Comparative Study of Machine Learning Models for Textual Medical Notes Classification

Yan Zhang [*,†,‡] , Huynh Trung Nguyen Le [†,‡] , Nathan Lopez [†,‡] , Kira Phan [†,‡]

Posted Date: 18 November 2025

doi: 10.20944/preprints202511.1173.v1

Keywords: medical text classification; machine learning; clinical notes analysis; logistic regression; electronic health records

 doi:10.20944/preprints202511.1173.v1

Article

Comparative Study of Machine Learning Models for Textual Medical Notes Classification

Yan Zhang *,†,‡ and Huynh Trung Nguyen Le †,‡ and Nathan Lopez †,‡ and Kira Phan †,‡

School of Computer Science and Engineering, California State University San Bernardino
* Correspondence: yan.zhang@csusb.edu; Tel.: +1-909-537-5333
† Current address: 5500 University Parkway, San Bernardino, CA, 92407, USA.
‡ All authors contributed differently to this work.

Abstract

The expansion of electronic health records (EHRs) has generated a large amount of unstructured textual data, such as clinical notes and medical reports, which contain diagnostic and prognostic information. Effective classification of these textual medical notes is critical for improving clinical decision support and healthcare data management. This study presents a comparative analysis of four traditional machine learning algorithms, Random Forest, Logistic Regression, Multinomial Naive Bayes, and Support Vector Machine, for multiclass classification of medical notes into four disease categories: Neoplasms, Digestive System Diseases, Nervous System Diseases, and Cardiovascular Diseases. A dataset containing 9,633 labeled medical notes was preprocessed through text cleaning, lemmatization, stop-word removal, and vectorization using term frequency–inverse document frequency (TF-IDF) representation. Each model was tuned using grid search and cross validation to optimize classification performance. Evaluation metrics, including accuracy, precision, recall, and F1-score, were used to assess model performance. The results indicate that Logistic Regression achieved the highest overall accuracy (0.83), followed closely by Random Forest, Support Vector Machine and Naive Bayes (0.80 each). These findings confirm that traditional machine learning models remain robust, interpretable, and computationally efficient tools for textual medical note classification.

Keywords: medical text classification; machine learning; clinical notes analysis; logistic regression; electronic health records

1. Introduction

The growing digitization of healthcare systems has led to the exponential accumulation of clinical data in electronic health records (EHRs) [1,2]. Among these records, textual medical notes, such as physicians' observations, discharge summaries, and clinical reports, contain valuable information that can reflect patients' conditions and diagnoses. However, the unstructured and heterogeneous nature of such textual data poses challenges for automated processing and analysis [3]. Efficiently classifying textual medical notes into relevant disease categories can facilitate improved clinical decision making and support healthcare management systems through faster retrieval and knowledge extraction.

In recent years, Machine Learning (ML) approaches have shown promise in transforming unstructured clinical text into actionable insights. By learning patterns and associations from labeled datasets, ML algorithms can help automatically categorize medical documents according to disease types. This capability is important for developing intelligent healthcare applications such as predictive diagnostic tools and population health monitoring frameworks [4]. Selecting an appropriate classification model remains a challenge, as different algorithms show varying degrees of performance depending on data characteristics, preprocessing methods, and feature representations.

Many studies have investigated text classification methods in medical and clinical domains. Early approaches relied mainly on rule-based systems and manual feature engineering, which required

 doi:10.20944/preprints202511.1173.v1

domain expertise and were limited in scalability [5,6]. With advancements in natural language processing (NLP) and Artificial Intelligence (AI), machine learning algorithms such as Naive Bayes, Logistic Regression, Support Vector Machines (SVMs), and ensemble methods like Random Forests have emerged as effective approaches for medical text classification. These algorithms differ in their learning strategies, underlying assumptions, and robustness to noise, which make comparative analysis essential for identifying their relative strengths and weaknesses in healthcare applications.

Despite notable progress, the comparative evaluation of traditional ML models on disease-specific medical note classification remains underexplored. Many prior works focus on binary classification tasks (e.g., disease vs. non-disease) or utilize domain-specific corpora with limited generalizability. Moreover, while deep learning models have gained more attention, they often require large-scale datasets and substantial computational resources, which may not be feasible for smaller healthcare institutions [6]. Therefore, a systematic comparison of conventional ML algorithms on a moderately sized, multiclass textual medical dataset remains a valuable contribution to the field.

This research aims to address this gap by conducting a comparative study of four widely used machine learning models, i.e., Random Forest, Logistic Regression, Naive Bayes, and Support Vector Machine, for classifying textual medical notes. The dataset employed contains over 9,000 labeled clinical notes categorized into four major disease classes: Neoplasms, Digestive System Diseases, Nervous System Diseases, and Cardiovascular Diseases. Each model is trained and fine-tuned using appropriate preprocessing and feature extraction techniques to optimize classification performance. The study evaluates and contrasts these models using standard metrics such as accuracy, precision, recall, and F1-score to determine their relative effectiveness.

The key contributions of this paper can be summarized as follows:

- A comparative performance analysis of four established machine learning algorithms on a multi-class disease classification task.
- Empirical insights into the strengths, limitations, and suitability of traditional ML approaches for clinical text analytics.

The remainder of this paper is organized as follows. Section 2 reviews related work on medical text classification and prior comparative studies. Section 3 describes the dataset and preprocessing procedures used in this study. Section 4 details the methodologies of the selected machine learning models. Section 5 presents and discusses the experimental results. Finally, Section 6 concludes the paper and outlines directions for future research.

2. Related Work

The growing availability of electronic health records (EHRs) and digitized clinical narratives has accelerated research on machine learning (ML) and natural language processing (NLP) applications in healthcare [7]. Over the past decade, numerous studies have explored how ML techniques can extract, classify, and interpret meaningful information from unstructured medical text, such as clinical notes and discharge summaries. Existing studies range from broad surveys that summarize progress and challenges in the field to empirical investigations comparing traditional supervised algorithms, and more recently, deep learning and hybrid frameworks capable of modeling complex semantic patterns.

2.1. Surveys on Machine Learning for Clinical and Medical Text

Several surveys have examined the increasing integration of ML within healthcare and clinical NLP. Spasic et al. conducted a systematic review of 110 studies applying ML to clinical text and identified text classification as the most common NLP task in healthcare [4]. Authors found that most datasets were small and institution-specific, limiting model generalizability, and emphasized the annotation bottleneck as a critical challenge for supervised learning [4]. Strategies such as active learning, distant supervision, and crowdsourcing were discussed as potential solutions to reduce manual labeling costs.

 doi:10.20944/preprints202511.1173.v1

Mustafa et al. surveyed the emerging field of Automated Machine Learning (AutoML) in healthcare, highlighting its potential for clinical note analysis [8]. Although AutoML has shown promise in structured data settings, its application to unstructured medical text remains underdeveloped. The authors noted key barriers including data heterogeneity, privacy concerns, and model interpretability, concluding that an AutoML platform for clinical notes could greatly enhance scalability and reduce human effort in ML-based healthcare solutions [8].

Kino et al. provided a scoping review of ML applications to the social determinants of health (SDH) [9]. Reviewing 82 studies published before 2020, they observed that most used predictive ML models on structured survey data, with limited exploration of unstructured sources such as clinical narratives. The authors underscored the broader expansion of ML into health research and emphasized the need for interpretable, transparent, and well-validated approaches when applying ML to clinical textual data.

Kadhim offered a comprehensive overview of supervised ML techniques for text classification, detailing the standard pipeline of data preprocessing, feature extraction, and model evaluation [10]. The review compared algorithms such as Naive Bayes (NB), Support Vector Machines (SVM), and k-Nearest Neighbors (k-NN), noting that TF–IDF weighting schemes significantly enhance classification accuracy.

These reviews establish the theoretical and methodological foundation for applying ML to unstructured health text and motivate the empirical comparisons undertaken in the present study.

2.2. Traditional Machine Learning Approaches for Medical Text Classification

A broad range of studies demonstrate that traditional supervised learning remains effective for medical text classification and disease prediction. Weng et al. developed a machine learning–based NLP framework for medical subdomain classification of clinical notes [11]. Using cTAKES and UMLS features, the authors compared SVMs with convolutional recurrent neural networks and showed that SVMs offered comparable accuracy with better interpretability, validating their utility for cross-institutional applications [11].

López-Úbeda et al. proposed an ML-based system for automatic classification of radiological protocols using a corpus of 700,000 CT and MRI reports [12]. Several NLP-driven classifiers were evaluated, including SVM, Random Forest, neural networks, and transfer-learning approaches. The system achieved high accuracy and has since been implemented as a clinical decision-support tool, demonstrating the practical value of ML for workflow optimization.

Tiwari et al. examined multiclass disease prediction using Random Forest, SVM, Naive Bayes, and Decision Tree algorithms on a symptom dataset [13]. Both Random Forest and Decision Tree achieved the highest accuracy (≈99%), whereas Naive Bayes yielded the lowest (≈86%), confirming the effectiveness of ensemble and tree-based methods for healthcare classification tasks.

Sung et al. applied supervised ML and text mining to automated phenotyping of ischemic stroke using 4,640 patient EMRs [14]. The integration of structured variables with textual data improved classification, and decomposing the multiclass problem into binary subtasks further enhanced performance. Their findings highlight the potential of ML to replace manual annotation in disease phenotyping.

Rabby and Berka investigated multi-class classification of COVID-19 biomedical research papers using ten ML algorithms and eleven feature configurations [15]. They found that TF–IDF features of abstracts yielded the highest accuracy, with Random Forest and BERT models performing best, demonstrating the versatility of traditional ML for biomedical document classification.

Gupta et al. developed an NLP pipeline to automatically identify immune-related adverse events (irAEs) from unstructured oncology notes [16]. Employing keyword filtering, TF–IDF, and BioWordVec embeddings as input for Logistic Regression, SVM, Random Forest, CNN, and Bi-LSTM models, they achieved an F1 score of 0.75 and AUC of 0.85, demonstrating that classical ML methods augmented with embeddings can automate complex clinical annotation tasks.

 doi:10.20944/preprints202511.1173.v1

Gao et al. introduced KeyClass, a weakly supervised framework for assigning ICD-9 codes to unstructured clinical notes without manual labeling [17]. Tested on the MIMIC-III dataset, KeyClass achieved performance comparable to supervised models trained on thousands of labeled samples, underscoring the promise of weak supervision for scalable medical text classification.

Lenivtceva et al. explored multi-label classification of 11,671 Russian medical notes [18]. The authors compared several algorithms and proposed classifier-chain ensembles to capture inter-label dependencies, achieving notable performance gains and illustrating the strength of ensemble strategies for complex medical text tasks.

These studies confirm that traditional ML models, particularly SVM, Logistic Regression, and Random Forest, are able to offer robust, interpretable, and computationally efficient baselines for medical text classification.

2.3. Advances and Extensions Using Deep Learning and Hybrid Methods

While traditional algorithms remain effective, recent research increasingly apply deep learning and hybrid NLP approaches to capture the semantic richness of clinical text. da Silva et al. evaluated machine learning and deep learning models for oncology clinical notes, comparing Logistic Regression, Random Forest, Decision Tree, k-NN, Multilayer Perceptron (MLP), and LSTM networks on 3,308 documents [19]. Preprocessing raised mean accuracy from 26% to 93.9%, with the MLP model achieving the best F1 score (93.6%), demonstrating the influence of text normalization on performance.

Goodrum et al. developed a framework to classify scanned EHR documents into clinically relevant and non-relevant categories using OCR-extracted text [20]. A ClinicalBERT model achieved an accuracy of 0.973, highlighting the power of transformer architectures for document-level clinical classification.

Lu et al. compared seven deep learning models including CNN, RNN, GRU, LSTM, Bi-LSTM, Transformer encoders, and BERT, for discharge note classification under varying class-imbalance conditions [21]. Transformer encoders yielded the best results overall, whereas CNNs achieved similar accuracy with shorter training time, suggesting a practical balance between computational efficiency and predictive accuracy.

These studies reflect a gradual evolution from traditional ML pipelines toward deep neural and hybrid models that exploit pre-trained embeddings and transformer architectures to enhance semantic understanding. However, they also reveal that well-tuned traditional algorithms can offer comparable performance with greater interpretability and lower computational demands which are valuable in clinical settings.

Across surveys, traditional models, and modern deep learning methods, the literature demonstrates the maturity and adaptability of ML for clinical and medical text classification. The challenges, such as data sparsity, annotation costs, and the trade-off between interpretability and complexity, continue to shape the field. Building on these insights, the present study contributes by conducting a comparative evaluation of four traditional ML algorithms for multiclass classification of textual medical notes, thereby building empirical benchmarks to guide future research that may integrate deep learning, weak supervision, or AutoML techniques for enhanced performance.

3. Textual Medical Note Dataset and Preprocessing

3.1. Data Collection

The dataset used in this study was obtained from Kaggle, a publicly available machine learning repository [22]. The dataset contains textual medical notes derived from published clinical case descriptions and research summaries, representing five disease-related categories: Digestive System Diseases, Cardiovascular Diseases, Neoplasms, Nervous System Diseases, and General Pathological Conditions [22]. Each record consists of a paragraph describing a medical condition, patient case, or experimental study, annotated with a corresponding disease class label.

 doi:10.20944/preprints202511.1173.v1

The full dataset contains 28,880 records, divided evenly into 14,438 training and 14,442 testing entries. Only the training subset includes class labels, while the test subset is unlabeled and intended for prediction tasks. For this study, we utilize exclusively the 14,438 labeled training records to train, validate, and evaluate our models.

Each note shows a narrative style with a vocabulary that includes medical terminology, clinical findings, and disease-related terminology, making it well-suited for text classification experiments. The initial dataset distribution across the five classes was moderately imbalanced, with a dominant portion of samples labeled as General Pathological Conditions. Since this category represents a broad and heterogeneous group encompassing multiple disease systems, it was excluded from subsequent experiments to ensure more coherent and specific class boundaries. After excluding 4,805 records labeled as General Pathological Conditions, the resulting dataset consisted of 9,633 labeled medical notes across four disease classes. Table 1 outlines the distribution of medical notes across four disease classes after data cleaning.

Table 1. Distribution of medical notes across four disease classes after data cleaning.

Disease Class	Label	Number of Records	Percentage (%)
Neoplasms	1	3,163	32.8
Digestive System Diseases	2	1,494	15.5
Nervous System Diseases	3	1,925	20.0
Cardiovascular Diseases	4	3,051	31.7
Total	–	**9,633**	**100.0**

The average length of a raw medical note was approximately 187 words.

3.2. Data Preprocessing

Comprehensive text preprocessing was applied to ensure the dataset was clean, standardized, and machine-readable. The following steps were performed sequentially:

- Text cleaning. All text was converted to lowercase to eliminate case-based redundancy. Punctuation marks, numerical symbols, special characters, and non-alphabetic tokens were removed. Stop words (e.g., the, is, and, of) were filtered out to reduce noise.
- Tokenization and lemmatization. Each document was tokenized into individual words. Tokens were then lemmatized, that is, reduced to their base or dictionary form, to minimize inflectional variations (e.g., "studies" → "study," "patients" → "patient"). After these operations, the average document length was reduced from 187 words to approximately 111 words.
- To understand lexical distribution, word frequency analyses were generated for each class.
  - For Neoplasms (Class 1), the top frequent terms included patient, tumor, case, treatment, lesion, and disease.
  - For Digestive System Diseases (Class 2), the top frequent terms were patient, treatment, disease, study, group, and associated.
  - For Nervous System Diseases (Class 3), the top frequent terms are were patient, treatment, study, pain, group, and effect.
  - For Cardiovascular Diseases (Class 4), the top frequent terms were patient, blood pressure, coronary artery, left ventricular, myocardial infarction, and treatment.
- Text vectorization. The cleaned and lemmatized corpus was converted into a numerical format using vectorization techniques. The resulting document-term matrix contained 27,609 unique tokens, yielding a feature space of (9,633 × 27,609) for subsequent machine learning analysis.
- Dataset partitioning. The dataset was randomly split into training and testing subsets using a 90–10 ratio to enable model training and unbiased evaluation. This resulted in 8,669 records in the training set and 964 records in the test set.

 doi:10.20944/preprints202511.1173.v1

The following sample demonstrates the transformation of a raw medical note into its cleaned and lemmatized form. The original medical note is *"Duodenal-caval fistula. Duodenal-caval fistula is a rare, often lethal disease that requires prompt diagnosis and surgical correction. A case of duodenal-caval fistula due to duodenal ulceration is presented and discussed."* The preprocessed medical note is *"duodenal caval fistula duodenal caval fistula rare often lethal disease requires prompt diagnosis surgical correction case duodenal caval fistula due duodenal ulceration presented discussed"*. These preprocessing steps standardized the dataset, reduced sparsity, and prepared it for downstream feature extraction and classification modeling.

4. Methodologies

This section presents the machine learning algorithms used to classify textual medical notes into four disease categories. To identify the most effective traditional learning algorithm for this task, four widely used classifiers, Random Forest (RF), Naive Bayes (NB), Logistic Regression (LR), and Support Vector Machine (SVM), were implemented and compared under consistent preprocessing and feature extraction conditions. Each model represents a distinct learning paradigm: probabilistic reasoning in NB, linear discriminative modeling in LR, margin maximization in SVM, and ensemble-based decision aggregation in RF. All models were trained using the same vectorized representation of the preprocessed medical notes and optimized through hyperparameter tuning to achieve the best classification performance. The following subsections describe the theoretical foundations, parameter configurations, and implementation details of each model.

4.1. Random Forest Classification

The Random Forest (RF) algorithm is an ensemble-based learning method that constructs multiple decision trees during training and outputs the class predicted by the majority of trees [23]. By combining bootstrap aggregation (bagging) and random feature selection, Random Forest reduces overfitting while maintaining strong predictive accuracy [24]. Each tree is trained on a random subset of samples, and at each node split, a random subset of features is considered. This dual randomness enhances model diversity and stability, making Random Forest suited for high-dimensional and sparse text datasets [25].

In this study, the Random Forest classifier was applied to the vectorized medical notes to classify documents into four disease categories. The feature space consisted of 27,609 unique tokens derived from the preprocessed corpus. The algorithm recursively partitions the feature space to minimize impurity, measured using the Gini index. Its ability to handle large vocabularies without explicit feature selection makes it an appropriate choice for textual data.

Hyperparameters were optimized using a grid search with 5-fold cross-validation to achieve a balance between accuracy and generalization. The parameters tuned included the number of trees (100–500), maximum tree depth, the minimum number of samples required to be at a leaf node, and the minimum samples required for node splits. Model performance was assessed using the weighted average F1-score, which provides balanced sensitivity across all disease classes.

Random Forest was selected for its robustness to noise, interpretability of feature importance, and capacity to model nonlinear relationships in textual data. Its ensemble structure also identifies key discriminative medical terms, supporting interpretability and transparency in clinical applications.

4.2. Naive Bayes

The Naive Bayes (NB) classifier is a probabilistic model based on Bayes' Theorem, which assumes conditional independence among features given the class label [26,27]. Despite this simplifying assumption, it remains effective and computationally efficient for text classification because it models word-occurrence probabilities directly [28]. In this study, NB was applied to estimate the likelihood that a medical note belongs to one of four disease categories using the distribution of words within each note.

 doi:10.20944/preprints202511.1173.v1

Given a document $d = \{w_1, w_2, \ldots, w_n\}$ and a class c, the posterior probability is expressed as:

$$P(c|d) \propto P(c) \prod_{i=1}^{n} P(w_i|c), \tag{1}$$

where $P(c)$ is the prior probability of a class and $P(w_i|c)$ represents the likelihood of observing word w_i in that class [26]. The class with the highest posterior probability is then assigned as the predicted label.

The Multinomial Naive Bayes (MNB) variant was employed, as it is well suited for count-based representations such as term frequency or TF–IDF vectors derived from the medical notes. Additive Laplace smoothing was used to handle zero probabilities for rare or unseen words, improving generalization. The model was trained on 9,633 preprocessed records for robust performance estimation. To optimize the smoothing parameter α, a grid search with 5-fold cross-validation were conducted over the range $0.01 \leq \alpha \leq 1.0$, using the weighted F1-score as the selection metric.

Naive Bayes was chosen for its simplicity, speed, and interpretability which make it a strong baseline model for medical text classification.

4.3. Logistic Regression

The Logistic Regression (LR) model is a linear classifier that estimates the probability of a document belonging to a specific class using a logistic (sigmoid) function [29]. Logistic regression is a discriminative approach that directly learns the decision boundary between classes by maximizing the likelihood of the observed labels [30]. It is well suited for high-dimensional, sparse datasets like textual representations, where individual tokens serve as features.

Given a document represented by a feature vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the probability that it belongs to class c is expressed as

$$P(c|\mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x} + b)}}, \tag{2}$$

where $\mathbf{w}$ denotes the model weights and b is the bias term [30]. For multiclass problems, a softmax extension is applied to ensure all class probabilities sum to one.

In this study, logistic regression was implemented using the one-vs-rest (OvR) strategy, where an independent binary classifier was trained for each disease class. The input features were derived from the TF–IDF document–term matrix generated during preprocessing. Model parameters were optimized through L2 (ridge) regularization to control overfitting, and hyperparameter tuning was performed using grid search with 5-fold cross-validation. Regularization type and solver choice were further optimized by testing L1, L2, and elasticnet penalties with solvers including ‘lbfgs‘, ‘liblinear‘, and ‘saga‘. Model performance was evaluated using the weighted averaged F1-score.

Logistic Regression was selected for its interpretability, scalability, and ability to produce well-calibrated probability estimates, making it a reliable model for textual medical note classification.

4.4. Support Vector Machine Classification

Support Vector Machine (SVM) is a discriminative learning algorithm that seeks an optimal separating hyperplane to maximize the margin between data points of different classes [31]. It is effective for text classification, where data are typically high-dimensional and sparse—conditions under which linear SVMs perform well [32]. By identifying a subset of critical data points, known as support vectors, the model defines the decision boundary that best separates classes in the feature space.

Given a training set of labeled samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$, where $\mathbf{x}_i \in \mathbb{R}^n$ represents the feature vector and $y_i \in \{-1, +1\}$ denotes the class label, the SVM optimization problem can be formulated as:

$$\min_{\mathbf{w}, b, \xi_i} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{m} \xi_i$$

 doi:10.20944/preprints202511.1173.v1

$$\text{s.t. } y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \tag{3}$$

where $\mathbf{w}$ is the weight vector, b is the bias term, ξ_i are slack variables, and C is the regularization parameter balancing margin maximization and classification error [33].

In this study, a linear SVM was employed due to its scalability and strong performance on large text corpora. The model was trained on a TF–IDF document–term matrix containing 27,609 features. Hyperparameters, including the regularization parameter C and kernel type, were tuned using grid search with 5-fold cross-validation. Linear, polynomial, radial basis function (RBF), and sigmoid kernels were tested, with performance evaluated using the weighted averaged F1-score to account for class imbalance. The final configuration, based on the highest cross-validated F1-score, achieved a balanced trade-off between accuracy and generalization, confirming SVM's robustness and interpretability for medical text classification.

4.5. Evaluation Metrics

To evaluate the performance of the four machine learning models, a set of widely accepted metrics was employed. These metrics provide complementary perspectives on predictive accuracy, robustness across classes, and reliability when applied to multiclass textual data. The primary metrics used in this study include accuracy, precision, recall, and the F1-score, each computed for individual class labels and then aggregated to summarize overall model performance.

For a given class label $c \in \{1, 2, \ldots, C\}$, where C represents the total number of categories, and TP_c, FP_c, and FN_c denote true positives for class c, false positives for class c, and false negatives for class c, respectively. The following definitions apply:

- *Precision for class c* quantifies the proportion of correctly predicted samples of class c among all samples predicted as class c:

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

- *Recall for class c* (or sensitivity) measures the proportion of actual samples of class c that were correctly identified:

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

- *F1-score for class c* is the harmonic mean of precision and recall:

$$\text{F1-score}_c = 2 \times \frac{\text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}. \tag{6}$$

- *Accuracy* measures the overall proportion of correctly classified samples across all classes:

$$\text{Accuracy} = \frac{\sum_{c=1}^{C} TP_c}{|\text{All Samples}|}. \tag{7}$$

Since this study involves four disease categories, macro averaged and weighted averaged versions of these metrics were used. Macro-averaging treats all classes equally, regardless of their size, while weighted averaging assigns weights proportional to the number of instances per class, providing a more realistic assessment under class imbalance.

Additionally, confusion matrices were analyzed to visualize class-level performance, identify common misclassifications, and detect semantic overlaps among disease categories. It provides a detailed view of the classifier's performance by comparing predicted versus actual class labels. The diagonal elements in a confusion matrix represent correctly classified medical notes, while off-diagonal values indicate misclassifications. These evaluation metrics provide a comprehensive understanding of model performance, enabling a fair comparison among the Random Forest, Naive Bayes, Logistic Regression, and Support Vector Machine classifiers.

 doi:10.20944/preprints202511.1173.v1

5. Experimental Results and Analysis

This section presents the experimental results obtained from the comparative evaluation of the four machine learning models (Random Forest, Multinomial Naive Bayes, Logistic Regression, and Support Vector Machine) applied to classifying textual medical notes into four disease categories. All experiments were conducted using the scikit-learn library in Python, which provides reliable and standardized implementations of machine learning algorithms and evaluation tools [34]. Each model was trained and optimized using the best hyperparameter configurations determined through cross-validation, as detailed in Section 4.1-4.4. The experiments were conducted on the preprocessed dataset containing 9,633 labeled records. The evaluation focuses on multiple performance metrics, including accuracy, precision, recall, and F1-score, computed in both macro- and weighted-averaged forms. The following subsections present detailed results for each model, followed by a comparative analysis highlighting their relative strengths, weaknesses, and suitability for medical text classification tasks.

5.1. Random Forest Classification Model Evaluation

The Random Forest (RF) classifier was trained and optimized through grid search with 5-fold cross-validation to identify the most effective parameter configuration. The best performing model was obtained with the following parameters: `bootstrap = True`, `max_depth = None`, `max_features = 'sqrt'`, `min_samples_leaf = 4`, `min_samples_split = 2`, and `n_estimators = 400`. This configuration allows the model to grow deep trees with random feature subsets while maintaining generalization through leaf size regularization.

The confusion matrix in Table 2 shows how the Random Forest model's predictions are distributed across the four disease categories on the test set.

Table 2. Confusion matrix for the Random Forest classifier.

	Pred 1	**Pred 2**	**Pred 3**	**Pred 4**
True 1 Neoplasms	290	10	12	9
True 2 Digestive	32	87	14	12
True 3 Nervous System	30	2	125	38
True 4 Cardiovascular	14	5	13	271

We can see that 290 notes of Neoplasms (Class 1), 87 notes of Digestive System Diseases (Class 2), 125 notes of Nervous System Diseases (Class 3), and 271 notes of Cardiovascular Diseases (Class 4) were correctly classified. While 32 notes of Digestive System Diseases (Class 2) and 30 notes of Nervous System Diseases (Class 3) were misclassified into Neoplasms (Class 1), 38 notes of Nervous System Diseases (Class 3) were misclassified into Cardiovascular Diseases (Class 4), likely due to overlapping terminology in clinical text.

Table 3 summarizes the classification report, which includes class-wise precision, recall, F1-score, and support.

Table 3. Performance metrics for the Random Forest classifier.

Class	**Precision**	**Recall**	**F1-score**	**Support**
1 Neoplasms	0.79	0.90	0.84	321
2 Digestive	0.84	0.60	0.70	145
3 Nervous System	0.76	0.64	0.70	195
4 Cardiovascular	0.82	0.89	0.86	303
Accuracy			**0.80**	964
Macro avg	0.80	0.76	0.77	964
Weighted avg	0.80	0.80	0.80	964

Among all categories, Neoplasms (Class 1) and Cardiovascular Diseases (Class 4) achieved the highest recall values (0.90 and 0.89), showing that the model effectively identified these disease types.

 doi:10.20944/preprints202511.1173.v1

In contrast, Digestive System Diseases (Class 2) and Nervous System Diseases (Class 3) had lower recall values (0.60 and 0.64), reflecting greater confusion with other classes. The high precision for Digestive System Diseases (0.84) and Cardiovascular Diseases (0.82) demonstrates its reliability in avoiding false positives, while the strong recall for Neoplasms highlights its ability to correctly identify oncological notes.

Overall, the Random Forest classifier performed competitively across all metrics, establishing a strong baseline for textual medical note classification.

5.2. Naive Bayes Model Evaluation

The Multinomial Naive Bayes (MNB) classifier was evaluated after hyperparameter tuning using 5-fold cross-validation, which identified the optimal settings as $\alpha = 0.1$ and `fit_prior = False`. This configuration provided the best balance between bias and variance, preventing over-smoothing while allowing the model to rely primarily on the word distribution within each disease class. The final model was trained using these optimal parameters and evaluated on the held-out test set of 964 medical notes.

Table 4 presents the confusion matrix of the Naive Bayes classifier, which illustrates how the model classified the four disease categories based on word-occurrence probabilities.

Table 4. Confusion matrix for the Multinomial Naive Bayes classifier.

	Pred 1	**Pred 2**	**Pred 3**	**Pred 4**
True 1 Neoplasms	255	29	27	10
True 2 Digestive	20	109	6	10
True 3 Nervous System	14	5	146	30
True 4 Cardiovascular	8	12	22	261

The matrix shows that Cardiovascular Diseases (Class 4) and Neoplasms (Class 1) were predicted most accurately, with relatively few misclassified samples. In contrast, Digestive System Diseases (Class 2) and Nervous System Diseases (Class 3) show moderate confusion with other categories, likely because of overlapping medical terminology and common clinical symptoms in the textual data.

Table 5 lists the classification report, which includes class-wise precision, recall, F1-score, and support.

Table 5. Performance metrics for the Multinomial Naive Bayes classifier.

Class	**Precision**	**Recall**	**F1-score**	**Support**
1 Neoplasms	0.86	0.79	0.83	321
2 Digestive	0.70	0.75	0.73	145
3 Nervous System	0.73	0.75	0.74	195
4 Cardiovascular	0.84	0.86	0.85	303
Accuracy			**0.80**	964
Macro avg	0.78	0.79	0.78	964
Weighted avg	0.80	0.80	0.80	964

The classification report provides a quantitative summary of model performance. The model achieved an overall accuracy of 0.80. Neoplasms (Class 1) and Cardiovascular Diseases (Class 4) achieved the highest precision (0.86 and 0.84, respectively) and recall (0.79 and 0.86), demonstrating that the Naive Bayes model effectively distinguishes these well-defined disease categories. In contrast, Digestive System Diseases (Class 2) and Nervous System Diseases (Class 3) yielded slightly lower recall values (0.75 each), suggesting moderate misclassification between these two categories. Naive Bayes obtained a macro-averaged precision of 0.78, recall of 0.79, and F1-score of 0.78, while the weighted averages were all approximately 0.80.

Overall, the Multinomial Naive Bayes classifier produced balanced results with strong overall accuracy and consistent weighted averages across all metrics. Its efficiency and simplicity make it a

 doi:10.20944/preprints202511.1173.v1

reliable baseline model for multiclass medical note classification, particularly when computational efficiency and interpretability are desired.

5.3. Logistic Regression Model Evaluation

The Logistic Regression (LR) classifier was evaluated after extensive hyperparameter optimization using 5-fold cross-validation. The best configuration was obtained with the parameters: `C = 1`, `class_weight = None`, `l1_ratio = 0.5`, `max_iter = 1000`, `multi_class = 'multinomial'`, `penalty = 'elasticnet'`, and `solver = 'saga'`. This combination of elastic-net regularization and the SAGA solver provided a balanced control of both L1 and L2 penalties, enabling the model to handle sparse features while maintaining good generalization. The multinomial formulation was chosen to directly optimize the cross-entropy loss for the four disease categories.

The confusion matrix in Table 6 summarizes the classifier's predictions across the four disease categories on the test set.

Table 6. Confusion matrix for the Logistic Regression classifier.

	Pred 1	**Pred 2**	**Pred 3**	**Pred 4**
True 1 Neoplasms	280	13	22	6
True 2 Digestive	24	102	13	6
True 3 Nervous System	21	1	149	24
True 4 Cardiovascular	9	7	20	267

Most Neoplasms (Class 1) and Cardiovascular Diseases (Class 4) notes were correctly identified, demonstrating the model's effectiveness in detecting strong class-specific textual patterns. Moderate confusion occurred between Digestive System Diseases (Class 2) and Nervous System Diseases (Class 3), likely due to shared terminology and overlapping clinical contexts in medical narratives.

The detailed performance metrics are shown in Table 7, presenting class-wise precision, recall, F1-score, and support.

Table 7. Performance metrics for the Logistic Regression classifier.

Class	**Precision**	**Recall**	**F1-score**	**Support**
1 Neoplasms	0.84	0.87	0.85	321
2 Digestive	0.83	0.70	0.76	145
3 Nervous System	0.73	0.76	0.75	195
4 Cardiovascular	0.88	0.88	0.88	303
Accuracy			**0.83**	964
Macro avg	0.82	0.81	0.81	964
Weighted avg	0.83	0.83	0.83	964

The classification report shows the Logistic Regression classifier achieved an overall accuracy of 0.83. Cardiovascular Diseases (Class 4) and Neoplasms (Class 1) obtained the highest recall values of 0.88 and 0.87, respectively, indicating strong discriminative capability for these categories. Neoplasms (Class 1) also performed strongly with an F1-score of 0.85, while Digestive System Diseases (Class 2) and Nervous System Diseases (Class 3) showed slightly lower recall values (0.70 and 0.76, respectively), suggesting partial overlap in their textual characteristics. Logistic Regression achieved a macro-averaged precision of 0.82, recall of 0.81, and F1-score of 0.81, with weighted averages of 0.83 across all metrics.

Overall, the Logistic Regression model demonstrated robust and consistent performance across all categories. Its high accuracy and balanced precision-recall scores confirm its effectiveness for multiclass text classification of medical notes. The model's interpretability and well-calibrated probability estimates further highlight its suitability for clinical NLP tasks, establishing it as the best-performing approach among the four models evaluated.

 doi:10.20944/preprints202511.1173.v1

5.4. Support Vector Machine Classification Model Evaluation

The Support Vector Machine (SVM) classifier was trained and optimized using grid search with 5-fold cross-validation. The optimal configuration was identified as `C = 1`, `class_weight = None`, and `kernel = 'linear'`. The linear kernel was selected because it provided the best balance between classification accuracy and computational efficiency, particularly for the high-dimensional sparse feature space derived from the vectorized medical text. The regularization parameter ($C = 1$) offered a suitable trade-off between maximizing the margin and minimizing classification error, while maintaining stable convergence across folds.

Table 8 presents the confusion matrix, which summarizes the model's predictions across the four disease categories on the test set.

Table 8. Confusion matrix for the Support Vector Machine classifier.

	Pred 1	**Pred 2**	**Pred 3**	**Pred 4**
True 1 Neoplasms	261	24	26	10
True 2 Digestive	25	102	9	9
True 3 Nervous System	21	3	146	25
True 4 Cardiovascular	13	9	18	263

Most Neoplasms (Class 1) and Cardiovascular Diseases (Class 4) were correctly identified, while moderate confusion occurred between Digestive System Diseases (Class 2) and Nervous System Diseases (Class 3). This overlap likely results from shared medical terminology and similar contextual patterns in the clinical text.

The performance metrics in Table 9 reveal balanced precision, recall, and F1-scores across all categories.

Table 9. Performance metrics for the Support Vector Machine classifier.

Class	**Precision**	**Recall**	**F1-score**	**Support**
1 Neoplasms	0.82	0.81	0.81	321
2 Digestive	0.74	0.70	0.72	145
3 Nervous System	0.73	0.75	0.74	195
4 Cardiovascular	0.86	0.87	0.86	303
Accuracy			**0.80**	964
Macro avg	0.79	0.78	0.78	964
Weighted avg	0.80	0.80	0.80	964

The classification report indicates that the model achieved an overall accuracy of 0.80 on the test set. Cardiovascular Diseases (Class 4) achieved the highest precision and recall (0.86 and 0.87), demonstrating that the SVM effectively identified this category with minimal misclassification. Neoplasms (Class 1) also performed well, with an F1-score of 0.81. Nervous System Diseases (Class 3) attained a recall of 0.75, whereas Digestive System Diseases (Class 2) showed slightly lower recall (0.70), with some overlap into the Neoplasms category, which is an expected outcome given the linguistic similarity of pathological terms across gastrointestinal and neoplastic conditions.

Overall, the Support Vector Machine classifier showed stable performance across all metrics. Its ability to handle high-dimensional, sparse textual features made it particularly suitable for this dataset. The results confirm that SVM offers a robust balance between predictive accuracy and generalization, establishing it as a competitive model for textual medical note classification.

5.5. Model Comparison and Discussion

The comparative performance analysis of the four machine learning models, i.e., Random Forest (RF), Multinomial Naive Bayes (MNB), Logistic Regression (LR), and Support Vector Machine (SVM), revealed consistent and competitive results across all approaches. Each algorithm achieved a strong level of predictive accuracy on the test set, demonstrating that traditional machine learning techniques

 doi:10.20944/preprints202511.1173.v1

remain effective for multiclass medical text classification when supported by comprehensive feature preprocessing.

Table 10 summarizes the overall classification accuracy of the four models.

Table 10. Accuracy comparison of the four machine learning models.

Model	Accuracy	Weighted-average F1
Random Forest (RF)	0.80	0.80
Multinomial Naive Bayes (MNB)	0.80	0.80
Logistic Regression (LR)	0.83	0.83
Support Vector Machine (SVM)	0.80	0.80

Among the evaluated models, Logistic Regression achieved the highest overall accuracy (0.83) and the best weighted F1-score, indicating that it effectively captured linear decision boundaries within the vectorized text representation. The use of elastic-net regularization allowed the model to balance sparsity and complexity, reducing overfitting while maintaining interpretability.

Random Forest, Support Vector Machine, and Multinomial Naive Bayes classification models achieved similar accuracies (0.80), demonstrating stable and reliable performance across all disease categories. Random Forest showed strong recall for Neoplasms and Cardiovascular Diseases, confirming its ability to model nonlinear relationships and handle noisy textual features. The linear SVM achieved balanced precision and recall by effectively separating high-dimensional sparse features using its margin-maximization principle. The Multinomial Naive Bayes classifier, though based on a simplifying independence assumption, performed competitively and remained the most computationally efficient, making it well suited for large-scale or real-time clinical applications.

Across all models, the best performance was observed for Cardiovascular Diseases and Neoplasms, where domain-specific terminology (e.g., “myocardial infarction,” “tumor,” “lesion”) provided clearer linguistic cues for classification. In contrast, Digestive System Diseases and Nervous System Diseases exhibited higher misclassification rates, likely due to overlapping vocabulary and symptom-related expressions such as “pain,” “treatment,” and “study.”

From a methodological perspective, Logistic Regression and SVM offered the best trade-off between accuracy, interpretability, and computational efficiency, while Random Forest provided valuable insight into feature importance. Overall, the comparable accuracy across all four models underscores the robustness of traditional machine learning methods for clinical text classification.

6. Conclusion and Future Work

This study conducted a comprehensive comparative analysis of four traditional machine learning algorithms, i.e., Random Forest, Multinomial Naive Bayes, Logistic Regression, and Support Vector Machine, for the classification of textual medical notes into four disease categories: Neoplasms, Digestive System Diseases, Nervous System Diseases, and Cardiovascular Diseases. Using a dataset of 9,633 preprocessed medical documents derived from publicly available clinical text sources, each model was trained and fine-tuned through systematic hyperparameter optimization and cross-validation. The evaluation employed standard metrics such as accuracy, precision, recall, and F1-score to ensure a robust and objective performance assessment.

The experimental results demonstrate that all four models performed effectively, achieving accuracies in the range of 0.80–0.83, thereby confirming the suitability of traditional machine learning methods for multiclass medical text classification tasks. Among them, Logistic Regression achieved the highest accuracy and F1-score (both 0.83), followed closely by the Random Forest, Multinomial Naive Bayes, and Support Vector Machine, which all reached accuracies of 0.80. Overall, the findings reveal that when supported by careful preprocessing, feature engineering, and parameter tuning, traditional machine learning models can achieve strong and interpretable performance in classifying medical notes without requiring complex deep learning architectures.

 doi:10.20944/preprints202511.1173.v1

While the results are promising, several avenues remain open for future work. First, the incorporation of semantic embedding techniques such as Word2Vec, GloVe, or BioBERT could improve the models' ability to capture contextual relationships and subtle linguistic nuances in clinical text. Second, exploring transformer-based models like BERT or ClinicalBERT may further enhance classification accuracy by leveraging deeper contextual understanding. Third, expanding the dataset to include a wider range of diseases and real-world electronic health record (EHR) notes would improve generalizability.

In conclusion, this research provides valuable empirical evidence that traditional machine learning methods, when properly optimized, remain powerful tools for medical text classification. The study lays a solid foundation for future integration of linguistic representation learning and explainable modeling in clinical natural language processing applications.

Author Contributions: Conceptualization, Y.Z.; methodology, Y.Z.; investigation, H.L.; data curation, H.L.; experiments, H.L., N.L. and K.P.; writing—original draft preparation, Y.Z.; writing—review and editing, Y.Z.; supervision, Y.Z; All authors have read and agreed to the published version of the manuscript.

Funding: The research did not receive external funding.

Data Availability Statement: The original dataset is publicly available on Kaggle. The pre-processed dataset is available upon request from the authors.

Acknowledgments: The authors gratefully acknowledge support from the California State University, San Bernardino (CSUSB) College of Natural Sciences (CNS) Proactive Approaches for Training Hispanics in STEM (PATHS) program, which provided funding for three undergraduate researchers during summer 2025. This material is based upon work supported by the National Science Foundation under Grant No. 2322436.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Badawy, M.; Ramadan, N.; Hefny, H.A. Big data analytics in healthcare: data sources, tools, challenges, and opportunities. *Journal of Electrical Systems and Information Technology* **2024**, *11*, 63.
2. Tang, A.S.; Woldemariam, S.R.; Miramontes, S.; Norgeot, B.; Oskotsky, T.T.; Sirota, M. Harnessing EHR data for health research. *Nature Medicine* **2024**, *30*, 1847–1855.
3. Tayefi, M.; Ngo, P.; Chomutare, T.; Dalianis, H.; Salvi, E.; Budrionis, A.; Godtliebsen, F. Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews: Computational Statistics* **2021**, *13*, e1549.
4. Spasic, I.; Nenadic, G. Clinical text data in machine learning: systematic review. *JMIR Medical Informatics* **2020**, *8*, e17984.
5. Wilcox, A.B.; Hripcsak, G. The role of domain knowledge in automating medical text Report classification. *Journal of the American Medical Informatics Association (JAMIA)* **2003**, *10*, 330.
6. Li, Q.; Peng, H.; Li, J.; Xia, C.; Yang, R.; Sun, L.; Yu, P.S.; He, L. A survey on text classification: from traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)* **2022**, *13*, 1–41.
7. Khalate, P.; Gite, S.; Pradhan, B.; Lee, C.W. Advancements and gaps in natural language processing and machine learning applications in healthcare: a comprehensive review of electronic medical records and medical imaging. *Frontiers in Physics* **2024**, *12*, 1445204.
8. Mustafa, A.; Rahimi Azghadi, M. Automated machine learning for healthcare and clinical notes analysis. *Computers* **2021**, *10*, 24.
9. Kino, S.; Hsu, Y.T.; Shiba, K.; Chien, Y.S.; Mita, C.; Kawachi, I.; Daoud, A. A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects. *SSM Population Health* **2021**, *15*, 100836.
10. Kadhim, A.I. Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review* **2019**, *52*, 273–292.
11. Weng, W.H.; Wagholikar, K.B.; McCray, A.T.; Szolovits, P.; Chueh, H.C. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC Medical Informatics and Decision Making* **2017**, *17*, 155.

12. López-Úbeda, P.; Diaz-Galiano, M.C.; Martin-Noguerol, T.; Luna, A.; Urena-Lopez, L.A.; Martin-Valdivia, M.T. Automatic medical protocol classification using machine learning approaches. *Computer Methods and Programs in Biomedicine* **2021**, *200*, 105939.
13. Tiwari, P.; Upadhyay, D.; Pant, B.; Mohd, N. Multiclass classification in machine learning algorithms for disease prediction. In Proceedings of the International Conference on Advanced Informatics for Computing Research. Springer, 2021, pp. 102–111.
14. Sung, S.F.; Lin, C.Y.; Hu, Y.H. EMR-based phenotyping of ischemic stroke using supervised machine learning and text mining techniques. *IEEE Journal of Biomedical and Health Informatics* **2020**, *24*, 2922–2931.
15. Rabby, G.; Berka, P. Multi-class classification of COVID-19 documents using machine learning algorithms. *Journal of Intelligent Information Systems* **2023**, *60*, 571–591.
16. Gupta, S.; Belouali, A.; Shah, N.J.; Atkins, M.B.; Madhavan, S. Automated identification of patients with immune-related adverse events from clinical notes using word embedding and machine learning. *JCO Clinical Cancer Informatics* **2021**, *5*, 541–549.
17. Gao, C.; Goswami, M.; Chen, J.; Dubrawski, A. Classifying unstructured clinical notes via automatic weak supervision. In Proceedings of the Machine Learning for Healthcare Conference. PMLR, 2022, pp. 673–690.
18. Lenivtceva, I.; Slasten, E.; Kashina, M.; Kopanitsa, G. Applicability of machine learning methods to multi-label medical text classification. In Proceedings of the International Conference on Computational Science. Springer, 2020, pp. 509–522.
19. da Silva, D.P.; Fröhlich, W.d.R.; Schwertner, M.A.; Rigo, S.J. Clinical Oncology Textual Notes Analysis Using Machine Learning and Deep Learning. In Proceedings of the Brazilian Conference on Intelligent Systems. Springer, 2023, pp. 140–153.
20. Goodrum, H.; Roberts, K.; Bernstam, E.V. Automatic classification of scanned electronic health record documents. *International Journal of Medical Informatics* **2020**, *144*, 104302.
21. Lu, H.; Ehwerhemuepha, L.; Rakovski, C. A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. *BMC Medical Research Methodology* **2022**, *22*, 181.
22. Kaggle. Medical Text, 2019. Accessed: June 8, 2025.
23. Breiman, L. Random forests. *Machine Learning* **2001**, *45*, 5–32.
24. Rigatti, S.J. Random forest. *Journal of Insurance Medicine* **2017**, *47*, 31–39.
25. Sun, Y.; Li, Y.; Zeng, Q.; Bian, Y. Application research of text classification based on random forest algorithm. In Proceedings of the 2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE). IEEE, 2020, pp. 370–374.
26. Murphy, K.P.; et al. Naive bayes classifiers. *University of British Columbia* **2006**, *18*, 1–8.
27. Webb, G.I.; Keogh, E.; Miikkulainen, R. Naïve Bayes. *Encyclopedia of Machine Learning* **2010**, *15*, 713–714.
28. Singh, G.; Kumar, B.; Gaur, L.; Tyagi, A. Comparison between multinomial and Bernoulli naïve Bayes for text classification. In Proceedings of the 2019 International conference on automation, computational and technology management (ICACTM). IEEE, 2019, pp. 593–596.
29. Pampel, F.C. *Logistic Regression: A Primer*; Number 132, Sage Publications, 2020.
30. Shah, K.; Patel, H.; Sanghvi, D.; Shah, M. A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augmented Human Research* **2020**, *5*, 12.
31. Sarkar, A.; Chatterjee, S.; Das, W.; Datta, D. Text classification using support vector machine. *International Journal of Engineering Science Invention* **2015**, *4*, 33–37.
32. Dadgar, S.M.H.; Araghi, M.S.; Farahani, M.M. A novel text mining approach based on TF-IDF and Support Vector Machine for news classification. In Proceedings of the 2016 IEEE International Conference on Engineering and Technology (ICETECH). IEEE, 2016, pp. 112–116.
33. Mammone, A.; Turchi, M.; Cristianini, N. Support vector machines. *Wiley Interdisciplinary Reviews: Computational Statistics* **2009**, *1*, 283–289.
34. Scikit-learn Developers. scikit-learn: Machine Learning in Python, 2025. Accessed: June 10, 2025.