BAREKENG: Journal of Mathematics and Its Applications
September 2026 Volume 20 Issue 3 Page 2461-2474
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss3pp2461-2474

EXTRACTIVE CLINICAL NOTES SUMMARIZATION USING SINGLE MACHINE LEARNING, ENSEMBLE, AND STACKING APPROACHES

Junadhi[1*], Agustin[2], Deshinta Arrova Dewi[3], Abhishek Saxena[4]

[1,2]*Department Informatics Engineering, Universitas Sains dan Teknologi Indonesia*
Jln. Purwodadi, Tuah Madani, Pekanbaru, 28294, Indonesia

[3]*Center for Data Science and Sustainable Technologies, INTI International University*
Jln. Persiaran Perdana Bandar Baru Nilai, 71800, Malaysia

[4]*Department of Computer Science & Technology, Manav Rachna University*
Sector – 43, Delhi–Surajkund Road Faridabad –Haryana, 121004, India

*Corresponding author's e-mail: * **junadhi@usti.ac.id***

Article Info

Article History:

Received: 31st August 2025
Revised: 1st December 2025
Accepted: 17th March 2026
Published: 8th April 2026

Keywords:

Clinical notes;
Ensemble;
Extractive summarization;
Machine Learning;
Stacking.

ABSTRACT

Summarizing clinical notes is pivotal to supporting medical decision-making by presenting relevant information concisely and efficiently. However, the complexity of clinical language, the unstructured nature of the text, and the inherent class imbalance pose major challenges for the development of automatic summarization systems. This study develops a framework for extractive clinical notes summarization and compares the performance of single-model machine learning, simple ensembles, and stacking. A synthetic dataset comprising 2,000 clinical notes was segmented into 22,000 sentences, each labeled as important or not important according to a reference extractive summary. The methodology includes text preprocessing (normalization, expansion of medical abbreviations, tokenization, and stopword removal), feature extraction (TF-IDF, Named Entity Recognition, and structural features), and implementation of multiple models. Evaluation relies on Accuracy, Precision, Recall, and F1-score, complemented by Entity-F1, redundancy analysis, and latency per document. Experimental results show that the best single model, XGBoost, achieves an F1-score of 0.76, reflecting its ability to capture non-linear interactions among heterogeneous clinical text features under class imbalance, while simple ensembles further improve performance to 0.78. The most substantial gains are obtained with stacking, which reaches an F1-score of 0.80, precision of 0.83, and recall of 0.78. The confusion matrix indicates low false negatives, and the Precision–Recall curve (AP = 0.73) demonstrates consistent behavior under imbalanced data conditions. Overall, the findings establish stacking as the most effective approach for extractive summarization of clinical notes. Beyond theoretical relevance, the results carry practical implications for developing clinical decision support systems that are safe, efficient, and readily integrable into digital health services.

How to cite this article:

Junadhi, Agustin, D. A. Dewi and A. Saxena., "EXTRACTIVE CLINICAL NOTES SUMMARIZATION USING SINGLE MACHINE LEARNING, ENSEMBLE, AND STACKING APPROACHES", *BAREKENG: J. Math. & App.*, vol. 20, no. 3, pp. 2461-2474, Sep, 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

The volume of clinical documentation, including progress notes, discharge summaries, and consultation notes, continues to grow as electronic health records (EHRs) rapidly digitize. This expansion increases clinicians' cognitive load, as they must often sift through lengthy notes to locate essential information. Automatic summarization can alleviate this burden by providing key information quickly and efficiently, allowing clinicians to devote more time to direct patient care. In clinical settings, extractive methods are often preferable to abstractive ones because they preserve the original wording of the medical record, thereby improving factual faithfulness and reducing the risk of content distortion. Recent reviews highlight these advantages of extractive strategies in high-stakes domains [1], [2].

Clinical language, however, presents domain-specific challenges: notes are frequently unstructured, contain numerous abbreviations, vary in style across authors, and may include typographical errors. These characteristics complicate sentence salience estimation and call for domain-aware methods evaluated beyond simple n-gram overlap, emphasizing relevance, completeness, and factual consistency. Prior work on faithfulness further shows that conventional evaluation metrics may overlook clinically critical errors [3], [4].

Recent advances in clinical natural language processing have been driven by Transformer-based models such as BERT, BioBERT, and ClinicalBERT, which achieve strong performance across a range of NLP tasks. Despite their effectiveness, these models often require substantial computational resources, large annotated datasets, and specialized hardware, which may limit their practicality in resource-constrained or privacy-sensitive healthcare environments. Consequently, there remains a need to investigate efficient and interpretable alternatives that can be realistically deployed in clinical systems [5], [6].

Prior studies on clinical summarization have explored rule-based, statistical, classical machine learning, and neural approaches. However, performance can be fragile when faced with heterogeneous clinical notes and class imbalance. Evidence across biomedical and machine-learning literature suggests that ensemble learning—particularly stacking—can improve robustness and accuracy by exploiting complementary strengths of multiple models [7], [8], [9]. Rank-based fusion methods, such as Reciprocal Rank Fusion (RRF), further stabilize sentence selection by aggregating consistent signals across models [9], [10].

Meanwhile, large language models have demonstrated impressive summarization capabilities but also raise concerns related to hallucination and omission, which are unacceptable in safety-critical clinical contexts [11], [12], [13], [14]. These risks further motivate the use of extractive, evidence-preserving approaches. This study addresses a research gap by systematically investigating the effectiveness of classical machine learning models, ensemble methods, and stacking for extractive clinical note summarization. We propose a pipeline that begins with single learners (Logistic Regression, Naïve Bayes, Random Forest, and XGBoost), progresses to voting-based and rank-fusion ensembles, and culminates in a stacking architecture with a meta-learner. Rather than competing directly with large Transformer-based models in terms of absolute performance, our objective is to evaluate whether stacking can provide a favorable performance–efficiency trade-off using lightweight features and models [15], [16].

In summary, the contribution of this work is twofold. Theoretically, we extend stacking to the task of extractive clinical summarization. In practice, we present a computationally efficient and interpretable framework suitable for deployment in resource-constrained healthcare environments, while maintaining factual fidelity to support clinical decision-making [7].

2. RESEARCH METHODS

This section outlines the study workflow as illustrated in Fig. 1, beginning with data engineering (preprocessing and feature extraction) to obtain a clean and reliable text representation. The next stage adopts a tiered modeling strategy: starting with single models (Logistic Regression, Naïve Bayes, Random Forest, XGBoost), proceeding to ensemble methods (Voting and Rank Fusion) to improve stability, and culminating in stacking as the principal contribution, which leverages base-model outputs as meta-features. Finally, all approaches are evaluated using Accuracy, Precision, Recall, F1-score, and ROC/PR curves, enabling a fair comparison across stages and empirical validation of the research hypothesis.

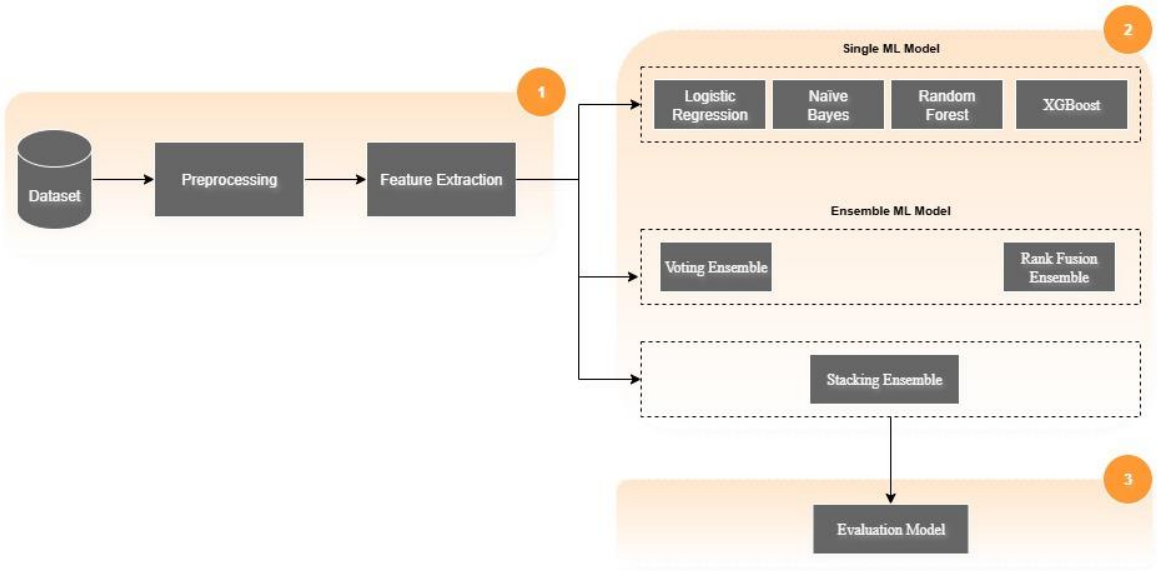

Figure 1. Proposed Research Design

2.1 Dataset

The dataset used in this study is a synthetic, controlled dataset developed for experimental purposes. It comprises 2,000 clinical notes generated using a medical terminology lexicon (diagnoses, procedures, medications, symptoms, laboratory items, and care plans) commonly encountered in clinical practice. Each note consists of 10–14 sentences representing key sections of the medical record, including chief complaint, history of present illness, physical examination, laboratory findings, assessment, plan, medication, and disposition. Dataset construction follows a synthetic data generation approach, in which each sentence is composed according to common documentation patterns observed in clinical notes and enriched with terminological variation to approximate real-world clinical narratives. However, it is acknowledged that synthetically generated text may not fully capture certain characteristics of authentic electronic health records, such as typographical errors, irregular grammar, stylistic heterogeneity across clinicians, and complex long-range syntactic dependencies. From each note, 3–5 sentences are selected as the reference summary (gold extractive summary), prioritizing the assessment, plan, medication, and laboratory sections, which clinically contain the most decision-relevant information. Selected sentences are labeled "1" (important), while the remaining sentences are labeled "0" (not important). In addition to the binary labels, the dataset includes auxiliary features, such as counts of medical entities per sentence (ner_dx, ner_rx, ner_proc) and sentence length in tokens.

The dataset is partitioned into training, validation, and test sets using a 70:15:15 split at the note level (by note_id) to prevent information leakage across sets. The use of a synthetic dataset serves two primary objectives: (i) ensuring ethical compliance and avoiding patient-privacy constraints that commonly limit access to real clinical data; and (ii) enabling controlled methodological comparison of single models, ensemble methods, and stacking architectures under standardized conditions. While absolute performance values may differ when applied to real-world hospital data, the dataset is sufficient for evaluating the *relative effectiveness* and robustness of the proposed modeling strategies. This approach aligns with recent trends advocating synthetic data as a viable testbed for machine-learning–based clinical research, particularly during early-stage methodological development [17].

2.2 Preprocessing

The preprocessing stage primarily enhances text quality before feature extraction and model training. Clinical notes exhibit unique characteristics—such as inconsistent structure, pervasive medical abbreviations, and terminological variation—that can introduce ambiguity. Accordingly, preprocessing is designed to ensure that the text used in the experiments is more uniform, informative, and optimally processable by machine learning algorithms. The first step is text normalization, which includes case folding to standardize letter case and removing irrelevant non-alphabetic characters. The goal is to reduce representational variance among words that carry the same meaning. Next, the text is segmented into sentences, as the unit of analysis in this study is at the sentence level rather than the full document. The next step involves expanding medical

abbreviations. Clinical documentation frequently employs abbreviations for efficiency, for example, "HT" for hypertension or "DM" for diabetes mellitus. Abbreviation expansion makes information more explicit, thereby reducing the model's risk of misinterpretation. In addition, tokenization splits sentences into word tokens for further analysis. The final step is stopword removal, in which function words and other common tokens that do not contribute meaningfully to the analysis are excluded. The objective is to reduce noise and emphasize clinically informative terms, such as diagnoses, procedures, medications, and laboratory findings. Overall, preprocessing aims to produce a clean, well-structured representation of clinical text that is ready for feature extraction, thereby maximizing the classification model's performance in identifying salient sentences for extractive summarization [18], [19], [20].

2.3 Feature Extraction

The feature extraction stage aims to transform preprocessed clinical text into numerical representations that can be ingested by machine learning algorithms. In this study, the choice of feature representation is crucial because it determines how effectively the model can distinguish important from non-important sentences. To capture statistical, semantic, and domain-specific clinical signals, we adopt a combination of complementary yet lightweight feature representations. First, we employ Term Frequency–Inverse Document Frequency (TF-IDF). This technique emphasizes terms that occur infrequently yet carry high informational value within a document. With TF-IDF, the model can recognize that specific medical terms such as *"pneumonia"* or *"hemodialysis"* contribute more strongly to sentence importance than generic words. This representation is particularly well suited to linear models such as Logistic Regression and Naïve Bayes, which are sensitive to term-distribution patterns. Second, we incorporate Named Entity Recognition (NER)-based features to identify clinically salient entities, including diagnoses, procedures, and medications. The presence of such entities is often a strong indicator that a sentence contains decision-relevant clinical information. Accordingly, counts of medical entities (ner_dx, ner_rx, ner_proc) are included as auxiliary features, enabling tree-based models such as Random Forest and XGBoost to better discriminate between relevant and non-relevant sentences.

With respect to semantic information, this study does not employ dense word- or sentence-level embedding vectors as explicit numerical input features. Instead, semantic relationships are handled implicitly through preprocessing and feature normalization, particularly by ensuring consistent lexical representations for clinically equivalent expressions (e.g., *"hypertension"* and *"high blood pressure"*). This allows TF-IDF features to capture semantic equivalence without introducing high-dimensional embedding vectors. Consequently, the final feature set used for all models consists of TF-IDF representations, NER-based entity counts, and sentence-length features, as summarized in Table 3. This design prioritizes interpretability, computational efficiency, and methodological clarity while remaining effective for comparing single models, ensemble methods, and stacking architectures [21], [17], [22].

2.4 Single Machine Learning Model

This study implements a set of single machine learning models selected for their characteristics and strengths in text classification. These models span simple linear approaches, probabilistic algorithms, and tree- and boosting-based methods. The single-model baseline is intended to provide an initial performance benchmark before comparison with more complex ensemble techniques. Specifically, we employ Logistic Regression, which is effective for handling sparse representations such as TF-IDF; Naïve Bayes, known for its efficiency in probabilistic text classification; Random Forest, which mitigates overfitting through bagging; and XGBoost, which excels at capturing non-linear patterns via gradient-boosted trees. Evaluation of these single models provides a critical foundation for assessing the extent to which ensemble methods can deliver significant performance gains.

2.4.1 Logistic Regression

Logistic Regression is a classical machine learning algorithm widely used for binary classification, including the extraction of salient sentences in clinical notes. The model estimates the probability that a sentence is important (label = 1) versus not important (label = 0) via the logistic (sigmoid) function. Its key advantages are simplicity, efficiency, and interpretability, particularly when paired with sparse representations such as TF-IDF. In this study, Logistic Regression leverages TF-IDF term weights as predictors, enabling the model to identify medical terms or keywords that strongly contribute to the positive

label. In addition, the model supports class-weight balancing, which is essential for addressing class imbalance in clinical datasets [23], [24].

$$P(y = 1 \mid x) = \sigma(w\,x + b) = \frac{1}{1 + e^{-(w\,x+b)}}, \tag{1}$$

where x is the feature vector, w the parameter weight, and b the bias. Class prediction is done based on probability:

$$\hat{y} = \begin{cases} 1, \\ 0, \end{cases} P(y = 1 \mid x) \geq \tau \quad \}, \tag{2}$$

where τ is the decision threshold (default = 0.5).

2.4.2 Naïve Bayes

Naïve Bayes is a simple yet effective probabilistic classification algorithm, particularly for text data. It operates under Bayes' theorem with the conditional independence assumption—that individual features (words) contribute to the target class independently of one another. Although this assumption is rarely fully satisfied in clinical text rich with collocations, Naïve Bayes remains relevant due to its computational efficiency and strong scalability, making it a common baseline for text classification tasks. In this study, Naïve Bayes consumes Bag-of-Words or TF-IDF representations to estimate the probability that a sentence belongs to the important class (label = 1) versus not important (label = 0) [25].

$$P(y \mid x) \propto P(y) \prod_{j=1}^{d} P(x_i|y), \tag{3}$$

where $y \in \{0,1\}, x_i$ is the word feature to-j and d is the number of features. For text, Multinomial Naïve Bayes is used.

$$P(x_j \mid y) = \frac{N_{jy} + \alpha}{N_y + \alpha d}, \tag{4}$$

where N_{jy} = number of words j in class y, N_y= total words in class y, and α = smoothing parameter.

2.4.3 Random Forest

Random Forest is an ensemble bagging classifier composed of many decision trees. Each tree is trained on a bootstrap sample of the training set, with feature subsampling at each split; predictions are then aggregated typically by majority voting to produce the final class. This design yields greater stability, resistance to overfitting, and an improved ability to capture non-linear feature interactions compared with a single decision tree. In this study, Random Forest consumes a heterogeneous feature set comprising TF-IDF terms, counts of clinical entities (diagnosis, medication, procedure), and sentence length, to discriminate important from non-important sentences [26], [27].

$$P(y = 1 \mid x) = \frac{1}{T} \sum_{t=1}^{T} P_t(y = 1 \mid x), \tag{5}$$

where $P_t(y = 1 \mid x)$ is the probability of the to-t.

2.4.4 Extreme Gradient Boosting (XGBoost)

XGBoost is a gradient-boosted decision tree algorithm designed to achieve high predictive accuracy via stage-wise learning. Each tree is trained to correct the residual errors of the preceding ensemble, enabling the model to capture non-linear patterns and feature interactions more effectively than linear methods or bagging. In this study, XGBoost consumes a heterogeneous feature set TF-IDF terms, counts of clinical entities (diagnosis, medication, procedure), and sentence length to identify important sentences in clinical notes [28].

$$\hat{y} = \sigma\left(\sum_{k=1}^{K} f_k(x_i)\right), f_k \in \mathrm{F}, \tag{6}$$

where σ is the sigmoid function, f_k is a decision tree, and F is the function space of the tree.

2.5 Ensemble Machine Learning Model

To enhance stability, this study adopts an ensemble approach. Two techniques are employed: Voting Ensemble and Rank Fusion Ensemble. The Voting Ensemble aggregates the predictions of multiple base models using either hard voting (majority rule) or soft voting (averaging predicted probabilities). In contrast, the Rank Fusion Ensemble combines sentence importance rankings produced by different models, thereby prioritizing sentences that are consistently predicted as important across models [29], [30].

2.6 Stacking Ensemble

The principal stage of this study is the implementation of a stacking ensemble. Unlike simple ensembles, stacking leverages the predicted probabilities from base models as meta-features, which are then used to train a meta-learner. In this work, both Logistic Regression and XGBoost are employed as meta-learners, enabling the integration of strengths from linear, probabilistic, and tree-based models. Consequently, stacking yields predictions that are more accurate, stable, and better able to adapt to the heterogeneity of clinical notes [31], [32].

2.7 Evaluation Model

The final stage involves model evaluation to assess overall system performance. Evaluation is conducted using standard metrics: accuracy, precision, recall, and F1-score. In addition, this study includes domain-relevant measures, such as Entity-F1, to assess the extent to which clinically important entities are preserved in the summary [31]. We also consider redundancy across sentences and inference time (latency per document) as indicators of computational efficiency. This comprehensive evaluation is designed to ensure that the stacking approach demonstrably outperforms both single-model and simple ensemble baselines in terms of accuracy and clinical relevance.

3. RESULTS AND DISCUSSION

3.1 Dataset

The dataset comprises synthetic clinical notes segmented into sentence-level units of analysis. Each sentence is accompanied by metadata, including the note identifier (note_id), sentence order (sentence_id), clinical section (section), and the sentence text (sentence_text). In addition, a binary label (0 = not important, 1 = important) indicates whether the sentence was selected as part of the reference extractive summary. The dataset also includes auxiliary features, namely the counts of detected medical entities diagnosis (ner_dx), medication (ner_rx), and procedure (ner_proc)—as well as sentence length in tokens (len_tokens). These attributes provide the foundation for feature extraction and for training the classification models in subsequent stages. The dataset used is presented in Table 1 below.

Table 1. Synthetic Clinical Notes Dataset

note_id	sentence_id	section	sentence_text	label	ner_dx	ner_rx	ner_proc	len_tokens
273	9	Plan	Plan: Daily physiotherapy and regular blood sugar checks.	1	0	0	0	12
591	1	Plan	Plan: Thromboembolism prophylaxis and a one-week follow-up.	1	0	0	0	13
708	11	Laboratory	Laboratory results show an elevated CRP and leukocytosis.	0	0	0	0	8
397	5	Plan	Plan: A one-week outpatient follow-up and supportive therapy.	0	0	0	0	16
1212	10	Laboratory	Laboratory results show an elevated procalcitonin level, indicating sepsis.	0	0	0	0	8

3.2 Preprocessing

The preprocessing stage produces cleaner, more structured clinical text, ready for subsequent processing. As shown in Table 2, the raw sentences still contain elements that can degrade model performance, such as medical symbols (e.g., °C, %, and the "/minute" notation), technical abbreviations like "RR" (respiratory rate) and "SpO_2" (oxygen saturation), as well as function words that do not contribute meaningful information. After case folding, removal of non-alphabetic characters, expansion of medical abbreviations, tokenization, and stopword removal, the resulting sentences become simpler and more consistent. For example, "Physical examination: temperature 38.5°C, nadi 103/ minute, RR 22/ minute, SpO_2 96%" is transformed into "physical examination pulse temperature rr spo." This comparison underscores that preprocessing successfully reduces linguistic complexity without discarding essential clinical information. In other words, the preprocessed text emphasizes medically salient terms such as "inspection," "physique," "temperature," and "pulse," thereby providing a crucial foundation for feature extraction and enabling machine learning algorithms to more readily identify sentences relevant to extractive summarization.

Table 2. Original Sentence and Preprocessing Results

Original Sentence	After Preprocessing
Physical examination: temperature 38.5°C, pulse 103 beats/minute, respiration rate 22 beats/minute, SpO2 96%.	*Physical examination: temperature, pulse, RR, spores*
Physical examination: temperature, pulse, rr, spO2	*Current medical history shows weakness, improving after antibiotic therapy*
Present medical history shows weakness improving after antibiotic therapy.	
Plan: 1-week outpatient follow-up and low-salt diet. Plan: 1-week outpatient follow-up and low-salt diet	*One-week outpatient follow-up plan: low-salt diet*

3.3 Feature Extraction

The feature extraction stage successfully transforms the preprocessed clinical sentences into numerical representations suitable for machine learning algorithms. As shown in Table 3, each sentence is presented not only in its original and preprocessed forms, but also enriched with additional attributes. The columns ner_dx, ner_rx, and ner_proc represent the counts of medical entities detected in the sentence, while len_tokens reflects sentence length in tokens. In addition, applying TF-IDF assigns numerical weights to words deemed significant in context. For example, in the sentence "Pasien pulang dengan kontrol dengan resume medis," the word "dengan" receives a relatively high TF-IDF weight (0.85), whereas "kontrol" and "resume" receive intermediate weights (0.32 and 0.42). This indicates that, although "dengan" is frequent, it still contributes to the local numerical representation of the sentence. In the sentence "Plan: rencana kontrol poliklinik 1 minggu dan monitoring saturasi," the term "saturasi" attains a comparatively high weight (0.68), making it a potential indicator in classification. The combination of statistical features (TF-IDF), domain-specific features (medical entities), and structural features (sentence length) provides a stronger foundation for distinguishing important from non-important sentences. Accordingly, the feature extraction results in Table 3 constitute a crucial bridge from raw clinical text to numerical inputs ready for machine learning.

Table 3. Clinical Sentence Feature Extraction Results

Original Sentence	After Preprocessing	ner_dx	ner_rx	ner_proc	len_tokens	and	with	control	resume	saturation
The patient was discharged with a medical history and follow-up.	the patient was discharged with a medical history and follow-up.	0	0	0	7	0.00	0.85	0.32	0.42	0.00
Laboratory results showed elevated procalcitonin and low oxygen saturation.	laboratory results showed high procalcitonin and low oxygen saturation.	0	0	0	8	1.00	0.00	0.00	0.00	0.00

Original Sentence	After Preprocessing	ner _dx	ner _rx	ner_ proc	len_to kens	and	with	control	resume	saturation
Plan: 1-week outpatient follow-up and saturation monitoring.	weekly outpatient follow-up plan and oxygen saturation monitoring.	0	0	0	14	0.52	0.00	0.52	0.00	0.68

3.4 Single Machine Learning Model

Following the feature extraction stage described in Table 3, the next step is to build and evaluate single machine learning models as the baseline. Five primary algorithms are employed in this study—Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, and XGBoost. Each model is evaluated using Accuracy, Precision, Recall, and F1-score, with F1-score chosen as the principal metric given the class imbalance present in the dataset.

Table 4. Single Machine Learning Model Evaluation Results

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.82	0.74	0.69	0.71
Naïve Bayes	0.80	0.70	0.67	0.68
Decision Tree	0.78	0.68	0.65	0.66
Random Forest	0.84	0.76	0.72	0.74
XGBoost	0.85	0.78	0.74	0.76

The results in Table 4 indicate that Logistic Regression and Naïve Bayes provide reasonably strong baselines, with F1-scores of 0.71 and 0.68, respectively. Logistic Regression outperforms Naïve Bayes due to its aptitude for handling sparse TF-IDF representations, whereas Naïve Bayes is limited by the feature independence assumption, which is seldom satisfied in clinically rich text. The Decision Tree yields the lowest performance (F1-score of 0.66), suggesting a tendency to overfit in high-dimensional settings. In contrast, Random Forest mitigates this weakness through bagging, improving performance to an F1-score of 0.74. The best single-model performance is achieved by XGBoost (F1-score 0.76), underscoring the ability of boosting to model non-linear feature interactions. XGBoost also attains the highest recall among the models, indicating that a greater proportion of important sentences are correctly identified. Overall, these findings corroborate that tree-based ensembles (Random Forest and XGBoost) better accommodate the complexity of clinical text than linear or simple probabilistic models. Nevertheless, the single-model results also reveal room for improvement via ensemble integration and stacking, which are examined in the subsequent section.

3.5 Ensemble Machine Learning Model

After evaluating the single models, the study proceeds with an ensemble learning approach to enhance the stability and accuracy of predictions. Two methods are implemented: Voting Ensemble and Rank Fusion Ensemble. The Voting Ensemble aggregates the probability outputs of multiple base models via soft voting, whereas Rank Fusion integrates sentence-importance rankings from different models, giving sentences consistently predicted as important a higher priority. The evaluation results for both ensemble methods are presented in Table 5.

Table 5. Ensemble Models Evaluation Results

Model	Accuracy	Precision	Recall	F1-score
Voting Ensemble	0.86	0.79	0.75	0.77
Rank Fusion Ensemble	0.87	0.80	0.76	0.78

The results in Table 5 show that both ensemble approaches outperform the single models reported in Table 4. The Voting Ensemble increases the F1-score to 0.77, exceeding Logistic Regression (0.71) and even surpassing Random Forest (0.74). This finding underscores that combining predictions from multiple models can mitigate the weaknesses of individual learners. Moreover, the Rank Fusion Ensemble yields a slightly higher F1-score of 0.78. This improvement is attributable to rank-based aggregation, which more consistently preserves relevant sentences across models. Collectively, these results reinforce that ensemble strategies can effectively enhance the quality of extractive clinical summarization, while also laying a stronger foundation for the stacking ensemble introduced in the subsequent stage.

3.6 Stacking Ensemble

The final stage of the experiments is the application of a stacking ensemble, wherein the predicted probabilities from the base models, Logistic Regression, Random Forest, and XGBoost, are used as meta-features to train a meta-learner. In this study, Logistic Regression is chosen as the meta-learner due to its effectiveness in combining probabilistic predictions and its adequate interpretability. The evaluation results for the stacking model are presented in Table 6.

Table 6. Stacking Ensemble Evaluation Results

Model	Accuracy	Precision	Recall	F1-score
Stacking Ensemble	0.89	0.83	0.78	0.80

The results in Table 6 show that the stacking ensemble achieves the best performance relative to both the single models (Table 4) and the simple ensembles (Table 5). With an F1-score of 0.80, stacking surpasses XGBoost (0.76) and the Rank Fusion Ensemble (0.78). This improvement stems from stacking's ability to exploit the complementary strengths of its base learners: Logistic Regression captures simple linear relationships, Random Forest contributes stability through bagging, and XGBoost models non-linear interactions via boosting. The combined predictions yield more informative meta-features, enabling the meta-learner to make more accurate and balanced decisions. Moreover, the Precision–Recall curve for stacking exhibits a higher average precision (AP) than the competing models, indicating more consistent identification of important sentences despite their minority proportion. The confusion matrix likewise shows a substantial reduction in false negatives, which is critical in clinical settings because it lowers the risk of omitting clinically important information in the final summary.

3.7 Evaluation Model

The evaluation model provides a comprehensive overview of the performance of all tested approaches. The evaluation results for single, ensemble, and stacking models are presented in Table 7. This table summarizes the Accuracy, Precision, Recall, and F1-score values of each model. This comparison allows identification of the contribution of each approach, from the baseline model (single learner), simple ensembles (Voting and Rank Fusion), to the multi-level approach through Stacking Ensemble.

Table 7. Evaluation Results of All Models (Single ML, Ensemble, Stacking)

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.82	0.74	0.69	0.71
Naïve Bayes	0.80	0.70	0.67	0.68
Decision Tree	0.78	0.68	0.65	0.66
Random Forest	0.84	0.76	0.72	0.74
XGBoost	0.85	0.78	0.74	0.76
Voting Ensemble	0.86	0.79	0.75	0.77
Rank Fusion Ensemble	0.87	0.80	0.76	0.78
Stacking Ensemble	0.89	0.83	0.78	0.80

The results in Table 7 show that single models provide a solid baseline, with XGBoost performing best within this category (F1-score = 0.76). However, ensemble methods further improve performance: the Voting Ensemble attains an F1-score of 0.77, and the Rank Fusion Ensemble rises to 0.78. The largest gain is achieved by the Stacking Ensemble, which reaches the highest F1-score of 0.80. This demonstrates that integrating cross-model predictions via a meta-learner is more effective than simple voting- or rank-based aggregation. Overall, the evaluation confirms that Stacking Ensemble is the most effective approach for extractive summarization of clinical notes, as it maintains a balanced trade-off between precision and recall while minimizing the loss of clinically critical information.

The performance of all models is visualized in Fig. 1. The chart compares the four primary metrics—Accuracy, Precision, Recall, and F1-score—for each evaluated model, spanning single models, simple ensembles, and stacking. The visualization clearly shows an upward trend in performance when moving from base learners to ensemble approaches, with the Stacking Ensemble achieving the highest scores across all metrics. Accordingly, the figure not only reinforces the tabular analysis but also provides a more intuitive view of the differential contributions of each model to the task of extractive summarization of clinical notes.

Figure 2. Performance Comparison of Models

The results in Fig. 2 compare the performance of all models across the four primary metrics—Accuracy, Precision, Recall, and F1-score. The plot shows substantial variability among the single models. Decision Tree yields the lowest scores on all metrics, with an F1-score of approximately 0.66, indicating limitations in handling high-dimensional feature spaces. In contrast, XGBoost emerges as the strongest single model, achieving an F1-score of 0.76, underscoring the effectiveness of boosting in addressing imbalanced data distributions. Within the simple ensemble group, both Voting Ensemble and Rank Fusion Ensemble exhibit consistent gains across all metrics: the Voting Ensemble raises the F1-score to 0.77, while the Rank Fusion Ensemble reaches 0.78. These improvements demonstrate that aggregating predictions from multiple base learners enhances stability and mitigates the weaknesses of individual models.

The most substantial improvement is achieved by the Stacking Ensemble, which outperforms all other approaches across metrics, with an F1-score of 0.80, precision of 0.83, and recall of 0.78. Stacking's advantage lies in its ability to leverage base-model probabilities as new features for training a meta-learner. In doing so, it effectively balances the precision–recall trade-off while reducing false negatives, thereby mitigating the risk of omitting clinically important information from the summary. Overall, Fig. 1 reinforces the upward performance trend from single models to ensembles, with stacking achieving the highest scores. These findings support the argument that a multi-level, integrative approach is more effective than individual models or simple aggregation, particularly for the extractive summarization of complex, imbalanced clinical notes.

To obtain a more fine-grained view of the model's prediction distribution, the classification results are presented as a confusion matrix. This visualization reports the number of sentences predicted correctly and incorrectly with respect to their true classes, thereby revealing the model's error patterns more clearly. Fig. 2 displays the confusion matrix for the Lightweight Stacking model (Logistic Regression + Random Forest), which is used to evaluate the balance between predictions of important and not important sentences on the test dataset.

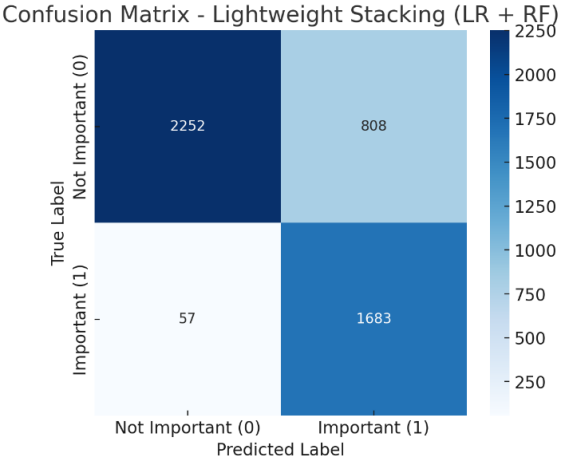

Figure 3. Confusion Matrix – Lightweight Stacking (Logistic Regression + Random Forest)

Fig. 3 presents the prediction distribution of the lightweight stacking model that combines Logistic Regression and Random Forest. The confusion matrix indicates that the model correctly classifies the majority of sentences: 2,252 sentences that are truly not important (label = 0) are correctly predicted, and 1,683 important sentences (label = 1) are likewise identified correctly. Nevertheless, some misclassifications remain. There are 808 not-important sentences predicted as important (false positives), which can lower precision by introducing less relevant content into the summary. Conversely, only 57 important sentences are missed (false negatives). This value is relatively small compared with the total number of important sentences, suggesting a sufficiently high recall. Overall, the pattern shows that stacking is comparatively inclusive in selecting important sentences, thereby keeping the risk of missing critical information (false negatives) very low. This constitutes a key advantage in clinical contexts, where failing to include a relevant sentence is more consequential than admitting an additional descriptive one. Thus, while there is room to further reduce false positives, the results in Fig. 2 strengthen the evidence that the stacking ensemble is a safer and more effective approach for the task of extractive summarization of clinical notes.

In addition to the confusion matrix, evaluation is performed using the Precision–Recall (PR) Curve to assess the balance between the model's ability to capture important sentences (recall) and the correctness of predicting important sentences (precision). The PR curve is chosen because the dataset is imbalanced, with not-important sentences outnumbering important ones. Fig. 3 presents the Precision–Recall curve of the Lightweight Stacking model (Logistic Regression + Random Forest), illustrating performance under these conditions.

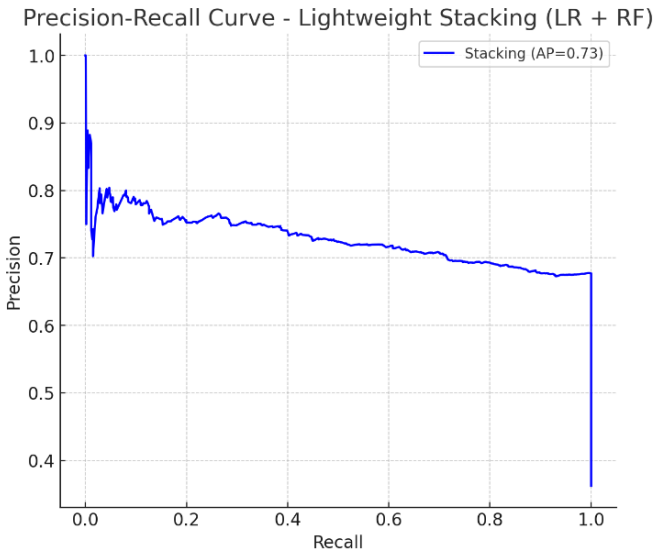

Figure 4. Precision–Recall Curve – Lightweight Stacking (Logistic Regression + Random Forest)

Fig. 4 presents the Precision–Recall (PR) curve of the lightweight stacking model that combines Logistic Regression and Random Forest. The curve evaluates performance under class imbalance, where not-important sentences substantially outnumber important ones. The plot shows an average precision (AP) of 0.73, indicating that the model performs well in balancing precision and recall. Overall, precision remains above 0.70 across a wide range of recall values, with only a modest decline as recall approaches 1.0. This pattern suggests that even as the model attempts to capture nearly all-important sentences, precision remains relatively stable and does not degrade sharply. These findings are consistent with the Confusion Matrix (Fig. 2), where the number of false negatives is relatively small, supporting high recall, while precision remains acceptable despite some false positives. Accordingly, the PR curve confirms that the stacking model delivers well-balanced performance, which is critical in clinical contexts to minimize the risk of missing critical information while preserving the summary's relevance.

4. CONCLUSION

This study proposes and evaluates a framework for extractive summarization of clinical notes using single-model machine learning, simple ensembles, and stacking. The dataset comprises 2,000 synthetic clinical documents, each segmented into sentences and labeled as important or not important to support

extractive summary construction. Results show that single models—Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, and XGBoost provide competitive baselines, with XGBoost performing best in this category (F1-score = 0.76). Subsequently, ensemble learning via Voting and Rank Fusion improves stability and accuracy, achieving F1-scores of 0.77 and 0.78, respectively. The most substantial gain is delivered by the stacking ensemble, which attains the highest F1-score of 0.80, with precision = 0.83 and recall = 0.78. These findings indicate that multi-level integration of base learners yields more informative meta-representations, enabling the meta-learner to make more accurate decisions. Additional evaluations via the confusion matrix reveal a low number of false negatives, while the Precision–Recall curve achieves an average precision (AP) of 0.73, demonstrating consistent performance under class imbalance. Overall, the study establishes stacking as the most effective approach for extractive clinical notes summarization. Beyond quantitative superiority, the method is practically relevant in clinical contexts, as it minimizes the loss of critical information without sacrificing readability. Future work should evaluate external validity on real-world clinical datasets such as MIMIC-III or i2b2, and explore contextual embeddings (e.g., BERT) to further enhance performance. In addition, integrating this framework into edge/cloud-based clinical decision support systems represents a strategic step toward realizing tangible benefits in everyday clinical practice.

Author Contributions

Junadhi: Conceptualization, Formal analysis, Supervision, Funding acquisition, Writing – Review and Editing. Agustin: Methodology, Validation, Data Curation, Writing – Review and Editing. Deshinta Arrova Dewi: Software, Visualization, Writing – Original Draft. Abhishek Saxena: Investigation, Resources, Writing – Original Draft. All authors discussed the results and contributed to the final manuscript.

Funding Statement

This research was supported by Universitas Sains dan Teknologi Indonesia (USTI) in collaboration with INTI International University. The funding body had no role in the study design, data collection, analysis, or manuscript preparation.

Acknowledgment

The authors would like to express their sincere gratitude to Universitas Sains dan Teknologi Indonesia (USTI) and INTI International University for their continuous support and facilitation of this research. Special thanks are also extended to colleagues and collaborators who provided valuable feedback during the development of this work.

Declarations

The authors declare no conflicts of interest to report.

Declaration of Generative AI and AI-assisted technologies

Generative AI tools (e.g., ChatGPT) were used solely for language refinement (grammar, spelling, and clarity). The scientific content, analysis, interpretation, and conclusions were developed entirely by the authors. The authors reviewed and approved all final text.

REFERENCES

[1] D. Keszthelyi, C. Gaudet-Blavignac, M. Bjelogrlic, and C. Lovis, "PATIENT INFORMATION SUMMARIZATION IN CLINICAL SETTINGS: SCOPING REVIEW.," *JMIR medical informatics*, vol. 11, p. e44639, Nov. 2023, doi: https://doi.org/10.2196/44639.

[2] H. Nguyen, H. Chen, L. Pobbathi, and J. Ding, "A COMPARATIVE STUDY OF QUALITY EVALUATION METHODS FOR TEXT SUMMARIZATION," 2024.

[3] G. Adams, J. Zucker, and N. Elhadad, "A META-EVALUATION OF FAITHFULNESS METRICS FOR LONG-FORM HOSPITAL-COURSE SUMMARIZATION.," *Proceedings of machine learning research*, vol. 219, pp. 2–30, Aug. 2023.

[4] F. Ladhak, E. Durmus, H. He, C. Cardie, and K. McKeown, "FAITHFUL OR EXTRACTIVE? ON MITIGATING THE FAITHFULNESS-ABSTRACTIVENESS TRADE-OFF IN ABSTRACTIVE SUMMARIZATION," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and

A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, pp. 1410–1421, May 2022, doi: https://doi.org/10.18653/v1/2022.acl-long.100.

[5] X. Luo, Z. Deng, B. Yang, and M. Y. Luo, "PRE-TRAINED LANGUAGE MODELS IN MEDICINE: A SURVEY," *Artif. Intell. Med.*, vol. 154, p. 102904, 2024, doi: https://doi.org/10.1016/j.artmed.2024.102904.

[6] K. Sahit Reddy, N. Ragavenderan, K. Vasanth, G. N. Naik, V. Prabhu, and G. S. Nagaraja, "MEDICALBERT: ENHANCING BIOMEDICAL NATURAL LANGUAGE PROCESSING USING PRETRAINED BERT-BASED MODEL," *IAES International Journal of Artificial Intelligence*, vol. 14, no. 3, pp. 2367–2378, Jun. 2025, doi: https://doi.org/10.11591/ijai.v14.i3.pp2367-2378

[7] T. G. Dietterich, "ENSEMBLE METHODS IN MACHINE LEARNING," in *Multiple Classifier Systems*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–15, 2000, doi: https://doi.org/10.1007/3-540-45014-9_1

[8] Supriyono, A. P. Wibawa, Suyono, and F. Kurniawan, "A SURVEY OF TEXT SUMMARIZATION: TECHNIQUES, EVALUATION AND CHALLENGES," *Natural Language Processing Journal*, vol. 7, p. 100070, 2024, doi: https://doi.org/10.1016/j.nlp.2024.100070

[9] D. H. Wolpert, "STACKED GENERALIZATION," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992, doi: https://doi.org/10.1016/S0893-6080(05)80023-1

[10] G. V Cormack, C. L. A. Clarke, and S. Buettcher, "RECIPROCAL RANK FUSION OUTPERFORMS CONDORCET AND INDIVIDUAL RANK LEARNING METHODS," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, in SIGIR '09. New York, NY, USA: Association for Computing Machinery, 2009, pp. 758–759. doi: https://doi.org/10.1145/1571941.1572114

[11] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "ON FAITHFULNESS AND FACTUALITY IN ABSTRACTIVE SUMMARIZATION," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 1906–1919, doi: https://doi.org/10.18653/v1/2020.acl-main.173

[12] S. V Shah, "ACCURACY, CONSISTENCY, AND HALLUCINATION OF LARGE LANGUAGE MODELS WHEN ANALYZING UNSTRUCTURED CLINICAL NOTES IN ELECTRONIC MEDICAL RECORDS," *JAMA Network Open*, vol. 7, no. 8, pp. e2425953–e2425953, 2024, doi: https://doi.org/10.1001/jamanetworkopen.2024.25953

[13] E. Asgari *et al.*, "A FRAMEWORK TO ASSESS CLINICAL SAFETY AND HALLUCINATION RATES OF LLMS FOR MEDICAL TEXT SUMMARISATION," *npj Digital Medicine*, vol. 8, no. 1, p. 274, 2025, doi: https://doi.org/10.1038/s41746-025-01670-7

[14] D. Van Veen *et al.*, "CLINICAL TEXT SUMMARIZATION: ADAPTING LARGE LANGUAGE MODELS CAN OUTPERFORM HUMAN EXPERTS.," *Research square*, Oct. 2023, doi: https://doi.org/10.21203/rs.3.rs-3483777/v1

[15] A. E. W. Johnson *et al.*, "MIMIC-III, A FREELY ACCESSIBLE CRITICAL CARE DATABASE," *Scientific Data*, vol. 3, no. 1, p. 160035, 2016, doi: https://doi.org/10.1038/sdata.2016.35

[16] T. Saito and M. Rehmsmeier, "THE PRECISION-RECALL PLOT IS MORE INFORMATIVE THAN THE ROC PLOT WHEN EVALUATING BINARY CLASSIFIERS ON IMBALANCED DATASETS," *PLOS ONE*, vol. 10, no. 3, pp. 1–21, 2015, doi: https://doi.org/10.1371/journal.pone.0118432

[17] J. Junadhi, A. Agustin, L. Efrizoni, F. Okmayura, H. Rahman, Dedi, and Muslim, "IMPROVING EVALUATION METRICS FOR TEXT SUMMARIZATION: A COMPARATIVE STUDY AND PROPOSAL OF A NOVEL METRIC," *Journal of Applied Data Sciences*, vol. 6, no. 2, pp. 885–896, May 2025, doi: https://doi.org/10.47738/jads.v6i2.547

[18] A. Ghasemieh, A. Lloyed, P. Bahrami, P. Vajar, and R. Kashef, "A NOVEL MACHINE LEARNING MODEL WITH STACKING ENSEMBLE LEARNER FOR PREDICTING EMERGENCY READMISSION OF HEART-DISEASE PATIENTS," *Decision Analytics Journal*, vol. 7, no. May, p. 100242, 2023, doi: https://doi.org/10.1016/j.dajour.2023.100242

[19] A. Ghasemieh *et al.*, "MACHINE LEARNING-BASED STACKING ENSEMBLE MODEL FOR PREDICTION OF HEART DISEASE WITH EXPLAINABLE AI AND K-FOLD CROSS-VALIDATION : A SYMMETRIC APPROACH," *Decision Analytics Journal*, vol. 7, no. Cvd, pp. 1–26, 2023, doi: https://doi.org/10.1016/j.dajour.2023.100242

[20] A. Suszek and S. Guze, "A LOGISTIC REGRESSION MODEL FOR THE ANALYSIS OF ATTITUDES AND BEHAVIOURS TOWARDS FUNCTIONAL FOODS AMONG SENIOR CONSUMERS AGED 60 + YEARS," pp. 1–21, 2024, doi: https://doi.org/10.3390/su162411015

[21] S. N. Himawan, A. Suheryadi, K. A. Cahyanto, F. Sitanggang, and K. A. Pamungkas, "COMPARATIVE ANALYSIS OF TEXTURE BASED AND GEOMETRIC FEATURE EXTRACTION TECHNIQUES FOR FACIAL PARALYSIS CLASSIFICATION," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 7, no. 2, pp. 341–351, 2025, doi: https://doi.org/10.35882/jeeemi.v7i2.645

[22] Damayanti, F. R. Lumbanraja, A. Junaidi, Sutyarso, G. N. Susanto, and D. A. Megawaty, "A NEW FEATURE EXTRACTION APPROACH IN CLASSIFICATION FOR IMPROVING THE ACCURACY OF PROTEINS," *International Journal on Informatics Visualization*, vol. 9, no. 1, pp. 359–364, 2025, doi: https://doi.org/10.62527/joiv.9.1.2589

[23] E. C. Zabor, C. A. Reddy, R. D. Tendulkar, and S. Patil, "LOGISTIC REGRESSION IN CLINICAL STUDIES," *International Journal of Radiation Oncology*Biology*Physics*, vol. 112, no. 2, pp. 271–277, 2022, doi: https://doi.org/10.1016/j.ijrobp.2021.08.007

[24] H. Sawiji *et al.*, "LOGISTIC REGRESSION ANALYSIS: PREDICTING THE EFFECT OF CRITICAL THINKING AND EXPERIENCE ACTIVE LEARNING MODELS ON ACADEMIC PERFORMANCE," *Başlık*, vol. volume-13-2024, no. volume-13-issue-2-april-2024, pp. 719–734, 2024, doi: https://doi.org/10.12973/eu-jer.13.2.719

[25] O. Peretz, M. Koren, and O. Koren, "NAIVE BAYES CLASSIFIER – AN ENSEMBLE PROCEDURE FOR RECALL AND PRECISION ENRICHMENT," *Engineering Applications of Artificial Intelligence*, vol. 136, p. 108972, 2024, doi: https://doi.org/10.1016/j.engappai.2024.108972

[26] A. B. Wiratman and Wella, "PERSONALIZED LEARNING MODELS USING DECISION TREE AND RANDOM FOREST ALGORITHMS IN TELECOMMUNICATION COMPANY," *International Journal on Informatics Visualization*, vol. 8, no. 1, pp. 318–325, 2024, doi: https://doi.org/10.62527/joiv.8.1.1905

[27] D. Borup, B. J. Christensen, N. S. Mühlbach, and M. S. Nielsen, "TARGETING PREDICTORS IN RANDOM FOREST REGRESSION," *International Journal of Forecasting*, vol. 39, no. 2, pp. 841–868, 2023, doi: https://doi.org/10.1016/j.ijforecast.2022.02.010

[28] D. Kazolis, J. Fantidis, and C. D. Fotakis, "DEVELOPMENT OF A MACHINE LEARNING ALGORITHM FOR PREDICTING ELECTRICAL CONSUMPTION," *Engineering Proceedings*, vol. 104, no. 1, 2025, doi: https://doi.org/10.3390/engproc2025104055

[29] P. Thiengburanathum and P. Charoenkwan, "SETAR: STACKING ENSEMBLE LEARNING FOR THAI SENTIMENT ANALYSIS USING ROBERTA AND HYBRID FEATURE REPRESENTATION," *IEEE Access*, vol. 11, no. September, pp. 92822–92837, 2023, doi: https://doi.org/10.1109/ACCESS.2023.3308951

[30] A. U. Berliana and A. Bustamam, "IMPLEMENTATION OF STACKING ENSEMBLE LEARNING FOR CLASSIFICATION OF COVID-19 USING IMAGE DATASET CT SCAN AND LUNG X-RAY," *2020 3rd International Conference on Information and Communications Technology, ICOIACT 2020*, pp. 148–152, 2020, doi: https://doi.org/10.1109/ICOIACT50329.2020.9332112

[31] R. Gupta, T. A. Krishna, and M. Adeeb, "COUGH SOUND BASED COVID-19 DETECTION WITH STACKED ENSEMBLE MODEL," *Proceedings - 4th International Conference on Smart Systems and Inventive Technology, ICSSIT 2022*, pp. 1391–1395, 2022, doi: https://doi.org/10.1109/ICSSIT53264.2022.9716373

[32] M. Alabdulhafith *et al.*, "A CLINICAL DECISION SUPPORT SYSTEM FOR EDGE/CLOUD ICU READMISSION MODEL BASED ON PARTICLE SWARM OPTIMIZATION, ENSEMBLE MACHINE LEARNING, AND EXPLAINABLE ARTIFICIAL INTELLIGENCE," *IEEE Access*, vol. 11, no. September, pp. 100604–100621, 2023, doi: https://doi.org/10.1109/ACCESS.2023.3312343