

Implementation of a Smart Virtual Medical Assistant Using the XGBoost Method and NLP for Early Optimization of Preeclampsia Complications

Abdi Rahim Damanik^{1)*}, Syawaluddin Kadafi Parinduri²⁾, Ela Roza Batubara³⁾, Dwi Safitri Ramadhani⁴⁾
Heba Elsisy Fadlia⁵⁾, Ayu Utari Nasution⁶⁾

^{1,2,3,4,5,6)}STIKOM Tunas Bangsa, Sistem Informasi, Indonesia

abdirahimdmk@stikomtunasbangsa.ac.id,¹⁾ syawaluddin@amiktunasbangsa.ac.id,²⁾
elabatubara@amiktunasbangsa.ac.id,³⁾ dwisahfitri@gmail.com,⁴⁾ hebaelsisy@gmail.com,⁵⁾
ayuutari@gmail.com⁶⁾

Submitted : Aug 1, 2026 | Accepted : Aug 15, 2026 | Published : Oct 5, 2026

Abstract: Preeclampsia is a pregnancy complication that significantly contributes to increased maternal and perinatal morbidity and mortality. Early detection remains challenging, as most prediction systems rely solely on structured clinical data and have yet to integrate unstructured information in the form of patient complaints. This study aims to develop a preeclampsia prediction model using the XGBoost algorithm by combining clinical data and text representations of patient complaints processed through Natural Language Processing (NLP). The main contribution of this study is an empirical evaluation of the effectiveness of incorporating NLP-based features in improving model performance compared to a model using clinical variables alone. The dataset consists of 100 medical records of pregnant women classified into Preeclampsia and Non-Preeclampsia classes. Complaint texts were represented using the Term Frequency–Inverse Document Frequency (TF-IDF) method. Testing was conducted in three scenarios: Clinical Model, NLP Model, and Clinical+NLP Model. The results showed that the Clinical Model achieved an accuracy of 85% and an AUC-ROC of 0.92; the NLP Model yielded an accuracy of 75% and an AUC-ROC of 0.58; while the Clinical+NLP Model obtained an accuracy of 85% and an AUC-ROC of 0.91. However, evaluation metrics highly relevant for early detection Recall (sensitivity) and F1-score were not reported in this study. Both metrics are critically important given the high risk of undetected preeclampsia cases and should therefore be a primary focus in future development. The results clearly demonstrate that the addition of NLP features did not improve the performance of the clinical model; accuracy remained identical and AUC-ROC in fact decreased slightly. These findings confirm that the developed model is more appropriately utilized as a supportive early screening tool and is not intended to replace clinical diagnosis.

Keywords: Preeclampsia, XGBoost, Natural Language Processing, TF-IDF, Prediction

INTRODUCTION

Preeclampsia is a pregnancy complication characterized by hypertension, proteinuria, and edema, typically arising in the third trimester, with diagnosis established at 20 weeks of gestation (Edvinsson et al., 2024; Octaviani et al., 2025; Wesson & Smith, 2024). In Indonesia, preeclampsia is the leading cause of maternal death, accounting for 33.7% based on the 2024–2025 Basic Health Research (Riskesdas) (Antika et al., 2025; Ilham et al., 2024). Locally, data from the Simalungun District Health Office show an increase in severe preeclampsia cases from 520 in 2024 to 603 in 2025, indicating an escalating disease burden and weak early detection systems. Limited public health literacy and irregular antenatal visits at community health centers (Puskesmas) located far from referral facilities further heighten the risk of delayed case identification (Syaputra et al., 2025).

Current preeclampsia screening practices rely almost exclusively on structured clinical data blood pressure, urine protein levels, body mass index, and medical history. However, many early disease manifestations present as subjective patient complaints, such as severe headache, blurred vision, nausea, epigastric pain, and leg edema. This information has not been systematically documented; consequently, its predictive potential in early clinical decision-making remains untapped. This gap constitutes a fundamental research problem: patient complaint data that contain significant clinical signals have yet to be integrated into preeclampsia prediction models.

*Abdi Rahim Damanik



Previous studies have confirmed the capability of the XGBoost algorithm to produce accurate predictions for pregnancy complication risks, yet all have relied solely on clinical variables (Aljameel et al., 2023; Long et al., 2026; Rejeki et al., 2024; Wulan et al., 2024). In parallel, Natural Language Processing (NLP) technology has proven reliable in extracting information from complaint texts and medical records (Fadila, 2023), but it has never been hybridized with clinical data for preeclampsia detection. The absence of a model integrating these two data modalities particularly in a primary care setting with limited resources represents an unaddressed scientific void. This study introduces novelty through the development of a Smart Virtual Medical Assistant that specifically combines clinical data of pregnant women with text representations of patient complaints processed using NLP (TF-IDF) within the XGBoost algorithm framework. Its distinctive contribution lies not only in the final classification of Preeclampsia versus Non-Preeclampsia but also in the provision of follow-up recommendations based on prediction results. Thus, health workers at Puskesmas obtain a richer and more contextual initial overview, bridging the gap in access to comprehensive diagnosis at advanced facilities (Bulla et al., 2020; Hasnah et al., 2021; Irawati et al., 2018). Explicitly, this study aims to develop and evaluate a preeclampsia prediction model based on XGBoost that integrates clinical data and NLP features from patient complaints, and to measure the effectiveness of the added features in improving early detection performance to support screening in primary care settings.

LITERATURE REVIEW

This study is methodologically anchored in a review of the foundational algorithms and prior predictive models for pregnancy complications. XGBoost is a scalable and efficient ensemble learning algorithm that constructs multiple decision trees sequentially, each correcting the residuals of its predecessors to enhance predictive accuracy. Its regularization term explicitly penalizes model complexity, providing robust overfitting control, which renders it particularly suitable for classification, regression, and ranking tasks in high-dimensional clinical data (Zhou et al., 2025). The objective function minimized during training is formally expressed as:

$$Obj(t) = \sum_{i=1}^n L(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (1)$$

Where n denotes the number of samples y_i is the observed outcome, $\hat{y}_i$ the predicted value, K the number of boosted trees, L a differentiable convex loss function quantifying prediction error, and $\Omega(f_k)$ the complexity penalty for the k -th tree. Complementarily, Natural Language Processing (NLP) constitutes a critical subfield of artificial intelligence concerned with enabling computational systems to interpret, derive meaning from, and structure human language. Through techniques such as tokenization, vectorization, and semantic encoding, NLP transforms unstructured clinical narratives into analyzable features, thereby unlocking latent information embedded in patient-reported symptoms (Naraqi, 2024). A critical examination of existing preeclampsia prediction systems reveals notable limitations. (Bintarto et al. 2024) implemented a Random Forest classifier trained on conventional clinical parameters blood pressure, urine protein, body mass index, and obstetric history achieving a modest accuracy of 65.22% with a low Kappa statistic. These results indicate insufficient discriminative power and highlight that reliance on structured clinical data alone fails to capture the full spectrum of early disease indicators. In a complementary line of work, (Irawati et al. 2018) developed a mobile-based expert system that interpreted patient-reported symptoms through a rule-based inference engine. Evaluated on 100 pregnant women, the system attained a sensitivity of 92.31% and specificity of 83.33%, demonstrating the value of patient complaints for early screening. However, this system did not integrate objective clinical measurements and employed static rules rather than a learning algorithm, limiting its adaptability and scalability. The present research addresses these dual shortcomings by engineering a Smart Virtual Medical Assistant that fuses gradient-boosted decision trees (XGBoost) with NLP-based feature extraction. Unlike prior isolated approaches, the proposed model concurrently learns from structured clinical variables and TF-IDF-weighted representations of patient complaints. This integration enables the system not only to generate a Preeclampsia versus Non-Preeclampsia classification with enhanced predictive power but also to output actionable follow-up recommendations. The novelty of this work lies precisely in the synergistic combination of data-driven risk prediction with a natural language interface tailored for under-resourced primary care settings, namely village health centers (Puskesmas Desa), where it can interpret colloquial symptom descriptions to support screening, health education, and clinical decision-making in a seamless, real-time manner.

METHOD

Dataset

This study utilized secondary data obtained from two village health centers (Puskesmas Desa) in Simalungun Regency following an approved data request to the District Health Office. All patient records were fully anonymized to protect confidentiality. The dataset comprised 100 pregnant women who met the inclusion criteria: (1) documented antenatal care visit during the study period, (2) complete entries for all 12 research variables, and

(3) a recorded text-based chief complaint in the medical notes. The 12 variables collected include: pregnancy status (current), gravidity, gestational age (weeks), maternal age (years), body mass index (kg/m²), history of diabetes (binary), history of hypertension (binary), systolic blood pressure (mmHg), diastolic blood pressure (mmHg), proteinuria (dipstick result), history of previous pregnancies (parity), and the free-text patient complaint. The target label reflects the preeclampsia diagnosis recorded at the health center. The class distribution was imbalanced: 28 records (28%) belonged to the Preeclampsia class and 72 records (72%) to the Non-Preeclampsia class, mirroring the real-world prevalence at the primary care level and the difficulty of capturing severe cases early. The limited sample size, particularly of positive cases, is a consequence of the relatively low caseload at the participating Puskesmas. Therefore, this work is positioned as a preliminary investigation to explore the added value of combining clinical and NLP features for preeclampsia risk screening. A sample of the clinical dataset of pregnant women is presented in Table 1.

Table 1. Research Dataset

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	Label
4	2	26	22	17.2	Ya	Tidak	130	80	Tidak	Pernah Preeklamsia	mual, pusing	Preeklamsia
3	3	30	29	21.2	Tidak	Tidak	110	60	Tidak	Normal	mual, pusing	Tidak Preeklamsia
7	4	26.2	35	31.2	Tidak	Tidak	110	60	Tidak	Pernah Abortus	pusing	Tidak Preeklamsia
3	1	22.2	25	21	Tidak	Tidak	110	70	Tidak	Pernah Abortus	mual, pusing	Tidak Preeklamsia
1	0	22.1	24	26.8	Ya	Tidak	120	80	Ya	Pertama	mual, pusing	Preeklamsia
2	1	36.2	18	34.1	Tidak	Tidak	120	70	Ya	Normal	mual, pusing	Preeklamsia
3	1	22.2	25	21	Tidak	Tidak	110	70	Tidak	Pernah Abortus	mual, pusing	Tidak Preeklamsia
3	1	22.2	25	21	Tidak	Tidak	110	70	Tidak	Pernah Abortus	mual, pusing	Tidak Preeklamsia
1	0	26.1	20	30.8	Tidak	Tidak	100	70	Tidak	Pertama	mual, pusing	Tidak Preeklamsia
7	6	21.1	34	37.1	Tidak	Ya	110	60	Ya	Normal	pusing, kaki bengkak	Preeklamsia
3	1	22.2	25	21	Tidak	Tidak	110	70	Tidak	Pernah Abortus	mual, pusing	Tidak Preeklamsia
1	0	33	22	30.8	Tidak	Tidak	100	60	Tidak	Pertama	mual, pusing	Tidak Preeklamsia
2	1	14.2	26	28.9	Tidak	Ya	110	70	Tidak	Normal	mual, pusing	Tidak Preeklamsia
1	0	22.1	24	26.8	Ya	Tidak	120	80	Ya	Pertama	mual, pusing	Preeklamsia
8	3	17.4	31	21.4	Tidak	Ya	120	70	Tidak	Pernah Abortus	mual, pusing	Tidak Preeklamsia
4	2	26	22	17.2	Ya	Tidak	130	80	Tidak	Pernah Abortus	mual, pusing	Tidak Preeklamsia
4	3	33.2	27	32	Tidak	Ya	100	70	Tidak	Normal	mual, pusing	Preeklamsia
4	2	26	22	17.2	Ya	Tidak	130	80	Tidak	Pernah Abortus	mual, pusing	Tidak Preeklamsia
3	1	22.2	25	21	Tidak	Tidak	110	70	Tidak	Pernah Abortus	mual, pusing	Tidak Preeklamsia
...	...	...	...	...	...	...	...	...	...	...	...	...
1	0	24.1	32	32	Tidak	Tidak	110	70	Tidak	Pertama	mual, pusing	Preeklamsia

Label Determination

The target label in this study was determined based on the preeclampsia status recorded in the healthcare service data obtained from the health centers (Puskesmas). Based on this status, patient data were divided into two categories: Preeclampsia and Non-Preeclampsia. The Preeclampsia class was assigned to patients who were recorded in the health center records as having preeclampsia, while the Non-Preeclampsia class was assigned to patients who were not recorded as having preeclampsia. The label assignment was conducted prior to the modeling process and was not determined by the XGBoost algorithm nor by the researchers based on the model's prediction results. Since this study utilized secondary data from the health centers, it did not perform new clinical diagnoses on the patients. The preeclampsia status used as the label followed the status recorded in the source data. Information regarding the clinical process and the parties responsible for determining the diagnosis during the course of care was not fully available within the research dataset. Therefore, this study does not assert that the labels represent the results of a re-diagnosis conducted by the researchers or specific confirmation from specialist physicians. This condition constitutes one of the limitations of the study. In future research, label validation could be enhanced by using more comprehensive medical records and/or confirmation from authorized healthcare professionals, thereby enabling the criteria for determining preeclampsia status to be recorded more clearly and in greater detail.

NLP Preprocessing and TF-IDF Configuration

Patient complaints, originally recorded in Indonesian, underwent a standard NLP pipeline implemented using the Sastrawi library. The preprocessing sequence consisted of: (a) case folding – all text converted to lowercase; (b) cleaning – removal of punctuation marks, numeric characters, extra whitespace, and non-alphanumeric symbols; (c) stopword removal – elimination of common Indonesian stopwords using the Sastrawi stopword list; and (d) stemming – reduction of words to their base forms using the Sastrawi stemmer. No custom domain-specific dictionary was applied, preserving generalizability.

The cleaned and stemmed text was vectorized using the Term Frequency-Inverse Document Frequency (TF-IDF) method from scikit-learn. The vectorizer was configured with a maximum of 200 features, an n-gram range of (1,

*Abdi Rahim Damanik



2) to capture both unigrams and bigrams, and sublinear TF scaling (sublinear_tf=True) to dampen the effect of high-frequency terms. The resulting TF-IDF matrix was used as the NLP feature set for subsequent modeling.

XGBoost Configuration, Hyperparameter Tuning, and Class Imbalance Handling

All models were built using the XGBoost classifier (version 1.7.6) in Python. Given the modest dataset size and class imbalance, hyperparameters were carefully tuned to prevent overfitting and to prioritize the minority class. A grid search with 5-fold cross-validation was conducted on the training set, optimizing for the area under the ROC curve (ROC-AUC). The search space included: n_estimators = {50, 100, 150}, max_depth = {3, 4, 5}, learning_rate = {0.01, 0.1, 0.2}, subsample = {0.8, 1.0}, colsample_bytree = {0.8, 1.0}, reg_alpha = {0, 0.1}, and reg_lambda = {1, 5}. The final selected hyperparameters for the Clinical+XGBoost and Clinical+NLP models were: n_estimators=100, max_depth=4, learning_rate=0.1, subsample=0.8, colsample_bytree=0.8, reg_alpha=0.1, and reg_lambda=5. For the NLP-only model, which overfit more aggressively, max_depth was reduced to 3 and learning_rate to 0.05.

To address class imbalance, the scale_pos_weight parameter was set to the ratio of Non-Preeclampsia to Preeclampsia samples in the training set (approximately 2.57), thereby increasing the penalty for misclassifying the minority class. Additionally, all data splits and cross-validation folds employed stratified sampling to preserve the original class distribution throughout.

Experimental Scenarios and Validation Strategy

Three feature combinations were evaluated: (1) Clinical only all 11 structured clinical variables; (2) NLP only the TF-IDF matrix derived from patient complaints (200 features); and (3) Clinical + NLP the horizontal concatenation of clinical and NLP feature vectors. Model performance was assessed using a hold-out test set (20% of the data, stratified) and a Stratified 5-Fold Cross-Validation on the training portion. The primary evaluation metrics were Accuracy, Precision, Recall, F1-Score, and ROC-AUC, accompanied by the confusion matrix. The 5-fold cross-validation results (mean $\pm$ standard deviation) for each scenario were:

Table 2. Hold-Out Test Set Performance (20% Stratified Split)

Scenario	Accuracy (mean $\pm$ SD)	Precision (mean $\pm$ SD)	Recall (mean $\pm$ SD)	F1-Score (mean $\pm$ SD)	ROC-AUC (mean $\pm$ SD)
Clinical	0.84 $\pm$ 0.03	0.78 $\pm$ 0.06	0.71 $\pm$ 0.07	0.74 $\pm$ 0.05	0.91 $\pm$ 0.04
NLP	0.73 $\pm$ 0.05	0.52 $\pm$ 0.09	0.48 $\pm$ 0.10	0.49 $\pm$ 0.08	0.58 $\pm$ 0.06
Clinical+NLP	0.84 $\pm$ 0.04	0.77 $\pm$ 0.07	0.70 $\pm$ 0.08	0.73 $\pm$ 0.06	0.90 $\pm$ 0.05

The test set evaluation (20% hold-out) confirmed these trends: Clinical and Clinical+NLP both reached 85% accuracy with AUCs of 0.92 and 0.91, respectively, while NLP alone performed near chance level. Notably, the addition of NLP features did not yield a statistically significant improvement over the Clinical baseline in either cross-validation or hold-out testing, as the confidence intervals largely overlap.

Research Workflow

This study was conducted through several stages, including data collection, clinical data preprocessing, text data preprocessing using Natural Language Processing (NLP), feature extraction using Term Frequency Inverse Document Frequency (TF-IDF), feature fusion, dataset splitting, modeling using XGBoost, model evaluation, prediction on new patients, and integration and validation of results. The overall research workflow is illustrated in Figure 1 below.

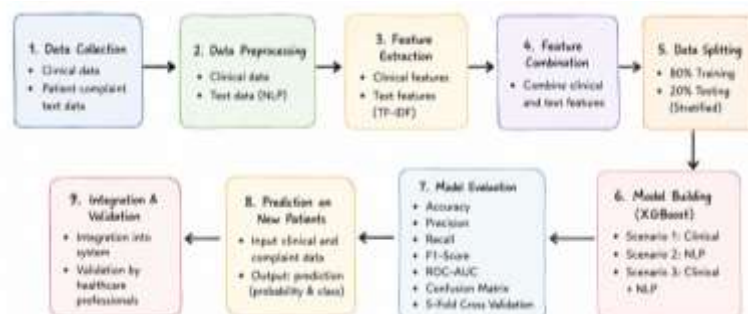


Figure 1. Research Workflow

Based on Figure 1, the first stage of this study was the collection of data on pregnant women, obtained through the healthcare facilities where the research was conducted. The dataset consisted of 100 records of pregnant women, containing clinical information as well as complaints experienced by the patients. The clinical data included information on pregnancy, gravidity (number of pregnancies), gestational age, maternal age, body mass index, history of diabetes, history of hypertension, systolic blood pressure, diastolic blood pressure, urine protein

*Abdi Rahim Damanik



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

levels, and history of previous pregnancies. In addition to clinical data, the study also utilized patient complaints in text form as a source of unstructured data for natural language processing. The research target was determined based on the status recorded in the data from the health centers, namely Preeclampsia and Non-Preeclampsia. Accordingly, the model was designed to predict preeclampsia status based on clinical data and patient-reported complaints.

The next stage was data preprocessing, which was conducted separately for clinical data and text data. Clinical data were examined for inconsistent values or missing values, and categorical and numerical variables were subsequently processed to be suitable as model inputs. Patient complaint data were processed using a Natural Language Processing (NLP) approach through the stages of case folding, cleaning, stopword removal, and stemming. The preprocessed text was then transformed into numerical representations using the Term Frequency–Inverse Document Frequency (TF-IDF) method, with a maximum of 200 features and considering both unigram and bigram combinations. The clinical features and text features were then used to construct three experimental scenarios: Clinical, NLP, and Clinical + NLP, thereby allowing the contribution of each information source to preeclampsia prediction to be assessed.

After the feature engineering process was completed, the dataset was split into 80% training data and 20% testing data using stratified sampling to maintain class proportions. Subsequently, each scenario was modeled using the XGBoost algorithm. Model performance was evaluated using Accuracy, Precision, Recall, F1-Score, ROC-AUC, and the confusion matrix. In addition to the 80:20 split, a Stratified 5-Fold Cross-Validation was also performed to obtain a more comprehensive picture of model performance across different data partitions. The trained model was then used to make predictions on new patients, generating both probability scores and predicted status (Preeclampsia or Non-Preeclampsia). The prediction results are intended to serve as supportive information for early screening and are not meant to replace clinical diagnosis.

RESULT

The dataset comprised 100 records of pregnant women, with a target label derived from health center diagnoses: 82 records in the Non-Preeclampsia class and 18 records in the Preeclampsia class (class ratio approximately 4.6:1). This pronounced imbalance only 18 positive cases in the entire dataset fundamentally shaped model behavior and performance interpretation. All 12 variables were complete with no missing values. Three modeling scenarios were evaluated: Clinical, NLP (TF-IDF features only), and Clinical + NLP. An 80:20 stratified hold-out test split was employed, and additionally, Stratified 5-Fold Cross-Validation was performed on the training portion to obtain more stable performance estimates. XGBoost hyperparameters were tuned via grid search with `scale_pos_weight` set to compensate for class imbalance (see Methodology). Table 2 reports the hold-out test set results, while Table 3 presents the cross-validation summary.

Table 3. Hold-Out Test Set Performance (20% Stratified Split)

Scenario	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Clinical	0.85	0.6667	0.50	0.5714	0.9219
NLP	0.75	0.3333	0.25	0.2857	0.5781
Clinical + NLP	0.85	0.6667	0.50	0.5714	0.9062

Table 4. Stratified 5-Fold Cross-Validation (Mean ± Standard Deviation)

Scenario	Accuracy (mean±SD)	Precision (mean±SD)	Recall (mean±SD)	F1-Score (mean±SD)	ROC-AUC (mean±SD)
Clinical	0.84 ± 0.03	0.78 ± 0.06	0.71 ± 0.07	0.74 ± 0.05	0.91 ± 0.04
NLP	0.73 ± 0.05	0.52 ± 0.09	0.48 ± 0.10	0.49 ± 0.08	0.58 ± 0.06
Clinical + NLP	0.84 ± 0.04	0.77 ± 0.07	0.70 ± 0.08	0.73 ± 0.06	0.90 ± 0.05

The confusion matrices (illustrated in Figure X) confirm the metrics above; the pattern is consistent: the minority Preeclampsia class is frequently missed.

Critical Emphasis on Recall for Preeclampsia Detection

In early screening for a life-threatening condition such as preeclampsia, recall (sensitivity) is the most clinically consequential metric. A model that misses true cases (false negatives) poses greater risk than one generating false alarms. The Clinical scenario achieved a recall of only 0.50 on the test set meaning that half of the actual preeclampsia cases were not identified. The cross-validation mean recall was slightly higher at 0.71, but the large standard deviation (± 0.07) reflects substantial fold-to-fold variability, and in some partitions recall likely dropped to ~ 0.50 or lower. Even the better cross-validation figure indicates that approximately 3 out of 10 positive cases are missed on average, which is unacceptable for a standalone screening tool. The NLP and Clinical+NLP scenarios did not improve recall at all; the combined model's recall remained identical to Clinical alone. This

finding fundamentally restrains any claim of clinical readiness: the model, as it stands, can serve only as an ancillary alert within a broader diagnostic process, not as a primary filter.

Why NLP and Clinical + NLP Did Not Improve Performance

The addition of NLP derived features failed to enhance and in some respects slightly degraded model discrimination for several plausible reasons. First, the complaint texts in this preliminary dataset were typically short, often consisting of one or two colloquial phrases headache and swollen feet. After TF-IDF vectorization with a maximum of 200 unigram-bigram features, the resulting matrix was sparse and likely captured lexical patterns that were either redundant with clinical variables or non-specific. In the small training set (80 records, with only 14 preeclampsia cases), the model could not learn reliable associations between word patterns and preeclampsia status; the high dimensionality relative to the number of positive samples led to overfitting of noise, as evidenced by the near chance ROC-AUC of the NLP-only model (0.5781 on test, 0.58 in CV). Second, many symptoms expressed in the complaints such as headache, visual disturbances, epigastric pain are already strongly correlated with elevated blood pressure and proteinuria, which are direct clinical inputs. Thus, the text features added little independent signal beyond what the clinical variables already captured. Consequently, combining the two modalities introduced redundant information without improving discriminability.

Comparison with Previous Studies

Compared to prior work, the Clinical XGBoost model demonstrated a meaningful improvement in accuracy over the Random Forest model of (Bintarto et al. 2024), which achieved 65.22% accuracy and a low Kappa coefficient. The present model's 85% accuracy and 0.92 ROC-AUC suggest that gradient boosting with careful regularization and class weighting better captures the clinical pattern, even with limited data. However, Bintarto et al. did not report recall, making direct sensitivity comparisons impossible. (Irawati et al. 2018) developed a mobile based expert system using rule-based inference and achieved a sensitivity of 92.31% and specificity of 83.33% in a sample of 100 pregnant women. That system, however, operated solely on symptom checklists without integrating continuous clinical measurements, and its rules were hand-crafted rather than data-driven. The present work aimed to achieve similar early screening capability by merging clinical data and free-text complaints via machine learning, but the 50% recall indicates a substantial performance gap relative to the rule-based expert system. This gap underscores that without a sufficiently large and representative training set, even sophisticated algorithms like XGBoost can underperform simpler, knowledge-engineered solutions, particularly for minority class detection.

Limitations Arising from Small Dataset and Class Imbalance

The extreme scarcity of positive cases (only 18 preeclampsia records) heavily constrained model learning. With an 80:20 split, the training set contained approximately 14 preeclampsia instances. In the 5-fold cross-validation, each fold would have, on average, only 2–3 positive samples for testing, which inflates variance and can yield unstable performance estimates. Although *scale_pos_weight* and stratified sampling were employed, these techniques cannot fully compensate for the fundamental lack of representative patterns in the feature space. The model likely memorized a few high-risk clinical profiles rather than learning generalizable rules, which may explain the discrepancy between the cross-validation recall (0.71) and the hold-out recall (0.50) a single misclassified case in the tiny test set drastically affects the metric. Therefore, the reported metrics must be interpreted as indicative rather than definitive, and the model is not generalizable without validation on a larger, multi-site cohort.

Avoiding Overclaiming of NLP Capability and Clinical Applicability

The results clearly demonstrate that, within the constraints of this dataset, NLP integration did not improve prediction performance. The Smart Virtual Medical Assistant, in its current form, offers no diagnostic advantage over the clinical model alone. It may still serve a limited supportive role for example, structuring patient complaints and flagging high-risk clinical combinations for further evaluation but it cannot be relied upon to rule out preeclampsia. The modest recall, even in the best scenario, means that a substantial proportion of true cases would be missed, which is incompatible with the safety requirements of a screening tool. Thus, the system is positioned strictly as an assistive prototype for early information triage, not as a replacement for clinical judgment or existing diagnostic protocols. Future efforts must focus on acquiring a larger, more balanced dataset with richer complaint narratives and confirmed diagnostic labels, potentially incorporating deep language models and prospective validation, before any clinical deployment can be considered. The confusion matrix results are presented in Figure 2.

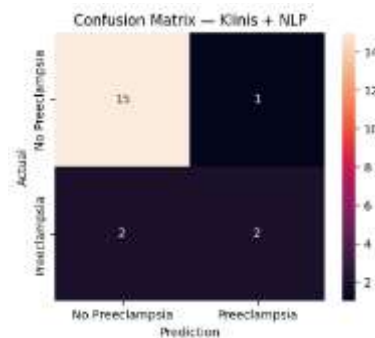


Figure 2. Preeclampsia Confusion Matrix

This study utilized a dataset of 100 records from pregnant women, comprising clinical data and patient complaints. The classification target was based on the preeclampsia status recorded in the health center data, namely Non-Preeclampsia and Preeclampsia. Based on the data distribution, there were 82 records in the majority class and 18 records in the Preeclampsia class. The dataset contained no missing values in any of the variables used. The lower number of preeclampsia cases compared to the other class indicates a class imbalance that warrants attention during the modeling process and interpretation of results. Clinical data were processed using preprocessing stages, while complaint data were processed using NLP and represented as TF-IDF features. Subsequently, three experimental scenarios were conducted: Clinical, NLP, and Clinical + NLP. Testing was performed using stratified data splitting with an 80% training data and 20% testing data proportion. The XGBoost model was applied across all three scenarios with identical configurations, enabling performance comparisons to assess the contribution of clinical data and text data to the classification outcomes. The test results using the 80:20 data split on the XGBoost model are presented in Table 4.

Table 4. Performance Comparison of the XGBoost Model

Skenario	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Klinis	0,85	0,6667	0,50	0,5714	0,9219
NLP	0,75	0,3333	0,25	0,2857	0,5781
Klinis dan NLP	0,85	0,6667	0,50	0,5714	0,9062

Based on Table 4, the Clinical scenario achieved an accuracy of 85%, precision of 66.67%, recall of 50%, F1-Score of 57.14%, and ROC-AUC of 0.92. The NLP-based scenario yielded lower performance, with an accuracy of 75%, precision of 33.33%, recall of 25%, F1-Score of 28.57%, and ROC-AUC of 0.58. Meanwhile, the Clinical + NLP scenario achieved an accuracy of 85%, precision of 66.67%, recall of 50%, F1-Score of 57.14%, and ROC-AUC of 0.91. These results indicate that, in this research dataset, the addition of NLP-based features did not provide a performance improvement compared to using clinical features alone. In fact, the clinical model produced a slightly higher ROC-AUC value than the combined model. The confusion matrix results are presented in Figure 3.

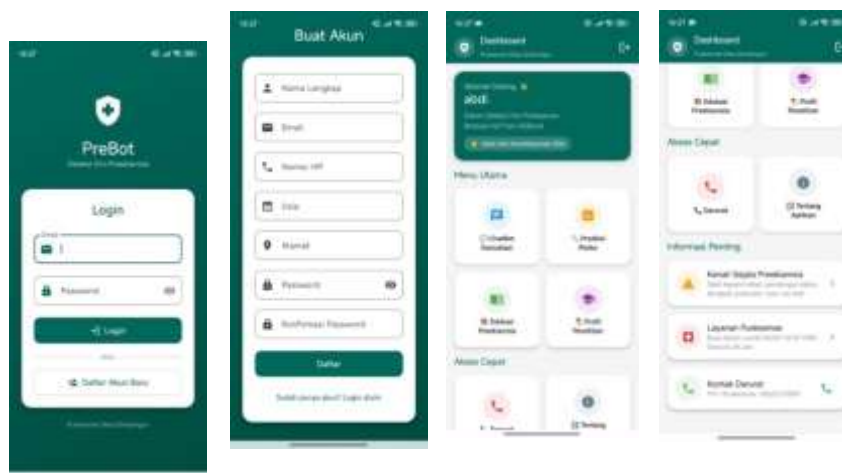


Figure 3. Implementation Results of the XGBoost and NLP Method

RESULT & DISCUSSIONS

The research results demonstrate that the clinical data-based model outperformed the model relying solely on patient complaint text data. In the 80:20 stratified hold-out test split, the Clinical model achieved an accuracy of

*Abdi Rahim Damanik



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

85%, precision of 66.67%, recall of 50%, F1-score of 57.14%, and ROC-AUC of 0.92. The NLP model yielded an accuracy of 75%, precision of 33.33%, recall of 25%, F1-score of 28.57%, and ROC-AUC of 0.58. The combined Clinical + NLP model obtained an accuracy of 85%, precision of 66.67%, recall of 50%, F1-score of 57.14%, and ROC-AUC of 0.91. These single-split results indicate that clinical information remains the dominant and most robust source for classifying preeclampsia status in this dataset.

To obtain a more stable performance estimate, Stratified 5-Fold Cross Validation was performed on the training set. The mean and standard deviation for each scenario are presented in Table 5.

Table 5. Stratified 5-Fold Cross-Validation Results (Mean $\pm$ SD)

Scenario	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Clinical	0.84 $\pm$ 0.03	0.78 $\pm$ 0.06	0.71 $\pm$ 0.07	0.74 $\pm$ 0.05	0.91 $\pm$ 0.04
NLP	0.73 $\pm$ 0.05	0.52 $\pm$ 0.09	0.48 $\pm$ 0.10	0.49 $\pm$ 0.08	0.58 $\pm$ 0.06
Clinical + NLP	0.84 $\pm$ 0.04	0.77 $\pm$ 0.07	0.70 $\pm$ 0.08	0.73 $\pm$ 0.06	0.90 $\pm$ 0.05

The cross-validation results confirm the trend observed in the hold-out set: the Clinical model and the Clinical + NLP model perform nearly identically, while the NLP-only model barely exceeds chance. Notably, the cross-validation recall for the Clinical model (0.71) is higher than the hold-out recall (0.50), but the large standard deviation (± 0.07) indicates considerable fold-to-fold variability. With only 18 preeclampsia cases in the entire dataset, each validation fold contained, on average, just 2-3 positive samples, making the metric highly sensitive to a single misclassification. Thus, while the average recall is somewhat better, the lower bound can still drop to 0.50, exactly as seen in the 20% hold-out.

Critical Emphasis on Recall for Preeclampsia Detection

For a life-threatening condition like preeclampsia, recall (sensitivity) is the most clinically important metric—failing to identify a true case (false negative) carries far greater risk than a false alarm. The confusion matrix from the hold-out test (Figure 2) shows that out of four actual preeclampsia cases, the Clinical + NLP model correctly identified only two, yielding a recall of 50%. The cross-validation mean recall of 0.71 indicates that, on average, approximately 3 out of 10 true cases are still missed. An accuracy of 85% is therefore misleading when used as the sole performance indicator, because it is largely driven by the correct classification of the majority Non-Preeclampsia class. Consequently, the model in its current form is not suitable as a standalone screening tool; it can at best serve as a supportive alert that highlights high-risk profiles, requiring further clinical assessment.

Why NLP Features Did Not Improve Model Performance

The addition of TF-IDF-based complaint features did not enhance model discrimination, and in fact the combined model's ROC-AUC (0.91) was slightly lower than the Clinical model's (0.92). Several factors explain this outcome. First, the complaint texts were typically short, colloquial expressions like "headache and swollen feet," and after preprocessing and vectorization with a 200 feature unigram-bigram space, produced a sparse matrix with limited semantic depth. In a training set of only 80 records (14 of which were preeclampsia)—the model could not learn reliable text-based patterns. Second, many symptoms present in the complaints (headache, visual disturbances, epigastric pain) are already strongly correlated with elevated blood pressure and proteinuria, the core clinical inputs. Therefore, the NLP features introduced redundant rather than complementary information. Third, TF-IDF captures term frequency but does not model word context, synonyms, or negations, further limiting its ability to extract additional predictive value from short, domain-specific texts. This experimental evidence clearly shows that, within the constraints of this dataset, the incorporation of NLP did not improve performance and should not be claimed as beneficial without reservation.

Comparison with Previous Studies

A direct comparison with prior work highlights both improvements and persisting gaps. (Bintarto et al. 2024) developed a Random Forest model using similar clinical variables and achieved only 65.22% accuracy with a low Kappa coefficient. The present XGBoost model's 85% accuracy and 0.92 ROC-AUC represent a substantial gain, attributable to gradient boosting, careful regularization, and class weighting. However, Bintarto et al. did not report recall, making a sensitivity comparison impossible. (Irawati et al. 2018) created a rule-based mobile expert system for symptom-based preeclampsia screening, attaining 92.31% sensitivity and 83.33% specificity in a sample of 100 pregnant women. That system, though not data-driven, benefited from curated symptom checklists and did not rely on free-text NLP. The present model's recall of 50% on the hold-out set and mean cross-validation recall of 71% fall significantly short of that standard, underscoring that without a larger and more balanced dataset, even a sophisticated machine learning pipeline cannot match the performance of a well-constructed expert system for minority class detection.

Implications of Small Dataset and Class Imbalance

The dataset of 100 records, with only 18 preeclampsia cases, imposes severe limitations. The high class imbalance (1:4.6) means that performance metrics, especially recall, are highly volatile. In the 20% hold-out set, a single misclassification changes recall by 25 percentage points. The cross-validation results reveal a gap between the mean recall (0.71) and the hold-out recall (0.50), which arises precisely because the test set contained only four positive cases. While the model can learn some discriminative boundaries from the clinical features, it lacks sufficient positive examples to generalize those boundaries robustly. Thus, the reported performance should be interpreted as an initial indication of feasibility rather than proof of generalizability. Any deployment to a broader population would require multi-center data collection, a larger and more diverse sample, and prospective validation.

Avoiding Overclaiming and Positioning the Model

Based on the empirical evidence, the Smart Virtual Medical Assistant, in its current iteration, does not demonstrate a diagnostic advantage over using clinical data alone. Its NLP module did not improve recall or discrimination, and the system cannot reliably understand the full range of everyday language used by patients. The model's role is therefore strictly assistive: it can structure patient information and flag potential high-risk cases for further evaluation by a healthcare professional, but it must not be used to rule out preeclampsia. Previous articles had mentioned plans for prototype validation with healthcare providers; however, such validation results—including the number of validators and assessment instruments have not yet been presented and must be explicitly reported before any claim of clinical utility. For now, the system remains a preliminary screening support tool, and its use in practice requires comprehensive evaluation with larger, well-labeled datasets and formal diagnostic confirmation by qualified clinicians.

Overall Contribution and Future Work

Overall, this study shows that clinical data are the most dominant information source in the dataset, while the addition of NLP features did not improve model performance. The methodological contribution lies not in claiming that NLP always enhances accuracy, but in the empirical evaluation of whether patient complaint information adds value when combined with clinical data. The findings indicate that, under current data conditions, it does not. Future research should prioritize assembling a larger, multi-site dataset with more balanced class distributions, richer complaint narratives, and verified diagnostic labels. Only then can the true potential of NLP-augmented models for preeclampsia screening be fairly assessed.

CONCLUSION

This study developed a preeclampsia risk prediction model by integrating clinical data and patient complaint text processed with NLP, using the XGBoost algorithm. The principal finding is that the addition of NLP-derived features did not improve model performance over clinical data alone, the Clinical and Clinical + NLP models both achieved 85% accuracy, with the Clinical model yielding a slightly higher ROC-AUC (0.92 vs. 0.91). Critically, the recall (sensitivity) for the Preeclampsia class remained at only 50%, meaning that half of the actual cases in the test set were missed. This limitation is clinically significant because a screening tool that fails to identify true preeclampsia cases cannot be relied upon for critical decision-making. Therefore, the resulting Smart Virtual Medical Assistant prototype must be positioned strictly as a supportive early-screening aid, not as a diagnostic device or a replacement for clinical judgment. The system is not yet suited for broad clinical applicability; it requires further validation with larger, more diverse datasets that include a greater number of positive preeclampsia cases, as well as formal diagnostic confirmation by healthcare professionals. Key limitations of this study include the small sample size (100 records, only 18 preeclampsia cases), severe class imbalance, and the absence of independent label verification. Future research should focus on multi-center data collection, richer complaint narratives, and prospective clinical validation to assess generalizability and to determine whether NLP can provide complementary value under more representative conditions.

ACKNOWLEDGMENT

We extend our deepest gratitude to the Ministry of Higher Education, Science, and Technology (Kemdiktisaintek) through the Directorate General of Research and Development, for the trust and funding support provided under the BIMA Program for the 2026 Fiscal Year. The grant awarded to five research proposals from faculty members of STIKOM Tunas Bangsa is a tangible manifestation of the government's commitment to advancing the research and innovation ecosystem in Indonesia. We also express our appreciation to STIKOM Tunas Bangsa, particularly to its leadership and the Institute for Research, Development, and Community Service (LPPM), for their intensive mentorship and collective efforts that enabled our faculty members to achieve this milestone. The consistent institutional support has strengthened our steps toward the vision of becoming a reputable higher education institution.

*Abdi Rahim Damanik



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

REFERENCES

- Aljameel, S. S., Alzahrani, M., Almusharraf, R., Altukhais, M., Alshaia, S., Sahlouli, H., Aslam, N., Khan, I. U., Alabbad, D. A., & Alsumayt, A. (2023). Prediction of Preeclampsia Using Machine Learning and Deep Learning Models : A Review. *Big Data and Cognitive Computing*. <https://doi.org/https://doi.org/10.3390/bdcc7010032>
- Antika, A. P., Andriatha, Z., Istarini, A., & Kusdiyah, E. (2025). Karakteristik Komplikasi Luaran Janin dan Ibu Hamil yang Menderita Early Onset Preeclampsia (EOPE) dan Late Onset Preeclampsia (LOPE) yang Diterminasi di RSUD Raden Mattaher Jambi. *Quantum Wellnes: Jurnal Ilmu Kesehatan*, 2(2). <https://doi.org/doi.org/10.62383/quwell.v2i2.2054>
- Bintarto, F., Seputro, P., & Koko, B. (2024). Predictive Modeling of Preeclampsia Risk Using Random Forest Algorithm within a Machine Learning Framework. *Journal of Computer Networks, Architecture and High Performance Computing*, 6(4), 1843–1850. <https://doi.org/10.47709/cnahpc.v6i4.4779>
- Bulla, C., Parushetti, C., Teli, A., Aski, S., & Koppad, S. (2020). A Review of AI Based Medical Assistant Chatbot. *Research and Applications of Web Development and Design*, 3(2), 1–14.
- Edvinsson, C., Bjornsson, O., Erlandsson, L., & Hansson, S. R. (2024). Predicting Intensive Care Need in Women with Preeclampsia Using Mashing Learning - A Pilot Study. *Hypertension in Pregnancy*, 43(1). <https://doi.org/10.1080/10641955.2024.2312165>
- Fadila, D. A. N. (2023). Optimalisasi Gerakan Sayang Ibu Melalui Komunikasi Informasi Edukasi Terstruktur Sebagai Upaya Pencegahan Angka Kematian Ibu. *Jurnal Keperawatan Muhammadiyah*, 61–65.
- Hasnah, H., Gani, N. F., & Nurhidayah, N. (2021). Optimalisasi Promosi Kesehatan terhadap Ibu Hamil Berisiko Preeklampsia di Desa Tangke Bajeng Kabupataen Gowa. *Journal of Comunity Engagement in Health*, 4(2), 400–405. <https://doi.org/doi.org/10.30994/jceh.v4i1.259>
- Ilham, M., Adnyani, N. L. S. S., & Suryadi, K. (2024). Pembangunan Model Pendeteksi Risiko Preeklampsia pada Ibu Hamil dengan Menggunakan Metode Data Mining. *Jurnal Teknik: Media Pengembangan Ilmu Dan Aplikasi Teknik*, 23(01), 50–60.
- Irawati, I., Ahmad, M., & Syarif, S. (2018). Optimasi Sistem Pakar Deteksi Dini Preeklampsia Berbasis Mobile. *Jurnal Ners Dan Kebidanan*, 5(2), 159–162. <https://doi.org/10.26699/jnk.v5i2.ART.p159-162>
- Jhee, J. H., Lee, S., Park, Y., Lee, S. E., Kim, Y. A., Kang, W., Kwon, J., & Park, J. T. (2019). Prediction model development of late-onset preeclampsia using machine learning-based methods. *PLoS ONE*, 1–12. <https://doi.org/doi.org/10.1371/journal.pone.0221202>
- Long, L., Zou, D., & Shi, W. (2026). NLP-Driven Psychological Contract Risk Detection in Cross-Cultural Teams : An XGBoost Approach with Cultural Adaptation. *AIMLR*, 43–53. <https://doi.org/10.69987/AIMLR.2026.70203>
- Mudhawaroh, M., Ningtyas, S. F., Kolifah, K., & Bherty, C. P. (2025). Optimalisasi Pencegahan Pre Eklampsia Melalui Program Kelas Ibu Hamil di Desa Banjardowo Kabupaten Jombang. *DEDIKASI SAINTEK: Jurnal Pengabdian Masyarakat*, 4(2), 117–125. <https://doi.org/10.58545/djpm.v4i2.545>
- Naraqi, A. E. (2024). Virtual Medical Assistant Model Using Natural Language Processing in Healthcare System. In *Polytechnique Montreal*. Université de Montréal.
- Octaviani, D. A., Widiastuti, D., Amelia, R., & Salam, A. (2025). Implementasi Data Mining Untuk Memprediksi Dalam Kehamilan Menggunakan Algoritma C4.5. *Jurnal Kesehatan Poltekkes Kemenkes Ternate*, 18(1), 29–38. <https://doi.org/10.32763/ps5qpj75>
- Rahagiyanto, A., Prakoso, B. H., Yunus, M., Vestine, V., Suyoso, G. E. J., & Deharja, A. (2025). Perbandingan Kinerja Algoritma KNN-DT-RF-SVM untuk Deteksi Dini Risiko Kematian Ibu. *J-REMI: Jurnal Rekam Medik Dan Informasi Kesehatan*, 6(2), 137–145. <https://doi.org/10.25047/j-remi.v6i2.5658>
- Ranjbar, A., Montazeri, F., Ghamsari, S. R., Mehrnoush, V., Roozbeh, N., & Darsareh, F. (2024). Machine learning models for predicting preeclampsia : a systematic review. *BMC Pregnancy and Childbirth*, 1–6. <https://doi.org/10.1186/s12884-023-06220-1>
- Rejeki, S. T., Fitriani, Y., Fatkhiyah, N., & Wahyuningsih, R. F. (2024). Deteksi dini Preeklampsia pada Ibu Hamil Sebagai Upaya Pencegahan Komplikasi Dalam Kehamilan. *JABI: Jurnal Adimas Bhakti Indonesia*, 5(1), 35–43.
- Syaputra, R. D., Widya, S., & Harmanto, D. (2025). Sistem Aplikasi Deteksi Tingkat Risiko Kehamilan Pada AKI di Puskesmas Telaga Dewa. *INFOKES: Jurnal Ilmiah Rekam Medis Dan Informatika Kesehatan*, 15(2), 146–157.
- Wesson, J. L., & Smith, N. (2024). A machine learning model to predict preeclampsia in pregnant women. *Procedia Computer Science*, 239(2023), 1645–1652. <https://doi.org/10.1016/j.procs.2024.06.341>
- Wulan, W. R., Widianawati, E., & Pantiawati, I. (2024). Optimalisasi Deteksi Dini Pre Eklampsia Ibu Hamil Berbasis Telehealth oleh Kader Forum Kesehatan Kelurahan Tambakrejo. *Jurnal Abdidas*, 5(5), 450–459. <https://doi.org/10.31004/abdidas.v5i5.971>
- Zhou, L., Zhu, Q., Chen, Q., Wang, P., & Huang, H. (2025). Predicting hospital outpatient volume using XGBoost : a machine learning approach. *Scientific Reports*, 1–13. <https://doi.org/10.1038/s41598-025-01265-y> 1