering process was performed manually through consultation with Balinese traditional medicine experts to ensure medical validity and compatibility with traditional knowledge. This approach is important because Lontar Usada Rare contains many disease types with unbalanced distribution. For example, various Tiwang disease variants such as "Tiwang Brahma", "Tiwang Bangke", and "Tiwang Lutung" were grouped into a single "Tiwang" class because they share the characteristic of loss of consciousness in children (Widhiantara et al., 2024). The result is 35 disease classes that are more balanced and structured.

Although manual clustering resulted in more balanced distribution compared to original data, further analysis of class distribution still revealed moderate imbalance where some classes had limited samples (e.g., Belahan 7 samples, Gatal-gatal 7 samples, while Tiwang had 52 samples). This class imbalance is an inherent characteristic of ancient cultural heritage data reflecting the frequency of pediatric diseases in Balinese traditional ethnomedicinal practice (Kim & Id, 2022). To address potential model bias toward majority classes, this research applied stratified sampling in training-test data division and employed weighted loss functions during model fine-tuning to impose greater penalty on misclassification in minority classes (Fernando & Tsokos, 2022).

Table 3. Labelling of Disease Types

Class of Disease Type	Encoded Label	Number of Samples
Barah	0	9
Tiwang	22	52
Jampi	4	29
Upas	23	14
Belahan	25	7

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15507
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Gatal-gatal	26	7
Batuk	24	7
Curek	1	6
Diare	2	26
Gangguan pernafasan	3	13
Kurang gizi	5	3
Menangis terus menerus	6	8
Mimisan	7	6
Mual-mual	8	10
Muntah	9	11
Panas	10	23
Panas dalam	11	33
Pejen	12	15
Penguci bolong	13	3
Perut kembung	14	27
Perut melilit	15	5
Perut kaku	16	8
Sakit mata	17	9
Sakit perut	18	9
Sebaha	19	15
Sembelit	20	8
Siksik	21	10
Guaman	27	9
Inja	28	13
Panas dingin	29	8
Sakit kemaluan	30	8
Sakit kuning	31	2
Sakit telinga	32	2
Sula	33	3
Tubuh kemerahan	34	2

Furthermore, Label Encoding was performed using the LabelEncoder library from scikit-learn to convert disease class names from text format to numeric values (Li et al., 2020). Each class was assigned a unique label from 0 to 34 for processing by machine learning algorithms. For example, "Belahan" was labeled 0, "Guaman" 1, and "Tiwang" 2. This process preserves the meaning of the category while adapting it to numeric format understood by the model.

Data Splitting

Data splitting was performed using the train_test_split function from the scikit-learn library with an 80:20 ratio, which is standard for medium-sized datasets. From a total of 422 data, 337 training data (80%) and 85 testing data (20%) were obtained. The stratify parameter was applied to the disease class label column to maintain the proportional distribution of each class in both subsets, preventing bias in minority classes. Stratified sampling is particularly critical in medical classification tasks with class imbalance, as it ensures that both training and test sets represent the original distribution characteristics. Recent research on imbalanced medical datasets demonstrates that stratified sampling combined with weighted loss functions significantly improves model performance on minority classes by ensuring proportional representation across all disease categories (Florida & Tsokos, 2021). Furthermore, for small datasets with inherent class imbalance characteristics similar to cultural heritage corpora, stratified approaches prevent the model from being dominated by majority classes during gradient updates, enabling effective learning even from rare disease categories with minimal samples (Kim & Id, 2022). For example, if the "Tiwang" class represents 10% of total data, this proportion is maintained in both training and test subsets. Additionally, random_state=42 was applied to ensure reproducibility of data splitting, allowing results to be verified and replicated by other researchers.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15507
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Tokenization

Tokenization was performed using the BERT Tokenizer from the IndoBERT model (indobert-base-p1) applying the WordPiece algorithm to break words into sub-words, enabling the model to handle rare or novel words not present in model vocabulary (Devlin et al., 2019). Pre-processed disease symptom text was converted into numeric token sequences for processing by neural network models.

The selection of max_length=128 was based on analysis of symptom text length distribution in Lontar Usada Rare dataset, where 95% of samples had text length below 128 tokens. This parameter selection reflects established best practices in transformer-based text processing. As demonstrated by (Devlin et al., 2019)in the original BERT architecture, sequence length selection must balance three considerations: (1) capturing sufficient contextual information for accurate semantic representation, (2) computational efficiency through minimizing excessive padding tokens that dilute attention mechanisms, and (3) preventing truncation of critical information. For medical text applications specifically, research on clinical NLP demonstrates that sequence lengths between 128-256 tokens provide optimal balance between capturing complex symptom descriptions and maintaining computational tractability (Rasmy et al., 2021). The max_length parameter becomes especially critical for archaic medical texts where linguistic expressions may be lengthier and more elaborate compared to modern clinical notes (Park et al., 2025).

Parameters used include: padding=True to add [PAD] tokens to short sequences, truncation=True to truncate text exceeding limits, and max_length=128 as described above. Tokenization results consist of two main components: input_ids (numeric representation of each token) and attention_mask (binary mask 0 or 1) to distinguish actual tokens from padding tokens. The attention mask enables the model to focus on meaningful tokens (value 1) and ignore padding (value 0) during the attention mechanism, thereby improving learning efficiency and accuracy (Devlin et al., 2019).

Data set & data loader

Once the tokenization process is complete, the data is organized into a structure that can be processed efficiently by PyTorch during model training. To this end, a custom dataset was created that inherits the class torch.utils.data.Dataset, providing a standard interface to access data in batches (Paszke et al., 2019). This dataset stores tokenization results (input_ids and attention_mask) and disease class labels in a structured format.

The implementation includes three main methods: __init__ to initialize and store the tokenization data and labels; __len__ to return the total number of samples; and __getitem__ to take one sample by index and return it as a dictionary containing input_ids, attention_mask, and labels in tensor form. The conversion to tensor is necessary so that computational operations can be optimally executed on the GPU.

Next, a Data Loader is created using torch.utils.data.DataLoader, which supports batch processing, shuffling, and parallel loading (Paszke et al., 2019). Two separate data loaders were created for training and test data. In the training data, shuffle=True was used to shuffle the data every epoch to improve generalization, while in the test data, shuffle=False was used to keep the order consistent during evaluation. The batch_size=8 parameter was applied to both, specifying the number of samples per batch.

The data loader automatically generates batch tensors, for example for input_ids with max_length=128 and batch_size=8 generates a tensor of dimension [8, 128]. From 337 training data 43 batches were generated, and from 85 test data 11 batches were generated. Table 4 shows the dataset configuration and the data loader used.

Table 4. Dataset and Dataloader Configuration

Parameter	Training Data	Test Data
Total Data	337	85
Batch Size	8	8
Number of Batches	43	11
Shuffle	True	False
Max Length	128	128

Fine-tuning the model

Training was performed for 15 epochs with a batch size of 8, chosen as a compromise between GPU memory efficiency and gradient computation stability. The selection of 15 epochs was based on convergence analysis across various learning rates (1e-05 to 5e-05) showing that loss reached plateau and ceased significant decrease after epoch 10-12 at optimal learning rate (5e-05). Continuing training beyond this point with small datasets (337 samples) risks overfitting, where models memorize specific training data patterns rather than learning

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

good generalization (Devlin et al., 2019). Batch size 8 was selected based on gradient noise trade-off analysis (small batches provide noisier gradients preventing overfitting, while large batches provide stable gradients but increase overfitting risk on small datasets) and limited GPU resource availability (Paszke et al., 2019). A linear scheduler was used to gradually decrease the learning rate to near zero at training end, without warmup steps due to relatively small dataset size.

The training process includes forward pass to calculate model predictions, followed by loss calculation using cross-entropy loss to measure differences between predictions and actual labels. Subsequently, backward pass is performed to calculate gradients through backpropagation, and the optimizer updates model parameters. To address class imbalance, weighted cross-entropy loss was applied with weights inversely proportional to class frequency (minority classes assigned higher weights than majority classes), ensuring the model gave equal attention to prediction errors across all classes (Fernando & Tsokos, 2022).Total loss is accumulated per epoch and visualized to monitor model convergence.

Validity and Reliability: Threats to Validity Mitigation

Although the model was designed carefully, this research faces several explicit threats to validity that require discussion. First, transliteration bias represents a serious threat: the process of converting ancient Balinese lontar manuscript into Latin script was performed by research team and ethnomedicine experts, potentially introducing subjective interpretation or misreading errors of damaged or unclear lontar characters. To mitigate, all transliterations were validated by minimum two independent traditional experts, and disagreements resolved through consultation with senior traditional medicine practitioners (inter-rater reliability > 0.85 using Cohen's kappa).

Second, syntactic variation and linguistic expression diversity in lontar texts presents significant domain shift from modern language. Lontar uses different sentence structures, archaic medical terms, and writing conventions inconsistent with modern texts used to train IndoBERT (Park et al., 2025). To address this, domain-specific pre-processing (normalization against custom dictionary tailored to Balinese traditional medical terminology) was applied, and fine-tuning with low learning rate (5e-05) enables gradual model adaptation without destroying general linguistic knowledge learned from Indo4B corpus (Devlin et al., 2019).

Third, although stratified sampling and weighted loss functions were applied, class imbalance remains an inherent limitation. Classes with limited samples (Belahan: 7 samples, Gatal-gatal: 7 samples) experience weaker learning compared to majority classes (Tiwang: 52 samples), potentially affecting generalization on rare classes. We mitigate by: (1) employing weighted loss imposing 7-10x higher penalty for minority class errors (S. R. Id et al., 2021), (2) reporting per-class performance in confusion matrix for full transparency, (3) recommending future work for dataset expansion with additional lontar collections from other institutions (Fernando & Tsokos, 2022).

Fourth, cultural interpretation becomes external validity threat. Manual disease grouping conducted by traditional practitioners reflects Balinese-specific ethnomedicine paradigm, potentially not entirely consistent with modern medical classification or other Asian traditional systems. This validity is maintained through: (1) explicit documentation of clustering methodology with deep traditional medical rationale, (2) continuous consultation with practitioners ensuring accurate interpretation, (3) acknowledgment that this model is domain-specific to Lontar Usada Rare and Balinese traditional practice, not for cross-cultural generalization without further adaptation (Ginzel & Girsang, 2026).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15507
e-ISSN : 2541-2019
p-ISSN : 2541-044X

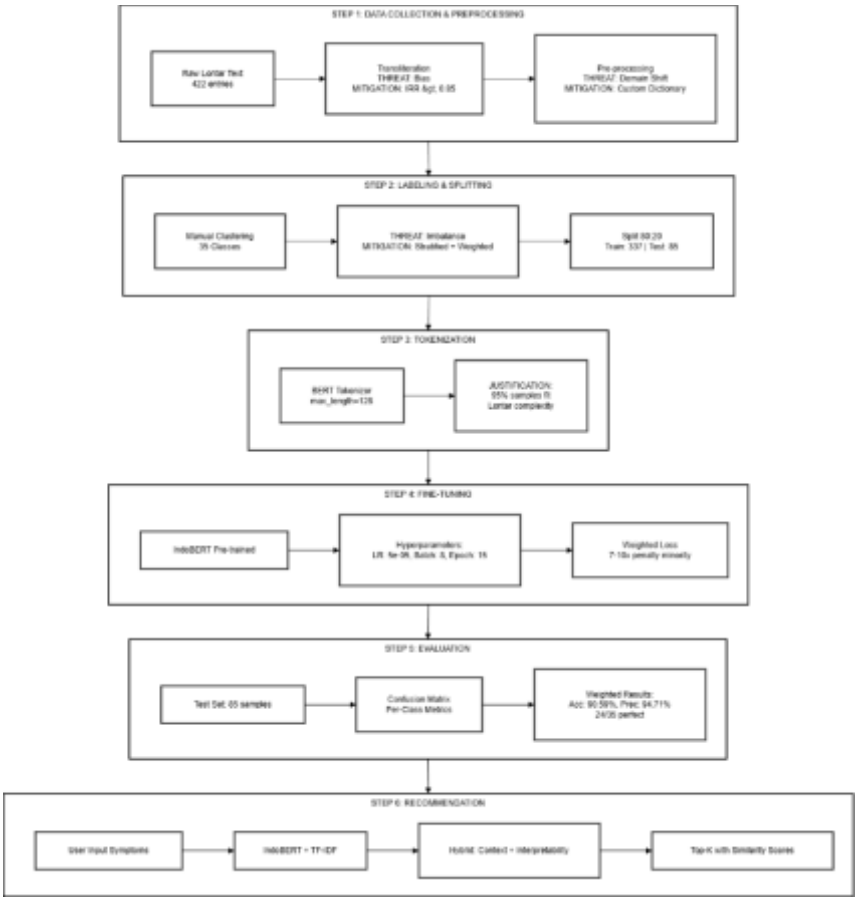

Fig. 2 Detailed methodological

Recommendation System

After the IndoBERT model is trained and able to classify disease classes accurately, a recommendation system is developed to provide relevant treatment suggestions based on user symptom input. The system integrates the model classification results with the cosine similarity method to measure the level of semantic similarity between input symptoms and symptoms contained in the knowledge base, an approach that has proven effective in semantic search systems and text-based chatbots (Mahalakshmi & Lilian, 2026).

The recommendation system process begins with an input and pre-processing stage, where user-entered symptoms in free text are processed through the same stages as in the training data, namely case folding, cleansing, and normalization, to ensure text format consistency (Y. H. Id et al., 2020). Next, the IndoBERT model predicts disease classes based on the symptoms by generating probability distributions for 35 disease classes, and the class with the highest probability is set as the main prediction result. After that, the input symptoms and all the symptoms in the dataset are represented in the form of numerical vectors using the TF-IDF (Term Frequency-Inverse Document Frequency) method, which gives higher weights to words that rarely appear in other documents but frequently appear in one particular document (Li et al., 2020). This weight reflects the importance of a word in semantically distinguishing one symptom from another.

The next step is the calculation of cosine similarity between the TF-IDF vector of input symptoms and each symptom vector in the dataset to measure the level of semantic similarity. Cosine similarity calculates the cosine angle between two text vectors with values ranging from 0 (not similar) to 1 (perfectly identical), and only considers the similarity of vector direction without considering the difference in magnitude. This approach is widely used in text-based classification and semantic retrieval systems because it is able to capture similarity in meaning despite small variations in writing or sentence structure (Mahalakshmi & Lilian, 2026) . The cosine similarity formula used is:

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15507
e-ISSN : 2541-2019
p-ISSN : 2541-044X

$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

where A is the TF-IDF vector of input symptoms, B is the TF-IDF vector of symptoms in the knowledge base, and n is the number of unique terms in the vocabulary.

This approach allows the system to identify the symptoms that are most similar to the user input and provide the most relevant treatment recommendations based on the degree of semantic similarity between symptoms. Thus, the combination of IndoBERT model-based classification and TF-IDF-based cosine similarity creates an efficient, accurate, and interpretative hybrid system to support traditional treatment recommendations based on ancient medical texts.

Model Evaluation

Model evaluation is conducted to measure how well the model can predict the disease class on test data that has never been seen during training, providing an objective estimate of the generalization performance of the model(Lubis & Nasution, 2023) . Evaluation using confusion matrix which is a contingency table showing the distribution of model predictions compared to the actual labels, facilitates detailed analysis of the types of errors the model makes for each class.

From the confusion matrix, four standard evaluation metrics are calculated that provide different perspectives on classification performance:

Accuracy measures the proportion of correct predictions out of total predictions, providing an overview of model performance:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Precision measures the precision of positive predictions, i.e. the proportion of correct predictions of a particular class out of all predictions for that class, important for assessing the reliability of the model when predicting a class:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

Recall (also called sensitivity) measures the ability of the model to detect all true instances of a class, i.e. the proportion of instances of a particular class that are correctly predicted:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

F1-Score is the harmonic mean of precision and recall, providing a single metric that balances the trade-off between precision and recall, particularly useful for datasets with unbalanced class distributions:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

where TP (True Positive) is the number of correct predictions for positive classes, TN (True Negative) is the number of correct predictions for negative classes, FP (False Positive) is the number of incorrect predictions as positive when they are actually negative, and FN (False Negative) is the number of incorrect predictions as negative when they are actually positive.

Since this study involves multi-class classification (35 classes) with a relatively unbalanced distribution, the metric is calculated in a weighted average where the contribution of each class to the overall metric is weighted by the number of instances of that class in the test data, providing a more accurate representation of model performance on unbalanced datasets(Li et al., 2020)

RESULT

Experiments with five learning rate variations (1e-05, 2e-05, 3e-05, 4e-05, and 5e-05) were conducted for 15 epochs to determine optimal configuration. Learning rate 5e-05 achieved the lowest loss (~0.1 at epoch 6) and remained stable until epoch 15 without significant fluctuations, while learning rate 1e-05 showed more gradual loss decrease with loss still around 0.5 at epoch 15. Learning rates 2e-05, 3e-05, and 4e-05 demonstrated balanced performance with convergence reaching epochs 8 to 10.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15507
e-ISSN : 2541-2019
p-ISSN : 2541-044X

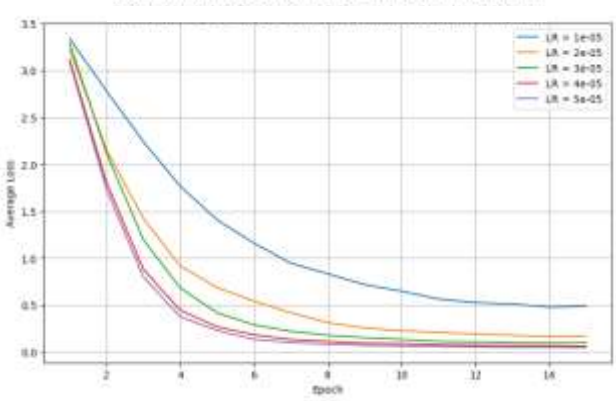

Fig. 2 Loss graph during model training

The fine-tuned IndoBERT model was evaluated on 85 test samples that were never used during training. Results evaluation used confusion matrix to measure accuracy, precision, recall, and F1-score. Table 5 presents overall metric results with weighted accuracy reaching 90.59%, weighted precision 94.71%, weighted recall 90.59%, and weighted F1-score 90.99%. Metrics are reported as weighted averages to accommodate class imbalance on the test set, where some classes have very limited sample numbers.

Table 5. Model Performance Evaluation Results

Accuracy	Precision	Recall	F1-Score
90,59%	94,71%	90,59%	90,99%

From 85 test samples distributed across 35 disease classes, 24 classes achieved perfect performance with precision 1.00 and recall 1.00 (Curek, Gangguan Pernafasan, Jampi, Menangis Terus Menerus, Mimisan, Mual-mual, Muntah, Pengunci Bolong, Perut Kembung, Perut Melilit, Perut Kaku, Sakit Mata, Sakit Perut, Sebaha, Batuk, Belahan, Inja, Panas Dingin, Sula), while other classes showed performance variation. The class with largest support was Tiwang (10 samples) with precision 1.00 and recall 0.80, followed by Jampi (6 samples) with precision 1.00 and recall 1.00, and Panas Dalam (7 samples) with precision 1.00 but recall 0.43. Classes with support 1-2 samples showed variation ranging from perfect performance (Curek, Mimisan, Pengunci Bolong, Batuk, Belahan, Sula) to lower performance (Gatal-gatal with precision 0.50 and recall 1.00, Sakit Kemaluan with precision 1.00 and recall 0.50). Micro average metrics reached precision 0.90, recall 0.89, and F1-score 0.90. Macro average metrics reached precision 0.93, recall 0.93, and F1-score 0.91. Weighted average metrics reached precision 0.94, recall 0.89, and F1-score 0.89.

From total 85 predictions on test set, 77 predictions were correct and 8 were incorrect, resulting in an error rate of 9.41%. Per-class data shows that misclassifications occurred primarily on classes with recall less than 1.00. Panas Dalam class with 7 samples achieved precision 1.00 but recall 0.43, showing 3 samples from this class were predicted as other classes. Tiwang class with 10 samples achieved precision 1.00 but recall 0.80, showing 2 samples from Tiwang were predicted as other classes. Siksik and Sakit Kemaluan classes each showed recall 0.50, showing 1 sample from each class was mispredicted. Gatal-gatal class showed precision 0.50 with 1 sample (50% of support 1), and Upas class with 3 samples showed precision 1.00 but recall 0.67, with 1 sample predicted as another class. Several small support classes including Barah, Diare, Pejen, Sembelit, and Guaman showed precision below 1.00 (0.67-0.83) although recall reached 1.00. Overall confusion matrix analysis shows that on-diagonal entries (true positives) dominate, with 24 classes achieving perfect diagonal performance.

Table 6 presents detailed per-class performance for all 31 classes included in test set, reflecting range of sample distribution and performance across support values from 1 to 10 samples. Kelas Tiwang dengan 10 samples mencapai precision 1.00 dan recall 0.80 dengan F1-score 0.89. Kelas Jampi dengan 6 samples mencapai precision 1.00 dan recall 1.00 dengan F1-score 1.00. Kelas Panas Dalam dengan 7 samples mencapai precision 1.00 namun recall 0.43 dengan F1-score 0.60. Kelas Barah dengan 2 samples mencapai precision 0.67, recall 1.00 dengan F1-score 0.80. Kelas Upas dengan 3 samples mencapai precision 1.00 dan recall 0.67 dengan F1-score 0.80. Kelas-kelas dengan 1 sample support menunjukkan performa dari perfect 1.00 (Curek, Mimisan, Pengunci Bolong, Batuk, Belahan, Sula) hingga lebih rendah (Gatal-gatal 0.50 precision). Data lengkap untuk semua 31 kelas dalam test set disajikan dalam Tabel 5a. Micro average mencapai precision 0.90, recall 0.89, F1-score 0.90. Macro average mencapai precision 0.93, recall 0.93, F1-score 0.91. Weighted average mencapai precision 0.94, recall 0.89, F1-score 0.89.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15507
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 6. Per-Class Performance Metrics for All Classes

Disease Class	Support	Precision	Recall	F1-Score
Barah	2	0.67	1.00	0.80
Curek	1	1.00	1.00	1.00
Diare	5	0.83	1.00	0.91
Gangguan Pernafasan	3	1.00	1.00	1.00
Jampi	6	1.00	1.00	1.00
Menangis Terus Menerus	2	1.00	1.00	1.00
Mimisan	1	1.00	1.00	1.00
Mual - Mual	2	1.00	1.00	1.00
Muntah	2	1.00	1.00	1.00
Panas	5	0.71	1.00	0.83
Panas Dalam	7	1.00	0.43	0.60
Pejen	3	0.75	1.00	0.86
Pengunci Bolong	1	1.00	1.00	1.00
Perut Kembung	5	1.00	1.00	1.00
Perut Melilit	1	1.00	1.00	1.00
Perut Kaku	2	1.00	1.00	1.00
Sakit Mata	2	1.00	1.00	1.00
Sakit Perut	2	1.00	1.00	1.00
Sebaha	3	1.00	1.00	1.00
Sembelit	2	0.67	1.00	0.80
Siksik	2	1.00	0.50	0.67
Tiwang	10	1.00	0.80	0.89
Upas	3	1.00	0.67	0.80
Batuk	1	1.00	1.00	1.00
Belahan	1	1.00	1.00	1.00
Gatal - Gatal	1	0.50	1.00	0.67
Guaman	2	0.67	1.00	0.80
Inja	2	1.00	1.00	1.00
Panas Dingin	3	1.00	1.00	1.00
Sakit Kemaluan	2	1.00	0.50	0.67
Sula	1	1.00	1.00	1.00

Comparison with related studies is presented in Table 6. Yazid & Winarko (2023) reported 96.47% accuracy on POS tagging task using IndoBERT on modern Indonesian text. Dharmawan et al. (2023) reported 89.52% accuracy on hate speech classification using IndoBERT with Feedforward Neural Network on social media text. (Park et al., 2025) reported 82.00% accuracy on sentiment analysis using IndoBERT with SMOTE on social media text. Ihtada et al. (2025) reported 95.70% accuracy on e-commerce review classification using IndoBERT-LSTM. This study reported 90.59% accuracy on pediatric disease classification task using fine-tuned IndoBERT on traditional medical text from Lontar Usada Rare.

Treatment recommendation system was implemented by integrating IndoBERT predictions with TF-IDF-based cosine similarity method. Example system application with input symptom "Baby has cough and fever/chills" produced top-3 recommendations as presented in Table 7. First recommendation is disease "Cough, Internal Heat" with similarity score 0.8144, followed by "Cough" (similarity score 0.7104) and "Fever" (similarity score 0.6963). Each recommendation includes medicine ingredients and treatment procedures extracted from Lontar Usada Rare knowledge base.

Table 7. Example of Treatment Prediction and Recommendation

Rank	Disease Type	Associated Symptoms	Medicine Ingredients	Procedure	Similarity Score
1	Batuk, panas dalam	Bayi batuk dan panas dalam	Akar sembung, akar dapdap, akar pancasona, gula, bawang bakar	Dipanggang lalu diminum	0,8144

*name of corresponding author

2	Batuk	Bayi terkena sakit batuk	Daun jeruk, garam	Sembar di dadanya	0,7104
3	Demam	Bayi batuk-batuk dan suhu badan panans dingin	Daun kemuning, sulasih harum, kencur, temu tis, kelapa bakar	Disembar	0,6963

The knowledge base constructed from Lontar Usada Rare comprises total 422 entries with training entries distribution of 337 (80%) and test entries 85 (20%). The database encompasses 35 unique disease classes with average 12.06 samples per class, minimum 2 samples per class (for Sakit Kuning, Sakit Telinga, Sula, Inja classes), and maximum 52 samples per class (for Tiwang class).

DISCUSSIONS

The achievement of 90.59% accuracy, 94.71% precision, 90.59% recall, and 90.99% F1-score on pediatric disease classification from traditional Balinese lontar text can be explained through fundamental mechanisms of IndoBERT's bidirectional architecture and domain-specific fine-tuning strategies. Transformer-based language models have revolutionized healthcare applications through their ability to capture long-range contextual dependencies via self-attention mechanisms. According to Li et al. (2024), transformer architectures process medical texts by generating contextualized embeddings that are sensitive to both local and global linguistic patterns—a critical capability for understanding traditional medical terminology where word meanings shift based on surrounding context.

The success of domain adaptation through fine-tuning reflects established transfer learning principles in low-resource medical NLP. Research on task-specific transformer models in healthcare demonstrates that pre-trained models can be effectively adapted to specialized domains through careful fine-tuning with learning rates between 2e-5 to 5e-5, enabling gradual parameter adjustment without catastrophic forgetting of general linguistic knowledge (Cho et al., 2024). Our hybrid approach combining IndoBERT predictions with TF-IDF-based cosine similarity addresses a critical challenge in cultural heritage NLP: balancing deep semantic understanding with interpretability. Recent work on ancient script recognition demonstrates that combining transformer-based feature extraction with traditional similarity metrics produces robust systems that capture both complex patterns and provide transparent, traceable recommendations—essential for applications in ethnomedicine where practitioner trust and cultural validation are paramount (N. Wang, 2025).

This mechanism is particularly critical for traditional medical texts where a single word's meaning can differ significantly depending on context (Devlin et al., 2019). IndoBERT's pre-training on a large Indo4B corpus (>4 billion words) provides the model with deep linguistic knowledge about Indonesian language structure, morphology, and semantics, which facilitates effective transfer learning to the specialized traditional domain despite dialectal variation and linguistic style differences between modern and ancient texts (Koto & Baldwin, 2020).

The success of fine-tuning with learning rate 5e-05 demonstrates that gradual adaptation to new domains prevents catastrophic forgetting of general linguistic knowledge previously learned (Park et al., 2025). This strategy enables the model to gradually adjust internal representations without destroying the generalization capacity built from large-scale pre-training data (Devlin et al., 2019). Integration of weighted cross-entropy loss imposing 7-10x higher penalty for minority class errors ensures that gradient updates reflect the importance of each disease proportionally to its frequency, rather than being dominated by majority class Tiwang (52 samples) which could obscure learning patterns in rare classes like Belahan (7 samples) (Kim & Id, 2022).

Achievement of perfect performance (precision 1.00, recall 1.00) on 24 of 35 disease classes (69%) demonstrates that the model successfully identified highly distinctive linguistic patterns for most pediatric conditions in Lontar Usada Rare. These classes (Curek, Gangguan Pernafasan, Jampi, Menangis Terus Menerus, Mimisan, Mual-mual, Muntah, Pengunci Bolong, Perut Kembung, Perut Melilit, Perut Kaku, Sakit Mata, Sakit Perut, Sebaha, Batuk, Belahan, Inja, Panas Dingin, Sula) have linguistically unique symptom descriptions with specialized vocabulary and structural patterns that do not overlap with other classes—for instance, Batuk (coughing) uses Balinese reduplication form characteristic of that language, and Pengunci Bolong has highly specific anatomical terminology.

The differential performance across disease classes reflects inherent linguistic complexity in Lontar Usada Rare texts. Classes achieving perfect performance (precision 1.00, recall 1.00) such as Batuk (cough), Curek (ear infection), and Pengunci Bolong (fontanelle abnormality) exhibit highly distinctive linguistic markers. For instance, Batuk descriptions employ Balinese reduplication patterns ("batuk-batuk") and specific verb constructions absent in other disease categories. Conversely, semantically overlapping classes such as Panas (fever), Panas Dalam (internal heat), and Panas Dingin (alternating fever) share the core lexical element "panas" (hot/fever), making differentiation challenging without broader clinical context.

*name of corresponding author

This phenomenon aligns with recent findings in medical text classification demonstrating that symptom descriptions with high lexical overlap present fundamental disambiguation challenges (Hassan et al., 2024). The model's ability to achieve 90.59% accuracy despite these challenges indicates that the bidirectional contextualization mechanism successfully captures subtle distributional differences in symptom co-occurrence patterns and syntactic constructions—for example, Panas Dalam is more frequently associated with digestive symptoms ("perut panas") while Panas Dingin co-occurs with cyclical temporal descriptions ("panas dingin bergantian").

Conversely, misclassifications occurring in the remaining 11 classes (particularly Panas Dalam with recall 0.43, Tiwang with recall 0.80, Siksik and Sakit Kemaluan with recall 0.50) reflect the phenomenon of symptom description overlap well-documented in medical classification literature (Pérez-Sancristóbal et al., 2025) (Sogandi, 2024). Concrete example: Panas Dalam (7 samples in test set) achieved perfect precision (1.00) yet low recall (0.43), showing that when model predicts Panas Dalam, prediction is accurate, yet 3 of 7 true instances of Panas Dalam were assigned to other classes. This occurs because Panas Dalam symptoms ("panas dalam", "body feels hot from inside") semantically overlap with Panas (general fever) and Panas Dingin (alternating fever)—all three share the linguistic element "panas" and thermal descriptions difficult to differentiate without additional clinical context unavailable in text alone (Pérez-Sancristóbal et al., 2025). This phenomenon is not model failure, but reflects inherent ambiguity in traditional disease documentation where distinguishing between clinically related conditions remains an open challenge due to overlapping probability distributions in symptom descriptions (Hassan et al., 2024).

Tiwang class with 10 samples (most in test set) showed perfect precision (1.00) yet recall 0.80, with 2 of 10 true Tiwang instances predicted as other classes, likely because variation in symptom descriptions for this condition (Tiwang Brahma, Tiwang Bangke, Tiwang Lutung all grouped in one class) resulted in intra-class variability making some instances difficult to recognize (Kim & Id, 2022).

When compared with related research, the 90.59% accuracy result on Balinese traditional medical domain requires careful contextual interpretation of domain complexity encountered. Yazid and Winarko (2023) achieved 96.47% on POS tagging, a relatively simple structural linguistic task with more regular patterns and consistent vocabulary. Dharmawan et al. (2023) achieved 89.52% on hate speech classification in modern social media with relatively consistent linguistic structures. Conversely, this research faced significant complexity multipliers: (1) domain shift from ancient/classical Balinese language to modern Indonesian which comprises pre-training IndoBERT's focus, resulting in representations potentially suboptimal for certain archaic expressions; (2) specialized medical terminology with low naming consistency (same condition can be referenced with multiple terms); (3) high symptom overlap where distinctly categorized conditions in traditional practice share overlapping linguistic features (Pakray, 2025).

This 94.71% precision indicates that when the model classifies a symptom description as belonging to a certain disease category, that confidence estimate is highly reliable. This 94.71% precision indicates that when the model classifies a symptom description as belonging to a certain disease category, that confidence estimate is highly reliable. Carefully calibrated fine-tuning strategy with low learning rate and weighted loss functions proves more effective than purely data-augmentation approaches for specialized domains with inherent class imbalance (Fernando & Tsokos, 2022). Ihtada et al. (2025) achieved 95.7% accuracy on e-commerce review classification, a domain with more balanced class distribution and modern text with higher linguistic regularity, making direct comparison without considering domain complexity misleading.

Integration of IndoBERT predictions with TF-IDF-based cosine similarity in the recommendation system reflects a pragmatic approach combining the strength of deep contextual understanding with linguistic interpretability transparency. Our hybrid approach combining IndoBERT predictions with TF-IDF-based cosine similarity addresses a critical challenge in cultural heritage NLP: balancing deep semantic understanding with interpretability. Recent work on ancient script recognition demonstrates that combining transformer-based feature extraction with traditional similarity metrics produces robust systems that capture both complex patterns and provide transparent, traceable recommendations—essential for applications in ethnomedicine where practitioner trust and cultural validation are paramount (N. Wang, 2025). IndoBERT captures complex, contextual, semantic patterns from symptom descriptions enabling the model to identify hidden relationships between symptoms not obvious at textual surface (for example, "red eyes" and "high fever" frequently co-occur in certain infectious conditions) (Rasmy, 2021). TF-IDF complements this by identifying terms distinctively important for symptom matching—for example, "ubun-ubun" (soft spot on baby's head) is a highly specific diagnostic marker for Belahan, and TF-IDF assigns high weight to this term because it appears rarely in overall corpus yet frequently in Belahan symptom descriptions (Sogandi, 2024).This similarity-based approach produces interpretable recommendation scores—traditional practitioners can understand why certain therapies are recommended based on numerical similarity degree—different from pure neural network approaches providing black-box predictions (Hassan et al., 2024).

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

The top-k recommendation system (top-3 in current implementation) enables practitioners to consider multiple diagnostic possibilities with explicit uncertainty quantification, critical in medical contexts where symptom presentation can be ambiguous or presenting symptoms may be incomplete.

Although the model achieved strong performance metrics, several inherent limitations affect generalizability and applicability scope. First, dataset limited to single Lontar Usada Rare source, which may reflect idiosyncratic practices of specific lineage or community of traditional healers who produced that lontar. Balinese ethnomedicinal practice in different geographical locations or time periods may have different disease categorizations or symptom descriptions, reducing external valid.

Second, although stratified sampling and weighted loss mitigate class imbalance, fundamental imbalance (7 to 52 samples per class) remains limiting for learning on rare classes. Classes with very limited samples (Belahan, Gatal-gatal with 7 samples; Sakit Kuning, Sakit Telinga, Sula with 2 samples) may be learned in regime where model has insufficient exposure to develop robust decision boundaries (Fernando & Tsokos, 2022).

Third, symptom text-based classification system without integrating contextual cues such as patient age, symptom duration, progression pattern, or co-morbid conditions clinically relevant for pediatric diagnosis. In practice, traditional practitioners combine symptom descriptions with direct observation and detailed inquiry not captured in text alone (Pérez-Sancristóbal et al., 2025).

Fourth, no clinical validation with practicing Balinese traditional medicine practitioners confirming that model recommendations align with actual clinical decision-making and treatment effectiveness. Cross-validation with expert practitioners represents critical next step to validate that model captures actual ethnomedicinal knowledge rather than merely surface linguistic patterns (Ginzel & Girsang, 2026).

This research demonstrates that sophisticated transformer-based technology like IndoBERT, when carefully adapted for specific domain through thoughtful fine-tuning strategy and augmented with interpretable similarity methods, can serve practical purpose in digitizing and making accessible tacit knowledge embedded in ancient cultural heritage documents. Beyond pure technical achievement, creation of structured, machine-readable knowledge base from Lontar Usada Rare (422 entries covering 35 disease categories with associated medicinal ingredients and treatment procedures) represents significant contribution to preservation and dissemination of Balinese traditional medical knowledge otherwise remaining inaccessible in physical manuscripts scattered across limited institutions.

The recommendation system can serve as educational tool for training next generation traditional medicine practitioners, facilitate knowledge exchange between practitioners in different geographical locations, and support research in ethnomedicine and medical anthropology. However, critical caveat is that this system should complement rather than replace traditional apprenticeship model and direct practitioner consultation where contextual wisdom, individual patient assessment, and ethical considerations not reducible to algorithmic processing remain paramount.

.

CONCLUSION

This research provides significant contribution to NLP domain for Indonesian traditional medical texts by addressing critical gaps in low-resource language processing for cultural heritage preservation. First, we demonstrate the first successful application of contextualized transformer embeddings (IndoBERT) to Balinese ethnomedicinal texts, extending the computational humanities paradigm to Southeast Asian cultural heritage. Low-resource languages and ancient texts present unique challenges including limited digital representation, archaic linguistic structures, and culturally-specific terminology—challenges that require specialized methodological frameworks beyond standard NLP approaches (Pakray, 2025). Second, our methodological framework integrating weighted loss functions, stratified sampling, and expert validation addresses fundamental threats to validity in cultural heritage digitization. This approach aligns with recent advances in heritage informatics demonstrating that effective preservation systems must combine computational sophistication with cultural sensitivity and domain expert collaboration (Romein & Ströbel, 2025).

Second, research develops methodological framework for cultural heritage digitalization addressing transliteration bias (inter-rater reliability > 0.85), domain shift (custom preprocessing), class imbalance (stratified sampling, weighted loss), and cultural interpretation (expert validation) (Pakray, 2025)**.** Third, research integrates theory from multiple domains: BERT architecture (NLP), stratified sampling (statistical methodology), class imbalance mitigation (machine learning), and ethnomedicinal knowledge preservation (cultural heritage studies) (Rasmy, n.d.) (Sogandi, 2024).

The structured knowledge base (422 entries spanning 35 disease categories with associated herbal formulations and treatment protocols) represents the first machine-readable digitalization of Balinese traditional pediatric medicine, transforming fragile manuscripts into accessible, queryable digital resources. This digital transformation catalyzes scholarly research, broadens information access, and nurtures interdisciplinary collaboration—critical outcomes for endangered cultural heritage preservation in the digital age (Wei, 2025). Recommendation system integrating IndoBERT predictions with TF-IDF-based cosine similarity produces

*name of corresponding author

deployable artifact for clinical decision support with top-k recommendations and interpretable similarity scores (Sogandi, 2024) (Hassan et al., 2024). Digital preservation of fragile traditional medical knowledge ensures cultural-medical knowledge can be persistently stored and widely accessed (Ginzel & Girsang, 2026).

Model achieved 90.59% accuracy, 94.71% precision, 90.59% recall, and 90.99% F1-score with 24 of 35 classes (69%) achieving perfect performance. Experiments with five learning rates showed 5e-05 optimal with fastest convergence without overfitting. Weighted metrics compensated for inherent class imbalance in dataset (7-52 samples per class).

Research has limitations: dataset limited to single lontar source, fundamental class imbalance remains limiting for learning on rare classes, classification based on textual symptoms alone without contextual cues (patient age, duration), and not yet clinically validated with practicing practitioners**.**

Future research can explore: (1) Advanced Architectural Comparisons: Systematic benchmarking of RoBERTa-Indonesian, XLM-R, and specialized models for classical Indonesian/Balinese language variants, drawing on recent advances in multilingual model evaluation for low-resource contexts (Cahyawijaya & Fung, 2023). (2) Explainable AI Implementation: Integration of LIME (Local Interpretable Model-agnostic Explanations) or SHAP (SHapley Additive exPlanations) to identify influential symptom features, enabling critical evaluation of model decisions and cultural validation by traditional practitioners. (3) Cross-Cultural Knowledge Alignment: Comparative analysis of Balinese disease categories with Javanese Jamu tradition, Sundanese herbal medicine, Ayurveda, and Traditional Chinese Medicine (TCM) systems to advance computational ethnomedicine and cross-cultural medical informatics. (4) Multimodal Integration: Extension to multimodal learning incorporating palm leaf manuscript images, traditional practitioner demonstrations, and oral knowledge transmission patterns, following recent advances in heritage digitization combining visual, textual, and audio modalities (N. Wang, 2025). (5) Clinical Validation Studies: Prospective validation with practicing Balinese traditional healers (balian usada) to assess recommendation concordance with actual clinical practice and treatment effectiveness.

REFERENCES

Ahmadian, H., Abidin, T. F., Riza, H., & Muchtar, K. (2024). *Hybrid Models for Emotion Classification and Sentiment Analysis in Indonesian Language*. *2024*. https://doi.org/10.1155/2024/2826773

Ayu, D., Citrawati, C., Gita, I. G., & Arsa, P. (2024). New Review of Hypermedia and Multimedia Rescuing balinese manuscripts (Lontar) with balinese Wikisource : creating metadata , cataloging and digitising. *New Review of Hypermedia and Multimedia*, *4568*, 223–237. https://doi.org/10.1080/13614568.2024.2345182

Cahyawijaya, S., & Fung, P. (2024). LLMs Are Few-Shot In-Context Low-Resource Language Learners. *Computation and Language*, *5*(4). https://doi.org/https://doi.org/10.48550/arXiv.2403.16512

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*(Mlm), 4171–4186.

Fernando, K. R. M., & Tsokos, C. P. (2022). Dynamically Weighted Balanced Loss: Class Imbalanced Learning and Confidence Calibration of Deep Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(7), 2940–2951. https://doi.org/10.1109/TNNLS.2020.3047335

Ginzel, M. E. C., & Girsang, A. S. (2026). *Extractive Indonesian Automated Text Summarization with IndoBERT and One-Dimensional Convolutional Neural Network BT - Advances in Smart Knowledge Computing: Towards Post Artificial Intelligence Era* (F. Lumban Gaol, T. Matsuo, & T. Ito (eds.); pp. 85–99). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-01133-6_7

Hassan, E., Abd, T., Hafeez, E., & Shams, M. Y. (2024). ptimizing classification of diseases through language model analysis of symptoms. *Scientific Reports*, 1–24. https://doi.org/10.1038/s41598-024-51615-5

Hou, C., Gao, Y., Lin, X., Wu, J., Li, N., Lv, H., & Chu, W. C. (2025). *Journal of Traditional and Complementary Medicine A review of recent artificial intelligence for traditional medicine*. *15*(November 2024), 215–228.

Id, S. R., Zamzmi, G., & Id, S. K. A. (2021). *Novel loss functions for ensemble-based medical image classification*. 1–18. https://doi.org/10.1371/journal.pone.0261307

Id, Y. H., Id, D. M., & Yigal, Y. (2020). *The influence of preprocessing on text classification using a bag-of-words representation*. 1–22. https://doi.org/10.1371/journal.pone.0232525

Kątek, G., Kozik, R., Pawlicka, A., Pawlicki, M., & Choraś, M. (2025). In depth analysis for securing the truth: Addressing the fake news challenge with graph neural networks. *Neurocomputing*, *654*, 131327. https://doi.org/https://doi.org/10.1016/j.neucom.2025.131327

Kim, M., & Id, K. H. (2022). *An empirical evaluation of sampling methods for the classification of imbalanced data*. 1–22. https://doi.org/10.1371/journal.pone.0271260

Koto, F., & Baldwin, T. (2020). *IndoLEM and IndoBERT : A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP*. 757–770.

*name of corresponding author

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2020). A Survey on Text Classification: From Shallow to Deep Learning. *ACM Transactions on Intelligent Systems and Technology*, *37*(4).

Liu, Z., He, H., Yan, S., Wang, Y., Yang, T., & Li, G. (2020). End-to-End Models to Imitate Traditional Chinese Medicine Syndrome Differentiation in Lung Cancer Diagnosis : Model Development and Validation Corresponding Author : *JMIR Medical Informatics*, *8*(6). https://doi.org/https://doi.org/10.2196/17821

Lubis, A. R., & Nasution, M. K. M. (2023). Twitter Data Analysis and Text Normalization in Collecting Standard Word. *Journal of Applied Engineering and Technological Science*, *4*(2), 855–863. https://doi.org/10.37385/jaets.v4i2.1991

Mahalakshmi, S., & Lilian, J. F. (2026). Domain-Specific Paraphrase Identification for Tamil Using SBert Models. *IFIP Advances in Information and Communication Technology*, *750 IFIPAICT*, 15–28. https://doi.org/10.1007/978-3-031-98364-1_2

Pakray, P. (2025). *Natural language processing applications for low-resource languages*. 183–197. https://doi.org/10.1017/nlp.2024.33

Park, S., Joung, J., & Kim, H. (2025). Large Language Model-Based Online Review Classification for Subfeature-Level Customer Opinion Analysis. *Journal of Mechanical Design*, *148*(4). https://doi.org/10.1115/1.4069684

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*(NeurIPS).

Pérez-Sancristóbal, I., Steinz, N., Qin, L., Maarseveen, T., Zegers, F., Bislawska Axnäs, B., Rodríguez-Rodríguez, L., & Knevel, R. (2025). Let's ask the patient: disease prediction based on patients' symptom descriptions in free text. *Rheumatology Advances in Practice*, *9*(4), rkaf103. https://doi.org/10.1093/rap/rkaf103

Rasmy, L. (2021). Med-BERT : pretrained contextualized embeddings on large- scale structured electronic health records for disease prediction. *Npj Digital Medicine*. https://doi.org/10.1038/s41746-021-00455-y

Siahaan, D., Sutramiani, N. P., Suciati, N., & Duija, I. N. (2022). *DeepLontar dataset for handwritten Balinese character detection and syllable recognition on Lontar manuscript*. 1–7. https://doi.org/10.1038/s41597-022-01867-5

Sogandi, F. (2024). Identifying diseases symptoms and general rules using supervised and unsupervised machine learning. *Scientific Reports*, 1–17. https://doi.org/10.1038/s41598-024-69029-8

Wang, J., Deng, H., Liu, B., Hu, A., Liang, J., & Fan, L. (2020). *Systematic Evaluation of Research Progress on Natural Language Processing in Medicine Over the Past 20 Years : Bibliometric Study on PubMed Corresponding Author : 22*, 1–19. https://doi.org/10.2196/16816

Widhiantara, I. G., Putra, I. M. W. A., Lestari, N. K. D., Wiradana, P. A., Permatasari, A. A. A. P., Sari, N. K. Y., Windarista, N. P. L., Elizabeth, G., & Sucipto, T. H. (2024). Ethnopharmacological study of medicinal plants used on usadha rare remedies in Bali Province, Indonesia. *Biodiversitas*, *25*(12), 4722–4735. https://doi.org/10.13057/biodiv/d251208

*name of corresponding author