INTERNATIONAL JOURNAL OF SOCIAL SCIENCE, EDUCATIONAL, ECONOMICS, AGRICULTURE RESEARCH AND TECHNOLOGY (IJSET) | ISSN (E): 2827-766X
Website: https://ijset.org/ | Volume 5 No. 12 (2026)

HYBRID INDOBERT AND RANDOM FOREST MODEL FOR ANAMNESIS INFORMATION EXTRACTION AND EMERGENCY DEPARTMENT TRIAGE CLASSIFICATION

Khairul Azman[1*], Rafiyandi[2], Mohammad Syafrullah[3]
[1,2] Computer Science, Faculty of Information Technology,
Budi Luhur University, Jakarta, Indonesia
E-mail: azman.khairul67@gmail.com[1*]; rafiyandi1905@gmail.com[2], mohammad.syafrullah@budiluhur.ac.id[3]

Received: 24/10/2026| **Revised:** 20/10/2026| **Accepted:** 02/11/2026| **Published:** 16/11/2026

Abstract

Accurate emergency department (ED) triage depends on timely interpretation of clinical information, yet Indonesian anamnesis records are often unstructured. This study developed a hybrid IndoBERT and Random Forest framework to extract clinical entities and classify patients into Green, Yellow, and Red triage categories. IndoBERT was fine-tuned for named entity recognition (NER) using a beginning, inside, outside annotation scheme. Extracted entities were converted into 27 features and combined with six vital-sign measurements and the Glasgow Coma Scale, producing 34 classifier inputs. After cleaning, the triage dataset contained 47,807 records; the NER dataset contained 29,870 unique sentences after deduplication. The NER model achieved a weighted F1-score of 0.9997. On 9,562 test records, Random Forest achieved an accuracy of 0.9662 and a weighted F1-score of 0.9666, exceeding the predefined performance thresholds. A web-based prototype demonstrated the integration of clinical data entry, entity extraction, and triage prediction. The framework provides a structured approach to using Indonesian clinical narratives for triage decision support. Broader external validation and prospective clinical evaluation are required before operational deployment.

Keywords: *Anamnesis; Emergency department triage; IndoBERT; Named entity recognition; Random Forest*

INTRODUCTION

The emergency department (ED) provides immediate care for patients with acute and potentially life-threatening conditions. Triage establishes treatment priorities by classifying patients according to the severity of their conditions. Accurate classification is essential for patient safety and the effective allocation of staff and facilities. Errors may delay treatment for critically ill patients and increase the risk of adverse outcomes (Johannessen, 2016; Soola et al., 2022). High patient volumes and time constraints place considerable demands on triage assessment and documentation. Decisions often depend on clinicians' experience and judgment, which can produce variation among practitioners (Johannessen, 2016; Soola et al., 2022). Electronic health records also require clinical information to be documented efficiently, accurately, and consistently (Ministry of Health of the Republic of Indonesia, 2021). These conditions create a need for methods that can convert free-text anamnesis into structured information to support triage decisions.

This study proposes a pipeline that extracts clinical entities from Indonesian ED anamnesis records through named entity recognition (NER), combines the extracted information with vital signs and Glasgow Coma Scale (GCS) scores, and predicts triage categories using Random Forest. NER provides an intermediate representation of information that would otherwise remain in unstructured text (IBM, 2023). Random Forest was selected for its robustness and ability to handle heterogeneous features (Breiman, 2001). The study also contributes an annotated Indonesian clinical NER dataset for research on ED anamnesis. The objectives were to develop an NER model for extracting relevant clinical information from Indonesian anamnesis, implement a Random Forest classifier using the extracted entities and clinical measurements, and evaluate the performance of the integrated framework. The broader aim was to support more accurate, consistent, and efficient triage assessment through the combined use of natural language processing (NLP) and machine learning.

LITERATURE REVIEW

Advances in artificial intelligence have created opportunities to support clinical decisions using both unstructured and structured data. NLP can extract information from patient narratives, while machine learning can use the resulting representations to classify patients by urgency (Rajkomar et al., 2019; IBM, 2023). These approaches have the potential to support faster and more consistent decision-making in emergency care (Alzubaidi et al., 2021).

Previous studies have used machine learning, sometimes combined with NLP, to predict clinical outcomes in emergency settings. Chang et al. (2022) compared five models for identifying low-severity patients with a short discharge length of stay. CatBoost achieved an AUC of 0.755 in internal validation, while XGBoost achieved 0.761 in external validation. Porto (2024) reviewed the use of nursing notes and chief complaints in triage prediction and reported benefits from combining NLP with algorithms such as XGBoost and deep neural networks. Yao et al. (2021) combined convolutional neural networks (CNNs), bidirectional gated recurrent units (BiGRUs), and attention, reporting an accuracy of 0.83 and an AUROC of 0.88 for hospital-admission prediction in the National Taiwan University Hospital dataset. Lee et al. (2023) reported an AUROC of 0.9974 for ED trauma mortality prediction using AdaBoost and an AdaBoost plus LightGBM ensemble.

Much of the earlier work used English-language medical datasets and clinical text directly, without an explicit intermediate step for extracting medical entities. Research on Indonesian ED anamnesis remains limited. Although previous work has investigated the combination of clinical narrative information and vital signs (Chang et al., 2022; Porto, 2024), a framework that converts Indonesian narratives into structured clinical entities could provide a useful basis for triage classification.

Two performance hypotheses were specified. H1 proposed that IndoBERT, fine-tuned on Indonesian ED anamnesis, would achieve a weighted NER F1-score of at least 0.93 on the test set despite variation in psychiatric and non-psychiatric terminology. H2 proposed that a Random Forest classifier combining NER-derived features, vital signs, and GCS would achieve test accuracy of at least 0.90 and a weighted F1-score of at least 0.93. The Results and Discussion section evaluates these criteria.

METHOD

Research Framework

This study developed an emergency department (ED) triage decision-support system using natural language processing (NLP), machine learning, anamnesis text, and patient clinical data. The research framework comprised two connected stages. The first extracted clinically relevant information from anamnesis text through IndoBERT-based named entity recognition (NER). The second combined the extracted information with patients' vital signs to classify triage levels using Random Forest. Figure 1 presents the research framework.

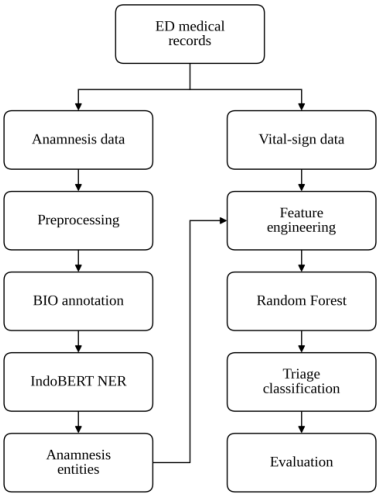

Figure 1. Research framework for anamnesis information extraction and ED triage classification

As shown in Figure 1, the process began with ED patient records containing anamnesis text and vital-sign data. The anamnesis text was preprocessed and annotated using the beginning, inside, outside (BIO) scheme to construct the NER dataset. This dataset was used to fine-tune IndoBERT for token classification. The trained model then identified and extracted clinical entities from the anamnesis text.

The clinical information extracted through NER was combined with patients' vital-sign data to construct input features for the Random Forest classifier. The classifier determined each patient's triage level from the symptoms described in the anamnesis and the patient's physiological condition. This approach incorporated information from unstructured clinical narratives alongside structured clinical measurements.

The research stages comprised data preparation, anamnesis text preprocessing, BIO annotation, IndoBERT training and inference, clinical entity extraction, integration with vital-sign data, feature construction, and Random Forest triage classification. The NER model and triage classifier were subsequently evaluated using appropriate metrics to assess their ability to extract clinical information and classify patient acuity.

Dataset

The dataset consisted of ED patient records containing anamnesis text, clinical measurements, and triage labels. Anamnesis text provided the unstructured information used for clinical entity extraction, while vital signs and Glasgow Coma Scale (GCS) scores provided structured information for triage classification. The classification target comprised three categories: GREEN, YELLOW, and RED.

The initial dataset contained 51,292 records. Exploratory analysis identified inconsistent triage labels, including GREEN and Green Zone, YELLOW and Yellow Zone, and RED and Red Zone, as well as records with missing labels (NaN). Six records were labeled Black Zone and were excluded because this category contained very few observations. Data cleaning and label standardization removed records with missing labels and Black Zone labels, then mapped the remaining label variants to GREEN, YELLOW, and RED.

Following these procedures, 47,807 records were retained for modeling: 17,987 GREEN records, 24,462 YELLOW records, and 5,358 RED records. YELLOW was the largest class, while RED was the smallest.

Table 1. Dataset distribution by triage class

Triage class	Number of records	Percentage
GREEN	17,987	37.63%
YELLOW	24,462	51.17%
RED	5,358	11.20%
Total	47,807	100%

A separate dataset was constructed for the NER task from anamnesis text that had undergone cleaning, normalization, and BIO annotation. Before deduplication, this dataset contained 39,976 anamnesis sentences and 409,624 tokens. Each token was labeled according to the predefined clinical entity schema, which covered physical and psychiatric symptoms. The annotated dataset provided the basis for fine-tuning IndoBERT for token classification.

The NER component served as an intermediate step for converting clinical information in anamnesis text into structured features. Entities extracted by the model were mapped to triage classification features and combined with vital-sign measurements and GCS scores.

Ethical Statement and Data Governance

The dataset was obtained from patient anamnesis records collected at three hospitals: RS Jiwa Aceh, RSUD dr. Fauziah Bireuen, and RSU Ali Kasim Gayo Lues. All records were anonymized before processing and annotation to remove patient-identifying information. Data cleaning, normalization, and BIO annotation were performed manually by the researcher and validated by medical personnel. No individually identifiable information was retained or used during model development, evaluation, or prototype implementation.

Data Preprocessing

Anamnesis Text Preprocessing

Preprocessing prepared the anamnesis text for clinical information extraction through NER. It reduced inconsistencies, standardized text representations, and prepared the data for annotation and IndoBERT training. The stages comprised data cleaning, label standardization, text normalization and standardization, case folding, and tokenization, as shown in Figure 2.

OPEN ACCESS

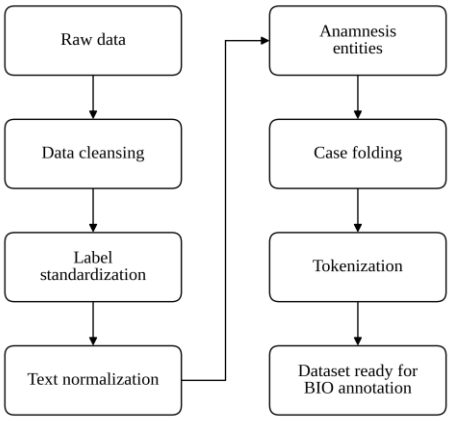

Figure 2. Data preprocessing stages

As shown in Figure 2, preprocessing began with data cleaning to identify and remove records unsuitable for modeling. Records without triage labels and the small number of Black Zone records were excluded to establish a consistently labeled dataset.

Label standardization harmonized the different names used for triage categories. GREEN and Green Zone were mapped to GREEN; YELLOW and Yellow Zone to YELLOW; and RED and Red Zone to RED. This procedure ensured that each record used a consistent class label.

The anamnesis text was then normalized and standardized to reduce variations in word forms and the spelling of clinical terms. Case folding converted the text to lowercase to produce a more uniform representation.

The final stage was tokenization, which divided anamnesis text into units that could be processed by the NLP model. Tokenization was performed according to IndoBERT's requirements and supported BIO annotation. For illustration, the Indonesian sentence "pasien mengeluh sesak napas sejak dua jam lalu" ("the patient reports shortness of breath for the past two hours") can be represented as the word tokens "pasien," "mengeluh," "sesak," "napas," "sejak," "dua," "jam," and "lalu." The resulting cleaned and standardized text was used to construct the NER dataset.

Vital Sign Preprocessing

In addition to anamnesis text, the study used structured clinical data comprising blood pressure, pulse rate, respiratory rate, body temperature, oxygen saturation, and GCS scores.

Clinical preprocessing ensured that each variable had a consistent format suitable for use as a numerical classification feature. Blood pressure readings recorded as combined systolic and diastolic values were separated into two variables. For example, a reading of 120/80 was represented by a systolic value of 120 and a diastolic value of 80.

Pulse rate, respiratory rate, body temperature, and oxygen saturation were represented numerically. GCS scores represented the patient's level of consciousness. The clinical component therefore comprised seven numerical features: systolic blood pressure, diastolic blood pressure, pulse rate, respiratory rate, body temperature, oxygen saturation, and GCS.

After preprocessing, the tokenized anamnesis text was used to construct the NER dataset, while the standardized clinical measurements were retained as numerical features. Information extracted through NER was subsequently combined with the vital-sign and GCS features to form the Random Forest classification dataset.

Named Entity Recognition

NER identified words and phrases describing symptoms and clinical conditions relevant to triage assessment. It converted information in unstructured anamnesis text into structured representations for feature extraction and engineering. The NER process comprised BIO annotation, dataset construction, IndoBERT training, and clinical entity extraction.

Entity Definition and BIO Annotation

Before model training, clinical entities in the anamnesis text were annotated at the token level using the BIO scheme. The B label marked the first token of an entity, the I label marked subsequent tokens belonging to the same entity, and the O label marked tokens outside an entity.

The annotated entities represented physical symptoms and psychiatric conditions relevant to triage. Examples included shortness of breath, bleeding, seizures, fainting, falls, accidents, trauma, fever, cough, aggression, restlessness, hallucinations, disorientation, and suicide attempts.

OPEN ACCESS

For example, in the Indonesian phrase "pasien mengeluh sesak napas" ("the patient reports shortness of breath"), "sesak" receives the label B-SESAK and "napas" receives I-SESAK, while the remaining tokens receive O. This annotation allows the model to learn entity boundaries and types from their surrounding context.

The annotated dataset initially contained 39,976 sentences and 409,624 tokens. After validation and deduplication, 29,870 unique sentences were retained and divided into training, validation, and test subsets.

IndoBERT Model

IndoBERT was used to extract clinical entities from Indonesian anamnesis text. Based on the bidirectional encoder representations from transformers (BERT) architecture, IndoBERT is pretrained on an Indonesian-language corpus. In this study, it was applied to token classification, assigning an entity label to each token in the anamnesis text.

The pretrained model was fine-tuned using the BIO-annotated anamnesis dataset. Fine-tuning adapted its pretrained language representations to the characteristics of clinical text and the entity categories used in this study.

NER Training

The NER dataset was divided into training, validation, and test subsets using an 80:10:10 ratio. The subsets contained 23,896, 2,987, and 2,987 sentences, respectively. The training set was used to fine-tune IndoBERT, the validation set to monitor training and select the best model, and the test set to evaluate performance on data not used during training.

The model was trained for 5 epochs with a batch size of 8 and a learning rate of 3×10^{-5}. During training, it learned associations among tokens, sentence context, and BIO labels to identify clinical entities in anamnesis text.

NER Inference and Entity Extraction

After fine-tuning, IndoBERT was used for inference on anamnesis text. The model predicted a BIO label for each token, after which tokens belonging to the same entity were combined to reconstruct the clinical entity.

The resulting list of entities for each anamnesis record was transformed into structured features. These NER-derived features were combined with vital-sign measurements and GCS scores for triage classification using Random Forest.

Feature Extraction and Engineering

Feature extraction and engineering converted the clinical entities identified in anamnesis text into structured inputs for the Random Forest classifier. Token-level NER outputs were transformed into features representing symptom presence, symptom counts, and indicators of potentially urgent conditions. These features were then combined with patients' vital signs and GCS scores.

Anamnesis-derived features were organized into psychiatric features, non-psychiatric features, and aggregate and red flag features. The psychiatric and non-psychiatric groups contained eight and 15 binary features, respectively, indicating whether particular symptoms were present in the text. Two aggregate features represented symptom counts, and two red flag features indicated potentially urgent conditions.

Psychiatric Features

The psychiatric features represented symptoms, behaviors, and mental conditions identified in the anamnesis. The eight indicators were aggression, raging behavior, restlessness, hallucinations, disorientation, decreased consciousness, threats to harm others, and suicide attempts.

Each psychiatric feature was encoded as 0 or 1. A value of 1 indicated that the corresponding symptom or condition was identified in the anamnesis, while 0 indicated that it was absent. Binary encoding provided a simple, interpretable numerical representation of the NER output for the Random Forest model.

For example, if an anamnesis record described aggressive and raging behavior, f_aggressive and f_rampage were each assigned a value of 1. Other psychiatric features were assigned 0 when their corresponding entities were not identified.

Non-Psychiatric Features

The non-psychiatric group contained 15 binary indicators representing cough, fever, shortness of breath, nausea, vomiting, abdominal pain, headache, dizziness, weakness, bleeding, seizures, fainting, falls, accidents, and trauma.

All non-psychiatric features were also encoded as binary variables. A value of 1 indicated that the corresponding symptom was identified by NER, while 0 indicated that it was absent. This representation enabled information from anamnesis text to be integrated with numerical clinical measurements during classification.

Aggregation and Red Flag Features

Aggregate features represented the number of symptoms identified in each group. The psy_feature_count feature recorded the number of psychiatric features assigned a value of 1 for a patient, while non_psychiatry_feature_count recorded the number of detected non-psychiatric features.

The two symptom-count features can be expressed mathematically as follows:

$$\text{psy_feature_count} = \sum_{i=1}^{8} f_{\text{psy,i}} \quad (1)$$

$$\text{non_psychiatry_feature_count} = \sum_{j=1}^{15} f_{\text{nonpsy,j}} \quad (2)$$

The first sum is taken over the eight psychiatric binary features, while the second is taken over the 15 non-psychiatric binary features.

Two binary red flag features, redflag_psychiatry and redflag_non_psychiatry, indicated conditions considered potentially urgent. The former represented severe psychiatric symptoms, while the latter represented potentially urgent medical conditions.

Anamnesis processing therefore produced 27 features: eight psychiatric features, 15 non-psychiatric features, two aggregate features, and two red flag features.

Clinical Feature Integration

The anamnesis-derived features were combined with seven clinical measurements: systolic blood pressure, diastolic blood pressure, pulse rate, respiratory rate, body temperature, oxygen saturation, and GCS.

These seven numerical clinical features were integrated with the 27 anamnesis-derived features at the patient level. Each row of the classification dataset represented a patient case combining symptom information from anamnesis with measurements of the patient's physiological condition.

The resulting feature vector contained 34 features, comprising 27 anamnesis-derived features and seven clinical features. It can be represented as follows:

$$X = [F_{\text{psychiatric}}, F_{\text{nonpsychiatric}}, F_{\text{aggregation}}, F_{\text{redflag}}, F_{\text{vital}}] \quad (3)$$

The psychiatric block contains eight binary features, the non-psychiatric block contains 15 binary features, the aggregate block contains two numerical features, the red flag block contains two binary features, and the clinical block contains seven numerical features.

The final structured dataset combined semantic information from anamnesis text with patient clinical measurements. It was used as input to the Random Forest classifier to predict GREEN, YELLOW, or RED triage categories.

Random Forest Classification

Random Forest was used to classify patient triage levels from features extracted from anamnesis records and clinical measurements. The algorithm was selected because it can accommodate both binary and numerical features and model nonlinear relationships among them. The input variables comprised features derived from anamnesis text, vital signs, and the Glasgow Coma Scale (GCS).

The classification dataset contained 34 features: 27 derived from anamnesis text and seven clinical features. The text-derived variables included psychiatric, non-psychiatric, aggregation, and red-flag features. The clinical variables were systolic blood pressure, diastolic blood pressure, pulse rate, respiratory rate, body temperature, oxygen saturation, and GCS. All 34 features were used to predict the patient's triage category.

Dataset Partitioning

The dataset was divided into training and test subsets to evaluate the model on data not used during training. The training subset was used to build the Random Forest classifier, while the test subset was reserved for assessing its predictive performance.

The split allocated 80% of the data to training and 20% to testing. Partitioning was performed after feature construction so that both subsets had the same feature structure. The test data were excluded from model fitting and used only for the final evaluation.

Random Forest Model Training

The Random Forest model was trained on the combined anamnesis and clinical features. The algorithm constructs multiple decision trees using different data samples and feature subsets. The predictions of the individual trees are then combined through majority voting to determine the final class.

Each decision tree learned relationships between patient triage levels, clinical measurements, and symptoms extracted from anamnesis text. Using multiple trees enabled the model to represent a wider range of data patterns than a single decision tree.

The model was trained using three target categories: Green, Yellow, and Red. The resulting classifier accepted a feature vector for a new patient and returned a predicted triage category.

OPEN ACCESS

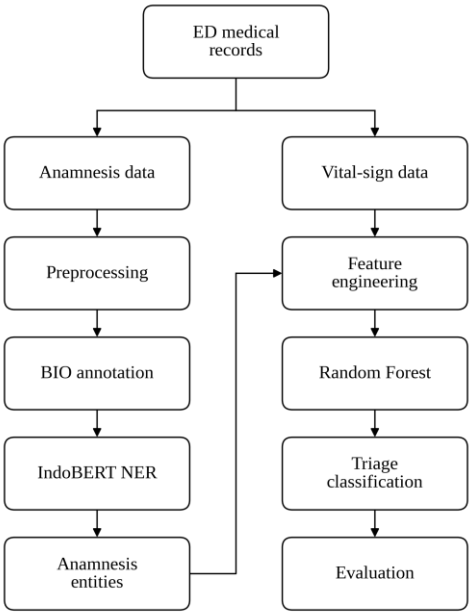

Figure 3. Classification process using Random Forest

Triage Level Prediction

During prediction, the test records were presented to the Random Forest model as feature vectors constructed during feature extraction and engineering. For each record, the model returned one of three triage categories: Green, Yellow, or Red.

The classification process can be expressed as follows:

$$\hat{y} = \mathrm{RF}(X) \quad (4)$$

where X is the patient feature vector containing 34 features, RF denotes the Random Forest model, and ŷ denotes the predicted triage category.

The predictions were compared with the actual triage labels in the test dataset to evaluate model performance. The following section reports the results using the relevant classification metrics.

RESULTS AND DISCUSSION

Named Entity Recognition Model Performance

Most clinical entities achieved precision, recall, and F1-scores of 1.0000. Shortness of breath, bleeding, and suicide attempt were among the entities recognized with perfect scores in the evaluation. Two entities showed slightly lower performance: decreased consciousness, with an F1-score of 0.9984, and weakness (lemas), with an F1-score of 0.9948. Weakness had the lowest score, indicating a small tendency to confuse this entity with semantically related complaints such as dizziness.

Table 2. Evaluation results for the named entity recognition model

Entity	Precision	Recall	F1-score	Support
Aggressiveness	1.0000	1.0000	1.0000	280
Cough	1.0000	1.0000	1.0000	289
Fever	1.0000	1.0000	1.0000	285
Disorientation	1.0000	1.0000	1.0000	278
Restlessness	1.0000	1.0000	1.0000	272
Hallucinations	1.0000	1.0000	1.0000	272
Fall	1.0000	1.0000	1.0000	262
Accident	1.0000	1.0000	1.0000	264

OPEN ACCESS

Entity	Precision	Recall	F1-score	Support
Seizures	1.0000	1.0000	1.0000	255
Weakness	0.9965	0.9931	0.9948	288
Rage	1.0000	1.0000	1.0000	283
Nausea	1.0000	1.0000	1.0000	281
Vomiting	1.0000	1.0000	1.0000	269
Abdominal pain	1.0000	1.0000	1.0000	285
Decreased consciousness	0.9967	1.0000	0.9984	305
Bleeding	1.0000	1.0000	1.0000	292
Fainting	1.0000	1.0000	1.0000	266
Dizziness	1.0000	1.0000	1.0000	281
Headache	1.0000	1.0000	1.0000	297
Shortness of breath	1.0000	1.0000	1.0000	286
Suicide attempt	1.0000	1.0000	1.0000	303
Threat to harm others	1.0000	1.0000	1.0000	302
Trauma	1.0000	1.0000	1.0000	286

These results indicate that IndoBERT effectively extracted clinical entities from Indonesian anamnesis text and transformed unstructured narratives into structured clinical information. The findings are consistent with previous work describing the contextual representations provided by transformer-based language models for named entity recognition (Devlin et al., 2019; IBM, 2023).

Token-Level and Entity-Level Evaluation

Because the overall NER performance was close to perfect, an additional evaluation examined whether the model correctly recognized entity boundaries as well as individual token labels. Table 3 compares token-level evaluation, in which each BIO-tagged token is assessed independently, with entity-level evaluation, in which a prediction is correct only when both the full entity span and its label match the ground truth.

Table 3. Token-level and entity-level NER evaluation

Evaluation level	Precision	Recall	F1-score
Token level	0.99995	0.99995	0.99995
Entity level	0.9997	0.9997	0.9997

The closely matching results at both evaluation levels indicate that the model consistently identified the start and end boundaries of multi-token clinical entities, including shortness of breath and decreased consciousness. This consistency reduces concern that the high F1-score resulted solely from a lenient token-level evaluation scheme.

Out-of-Distribution Evaluation

An intra-domain out-of-distribution (OOD) evaluation examined generalization to clinically similar but linguistically more varied sentences. Sentence embeddings were generated for all training and test sentences using a multilingual sentence-transformer model, and cosine similarity was calculated between each test sentence and the training set. The 10% of test sentences with the lowest maximum similarity to the training data were selected as the OOD subset. These were typically longer sentences containing multiple co-occurring entities. The subset was evaluated using the same pipeline as the in-distribution test.

OPEN ACCESS

Table 4. In-distribution and out-of-distribution NER performance

Entity	F1-score In-distribution	F1-score OOD	Change in F1-score
Weakness	0.9948	0.9888	−0.0060
Decreased consciousness	0.9984	1.0000	+0.0016
Other entities	≈ 1.0000	≈ 1.0000	≈ 0.0000

Performance remained stable for most entities, which were lexically explicit and unambiguous. Weakness (LEMAS) was the only entity with a notable decrease in F1-score, falling by 0.0060 as recall declined to 0.9778. This pattern is consistent with confusion between weakness and semantically related complaints in longer, more complex sentences. The in-distribution, token-level, and OOD evaluations did not reveal a marked deterioration in performance. However, the relatively homogeneous clinical domain means that some overfitting to domain-specific patterns cannot be excluded, and further cross-hospital or cross-domain testing is warranted.

Qualitative Error Analysis

A qualitative examination of predictions for individual anamnesis sentences complemented the quantitative evaluation. The model correctly identified explicit, frequently occurring entities and recognized multiple co-occurring or multi-token entities within a sentence. Two representative error cases were observed. In the first example, translated as “The patient appears confused and speaks incoherently,” the model correctly identified the term for confusion as DISORIENTATION but incorrectly assigned AGGRESSIVE followed by ANXIETY to the speech-related wording. This error illustrates the difficulty of representing complex behavioral descriptions as a coherent semantic unit. In the second example, translated as “The patient complained of chest pain since this morning,” the model predicted NYERI_PERUT (abdominal pain). The annotation scheme did not include a separate NYERI_DADA (chest pain) entity, and the model assigned the most frequent pain-related category encountered during training. This case illustrates a limitation in distinguishing pain location when the relevant category is absent from the annotation scheme and supports expanding the entity schema in future work.

Triage Classification Performance

The NER-derived features were combined with patients' vital signs and GCS. The Random Forest model was evaluated on 9,562 held-out test records, obtained through a stratified 20% split of the 47,807-record dataset. The model achieved the following results:

Accuracy: 0.9662
Macro F1-score: 0.9573
Weighted F1-score: 0.9666

Table 5. Triage classification performance

Triage category	Precision	Recall	F1-score	Support
Green	0.8976	0.9866	0.9400	1,573
Yellow	0.9499	0.9454	0.9477	2,710
Red	0.9979	0.9708	0.9842	5,279
Accuracy	-	-	0.9662	9,562
Macro average	0.9485	0.9676	0.9573	9,562
Weighted average	0.9678	0.9662	0.9666	9,562

The Random Forest model achieved an overall accuracy of 0.9662. For the Red category, representing the most clinically critical cases, precision was 0.9979, indicating that nearly all Red predictions were correct. High precision in this category is relevant to limiting false-positive over-triage and its demands on ED resources. Red recall was 0.9708, indicating that some Red cases were under-triaged. The Green category had high recall (0.9866) but lower precision (0.8976), reflecting the misclassification of some Yellow or Red cases as Green. The macro F1-score was 0.9573, while the weighted F1-score of 0.9666 was close to the overall accuracy. These aggregate results were high, although the lower precision for Green indicates variation in performance across categories.

OPEN ACCESS

The reported results met the predefined performance criteria for Hypothesis 2 (H2). Combining NER-derived features, vital signs, and GCS in a Random Forest classifier produced an accuracy of 0.9662 and a weighted F1-score of 0.9666, exceeding the respective thresholds of 0.90 and 0.93. The NER weighted F1-score of 0.9997 also exceeded the threshold of 0.93 specified in Hypothesis 1 (H1). Together, these findings suggest that clinical entities extracted through NER provide useful information for distinguishing patient acuity when combined with physiological indicators such as vital signs and GCS.

Confusion Matrix Analysis

The confusion matrix summarizes correct and incorrect predictions for each triage category.

Table 6. Confusion matrix

Actual / Predicted	Green	Yellow	Red	Total
Green	1,552	18	3	1,573
Yellow	140	2,562	8	2,710
Red	37	117	5,125	5,279

As shown in Table 6, most records in each category were classified correctly. Of the 1,573 Green cases, 1,552 were correctly classified, 18 were predicted as Yellow, and three were predicted as Red. Of the 2,710 Yellow cases, 2,562 were correctly classified, 140 were predicted as Green, and eight were predicted as Red. Of the 5,279 Red cases, 5,125 were correctly classified, 117 were predicted as Yellow, and 37 were predicted as Green.

Under-triage was the dominant error pattern. Yellow-to-Green errors (140 cases) and Red-to-Yellow or Red-to-Green errors (117 and 37 cases, respectively) were more frequent than over-triage in the opposite direction. Under-triage may delay treatment for patients with more severe conditions and therefore carries greater clinical risk. Despite the high overall scores, the model requires improvement in distinguishing cases near triage boundaries, including the Yellow–Red boundary. The system should consequently support clinical decisions while retaining medical judgment in the triage process (Porto, 2024).

Prototype Implementation

A prototype was developed to demonstrate the complete process from anamnesis and vital sign input to a triage recommendation. It uses a layered architecture comprising a web interface, an application backend, and a machine learning inference service. Medical personnel enter anamnesis text and vital signs through the interface. An Express.js backend validates the input, handles requests, and stores clinical data and prediction results in a PostgreSQL database. Validated data are forwarded to a stateless FastAPI inference service, which performs text preprocessing, NER-based entity extraction, feature aggregation, and triage classification using the trained Random Forest model. The predicted category is returned to the backend for storage and displayed through the interface. Separating data management from model inference supports maintenance and potential integration with hospital information systems (SIMRS), while retaining clinicians in the decision-making process.

The prototype illustrates how the deployment phase of the CRISP-DM methodology can be implemented as a clinical decision-support application with potential use in ED workflows (Wirth & Hipp, 2000).

OPEN ACCESS

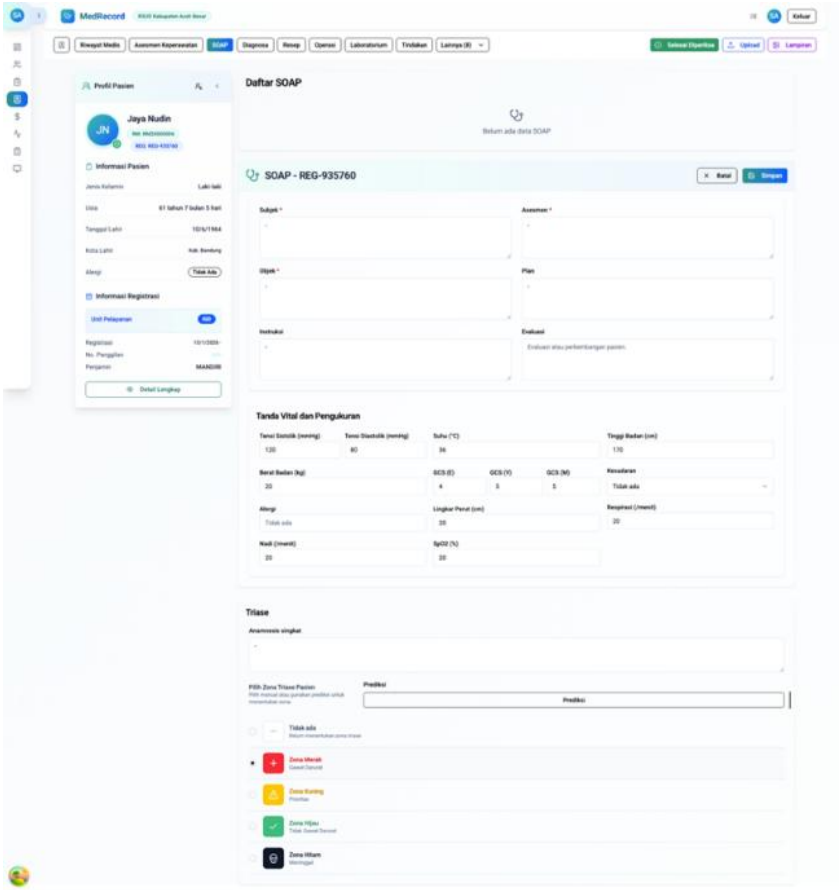

Figure 4. Prototype interface for clinical input and triage prediction. The original Indonesian interface contains SOAP notes, vital signs and measurements, anamnesis, and triage prediction panels.

Comparison with Previous Studies

The results were compared with those reported in earlier studies of machine learning-based triage systems to place the proposed approach in context.

Table 7. Comparison with previous studies

Study	Model or approach	Prediction target	Reported result
Yao et al. (2021)	CNN and BiGRU with attention	Hospital admission after an ED visit	Accuracy = 0.83; AUROC = 0.88 (NTUH dataset)
Chang et al. (2022)	CatBoost; XGBoost	Low-severity patients with short ED discharge length of stay	AUC = 0.755 (internal); 0.761 (external)
Lee et al. (2023)	AdaBoost; AdaBoost plus LightGBM	ED mortality among trauma patients	AUROC = 0.9974
Present study	IndoBERT NER and Random Forest	Green, Yellow, and Red triage categories	Accuracy = 0.9662; weighted F1-score = 0.9666

Table 7 places the proposed framework alongside related ED prediction studies. The studies used different datasets, prediction targets, and performance metrics, so their reported values do not establish comparative superiority. Combining clinical entities extracted through NER with vital signs allows the proposed classifier to consider both textual and physiological information.

The framework converts unstructured Indonesian anamnesis text into structured medical entities before triage classification. This intermediate representation provides clinically meaningful features derived from the context of patient narratives, whereas earlier approaches relied primarily on raw clinical text or structured variables.

OPEN ACCESS

The findings are consistent with a systematic review reporting benefits from combining NLP and machine learning for triage prediction (Porto, 2024).

IndoBERT provided contextual representations of Indonesian clinical narratives that supported the extraction of medically relevant information. The study therefore contributes a framework for converting unstructured clinical narratives into structured features for automated triage decision support, in addition to reporting classification performance.

Limitations and Future Work

Several limitations should be considered. First, the NER model recognizes only the categories defined in the annotation schema. Information outside these categories, such as specific pain locations, may be assigned to the closest available entity, as illustrated by the chest-pain example in the qualitative error analysis. Second, although records were collected from three hospitals, the dataset represents a limited range of clinical documentation styles, local terminology, and anamnesis-recording practices. Performance may differ in institutions with other documentation practices. Third, triage classification depends on the quality of the historical labels used as ground truth. Validation and hybrid labeling cannot fully eliminate subjective bias in the original clinical labels. Fourth, the prototype remains a simulated, standalone implementation. It has not been integrated with an operational hospital information system or evaluated by medical personnel in a live clinical workflow. Prospective evaluation in actual ED settings is therefore needed before clinical deployment.

Future research should expand the annotation schema to include more specific pain locations, symptom severity, and duration, and validate the framework using multicenter datasets with more diverse documentation styles. Reporting confidence intervals or standard deviations from cross-validation would provide additional information about performance uncertainty. Random Forest feature importance could improve interpretability, while statistical tests, such as McNemar's test, could support comparisons with baseline models when suitable paired predictions are available. Further work may investigate retrieval-augmented generation (RAG) and domain-specific large language models (LLMs) for clinical reasoning, alongside prospective evaluation and deployment of the prototype within ED workflows.

CONCLUSION

This study developed a framework combining NLP and machine learning to support ED triage classification using anamnesis text and vital sign data. The IndoBERT-based NER model extracted clinically relevant information from Indonesian anamnesis records. Random Forest achieved an accuracy of 0.9662 and a weighted F1-score of 0.9666 across the Green, Yellow, and Red categories, with particularly high precision for the Red category. These findings indicate that clinical information extracted from anamnesis records can support automated triage classification.

The principal contribution is an integrated framework combining Indonesian clinical entity extraction with machine learning-based triage classification. The annotated Indonesian ED dataset also provides a basis for further health informatics research. The framework has the potential to assist healthcare professionals in making faster, more consistent, and more objective triage assessments, particularly in EDs with high patient volumes.

Because the dataset was collected from a limited number of hospitals, the framework should be validated on larger and more diverse multicenter datasets. Future studies may also investigate advanced deep learning architectures and LLMs to improve clinical information extraction and triage prediction in real healthcare settings.

ACKNOWLEDGMENTS

The authors declare no conflict of interest regarding the publication of this paper. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

NOTATION

In Equations (1) and (2), the psychiatric and non-psychiatric binary indicators are summed to obtain symptom counts. In Equation (3), X denotes the 34-feature patient vector, comprising psychiatric, non-psychiatric, aggregate, red flag, and clinical feature blocks. In Equation (4), RF denotes the trained Random Forest classifier and the predicted label is Green, Yellow, or Red. Binary indicators and counts are dimensionless; clinical measurements retain their recorded units.

OPEN ACCESS

REFERENCES

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data, 8*, Article 53. https://doi.org/10.1186/s40537-021-00444-8

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Chang, Y.-H., Shih, H.-M., Wu, J.-E., Huang, F.-W., Chen, W.-K., Chen, D.-M., Chung, Y.-T., & Wang, C. C. N. (2022). Machine learning-based triage to identify low-severity patients with a short discharge length of stay in emergency department. *BMC Emergency Medicine, 22*, Article 88. https://doi.org/10.1186/s12873-022-00632-6

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

IBM. (2023, August 26). *What is named entity recognition?* https://www.ibm.com/think/topics/named-entity-recognition

Johannessen, L. E. F. (2016). How triage nurses use discretion: A literature review. *Professions and Professionalism, 6*(1). https://doi.org/10.7577/pp.1446

Lee, S., Kang, W. S., Kim, D. W., Seo, S. H., Kim, J., Jeong, S. T., Yon, D. K., & Lee, J. (2023). An artificial intelligence model for predicting trauma mortality among emergency department patients in South Korea: Retrospective cohort study. *Journal of Medical Internet Research, 25*, Article e49283. https://doi.org/10.2196/49283

Mariscal, G., Marbán, Ó., & Fernández, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review, 25*(2), 137–166. https://doi.org/10.1017/S0269888910000032

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., & Flach, P. (2021). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering, 33*(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

Ministry of Health of the Republic of Indonesia. (2021). *Cetak biru strategi transformasi digital kesehatan 2024* [Blueprint for the digital health transformation strategy 2024]. https://repository.kemkes.go.id/book/710

Porto, B. M. (2024). Improving triage performance in emergency departments using machine learning and natural language processing: A systematic review. *BMC Emergency Medicine, 24*, Article 219. https://doi.org/10.1186/s12873-024-01135-2

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

Soola, A. H., Mehri, S., & Azizpour, I. (2022). Evaluation of the factors affecting triage decision-making among emergency department nurses and emergency medical technicians in Iran: A study based on Benner's theory. *BMC Emergency Medicine, 22*, Article 174. https://doi.org/10.1186/s12873-022-00729-y

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the Fourth International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (pp. 29–39).

Yao, L.-H., Leung, K.-C., Tsai, C.-L., Huang, C.-H., & Fu, L.-C. (2021). A novel deep learning-based system for triage in the emergency department using electronic medical records: Retrospective cohort study. *Journal of Medical Internet Research, 23*(12), Article e27008. https://doi.org/10.2196/27008

OPEN ACCESS