Comparing XGBoost and LightGBM for Optimizing Health Content Categories

Nanda Oktaviana1), **Andrianingsih**2)*,
1)2)Sistem Informasi, Fakultas Teknologi Komunikasi dan Informatika, Universitas Nasional
1) nandaoktaviana2022@student.unas.ac.id, 2)andrianingsih@civitas.unas.ac.id

Submitted :Nov 5, 2025 | **Accepted** : Nov 27, 2025 | **Published** : Jan 01, 2026

Abstract: Indonesia's social media platforms contain large amounts of unverified health information. Research on Indonesian health-text mining still rarely focuses on disease-based classification, leaving a gap compared with studies that only address sentiment or general topic categorization. This study proposes a multi-class classification approach that uses IndoBERT embeddings combined with gradient-boosting classifiers (XGBoost and LightGBM) to categorize tweets into diabetes, hypertension, and heart disease. The dataset comprises 4,075 tweets collected from platform X (Twitter). Preprocessing involves text cleaning, anonymization, normalization, and the extraction of 768-dimensional IndoBERT embeddings. Experiments are conducted in Google Colab (Intel Xeon CPU, 13 GB RAM, optional NVIDIA T4 GPU) using stratified five-fold cross-validation.The best results are obtained by the IndoBERT × LightGBM pipeline, which achieves an accuracy of 0.8526 and a macro-averaged F1-score of 0.8527, outperforming the IndoBERT × XGBoost model (accuracy 0.8325 and macro F1-score 0.8326). Feature-importance analysis shows that contextual terms related to blood sugar, the heart, and blood pressure strongly influence the predictions. Overall, the proposed method provides an effective baseline for monitoring health-related text and supporting disease-oriented analytics in Indonesian-language social media.

Keywords: IndoBERT, LightGBM, XGBoost, health content classification, Indonesian text mining.

INTRODUCTION

Health misinformation has become a growing challenge in Indonesia's digital landscape, particularly concerning non-communicable diseases (NCDs) such as diabetes, heart disease, and hypertension. The rapid spread of health-related content on social media platforms such as X (formerly Twitter) blends credible medical information with unverified claims, complicating public understanding and diminishing health literacy (Suherman et al. 2023).

Conventional text representation methods such as Term Frequency–Inverse Document Frequency (TF–IDF) have been widely used in prior studies but remain limited in capturing linguistic context and meaning, especially in informal or code-mixed Indonesian text (Hindarto et al., 2025). Recent advances in natural language processing have introduced transformer-based models that can represent semantics contextually rather than through frequency counts. Among these, IndoBERT a pretrained BERT model adapted for the Indonesian language provides strong performance in text understanding tasks due to its ability to capture subword-level relationships and contextual dependencies (Koto et al. 2020) IndoBERT embeddings are particularly suited for analyzing short, noisy, and conversational text typical of social media content, making it a robust alternative to sparse vectorization for Indonesian-language data.

While deep contextual embeddings improve representation quality, selecting the appropriate classifier remains crucial for optimizing model performance. Ensemble-based learning methods such as XGBoost and LightGBM have demonstrated high accuracy, scalability, and efficiency across a wide range of classification tasks, including text and health informatics domains (Chen and Guestrin 2016). However, few comparative studies have examined the integration of IndoBERT embeddings with boosting algorithms for Indonesian-language health content classification. Most prior work has relied on simpler models or binary classification setups rather than multi-class frameworks involving multiple NCD categories. Consequently, the best-performing boosting method and optimal configuration that balance accuracy, interpretability, and computational efficiency remain unclear (Hindarto et al. 2025)

This study develops a supervised multi-class classification framework that integrates IndoBERT embeddings with XGBoost and LightGBM to categorize Indonesian-language health tweets into three disease topics: diabetes,

*name of corresponding author

heart disease, and hypertension. The dataset comprises social media posts containing disease-related keywords, which were cleaned, anonymized, normalized, and tokenized prior to embedding extraction using pretrained IndoBERT. The embeddings were then used as input to gradient boosting models trained through stratified five-fold cross-validation to maintain class balance. Evaluation metrics include accuracy, precision, recall, and macro-averaged F1-score, supported by confusion matrix analyses to interpret misclassification patterns.

To enhance interpretability, feature-importance analysis was employed to identify contextual features most influential in class assignment, highlighting linguistically meaningful tokens such as “blood sugar,” “heart,” and “blood pressure.” This interpretive layer ensures that model predictions remain transparent and explainable to both technical and non-technical stakeholders, aligning with best practices in explainable artificial intelligence (XAI) research.

LITERATURE REVIEW

Health misinformation is increasingly prevalent in Indonesia’s digital ecosystem, particularly regarding non-communicable diseases (NCDs) such as diabetes, heart disease, and hypertension. Social media platforms like X (formerly Twitter) allow health information to circulate rapidly, often blending verified medical insights with unsubstantiated claims. This mixture of credible and misleading content reduces public clarity and weakens health literacy (Suherman et al., 2023).

Traditional text-representation approaches such as TF–IDF have been widely adopted in previous Indonesian studies but struggle to capture contextual meaning, especially in short, informal, or code-mixed social media language (Hindarto et al., 2025). For instance, (Kaeren & Andrianingsih, 2025) employed Naive Bayes and Random Forest to analyse Indonesian-language user reviews of the LinkAja mobile-payment application, showing that conventional machine-learning pipelines can effectively perform sentiment classification while still relying on sparse text representations. Recent NLP developments offer more advanced representation techniques, particularly transformer-based models that encode semantics contextually. IndoBERT, a BERT variant trained on large Indonesian corpora, is highly effective for understanding nuanced linguistic structures by modeling subword patterns and contextual dependencies (Koto et al., 2020). Its ability to manage short, noisy, conversational Indonesian text makes IndoBERT embeddings substantially more robust than sparse frequency-based vectors.

Despite improvements in representation quality, the choice of classifier remains essential. Gradient-boosting methods such as XGBoost and LightGBM are well established for their strong predictive performance, computational efficiency, and capacity to generalize across diverse classification tasks (Chen & Guestrin, 2016). However, limited research has investigated the comparative performance of boosting algorithms when paired specifically with IndoBERT embeddings in multi-class Indonesian health-content classification settings. Most prior work focuses on binary sentiment tasks or simpler modeling pipelines, leaving a gap in understanding which boosting model is most effective for multi-topic disease categorization.

To address this gap, this study develops a supervised multi-class classification framework that integrates IndoBERT embeddings with XGBoost and LightGBM to categorize Indonesian-language health tweets into diabetes, hypertension, and heart-disease classes. The dataset containing 4,075 tweets undergoes cleaning, anonymization, normalization, and embedding extraction using pretrained IndoBERT. Model training is conducted through stratified five-fold cross-validation to ensure balanced evaluation across classes. Performance is assessed using accuracy, precision, recall, and macro F1-score, complemented by confusion matrix analyses to reveal misclassification patterns.

To strengthen interpretability, feature-importance analysis is applied to highlight the contextual features most influential in model decisions such as terms related to blood sugar, cardiac symptoms, and blood pressure. These insights support transparency and align with current guidelines on explainable artificial intelligence (XAI), ensuring that model outputs can be understood by both technical and non-technical stakeholders.

METHOD

This section describes the methodological framework for building and evaluating a multi-class categorization system for Indonesian-language health content on platform X (Twitter), focusing on three disease topics: Diabetes, Heart Disease, and Hypertension. The workflow was designed as an end-to-end pipeline, starting from data acquisition and ethical handling to text preprocessing, contextual embedding generation using IndoBERT, model training with XGBoost and LightGBM, hyperparameter optimization through stratified k-fold cross-validation, and final evaluation using macro-F1, accuracy, precision, recall, and confusion matrix analysis.

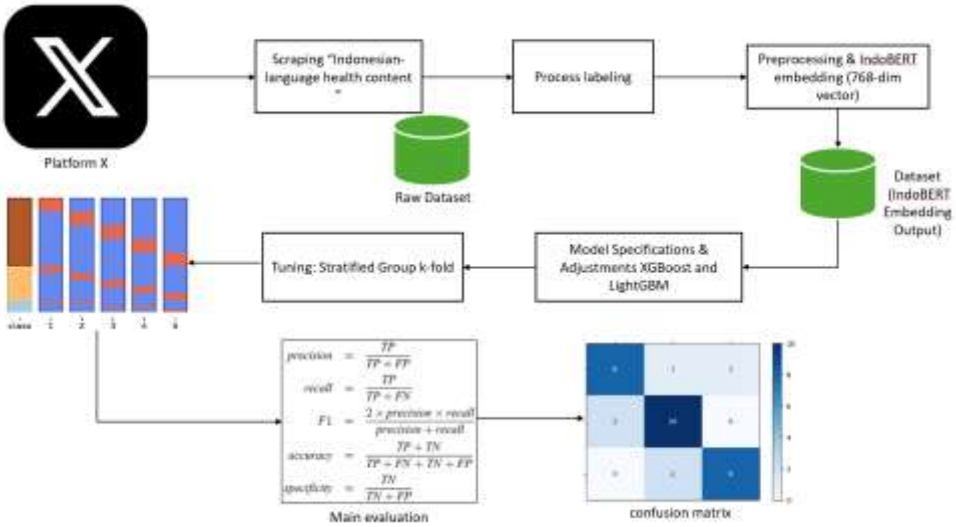
Platform X
Scraping "Indonesian-language health content"
Raw Dataset
Process labeling
Preprocessing & IndoBERT embedding (768-dim vector)
Dataset (IndoBERT Embedding Output)
Model Specifications & Adjustments XGBoost and LightGBM
Tuning: Stratified Group k-fold
precision = TP / (TP + FP)
recall = TP / (TP + FN)
F1 = 2 × precision × recall / (precision + recall)
accuracy = (TP + TN) / (TP + FN + TN + FP)
specificity = TN / (TN + FP)
Main evaluation
confusion matrix

Fig 1. Methodological pipeline for classifying Indonesian health-related tweets using IndoBERT embeddings combined with XGBoost and LightGBM

Figure 1 summarizes the IndoBERT-based model comparison workflow. Indonesian-language health content was collected from platform X (Twitter) and assigned to three target categories: Diabetes, Heart Disease, and Hypertension. The corpus was obtained using keyword and hashtag filtering for the three disease topics, and only public posts were retrieved. Usernames, URLs, and other identifying metadata were removed, while duplicate posts, uncommented retweets, and non-Indonesian tweets were excluded so that the final dataset remained relevant, anonymized, and compliant with research-ethics guidelines.

Each tweet was then labeled as Diabetes, Heart Disease, or Hypertension based on the presence of disease-related medical and general vocabulary as well as its sentence context. A concise annotation guideline was used to define the three classes, provide positive and negative examples, and specify how to handle borderline cases, thereby reducing subjectivity during manual labeling.

The textual data were preprocessed through case folding; normalization of URLs, mentions, emojis, and numbers; and the removal of excessive punctuation to reduce noise. Indonesian tokenization and stop-word removal were subsequently applied to eliminate non-informative function words, followed by light stemming or lemmatization to standardize morphological variants. Colloquial expressions and loanwords related to the target diseases (for example, informal references to blood sugar, blood pressure, cholesterol, or heart complaints) were normalized to their canonical Indonesian forms so that lexical overlap between classes could be minimized.

Instead of using sparse TF–IDF features, each cleaned tweet was transformed into a dense 768-dimensional contextual embedding using IndoBERT, a pretrained language model for Indonesian that captures semantic and morphological nuances at the subword level. These IndoBERT embeddings served as numerical input for the classification stage.

In the final stage, two gradient-boosting algorithms, XGBoost and LightGBM, were implemented as comparative classifiers. Both models were trained and tuned using stratified five-fold cross-validation to preserve class balance and reduce information leakage between training and validation folds. Hyperparameters were optimized through a combination of random and grid search, focusing on learning rate, tree depth, number of leaves, and regularization strength. Model performance was evaluated using accuracy, precision, recall, and macro-averaged F1-score, with confusion matrices used to visualize per-class misclassification patterns.

This workflow ensures that the comparative analysis of XGBoost and LightGBM is transparent, reproducible, and ethically compliant from data collection to model evaluation, and provides a scalable pipeline for Indonesian-language health content classification.

Dataset

Table 1. Sample dataset

Created at	Favorite Count	Text	URL
Thu Oct 09 2025	0	A sexually transmitted disease caused by the bacterium Treponema pallidum. It can attack the skin, genitals, and even the nervous system and heart if not treated promptly. Online consultation now: (link) #Syphilis #RajaSinga	https://pbs.twimg.com/amplify_video_thumb/1976122456961318914/img/y_iX6afjUpDrtU-W.jpg
Thu Oct 09 2025	0	Not many people know that gum infections can cause heart disease and even stroke. Thatâ€™s why Unilever Indonesia through Pepsodent is focusing on raising this awareness at BKGN 2025. (link)	https://pbs.twimg.com/media/G2yJSdIaEAA71KH.jpg
Thu Oct 09 0 2025	0	Plain water is a simple secret for a healthy life. It helps maintain blood pressure, smooth digestion, improve memory, care for skin, and prevent heart disease. Letâ€™s start the habit of drinking water today! (hashtags) (link)	https://pbs.twimg.com/media/G2xxXC4XYAAZX6K.jpg
Wed Oct 08 2025	4	@Roebanie The health and economic impacts of cigarettes are far greater than revenue from tobacco taxes	https://twitter.com/Roebanie/status/1734567890123456789
Wed Oct 08 2025	0	@an4rchyshooter Excessive instant noodle consumption can harm healthâ€"raising risks of heart disease, high blood pressure, and obesity due to high sodium, saturated fat, and preservatives. But it doesnâ€™t mean you must immediately...	https://twitter.com/an4rchyshooter/status/1734567890987654321

Table 1 presents a snapshot of the raw dataset containing Indonesian health-related tweets from platform X (Twitter). Each row corresponds to a single tweet and includes four main columns: Created at (timestamp), Favorite Count (user-engagement indicator), Text (tweet content, shown here in English translation), and URL (links to the original tweet or attached media/images). The sample texts are diverse—ranging from medical education (e.g., infection and cardiovascular risk information) to healthy-lifestyle reminders and public opinions on smoking—so they illustrate the breadth of discourse the model must handle. The Favorite Count column allows auxiliary analyses of whether more informative tweets attract higher interaction, while the URL column can be used to detect or filter media-bearing posts. Encoding artifacts such as "Letâ€™s" and "doesnâ€™t" are visible in some rows, highlighting the need for special-character normalization (UTF-8) during preprocessing so that the text is clean before being converted into IndoBERT embeddings for classification. In total, 4,075 candidate tweets were collected, and after cleaning and filtering, 3,811 labelled tweets were retained for modeling, as detailed in Table 2.

Table 2. Distribution of labelled tweets per disease category after preprocessing

Category	Number of tweets	Percentage (%)
Hypertension	1,285	33,72
Heart disease	1,273	33,40
Diabetes	1,253	32,88
Total	**3,811**	**100,00**

Table 2 reports the distribution of labelled tweets across the three target disease categories after cleaning and filtering. In total, 3,811 tweets were retained, consisting of 1,285 hypertension-related tweets (33.7%), 1,273 heart-disease tweets (33.4%), and 1,253 diabetes tweets (32.9%). The proportions are therefore close to one-third for each class, indicating only mild class imbalance. This near-balanced distribution is beneficial for supervised learning, because it reduces the risk that the classifiers overfit to a dominant category and allows the use of stratified five-fold cross-validation to obtain more reliable estimates of macro-averaged accuracy, precision, recall, and F1-score.

Table 3. Examples of tweet cleaning and text normalization

No	Raw tweet (before cleaning)	Cleaned text (after preprocessing)
1	A sexually transmitted disease caused by the bacterium Treponema pallidum. It can attack the skin, genitals, and even the nervous system and heart if not treated promptly. Online consultation now: (link) #Syphilis #RajaSinga	A sexually transmitted disease caused by the bacterium Treponema pallidum can attack the skin, genitals, nervous system, and heart if not treated promptly.
2	Not many people know that gum infections can cause heart disease and even stroke. Thatâ€™s why Unilever Indonesia through Pepsodent is focusing on raising this awareness at BKGN 2025. (link)	Not many people know that gum infections can cause heart disease and even stroke. That’s why Unilever Indonesia through Pepsodent is focusing on raising this awareness at BKGN 2025.
3	Plain water is a simple secret for a healthy life. It helps maintain blood pressure, smooth digestion, improve memory, care for skin, and prevent heart disease. Letâ€™s start the habit of drinking water today! (hashtags) (link)	Plain water is a simple secret for a healthy life. It helps maintain blood pressure, smooth digestion, improve memory, care for skin, and prevent heart disease. Let’s start the habit of drinking water today.
4	@Roebanie The health and economic impacts of cigarettes are far greater than revenue from tobacco taxes	The health and economic impacts of cigarettes are far greater than the revenue from tobacco taxes.
5	@an4rchyshooter Excessive instant noodle consumption can harm healthâ€”raising risks of heart disease, high blood pressure, and obesity due to high sodium, saturated fat, and preservatives. But it doesnâ€™t mean you must immediately...	Excessive instant noodle consumption can harm health, raising the risks of heart disease, high blood pressure, and obesity due to high sodium, saturated fat, and preservatives. But it doesn’t mean you must immediately stop eating them.

Table 3 presents several representative examples of raw tweets and their corresponding cleaned versions after preprocessing. The raw texts still contain typical social-media noise such as URLs, hashtags, user mentions, and UTF-8 encoding artifacts (e.g., Thatâ€™s, Letâ€™s, doesnâ€™t). In the cleaned texts, these elements are removed or corrected so that the sentences become grammatically coherent and easier to interpret. For instance, links and hashtags are deleted, user mentions at the beginning of the tweet are removed, and tokens such as Thatâ€™s and Letâ€™s are normalized into That’s and Let’s.

This cleaning step is crucial to ensure that the health-related information contained in each tweet is preserved while unnecessary noise is discarded. By converting noisy raw tweets into standardized text, the preprocessing pipeline produces higher-quality input for the IndoBERT embedding stage and, consequently, for the downstream XGBoost and LightGBM classifiers.

XGBoost

XGBoost implements gradient boosting over decision trees with an emphasis on speed and top-tier accuracy. Trees are added one after another, each one correcting the mistakes left by its predecessors. A learning rate applies shrinkage, so every new tree has a modest impact, which improves generalization. Both L1 and L2 penalties act as regularizations to keep the model from becoming overly complex. The algorithm supports sampling of rows and features to diversify trees and accelerate fitting. Missing values are handled automatically by assigning a default direction at split time. When features are high-dimensional and dense—as with IndoBERT embeddings—its split strategy that efficiently handles numerical feature spaces remains highly effective.

The most influential settings include the number of trees, maximum depth, minimum child weight, subsample ratio, column sampling per tree, and the alpha/lambda regularization terms. Lower learning rates usually require more estimators to achieve peak results. Depth and minimum child weight determine leaf size, helping tune the bias–variance trade-off. Typical practice uses stratified k-fold validation and selects models using macro-F1 in multi-class scenarios. Advantages include robustness, explainability through SHAP or feature importance, and practical training times on a single workstation. Downsides are higher memory needs with extremely many embedding dimensions and a tendency to overfit if regularization is insufficient. Overall, XGBoost makes a reliable baseline and a deployable choice for short-text classification, tabular prediction, and other tasks needing a careful balance of accuracy and efficiency.

1. The XGBoost model minimizes the following objective function:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i^{(t-1)}} + f_t(x_i)\right) + \Omega(f_t) \tag{1}$$

2. Where the regularization term is defined as:

$$\text{Obj}^{(t)} = \sum_{i=1}^{n} L\left(y_i, \widehat{y_i^{(t-1)}} + f_t(x_i)\right) + \Omega(f_t) \tag{2}$$

LightGBM

LightGBM is a tree-boosting library engineered to be fast and lightweight while maintaining competitive accuracy. Instead of splitting on raw feature values, it builds histograms that enable efficient split evaluation and optimized memory usage. Its best-first, leaf-wise growth strategy expands the most promising leaf at each step, allowing the model to capture complex, non-linear relationships even with relatively few trees—therefore, depth limits and sufficient min_data_in_leaf are essential to maintain generalization. Two major optimizations—Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB)—enhance computational efficiency, making LightGBM highly suitable for large, dense feature spaces such as IndoBERT embeddings. The algorithm natively supports missing-value handling, class weights for imbalanced data, early stopping, and monotonic constraints for domain-specific interpretability. Key hyperparameters include num_leaves, learning_rate, max_depth, min_data_in_leaf, feature_fraction, and bagging_fraction, together with L1/L2 regularization to control complexity. When carefully tuned, LightGBM delivers excellent accuracy, stability, and fast training times, demonstrating superior performance over XGBoost in this study, particularly when integrated with IndoBERT embeddings for short-text health content classification.Additive model (prediction as a sum of trees)

$$\widehat{y_i^{(T)}} = \sum_{t=1}^{T} f_t(x_i)\,, \qquad f_t \in \mathcal{F} \tag{3}$$

2. Optimal leaf value (result of minimizing second-order Taylor approximation loss)

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i+\lambda} \tag{4}$$

with gi = ∂ℓ/∂yi and hi = ∂2ℓ/∂yi2 in the previous iteration.

3. Split selection gain (criteria for dividing left–right nodes)

$$\text{Gain} = \frac{1}{2}\left(\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right) \tag{5}$$

with GL/R = ∑gi and HL/R = ∑hi on each side.

RESULT

The comparative evaluation of the two ensemble models, XGBoost and LightGBM, was conducted on the IndoBERT-embedded health-content dataset collected from platform X (Twitter), consisting of three disease categories: Diabetes, Heart Disease, and Hypertension. Both models were trained and validated using stratified five-fold cross-validation to ensure balanced class representation and robust generalization. Model performance was assessed using accuracy, macro-averaged precision, recall, and F1-score.

Table 4. Performance comparison of IndoBERT-based classifiers

Model	Accuracy (%)	Precision Macro (%)	Recall Macro (%)	F1-score Macro (%)
IndoBERT+XGBoost	83.25	83.29	83.25	83.26
IndoBERT+LightGBM	85.26	85.29	85.26	85.27

Table 4 presents the overall performance of the two IndoBERT-based classifiers. The IndoBERT × LightGBM pipeline achieved an accuracy of 85.26%, with macro precision, recall, and F1-score of 85.29%, 85.26%, and 85.27%, respectively. In comparison, IndoBERT × XGBoost obtained an accuracy of 83.25%, and macro precision, recall, and F1-score of 83.29%, 83.25%, and 83.26%. These macro-averaged metrics assign equal weight to each disease category, regardless of its sample size, and therefore provide a fair comparison under mild class imbalance. The results show that LightGBM consistently outperforms XGBoost by about 2 percentage points across all metrics, indicating that LightGBM can exploit the IndoBERT feature space more effectively for multi-class categorization of Indonesian health-related tweets.

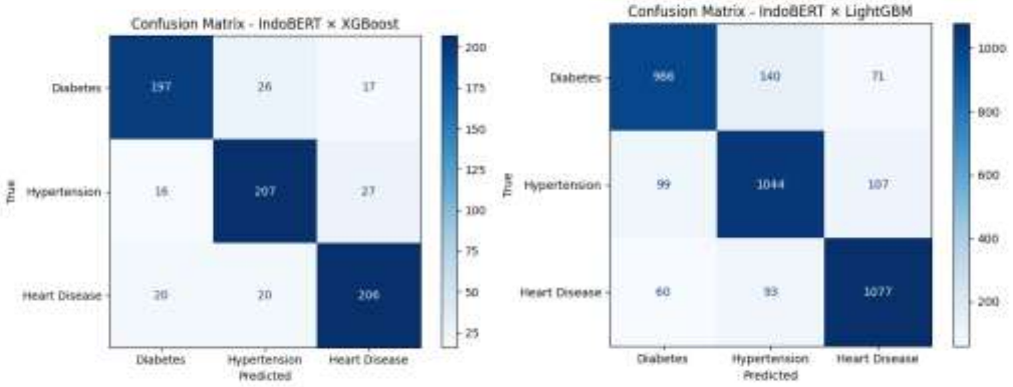

Figure 2. Confusion Matrices of IndoBERT × XGBoost and IndoBERT × LightGBM

Figure 2 illustrates the confusion matrices of the two ensemble models used in this study. The LightGBM model correctly classified most samples in all three categories—Diabetes, Heart Disease, and Hypertension—but produced a few misclassifications between Diabetes and Heart Disease, suggesting minor lexical overlap between related medical terms. In contrast, XGBoost achieved strong precision in the Heart Disease class but showed slightly lower recall in Hypertension, indicating a tendency to under-detect minority cases. Overall, LightGBM demonstrated more balanced performance across all categories, reducing the misclassification rate observed in XGBoost. These visual results are consistent with the numerical evaluation in Table 2, confirming that LightGBM yields higher accuracy and F1-score while maintaining classification stability.

Feature-importance analysis revealed that contextual keywords such as “blood pressure,” “sugar,” “cholesterol,” and “palpitations” were among the most influential predictors across both models. These terms represent the semantic anchors of the IndoBERT embeddings, showing that disease-related vocabulary significantly guides class assignment. The consistency of influential features across XGBoost and LightGBM supports the interpretability of the embedding-based approach, indicating that the models capture meaningful medical context rather than superficial word frequencies. This finding reinforces IndoBERT’s strength in representing nuanced Indonesian-language health discourse and its contribution to boosting model performance. Together, these results demonstrate that integrating linguistic embeddings with tree-based classifiers yields interpretable and domain-relevant predictions.

Overall, the ensemble-based models successfully classified Indonesian health-related tweets with high accuracy, interpretability, and robustness. The integration of IndoBERT embeddings enhanced semantic understanding, while LightGBM provided faster convergence and more consistent generalization under class imbalance. These findings confirm that LightGBM when combined with IndoBERT offers a reliable and explainable framework for health content classification in low-resource language settings. Furthermore, the results align with the study’s objective of developing a transparent and reproducible pipeline for automatic health-content categorization. In summary, the proposed IndoBERT + LightGBM framework demonstrates strong potential for practical deployment in public-health monitoring, issue tracking, and digital health literacy improvement in Indonesia.

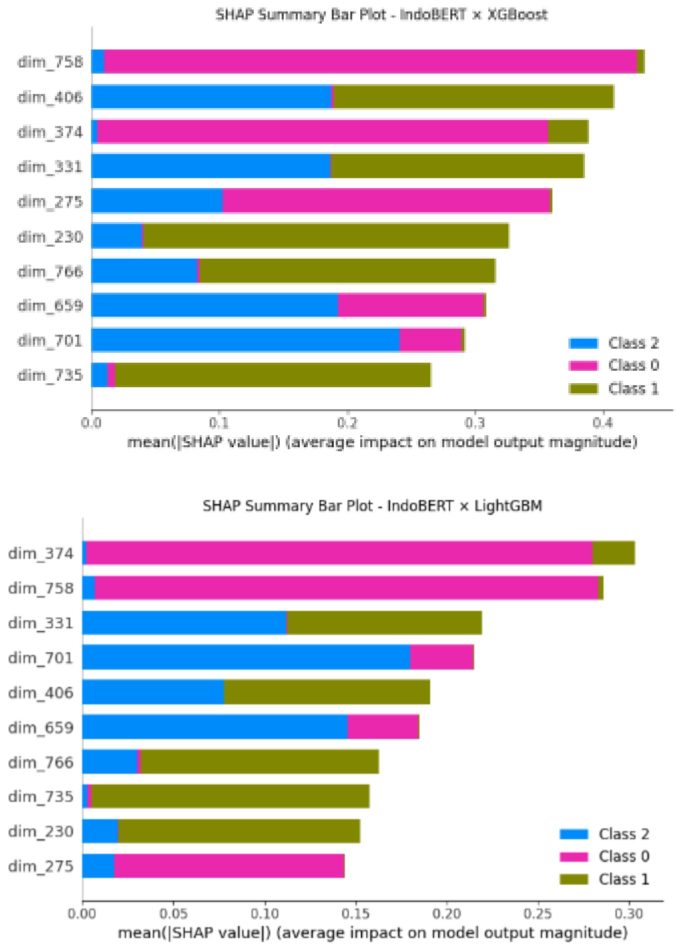

Figure 3. SHAP summary bar plots for IndoBERT-based XGBoost and LightGBM classifiers

Figure 3 shows the SHAP summary bar plots for the IndoBERT-based XGBoost and LightGBM classifiers. Each horizontal bar represents one IndoBERT embedding dimension, and its length reflects the mean absolute SHAP value, that is, the average impact of that dimension on the model output across all tweets. The coloured segments within each bar indicate the contributions for the three disease classes: heart disease, diabetes, and hypertension. Here, Class 0, Class 1, and Class 2 correspond to Heart Disease, Diabetes, and Hypertension, respectively. The plots suggest that both models rely on a similar subset of informative IndoBERT dimensions, but LightGBM assigns slightly higher importance to several of these dimensions, which is consistent with its higher macro-averaged accuracy and F1-score reported in Table 4.

DISCUSSIONS

The comparative analysis revealed that both LightGBM and XGBoost achieved strong performance in classifying Indonesian-language health content based on IndoBERT embeddings. LightGBM recorded the highest overall accuracy (0.8526) and macro-averaged F1-score (0.8527), indicating a superior ability to generalize across the three disease categories—Diabetes, Heart Disease, and Hypertension. XGBoost followed with 0.8325 accuracy and a macro-averaged F1-score of 0.8326, maintaining competitive yet slightly less stable results under mild class imbalance. This difference is consistent with their algorithmic design: LightGBM's leaf-wise tree growth allows deeper local optimization for dense IndoBERT embeddings, while XGBoost's level-wise expansion emphasizes more conservative generalization. These findings confirm that both models are suitable for IndoBERT-based text classification, with LightGBM providing a slightly more efficient and balanced performance.

Feature-importance analysis using SHAP revealed that IndoBERT embedding dimensions associated with contextual keywords such as "blood pressure", "sugar", "cholesterol", and "heart" were among the most influential predictors in both models. These dimensions encode medical semantics that distinguish disease-related discourse within Indonesian-language health content. The consistent appearance of such disease-specific contexts across the LightGBM and XGBoost SHAP plots demonstrates that the models learn meaningful linguistic representations rather than relying purely on shallow frequency statistics. This reinforces IndoBERT's strength in embedding

*name of corresponding author

nuanced health terminology and its contribution to improved model interpretability. From a linguistic standpoint, IndoBERT enhances contextual understanding by mapping medically relevant expressions across varying sentence structures.

Error analysis through confusion matrices shows that most tweets in each class are correctly classified, with residual misclassifications occurring primarily between conceptually related disease categories. Overlaps in health-related vocabulary—such as references to blood sugar, chest discomfort, or blood pressure—can cause tweets about cardiovascular risk factors to be assigned to Diabetes or Heart Disease interchangeably. These patterns suggest that additional lexical disambiguation, more fine-grained annotation guidelines, or class-weight adjustments could further stabilize performance across classes. Threshold calibration beyond the default 0.5 cutoff may also improve detection reliability for borderline cases, especially when tweets mention multiple risk factors or co-morbid conditions in a single sentence. Overall, the results confirm that LightGBM's adaptive structure yields stronger generalization for IndoBERT embeddings, while XGBoost remains a robust alternative for high-dimensional short-text classification.

Despite these promising results, several limitations must be noted. The dataset focuses solely on three disease categories from a single social-media platform, which limits generalization to other domains and communication channels. Informal expressions, mixed language, and lexical ambiguity present ongoing challenges for model accuracy. Future studies should extend the approach to larger and more diverse corpora, incorporate multilingual or domain-specific pretraining such as IndoBERTweet or mBERT, and include temporal analysis to track shifts in health-related discussions over time. Nevertheless, the IndoBERT + LightGBM framework demonstrates strong potential as an interpretable, scalable, and domain-relevant solution for automated health-content categorization in Indonesian-language social media.

CONCLUSION

This study demonstrates that ensemble learning combined with contextual language embeddings can reliably categorize Indonesian-language health content into disease-specific classes across digital platforms. Across head-to-head evaluations, the IndoBERT + LightGBM model achieved the highest performance (accuracy = 85.26%, macro-averaged F1 = 85.27%), while IndoBERT + XGBoost remained competitive (accuracy = 83.25%, macro-averaged F1 = 83.26), showing that both boosting frameworks are effective for short-text classification with high-dimensional linguistic features. Feature-importance analysis confirmed that the models learn meaningful linguistic structures rather than artifacts, identifying influential representations related to symptoms, medical terminology, and behavioural expressions that characterize discussions of diabetes, heart disease, and hypertension. These results validate a transparent, reproducible text-mining pipeline that transforms noisy social-media data into interpretable classifications, enabling its use for public-health monitoring, misinformation detection, and early situational awareness.

The practical benefits are immediate: agencies and health communicators can deploy IndoBERT + LightGBM for automatic content tagging and trend monitoring, use feature-importance insights to explain prediction outcomes, and adjust topic-level thresholds to reduce false alarms in ambiguous posts. Confusion-matrix patterns highlight which categories require further dataset balancing and annotation refinement, while SHAP-based interpretations provide linguistic indicators for improved data curation. Limitations also define clear directions for improvement: the dataset covers only three diseases, collected within a limited period, with mild class imbalance and contextual overlap across topics. Future research should expand coverage to additional diseases and time windows, apply cost-sensitive or focal loss functions to stabilise minority classes, and explore multilingual adaptation for regional variation. We also recommend scheduled model retraining, transparent logging of threshold changes, and continuous calibration against new data to maintain fairness and reliability. With these refinements, the IndoBERT + Boosting framework can evolve into an operational, explainable, and scalable decision-support system for monitoring and managing Indonesian digital health communication.

REFERENCES

Ahn, J. M., Kim, J., & Kim, K. (2023). Ensemble Machine Learning of Gradient Boosting (XGBoost, LightGBM, CatBoost) and Attention-Based CNN-LSTM for Harmful Algal Blooms Forecasting. *Toxins*, *15*(10), 608. https://doi.org/10.3390/toxins15100608

Chen, M., Wu, Y., Wingerd, B., Liu, Z., Xu, J., Thakkar, S., Pedersen, T. J., Donnelly, T., Mann, N., Tong, W., Wolfinger, R. D., & Bao, W. (2024). Automatic text classification of drug-induced liver injury using document-term matrix and XGBoost. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1401810

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Demirtürk, D., Mintemur, Ö., & Arslan, A. (2025). Optimizing LightGBM and XGBoost Algorithms for Estimating Compressive Strength in High-Performance Concrete. *Arabian Journal for Science and Engineering*. https://doi.org/10.1007/s13369-025-10217-7

Hindarto, D., Rachmadi, R. F., Hariadi, M., & Damastuti, F. A. (2025). Contextual Awareness System for Landslide Risk Recommendation in Crypto-Spatial. *2025 International Electronics Symposium (IES)*, 700–706. https://doi.org/10.1109/IES67184.2025.11161195

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. *Proceedings of the 28th International Conference on Computational Linguistics*, 757–770. https://doi.org/10.18653/v1/2020.coling-main.66

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Suherman, E., Hindarto, D., Makmur, A., & Santoso, H. (2023). Comparison of Convolutional Neural Network and Artificial Neural Network for Rice Detection. *Sinkron*, *8*(1), 247–255. https://doi.org/10.33395/sinkron.v8i1.11944

Ranković, N., Ranković, D., Ivanović, M., & Lukić, I. (2024). Explainable data mining model for hyperinsulinemia diagnostics. *Connection Science, 36*(1), 2325496. https://doi.org/10.1080/09540091.2024.2325496

Hindarto, D., Afarini, N., Informatika, P., Informasi, P. S., & Luhur, U. B. (2023). *COMPARISON EFFICACY OF VGG16 AND VGG19 INSECT CLASSIFICATION*. *6*(3), 189–195. https://doi.org/10.33387/jiko.v6i3.7008

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 757–770). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.66

Suherman, E., Hindarto, D., Makmur, A., & Santoso, H. (2023). Comparison of Convolutional Neural Network and Artificial Neural Network for Rice Detection. *Sinkron*, *8*(1), 247–255. https://doi.org/10.33395/sinkron.v8i1.11944

Pringandana, C. G. L., & Kusnawi. (2025). A comparative analysis of hyperparameter-tuned XGBoost and LightGBM for multiclass rainfall classification in Jakarta. *Jurnal Teknik Informatika (JUTIF)*, *6*(4), 2467–2483. https://doi.org/10.52436/1.jutif.2025.6.4.4965

Liu, Y., & Chen, Z. (2025). LightGBM-based human action recognition using sensors. *Sensors, 25*(12), 3704. https://doi.org/10.3390/s25123704

Kabir, J., & Chakraborty, A. (2024). Exploring Explainable Artificial Intelligence: A Comparative Analysis of Interpretability Techniques. *IJARCCE, 13*(3). https://doi.org/10.17148/IJARCCE.2024.13301

Kaeren, K., & Andrianingsih, A. (2025). Analisis sentimen aplikasi LinkAja di Google Play Store menggunakan algoritma Naïve Bayes dan Random Forest. Jurnal Riset dan Aplikasi Mahasiswa Informatika (JRAMI), *6*(2), 438–447. https://doi.org/10.30998/jrami.v6i02.13821