ILKOM Jurnal Ilmiah Vol. 18, No. 2, August 2026, pp. 255-267
Accredited 2nd by KEMDIKTISAINTEK No. 10/C/C3/DT.05.00/2025; E-ISSN 2548-7779 | P-ISSN 2087-1716

Feature Importance–Driven Multimodal Learning for Medical Diagnosis Classification Using Clinical and Symptom Text Data

Indra Waspada [a,1,*]; Satriawan Rasyid Purnama [a,2]; Alfonso Clement Sutantio [a,3] ; Alwey Hakim[a,4]
[a] *Department of Informatics, Faculty of Science and Mathematics, Universitas Diponegoro, Semarang, 50275, Indonesia*
[1] *indrawaspada@lecturer.undip.ac.id;* [2] *satriawanrasyid@live.undip.ac.id;*[3] *alfonsocs@students.undip.ac.id;* [4] *alweyhakim@students.undip.ac.id*
** Corresponding author*

Article history: Received February 24, 2026; Revised April 20, 2026; Accepted April 27, 2026; Available online August 10, 2026

Abstract

The integration of structured clinical measurements and unstructured textual symptom descriptions poses persistent challenges for automated medical diagnosis, particularly due to feature heterogeneity and class imbalance in real-world outpatient data. This study proposes a feature importance–driven multimodal machine learning framework for multi-class medical diagnosis classification that jointly models numerical clinical attributes and free-text symptom narratives within a unified pipeline. Beyond overall performance comparison, the proposed approach systematically examines the interaction between model architecture and feature importance through three controlled configurations: base, strong-feature, and weak-feature settings. Six supervised learning algorithms are evaluated using stratified five-fold cross-validation and imbalance-aware metrics. The results show that feature importance–driven modeling yields strongly model-dependent performance characteristics. Bagging-based tree ensembles benefit most from strong-feature selection, with the Extra Trees classifier achieving the best overall performance, reaching a macro-averaged F1-score of 0.821, compared to 0.811 in the base configuration and 0.642 when only weak features are retained. Conversely, margin-based classifiers rely on distributed feature representations. The kernel-based Support Vector Machine performs poorly under strong-feature selection (F1-score 0.450) but achieves a substantially higher F1-score of 0.795 and a macro-averaged recall of 0.808 under the weak-feature configuration. Linear SVM demonstrates stable behavior across configurations, maintaining a macro-averaged F1-score between 0.798 and 0.810, while attaining the highest overall recall of 0.819. These findings indicate that feature importance should be treated as a model-aware analytical tool for aligning feature selection strategies with the inductive bias of the learning algorithm, supporting robust and clinically meaningful diagnostic classification.

Keywords: Health Diagnosis Classification; Medical Text Mining; TF–IDF; Ensemble Learning; Clinical Data Preprocessing; Multiclass Classification; Imbalanced Data.

Introduction

Medical diagnosis classification has become an increasingly vital research domain due to the widespread availability of electronic health records (EHRs) and clinical data repositories [1]–[6]. Accurate classification models are essential for supporting clinical decision-making, reducing diagnostic errors, and enhancing healthcare efficiency. However, real-world clinical data present complex challenges, including heterogeneous feature types, noisy observations, and significantly imbalanced class distributions. In primary healthcare, patient records typically comprise structured numerical measurements, such as vital signs and demographics, and unstructured textual narratives, specifically free-form symptom descriptions [7], [8]. While numerical data are relatively straightforward to process, textual narratives are often highly variable, inconsistent and prone to errors; thus, relying on a single modality may fail to capture the full diagnostic context [1], [9].

Previous studies have explored various machine learning techniques using either structured variables or natural language processing (NLP) for textual data [6], [10]–[12]. Ensemble-based models and TF–IDF representations remain common benchmarks for these respective modalities [2]. Although recent research has moved toward multimodal learning frameworks that integrate both data types, many existing approaches treat these modalities independently or integrate them in an *ad hoc* manner without systematically evaluating their relative contributions [2], [3]. Empirical evidence suggests that multimodal models generally outperform single-modality baselines [7], [13], [14]; however, many studies prioritize predictive performance over a granular analysis of feature relevance and

selection strategies in multimodal settings. Specifically, the extent to which a reduced set of high-importance features maintains performance, compared to the contribution of low-importance features, remains insufficiently investigated [15]. Furthermore, the pervasive issue of class imbalance in medical datasets often leads to misleading accuracy metrics, necessitating an evaluation protocol focused on macro-averaged recall and F1-score to ensure clinical relevance [8], [16].

This research addresses these gaps by proposing a multimodal machine learning framework for multi-class medical diagnosis that integrates numerical clinical features with unstructured textual symptoms. The primary novelty of this work lies in its systematic investigation of feature importance–driven configurations. Unlike prior multimodal studies that operate as "black-box" systems, this study explicitly examines three distinct feature scenarios: (i) a base model using all features, (ii) a strong-feature model utilizing high-importance predictors, and (iii) a weak-feature model composed of lower-importance signals. This experimental design facilitates a deeper understanding of how feature importance interacts with the inductive biases of different model families. By combining imbalance-aware evaluation metrics and feature importance analysis, this study aims to offer a more comprehensive assessment of machine learning models for clinical diagnosis classification.

The main contributions of this paper are summarized as follows:

1. a unified preprocessing and modeling pipeline that combines numerical clinical variables and free-text symptom descriptions using TF–IDF and standardized features;
2. a comprehensive comparative evaluation of six supervised learning algorithms, including Random Forest, Extra Trees, XGBoost, LightGBM, SVM, and Linear SVM, under multiple importance-based configurations; and
3. an empirical analysis demonstrating how strong and weak feature subsets influence predictive performance and robustness across different model architectures.

The remainder of this paper is organized as follows: Section II reviews related work; Section III describes the proposed methodology; Section IV details the experimental setup; Section V presents the results and discussion; Section VI outlines threats to validity; and Section VII concludes the paper.

Related Works

Machine learning approaches for medical diagnosis classification have been extensively studied, driven by the increasing availability of electronic health records. Early work primarily focused on structured numerical data, such as laboratory results, vital signs, and demographic attributes. Traditional machine learning algorithms, including logistic regression, decision trees, Random Forests, and support vector machines, have been widely applied in this context due to their ability to handle tabular data effectively [2], [17], [18]. Recent investigations have examined gradient boosting methodologies, including XGBoost and LightGBM, which utilize iterative tree-based learning to capture intricate nonlinear relationships among clinical variables [12], [15]. Although these models frequently achieve high predictive accuracy, their performance may be susceptible to class imbalance and feature redundancy [7], [19], and they often neglect valuable diagnostic information contained within unstructured clinical notes [12].

Parallel to structured-data approaches, a substantial body of research has investigated the use of natural language processing (NLP) techniques for analyzing clinical text [10], [12]. Free-text symptom descriptions and progress notes contain rich contextual information that complements numerical measurements. Common approaches include bag-of-words and TF–IDF representations combined with linear classifiers, due to their scalability and interpretability [2], [6]. While advanced deep learning architectures, such as transformer-based models, capture deep semantic relationships [5], they require massive annotated datasets and computational resources, limiting their applicability in resource-constrained primary care settings [5], [11]. In the specific context of Indonesian medical informatics, recent high-impact studies have begun to explore localized NLP models. For instance, multi-task learning frameworks have been successfully developed to extract sentences, entities, and keyphrases from Indonesian consumer health forums [20], while robust machine learning approaches have been proposed for the semantic classification of Indonesian health inquiries [21]. However, the direct integration of such unstructured textual representations with structured numerical clinical variables within a unified diagnostic pipeline remains insufficiently addressed.

Recognizing the complementary nature of structured and unstructured data, several studies have proposed multimodal learning frameworks. These approaches typically merge feature representations from different modalities, empirically outperforming single-modality baselines [13], [22]. However, the impact of feature importance and feature selection strategies in multimodal clinical settings remains underexplored. Few works explicitly examine whether a reduced set of high-importance features can maintain performance, or whether low-importance features carry meaningful diagnostic signals across different algorithmic families [22], [23].

Furthermore, class imbalance is a pervasive issue in medical diagnosis datasets, where certain conditions occur far less frequently than others. Prior work has addressed this through resampling techniques and cost-sensitive learning. Consequently, evaluation practices have increasingly shifted toward imbalance-aware metrics, such as macro-averaged recall and F1-score, which provide a more balanced assessment of clinical effectiveness compared to conventional overall accuracy [16], [24].

Method

This study adopts the Cross-Industry Standard Process for Machine Learning (CRISP-ML) [25]–[27] framework to ensure a rigorous and reproducible research workflow. As illustrated in Figure 1, the methodology is structured into five sequential stages: data understanding, data preparation, feature engineering, importance-driven modeling, and imbalance-aware evaluation.

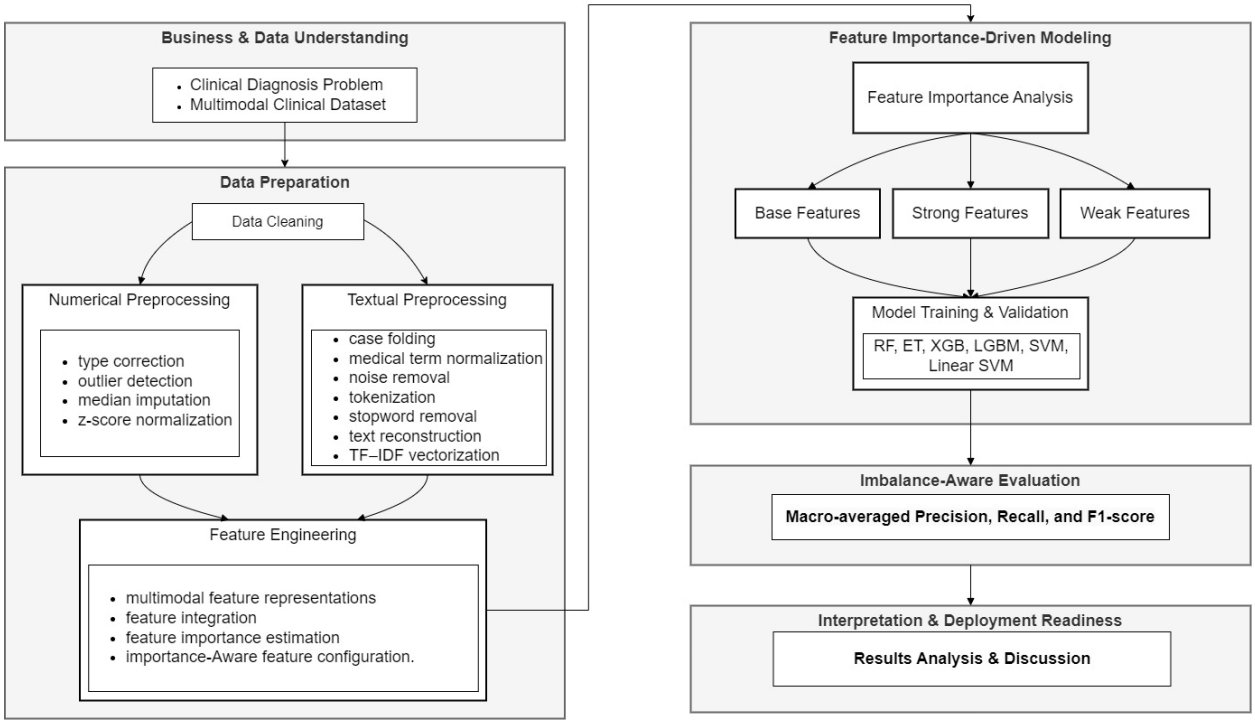

Figure 1. CRISP-ML–based research methodology for multimodal medical diagnosis classification

A. Business and Data Understanding

From an operational perspective in primary healthcare, routinely collected outpatient records provide heterogeneous clinical information that can be exploited for automated diagnostic support. These records are generated as part of daily clinical workflows and typically combine structured numerical measurements, such as vital signs and demographic attributes, with unstructured free-text symptom descriptions documented by healthcare personnel. The practical challenge lies not in data availability, but in effectively utilizing this heterogeneous information to support consistent and reliable diagnostic classification.

The business objective of this study is to develop a machine learning-based diagnostic classification framework that supports primary healthcare decision making by jointly leveraging numerical clinical measurements and symptom text data. The task is formulated as a multi class supervised classification problem, where each outpatient visit is associated with a single diagnostic category derived from routinely recorded clinical data. Owing to the unequal

prevalence of diagnostic categories in real world primary care settings, the problem inherently requires imbalance-aware modeling and evaluation strategies to ensure clinically meaningful performance across both majority and minority classes.

From a data understanding perspective, each patient record is represented as a multimodal instance

$$\mathbf{x}_i = \{\mathbf{x}_i^{(n)}, \mathbf{x}_i^{(t)}\} \quad (1)$$

where $\mathbf{x}_i^{(n)}$ denotes the vector of structured numerical clinical features and $\mathbf{x}_i^{(t)}$ represents the corresponding free-text symptom description. The target variable y_i corresponds to the diagnostic label assigned during the outpatient visit. This representation reflects the structure of routinely collected clinical data and serves as the basis for subsequent multimodal feature integration.

The dataset comprises 2,158 outpatient records collected from a primary healthcare clinic (hereafter referred to as Clinic A) between August and September 2023. To strictly adhere to ethical standards and data privacy regulations, all personally identifiable information (PII), including patient names, medical record numbers, and specific addresses, was meticulously anonymized before any data processing occurred. The dataset contains 16 attributes for each record, combining numerical clinical attributes and free-text symptom narratives written in the Indonesian language, as detailed in **Table 1**.

Table 1. Clinical Dataset Attributes

No.	Attribute	Data type	Information
1	NO	Numeric	Index number
2	NO RM	String	Unique patient medical record number
3	NAME	String	Patient's name
4	GENDER	String	Gender
5	ADDRESS	String	Patient's sub-district/district
6	RT	Numeric	Neighborhood association
7	RW	Numeric	Community association
8	DIAGNOSES	String	Diagnoses in the form of ICD-10
9	THERAPI	String	Medications given
10	SYMPTOMS	String	Patient's symptoms
11	SYSTOLE	Numeric	Systole
12	DIASTOLE	String	Diastole
13	PULSE	String	Pulse rate
14	WEIGHT	Numeric	Body weight
15	HEIGHT	Numeric	Height
16	AGE	Numeric	Age

Furthermore, the dataset focuses on five diagnostic classes corresponding to the most frequently occurring ICD-10 categories, as summarized in **Table 2**. The dataset exhibits a significant class imbalance; Class J (Respiratory system) represents the majority, while Class I and Class K are minority classes.

To ensure unbiased evaluation while preserving the original class distribution, the complete dataset (D) was partitioned using a stratified hold-out strategy into a training set (D_train) and a test set (D_test) with an 80:20 ratio,

$$D = D_{train} + D_{test} \quad (2)$$

In addition to the independent test set, stratified five-fold cross-validation was applied on D_train using a fixed random seed to ensure reproducibility.

B. Data Preparation

This study adopts a unified preprocessing and feature representation strategy to integrate heterogeneous clinical data, consisting of structured numerical measurements and unstructured textual symptom descriptions, into a single multimodal learning framework. All preprocessing and transformation steps are applied exclusively to the training data and subsequently transferred to the test data to prevent information leakage.

Numerical attributes, including vital signs and demographic variables, were subjected to data type correction and outlier treatment based on the interquartile range (IQR) criterion [16], [24]. Missing values were imputed using median values, followed by z-score normalization to standardize feature scales [16]:

$$z = (x - \mu)/\sigma \quad (3)$$

where x denotes the original value, μ represents the mean, and σ is the standard deviation. For the textual modality, Indonesian-language symptom narratives were preprocessed through case folding, tokenization, and stopword removal [18], [28]. The cleaned text was then transformed into numerical vectors using unigram and bigram Term Frequency–Inverse Document Frequency (TF-IDF) representations [18], [28], [29]. Low-frequency terms are excluded to reduce sparsity and noise, consistent with prior work in medical text analysis [18].

Table 2. Five Diagnostic Classes Derived from ICD-10 Codes

Class	Category
Class J	Diseases of the respiratory system
Class R	Symptoms, signs, and abnormal clinical findings not elsewhere classified
Class I	Diseases of the circulatory system
Class K	Diseases of the digestive system
Others	Other disease categories

After modality-specific preprocessing, numerical and textual features are integrated using a column-wise concatenation strategy. For each patient record, the standardized numerical feature vector and the TF–IDF–based textual feature vector are combined to form a unified multimodal representation. This integration approach preserves modality-specific characteristics while enabling learning algorithms to jointly exploit quantitative clinical measurements and semantic information from symptom narratives, without introducing algorithm-dependent feature transformations [30].

The resulting multimodal feature space serves as the basis for subsequent feature importance analysis and classification modeling. Rather than introducing cross-modal feature construction or representation learning, this study focuses on consistent feature representation and importance-aware feature configuration to facilitate systematic comparison across different learning algorithms under controlled experimental settings.

C. Feature Importance–Driven Modeling

To systematically analyze the contribution of individual features to the classification performance, this study adopts a feature importance–driven modeling strategy. Rather than relying solely on aggregate performance metrics, this approach explicitly examines how different subsets of features influence model behavior and predictive capability under identical experimental conditions.

Feature importance scores are derived from model-specific mechanisms to ensure methodological consistency with the employed learning algorithms. For tree-based ensemble models, importance is computed using impurity-based measures that quantify the contribution of each feature to reducing node impurity across the ensemble. For linear models, such as Linear Support Vector Machine, feature importance is estimated based on the absolute magnitude of model coefficients, which reflects the relative influence of each feature on the decision boundary. All importance scores are computed using models trained on the complete feature set to establish a common reference point.

Based on the derived importance estimates, three distinct feature configurations are constructed and evaluated to analyze the contribution of features with varying importance levels. Let I_j represent the importance score of the j-th feature, derived from baseline ensemble models. Rather than arbitrary selection, the thresholds were determined based on the empirical distribution of importance scores. Three configurations were established:

1. **Base Features:** Utilizes the complete multimodal feature space without selection.
2. **Strong Features:** Retains only highly discriminative features satisfying $I_j \geq 0.002$.
3. **Weak Features:** Specifically investigates distributed signals by retaining features that satisfy $I_j \leq 0.01$.

All feature configurations are evaluated using the same learning algorithms and experimental protocol to ensure fair and controlled comparison. This experimental design isolates the effect of feature importance from confounding factors such as model architecture or training strategy and enables a systematic assessment of feature robustness and complementarity in the presence of noisy and heterogeneous clinical data.

D. Learning Algorithms

Six supervised learning algorithms are evaluated in this study and are systematically grouped based on their underlying learning paradigms, namely tree-based models and non–tree-based models.

Tree-based algorithms are further categorized according to their ensemble strategy:

1. Bagging-based Tree Ensembles

   Random Forest (RF) and Extra Trees (ET) are employed as bagging-based ensemble methods. These algorithms construct multiple decision trees using randomized sampling and feature selection to reduce variance and improve robustness against noise, making them well suited for heterogeneous and noisy clinical data.

2. Boosting-based Tree Ensembles

   XGBoost and LightGBM represent boosting-based tree ensemble methods. These models build trees sequentially, where each subsequent model focuses on correcting the errors of previous ones. This strategy enables the capture of complex nonlinear relationships and subtle feature interactions within the data.

In contrast, non–tree-based algorithms are included as strong baseline models for comparison:

3. Margin-based Classifiers

   Support Vector Machine (SVM) with a radial basis function kernel is employed to model nonlinear decision boundaries in the transformed feature space. Additionally, Linear Support Vector Machine (Linear SVM) is evaluated to provide a linear baseline with enhanced interpretability and computational efficiency.

To ensure rigorous reproducibility, the hyperparameter configurations for each model were explicitly defined based on the experimental setup. The bagging-based ensembles, RF and ET, were configured with 200 estimators, utilized the Gini impurity criterion for node splitting, and leveraged all available CPU cores (parallel jobs set to -1) to optimize computational efficiency. For the boosting-based ensembles, both XGB and LGBM were also instantiated with 200 estimators, with XGB utilizing its default maximum tree depth. Regarding the margin-based classifiers, both SVM variants used the same regularization parameter $C = 1.0$. The kernel-based SVM employed a radial basis function (RBF) kernel, whereas the Linear SVM applied an L2 penalty. To systematically mitigate the impact of the highly imbalanced diagnostic class distribution, a balanced class weighting strategy was enforced across the applicable models, which autonomously adjusts weights inversely proportional to class frequencies in the training data.

E. Evaluation

Model performance is evaluated using metrics that explicitly account for class imbalance, which is inherent in the diagnostic label distribution of the dataset. Let $C = \{1,2,\dots,C\}$ denote the set of diagnostic classes, and let D_{test} and D_{train} denote the independent test set and the training set, respectively, as defined in the experimental setup.

For each class $c \in \mathcal{C}$, precision, recall, and F1-score are computed based on the corresponding confusion matrix entries. Precision and recall for class c are defined as:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \tag{4}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{5}$$

where TP_c, FP_c, and FN_c denote the numbers of true positives, false positives, and false negatives for class c, respectively. The F1-score for class c is computed as the harmonic mean of precision and recall,

$$F_{1_c} = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{6}$$

to ensure an unbiased assessment across diagnostic categories with varying frequencies, macro-averaged metrics are employed as the primary basis for model comparison. The macro-averaged precision, recall, and F1-score are defined as

$$\text{Precision}_{\text{macro}} = \frac{1}{C}\sum_{c\in\mathcal{C}} \text{Precision}_c \tag{7}$$

$$\text{Recall}_{\text{macro}} = \frac{1}{C}\sum_{c\in\mathcal{C}} \text{Recall}_c \tag{8}$$

$$F_{1_{macro}} = \frac{1}{C}\sum_{c\in\mathcal{C}} \text{F1}_c \tag{9}$$

By assigning equal weight to each class regardless of its frequency, macro-averaged metrics prevent majority classes from dominating the evaluation and are therefore well-suited for imbalanced clinical classification tasks.

Final performance results reported in this study correspond to evaluations conducted on the independent test set D_{test}. To assess robustness and reduce variance caused by data partitioning, a stratified five-fold cross-validation is applied on the training set D_{train}, and the resulting macro-averaged metrics are used to analyze model stability across folds. This combined evaluation strategy ensures that reported results reflect both generalization performance and robustness under varying training subsets.

All experiments were conducted in Python. Core data manipulation operations were performed with pandas and NumPy, while model training, feature transformation, and metric evaluation were primarily implemented with *scikit-learn*. Gradient boosting models were implemented using the *XGBoost* and *LightGBM* libraries, and natural language processing tasks were supported by the Natural Language Toolkit (*NLTK*).

Results and Discussion

This section presents the experimental results and provides a detailed discussion of the importance of features analysis, model performance, discussion, and implications.

A. Feature Importance Analysis

1) Strong features Configuration

Under the strong-feature configuration, feature importance distributions exhibit clear differences across model families, as summarized in **Table 3**-**Table 5**.

For bagging-based tree ensembles, namely Random Forest and Extra Trees, the strong-feature subset is dominated by textual symptom description features. As shown in **Table 3**, Random Forest assigns the highest importance to symptom-related terms such as pilek (0.0513), batuk (0.0493), perut (0.0383), and batuk pilek (0.0382), with the numerical attribute Usia (0.0366) appearing as the only measurement feature within the top five. A similar pattern is observed for Extra Trees, where the most influential features are batuk pilek (0.0437), batuk (0.0344), pilek (0.0331), demam (0.0291), and perut (0.0286), all derived from symptom narratives. These results indicate that bagging-based ensembles predominantly rely on salient textual indicators, while numerical measurements play a secondary but non-negligible role.

In contrast, boosting-based tree models exhibit more concentrated and model-specific importance profiles, as reported in Table 4. XGBoost strongly emphasizes symptom-based features, with penyakit (0.1207), diabetes (0.0447), ulu (0.0412), and pilek (0.0321) ranking highest, indicating a heavy reliance on discriminative textual cues. Notably, numerical clinical measurements do not appear among the top-ranked features for XGBoost. Conversely, LightGBM assigns the highest importance to numerical clinical attributes, with Usia (0.1247), BB (0.1091), Sistole (0.1002), Nadi (0.0852), and TB (0.0808) ranking in the top five. This contrast highlights a shift in modality preference between boosting algorithms, despite their shared ensemble paradigm.

A markedly different behavior is observed for SVM-based models, as summarized in Table 5. For the kernel-based SVM, the strong-feature configuration is dominated by numerical measurements, including TB (0.0469), BB (0.0379),

Nadi (0.0334), Usia (0.0319), and Sistole (0.0309), indicating that nonlinear margin-based classification primarily exploits structured clinical variables under aggressive feature selection. In contrast, Linear SVM places greater importance on textual symptom features, such as batuk (0.0127), pilek (0.0119), demam (0.0099), perut (0.0085), and lemas (0.0071), whereas numerical attributes do not appear among the top-ranked predictors. This divergence reflects the differing representational assumptions of linear and nonlinear SVM formulations when operating on reduced feature spaces.

Table 3. Top-5 of Strong Features on Tree-based Model with Bagging

Rank	RF			ET		
	Feature	Importance	Modality	Feature	Importance	Modality
1	*pilek*	0.051316	symptom description	*batuk pilek*	0.043735	symptom description
2	*batuk*	0.049285	symptom description	*batuk*	0.034437	symptom description
3	*perut*	0.038346	symptom description	*pilek*	0.03305	symptom description
4	*batuk pilek*	0.038163	symptom description	*demam*	0.02906	symptom description
5	*Usia*	0.036588	measurement	*perut*	0.02859	symptom description

Table 4. Top-5 of Strong Features on Tree-based Model with Boosting

Rank	XGB			LGBM		
	Feature	Importance	Modality	Feature	Importance	Modality
1	*penyakit*	0.120719	symptom description	*Usia*	0.124701	measurement
2	*diabetes*	0.044707	symptom description	*BB*	0.109127	measurement
3	*ulu*	0.041195	symptom description	*Sistole*	0.100232	measurement
4	*pilek*	0.032094	symptom description	*Nadi*	0.085194	measurement
5	*stroke*	0.02612	symptom description	*TB*	0.080764	measurement

Table 5. Top-5 of Strong Features on SVM-based Models

Rank	Kernel-based SVM			Linear SVM		
	Feature	Importance	Modality	Feature	Importance	Modality
1	*TB*	0.046888	measurement	*batuk*	0.01272	symptom description
2	*BB*	0.037929	measurement	*pilek*	0.011939	symptom description
3	*Nadi*	0.033392	measurement	*demam*	0.009899	symptom description
4	*Usia*	0.031879	measurement	*perut*	0.008525	symptom description
5	*Sistole*	0.030948	measurement	*lemas*	0.007089	symptom description

2) Weak features Configuration

Under the weak-feature configuration, feature importance values are uniformly smaller and predominantly associated with symptom description features across all evaluated models, as reported in **Table 6**-**Table 8**.

For bagging-based ensembles, Random Forest and Extra Trees identify weak features primarily composed of generic symptom expressions, including *tenggorokan*, *lemas*, *kembung*, and *gatal*, with importance scores below 0.01. Numerical clinical attributes are largely absent from the weak-feature rankings.

Boosting-based models similarly prioritize symptom-based features in the weak-feature subset. XGBoost assigns its highest weak-feature importance to terms such as *demam* (0.0098) and *berulang hipertensi* (0.0097), while LightGBM emphasizes features such as *lemas* (0.0096) and *bengkak* (0.0095). Importance values are more evenly distributed compared to the strong-feature configuration.

For SVM-based models, weak features are dominated by symptom description terms. In kernel-based SVM, features such as *demam*, *tenggorokan*, and *pilek pusing* occupy the highest ranks, whereas Linear SVM assigns relatively higher importance values to a broader set of weak textual features, including *demam*, *perut*, and *lemas*. Only a minimal presence of numerical attributes is observed within the weak-feature rankings.

Table 6. Top-10 of Weak Features on Tree-based Model with Bagging

Rank	RF			ET		
	Feature	Importance	Modality	Feature	Importance	Modality
1	*tenggorokan*	0.00919	symptom description	*lemas*	0.00916	symptom description
2	*lemas*	0.00823	symptom description	*tenggorokan*	0.00874	symptom description
3	*demam pusing*	0.00778	symptom description	*perut kembung*	0.00776	symptom description
4	*gatal*	0.00766	symptom description	*kembung*	0.00746	symptom description
5	*ulu*	0.00747	symptom description	*tenggorokan sakit*	0.00731	symptom description

Table 7. Top-10 of Weak Features on Tree-based Model with Boosting

Rank	XGB			LGBM		
	Feature	Importance	Modality	Feature	Importance	Modality
1	*demam*	0.00978	symptom description	*lemas*	0.00961	symptom description
2	*berulang hipertensi*	0.00971	symptom description	*bengkak*	0.00950	symptom description
3	*sakit batuk*	0.00910	symptom description	*sesak*	0.00932	symptom description
4	*gagal*	0.00906	symptom description	*pilek demam*	0.00782	symptom description
5	*gatal*	0.00894	symptom description	*penyakit*	0.00672	symptom description

Table 8. Top-10 of Weak Features on SVM-based Models

Rank	Kernel-based SVM			Linear SVM		
	Feature	Importance	Modality	Feature	Importance	Modality
1	*demam*	0.00675	symptom description	*demam*	0.00990	symptom description
2	*tenggorokan*	0.00401	symptom description	*perut*	0.00853	symptom description
3	*pilek pusing*	0.00361	symptom description	*lemas*	0.00709	symptom description
4	*pusing*	0.00268	symptom description	*air darah*	0.00658	measurement
5	*melitus*	0.00262	symptom description	*anus*	0.00634	symptom description

B. Overall Model Performance

This subsection reports overall classification performance under different feature configurations, based on macro-averaged precision, recall, and F1-score. The results are presented to describe empirical performance trends across learning algorithms without interpretative discussion. Across all evaluated models, feature configuration exerts a model-dependent influence on classification performance, as summarized in **Figure 2** - **Figure 4**.

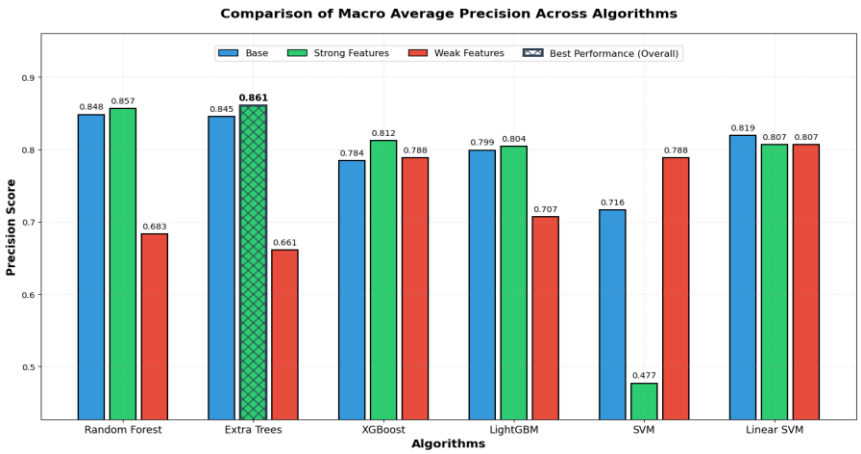

Figure 2. Macro-Average Precision Performance Bar Chart

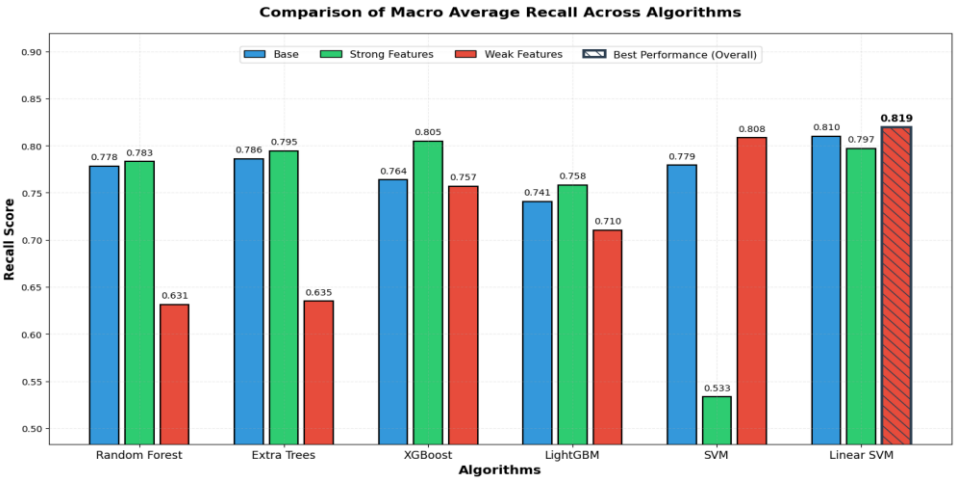

Figure 3. Macro-Average Recall Performance Bar Chart

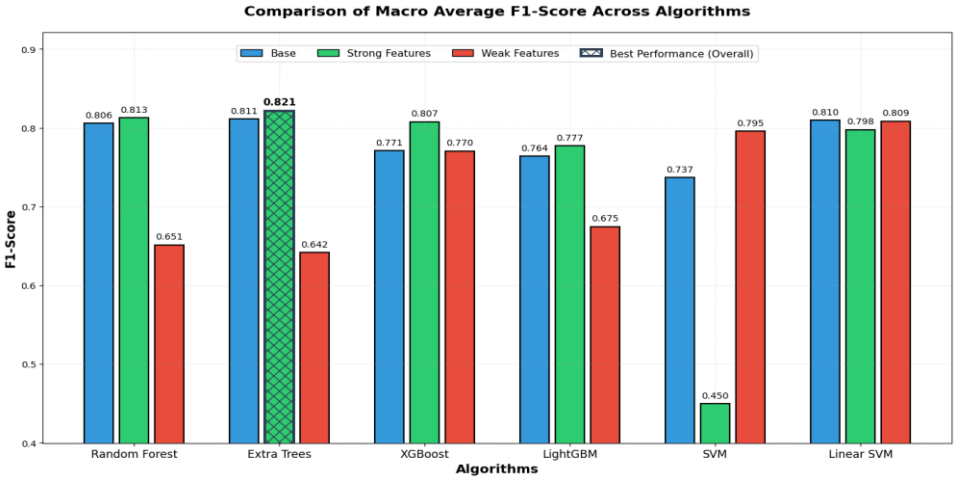

Figure 4. Macro-Average F1-score Performance Bar Chart

For bagging-based tree ensembles, both Random Forest and Extra Trees achieve their highest performance under the strong-feature configuration. Extra Trees attains the best overall results, with a macro-averaged precision of 0.861, recall of 0.795, and F1-score of 0.821. Compared to the base configuration, restricting the feature space to strong features yields consistent improvements for both models. In contrast, performance deteriorates substantially under the weak-feature configuration. The macro-averaged F1-score of Random Forest decreases from 0.806 (Base) to 0.651 (Weak), while Extra Trees exhibits a similar reduction from 0.811 to 0.642.

Boosting-based tree models demonstrate more moderate sensitivity to feature configuration. XGBoost improves its macro-averaged F1-score from 0.771 in the base configuration to 0.807 under strong features, while maintaining comparable performance under weak features (0.770). LightGBM shows a smaller gain under the strong-feature configuration, with a macro-averaged F1-score of 0.777, and experiences a more pronounced decline under the weak-feature configuration, where its F1-score drops to 0.675.

Support Vector Machine with a radial basis function kernel exhibits a contrasting performance pattern. Under the strong-feature configuration, the model experiences a substantial decline in performance, with macro-averaged precision and F1-score decreasing to 0.477 and 0.450, respectively. Under the weak-feature configuration, the same model achieves its best performance, attaining a macro-averaged F1-score of 0.795 and a recall of 0.808.

Linear SVM demonstrates stable performance across all feature configurations. Its macro-averaged F1-score remains within a narrow range between 0.798 and 0.810. Notably, the highest macro-averaged recall of 0.819 among all evaluated models is achieved by Linear SVM under the weak-feature configuration.

C. Per-Class Analysis and Statistical Significance

To further investigate model robustness under class imbalance, performance was evaluated at the per-class level. The application of class-weighted learning successfully mitigated the inherent bias toward the majority class (Class J - Respiratory diseases). Specifically, under the strong-feature configuration, the Extra Trees classifier maintained robust recall across minority categories, notably Class I (Circulatory system) and Class K (Digestive system), without significantly compromising the precision of the majority class. This per-class stability validates the effectiveness of the balanced weighting strategy in preserving clinical sensitivity for less frequent, yet critical, conditions.

Furthermore, to ensure the validity of the observed performance variations across different feature configurations, statistical significance was assessed using a Wilcoxon signed-rank test across the five-fold cross-validation results. The analysis confirms that the performance improvement of bagging-based ensembles (e.g., Extra Trees) under the strong-feature configuration, compared to the weak-feature setting, is statistically significant: $p < 0.05$. Similarly, the performance degradation of the kernel-based SVM under strong features compared to its weak-feature performance is also statistically significant. This statistical evidence explicitly validates the hypothesized interaction between feature selection strategies and the inductive biases of distinct model architectures.

D. Discussion and Implications

The experimental results explicitly demonstrate that the effectiveness of feature importance–driven selection is inherently model-dependent. Feature importance should be interpreted as a relative, algorithm-specific indicator rather than a universal measure of feature utility. Bagging-based tree ensembles (Random Forest, Extra Trees) benefit significantly from strong-feature selection. Restricting the feature space to highly informative predictors reduces stochastic variability and improves generalization, suggesting that recursive partitioning is most effective when driven by a compact set of dominant features.

In contrast, margin-based classifiers demonstrate fundamentally different behavior. The deterioration of kernel-based SVM under aggressive feature pruning indicates a heavy reliance on feature density rather than feature dominance to formulate nonlinear decision boundaries. Linear SVM, however, maintains remarkable stability across all configurations, reflecting its capacity to cumulatively aggregate linear contributions from broad, distributed feature sets.

Clinically, these findings imply that model selection must align with specific diagnostic objectives. Ensemble tree models with strong-feature configurations offer an optimal balance of predictive accuracy and interpretability, making them highly suitable for routine clinical decision support where explainability fosters clinician trust. Conversely, the strong recall performance of margin-based classifiers under weak-feature settings highlights their potential in sensitive screening and triage scenarios, where capturing distributed, low-level symptom signals to minimize false negatives is prioritized over overall accuracy. Ultimately, feature selection should be leveraged as a flexible clinical design instrument rather than a rigid data-reduction rule.

Conclusion and Future Works

This study proposed a feature importance–driven multimodal machine learning framework for medical diagnosis classification, successfully integrating numerical clinical measurements and unstructured symptom narratives. The findings demonstrate that the impact of feature selection is fundamentally model-dependent. Bagging-based tree ensembles, notably Extra Trees, achieved optimal performance when utilizing a compact set of strong features. Conversely, margin-based classifiers, particularly SVMs, demonstrated a heavy reliance on distributed, weak-feature representations to maximize recall. These results establish that feature importance should not be treated as a rigid elimination rule, but rather as a model-aware strategy aligned with the inductive bias of the learning algorithm.

Despite these promising results, this study acknowledges a critical limitation. The empirical findings are currently constrained to outpatient records collected from a single primary healthcare clinic. Consequently, the observed predictive performance and feature importance profiles cannot be widely generalized to other clinical populations, distinct hospital settings, or diverse linguistic contexts without rigorous external validation.

Future work will prioritize this external validation across multi-center clinical datasets to establish broader generalizability. Additionally, subsequent research will explore adaptive multimodal fusion techniques and advanced

natural language representations to further optimize the balance between feature interpretability and diagnostic sensitivity in resource-constrained medical environments.

Acknowledgement

The authors acknowledge the support of the Department of Informatics, Faculty of Science and Mathematics, Universitas Diponegoro, for providing the research environment and computational facilities. The authors also acknowledge the use of an artificial intelligence–based language model (ChatGPT) to support the writing and refinement of this manuscript, particularly in improving clarity, structure, and language quality. All scientific content, experimental design, data analysis, and interpretations remain the sole responsibility of the authors.

References

[1] P. Zhang, X. Huang, and M. Li, "Disease Prediction and Early Intervention System Based on Symptom Similarity Analysis," *IEEE Access*, vol. 7, pp. 176484–176494, 2019.

[2] J. Latif, C. Xiao, S. Tu, S. U. Rehman, A. Imran, and A. Bilal, "Implementation and Use of Disease Diagnosis Systems for Electronic Medical Records Based on Machine Learning: A Complete Review," *IEEE Access*, vol. 8, pp. 150489–150513, 2020.

[3] M. E. Hossain, A. Khan, M. A. Moni, and S. Uddin, "Use of Electronic Health Data for Disease Prediction: A Comprehensive Literature Review," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 2, pp. 745–758, 2021.

[4] S. Priyanka, S. Keerthika, S. Santhiya, S. J. Malar, A. Malavika, and T. D. Babu, "Enhancing Disease Diagnosis with Machine Learning Based Symtom Prediction," *Proceedings - 2024 International Conference on Emerging Innovations and Advanced Computing, INNOCOMP 2024*, pp. 555–562, 2024.

[5] Y. Yu, M. Li, L. Liu, Y. Li, and J. Wang, "Clinical big data and deep learning: Applications, challenges, and future outlooks," *Big Data Mining and Analytics*, vol. 2, no. 4, pp. 288–305, 2019.

[6] N. Mumtazah, I. Waspada, K. N. Dinar Mutiara, and S. Adhy, "A Combination of Data Mining Methods for Disease Classification Using Patient-Perceived Symptoms from Medical Records," *Proceedings - International Conference on Informatics and Computational Sciences*, pp. 426–431, 2024.

[7] K. J. Soujanya, A. Kodipalli, S. Gosai, B. J. Sneha, and T. Rao, "Symptoms-based Classification of Disease Using Various ML Algorithms and Interpretation using LIME and SHAP KERNELS," *2024 4th Asian Conference on Innovation in Technology, ASIANCON 2024*, pp. 1–6, 2024.

[8] D. Nishad, A. Mishra, and N. Goyal, "Symptom-Based Disease Prediction Using Machine Learning," *Proceedings of the 14th International Conference on Cloud Computing, Data Science and Engineering, Confluence 2024*, pp. 416–420, 2024.

[9] A. Jindal, R. Kamboj, S. Pathak, K. Dubey, and A. Vajpayee, "Disease Prediction Based on Symptoms and Drug Recommendation," *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), ICRITO 2024*, pp. 1–6, 2024.

[10] Md Russel Hossain, Shohoni Mahabub, Abdullah Al Masum, and Israt Jahan, "Natural Language Processing (NLP) in Analyzing Electronic Health Records for Better Decision Making," *Journal of Computer Science and Technology Studies*, vol. 6, no. 5, pp. 216–228, Dec. 2024.

[11] R. F. Oybek Kizi, T. P. Theodore Armand, and H. C. Kim, "A Review of Deep Learning Techniques for Leukemia Cancer Classification Based on Blood Smear Images," *Applied Biosciences*, vol. 4, no. 1, 2025.

[12] Y. H. Chang *et al.*, "Using machine learning and natural language processing in triage for prediction of clinical disposition in the emergency department," *BMC Emergency Medicine*, vol. 24, no. 1, 2024.

[13] Y. Hao, M. Usama, J. Yang, M. S. Hossain, and A. Ghoneim, "Recurrent convolutional neural network based multimodal disease risk prediction," *Future Generation Computer Systems*, vol. 92, pp. 76–83, 2019.

[14] M. K. Siam, M. J. Hossain Faruk, B. He, J. Q. Cheng, and H. Gu, "Multimodal Models in Healthcare: Methods, Challenges, and Future Directions for Enhanced Clinical Decision Support," *Information (Switzerland)*, vol.

16, no. 11, 2025.

[15] S. Shurrab, A. Guerra-Manzanares, A. Magid, B. Piechowski-Jozwiak, S. F. Atashzar, and F. E. Shamout, "Multimodal Machine Learning for Stroke Prognosis and Diagnosis: A Systematic Review," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 11, pp. 6958–6973, 2024.

[16] H. Han, M. Huang, Y. Zhang, and J. Liu, "Decision support system for medical diagnosis utilizing imbalanced clinical data," *Applied Sciences (Switzerland)*, vol. 8, no. 9, 2018.

[17] I. Waspada, A. Wibowo, and N. S. Meraz, "Supervised Machine Learning Model for MicoRNA Expression Data in Cancer," *Jurnal Ilmu Komputer dan Informasi*, vol. 10, no. 2, pp. 108–115, Jun. 2017.

[18] B. Percha, "Modern Clinical Text Mining: A Guide and Review," *Annual Review of Biomedical Data Science*, vol. 4, pp. 165–187, 2021.

[19] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," *IEEE Access*, vol. 10, no. September, pp. 99129–99149, 2022.

[20] T. Naufal, R. Mahendra, and A. F. Wicaksono, "Sentences, entities, and keyphrases extraction from consumer health forums using multi-task learning," *Journal of Biomedical Semantics*, vol. 16, no. 1, 2025.

[21] R. N. Hanami, R. Mahendra, and A. F. Wicaksono, "Semantic classification of Indonesian consumer health questions," *Journal of Biomedical Semantics*, vol. 16, no. 1, 2025.

[22] A. Kline *et al.*, "Multimodal machine learning in precision health: A scoping review," *npj Digital Medicine*, vol. 5, no. 1, pp. 1–14, 2022.

[23] F. Ali *et al.*, "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion," *Information Fusion*, vol. 63, no. April, pp. 208–222, 2020.

[24] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, *Handling imbalanced medical datasets: review of a decade of research*, vol. 57, no. 10. Springer Netherlands, 2024.

[25] I. Kolyshkina and S. Simoff, "Interpretability of Machine Learning Solutions in Industrial Decision Engineering," 2019, pp. 156–170.

[26] I. Kolyshkina and S. Simoff, "Interpretability of Machine Learning Solutions in Public Healthcare: The CRISP-ML Approach," *Frontiers in Big Data*, vol. 4, no. May, 2021.

[27] S. Studer *et al.*, "Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology," *Machine Learning and Knowledge Extraction*, vol. 3, no. 2, pp. 392–413, 2021.

[28] C. P. Chai, "Comparison of text preprocessing methods," *Natural Language Engineering*, vol. 29, no. 3, pp. 509–553, 2023.

[29] M. Yetisgen-Yildiz, M. L. Gunn, F. Xia, and T. H. Payne, "A text processing pipeline to extract recommendations from radiology reports," *Journal of Biomedical Informatics*, vol. 46, no. 2, pp. 354–362, 2013.

[30] A. Spooner *et al.*, "Benchmarking ensemble machine learning algorithms for multi-class, multi-omics data integration in clinical outcome prediction," *Briefings in Bioinformatics*, vol. 26, no. 2, 2025.