

# Medical Transcripts Classification

Group 31

Veer Sanghavi  
ASU ID: 1233263995  
[vrsangha@asu.edu](mailto:vrsangha@asu.edu)

Tejaswini Kulkarni  
ASU ID: 1233869691  
[tkulkar8@asu.edu](mailto:tkulkar8@asu.edu)

Mansi Nandkar  
ASU ID: 1233870562  
[mnandkar@asu.edu](mailto:mnandkar@asu.edu)

Bennet Meneze  
ASU ID: 1234209251  
[bmeneze2@asu.edu](mailto:bmeneze2@asu.edu)

**Abstract**— Clinical text classification is an important task in healthcare, where the ability to accurately categorize unstructured textual data into medical specialties can help with improvements in patient care and healthcare management. This project focuses on classifying medical transcripts by leveraging Sequential Forward Selection (SFS). SFS is pivotal in enhancing classification performance by ensuring that models are trained on the most informative features, thereby improving both speed and accuracy. The methodology encompasses several critical steps, starting from preprocessing textual data, such as tokenization, stop-word removal, and lemmatization, to transforming the text into analysable formats. After preprocessing, SFS identifies optimal feature subsets, which are used to train machine learning models, including Logistic Regression, Support Vector Machines (SVM), and Gradient Boosting Trees. Evaluation metrics such as accuracy, precision, recall, and F1-score are employed to assess model performance. The results of this study highlight the effectiveness of the proposed approach, achieving high accuracy for binary classifications. However, challenges arise in multiclass settings, particularly due to the limited dataset size and overlapping textual patterns between classes. This project underscores the need for larger, more granular datasets and innovative methods to address these issues. By advancing the field of clinical text classification, this work contributes to more accurate and efficient healthcare decision-making processes.

**Keywords**— *Sequential Forward Selection (SFS), Feature Selection, Categorical Boosting, Gradient Boosting Trees, Random Forest, Logistic Regression, Support Vector Machines (SVM), Clinical Text Classification.*

## I. INTRODUCTION

### A. Background

Medical text data generated during clinical operations contains a wealth of information crucial for decision-making. However, the majority of this data is unstructured, making it challenging to process and analyze. Clinical text classification transforms this data into structured forms, facilitating its use in various healthcare applications, such as automated diagnosis, electronic health records management, and patient care optimization.

### B. Problem and Importance

Despite significant advancements in Natural Language Processing (NLP) and machine learning, clinical text classification remains a challenging task due to several factors, including high-dimensional data, overlapping features across classes, and domain-specific language complexities. This project's focus on SFS addresses these challenges by optimizing feature selection, enabling more efficient and accurate classifications.

### C. Existing Literature

The field of clinical text classification has seen substantial research contributions, particularly in its intersection with Natural Language Processing (NLP) and feature selection methods. The goal of extracting meaningful insights from unstructured textual data has driven advancements across diverse methodologies, including rule-based approaches, deep learning models, and statistical machine learning techniques. This section reviews key contributions relevant to this project.

One early approach to clinical text classification relied heavily on rule-based systems, leveraging domain knowledge to create deterministic systems for classification. For example, Yao et al. (2019) presented a framework combining rule-based features with knowledge-guided convolutional neural networks. Their approach integrated medical ontologies to improve classification precision in tasks such as disease identification, showcasing how domain-specific information can enhance the performance of machine learning systems in healthcare contexts.

Deep learning has revolutionized many aspects of health informatics, including clinical text analysis. Ravi et al. (2017) highlighted the growing use of deep neural networks in tasks ranging from patient outcome prediction to diagnostic support. These models, including Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), are particularly effective in capturing

the sequential nature of clinical text data. However, such models often require large datasets and significant computational resources, limiting their applicability in resource-constrained environments.

The integration of semi-supervised learning techniques has been explored to address the challenge of limited labeled data in medical text classification. Vijay Garla et al. (2013) demonstrated the application of Laplacian SVMs to cancer case management, utilizing both labeled and unlabeled data. This approach showed improved performance in domains where annotated data is scarce, which is a common limitation in healthcare datasets.

Feature selection remains a pivotal step in clinical text classification workflows. Sequential Forward Selection (SFS), in particular, has been extensively studied for its ability to identify critical features while reducing dimensionality. Marcano-Cedeño et al. (2010) combined SFS with Artificial Metaplasticity Neural Networks, demonstrating the effectiveness of this approach in reducing overfitting and improving generalization performance. This work underscores the versatility of SFS in handling complex datasets across different domains, including healthcare.

Furthermore, information-theoretic algorithms for feature selection have also been explored. For instance, Last et al. (2002) proposed methods based on entropy measures to identify the most informative features in textual datasets. These techniques aim to balance the trade-off between reducing dataset complexity and retaining critical predictive elements.

Another cornerstone of clinical text classification is the preprocessing stage, where raw data is transformed into a machine-readable format. Lance De Vine et al. (2014) introduced methods leveraging neural language models to compute medical semantic similarity. These techniques enhance the quality of features derived from text, improving downstream classification performance.

Preprocessing steps like tokenization, stop-word removal, and lemmatization are widely acknowledged as essential for converting unstructured text into a structured format. Techniques such as TF-IDF (Term Frequency-Inverse Document Frequency) vectorization and Part-of-Speech (POS) tagging are particularly effective in emphasizing the importance of domain-specific terms while reducing noise in the data.

The applications of clinical text classification extend to several critical areas in healthcare. Uzuner et al. (2008) explored the identification of patient smoking status from discharge summaries, highlighting the potential of automated systems to analyze and extract relevant patient information. Similarly, Demner-Fushman et al. (2009) examined how NLP techniques can support clinical decision-making, further demonstrating the transformative impact of automated text analysis in improving healthcare delivery.

While the above methods demonstrate significant advancements, clinical text classification remains challenging due to data limitations, high dimensionality, and domain-specific complexities. Many studies emphasize the importance of increasing dataset size and quality, developing methods that generalize well across diverse medical texts, and integrating domain knowledge into feature engineering processes.

The methodologies reviewed above directly influence the approach taken in this project. By adopting Sequential Forward Selection as a feature selection mechanism, the project aligns with prior research demonstrating its utility in dimensionality reduction and performance optimization. Furthermore, the use of preprocessing techniques and evaluation metrics discussed in the literature provides a strong foundation for building and assessing the classification models.

#### *D. System Overview*

The proposed system for clinical text classification is a robust and modular framework designed to handle the challenges of unstructured medical data, dimensionality reduction, and class imbalance while maintaining high classification accuracy. The system integrates advanced preprocessing techniques, feature selection methods, and machine learning algorithms to build a reliable pipeline capable of classifying medical transcripts into predefined medical specialties.

Its primary objective is to transform raw medical transcription data into meaningful and actionable classifications. To achieve this, the design philosophy focuses on simplicity, scalability, and adaptability. Each component is structured to address specific challenges inherent in clinical text classification, such as noise in the data, imbalanced class distributions, and overlapping descriptors between medical specialties.

The system preprocesses medical transcription data, selects relevant features using SFS, and evaluates machine learning models like Logistic Regression, SVM, and Gradient Boosting. It measures performance through accuracy, precision, recall, and F1-score.

#### *E. Data Collection*

The dataset is sourced from Kaggle and comprises medical transcriptions scraped from mtsamples.com. It includes textual descriptions, medical specialties, sample names, and keywords.

#### *F. Components of the ML System*

- Preprocessing: Tokenization, stop-word removal, and lemmatization.
- Feature Selection: Employing SFS to identify key attributes.
- Model Building: Training and testing various classifiers.

- Evaluation: Using confusion matrices and classification metrics.

#### G. Experimental Results

Preliminary experiments show high accuracy for binary classifications but performance drops with increased class granularity due to dataset limitations.

Sequential Forward Selection (SFS) proved instrumental in identifying the most relevant features for classification. By iteratively adding features that maximized model performance, SFS reduced the feature set from several thousand potential attributes to a manageable and effective subset. This not only improved computational efficiency but also mitigated the risk of overfitting.

The system achieved notable success in binary classification tasks, where the objective was to distinguish between two medical specialties. The success of binary classification can be attributed to the relatively clear distinction between classes when only two categories are considered. Features selected by SFS effectively captured the unique language patterns specific to each class.

## II. DEFINITIONS AND PROBLEM STATEMENT

### A. Data

The dataset comprises structured textual records, each categorized under medical specialties, as shown in Table 1. The dataset used in this study was obtained from Kaggle, containing medical transcriptions categorized under different specialties. While the dataset provided a good starting point, it presented several challenges:

1. Imbalanced Classes: Some specialties had a significantly higher number of records compared to others. For instance, classes such as "Surgery" and "Consultation and History" were overrepresented, whereas "Cardiovascular/Pulmonary" had limited examples.
2. Overlapping Text Descriptions: Many transcriptions contained overlapping language, such as references to surgical procedures across multiple specialties. This made distinguishing between classes like "Cardiovascular/Pulmonary" and "Surgery" particularly challenging.
3. Small Dataset Size: The relatively small number of records limited the generalizability of the results, particularly for multiclass classification.

TABLE 1 DATASET DESCRIPTION

Column Names	Missing Values	Missing Value %	Unique Values	Column Definition
Description	0	0	2348	Short Description of Transcription
Medical Speciality	0	0	40	Medical Speciality Classification of Transcription
Sample Name	0	0	2377	Transcription Title
Transcription	33	1	2358	Sample Medical Transcriptions
Keywords	1068	21	3848	Relevant Keywords from Transcriptions

### B. Prediction Target

The primary objective is to classify medical transcripts into specific specialties.

### C. Variables

Key variables include text descriptions, extracted features, and corresponding labels (medical specialties).

### D. Problem Statement

- Given: A dataset of medical transcripts.
- Objective: Classify texts into medical specialties using SFS and machine learning models.
- Constraints: Dataset size, imbalanced classes, and class overlaps.

### III. OVERVIEW OF THE PROPOSED APPROACH

The proposed system utilizes SFS to optimize feature selection, reducing dimensionality while retaining classification efficacy. By combining efficient preprocessing with advanced classification algorithms, the system aspires to deliver precise results. The approach used integrates advanced Natural Language Processing (NLP) techniques with feature selection methodologies and machine learning algorithms to build an efficient and scalable clinical text classification system. The design emphasizes modularity and adaptability, ensuring that each component can be refined and enhanced independently based on performance metrics and domain-specific requirements. The comprehensive pipeline aims to transform raw, unstructured medical transcription data into accurate predictions of medical specialties, addressing challenges such as imbalanced datasets, overlapping class definitions, and high-dimensional feature spaces.

The pipeline of the project architecture consists of the following key stages, each carefully structured to optimize the overall performance of the classification system:

#### 1. Data Cleaning

This step focused on preparing the raw dataset by removing inconsistencies and improving the quality of the input data:

- Null Value Removal: Any records with missing values are eliminated to ensure data completeness.
- Filtering Classes with Low Representation: Classes with less than 10% frequency of the highest frequent medical specialty are removed to reduce class imbalance and improve model focus.
- Length-Based Filtering: Records with transcription lengths below 50 characters are discarded, as they often lack sufficient information for accurate classification.

#### 2. Preprocessing

Text preprocessing converts raw medical text into a format suitable for machine learning by reducing noise and emphasizing meaningful content:

- Stop-Word Removal: Common and irrelevant words (e.g., "and," "the") are removed to enhance the focus on medical terms.
- Tokenization: The text is divided into individual tokens (words or phrases) for feature extraction.
- POS Tagging and Filtering for Nouns: Part-of-Speech tagging is performed to focus on nouns, which often carry the most critical medical information.
- Lemmatization: Words are reduced to their root forms to avoid redundancy and improve consistency (e.g., "diagnoses" becomes "diagnose").
- Duplicate Removal: Duplicate records are identified and removed to prevent over-representation of certain patterns.
- Vectorization: Text is transformed into numerical representations using vectorization techniques like Term Frequency-Inverse Document Frequency (TF-IDF), which captures the importance of words in the context of the dataset.

#### 3. Model Data Preparation

Once the data is preprocessed, it is prepared for model training:

- Train-Test Split: The dataset is divided into training and testing subsets to evaluate the generalizability of the models.
- Undersampling: In cases of imbalanced datasets, undersampling is used to reduce the dominance of overrepresented classes, ensuring a more balanced training process.

#### 4. Model Building

Multiple machine learning models are trained to evaluate their suitability for the classification task. These models include:

- Logistic Regression: A simple yet interpretable linear model for classification.
- Support Vector Machines (SVM): A model that identifies the optimal hyperplane to separate classes, suitable for high-dimensional datasets.
- Random Forest: An ensemble learning method that combines decision trees for robust classification.
- Gradient Boosting Trees: An iterative boosting method that refines predictions by focusing on residual errors.
- Categorical Boosting: A specialized model designed to handle categorical data and imbalanced distributions, often yielding high accuracy.

## 5. Model Evaluation

The trained models are assessed using evaluation metrics and visual tools:

- Confusion Matrix: A detailed view of classification errors for each specialty, highlighting areas for improvement.
- Classification Report: Metrics like accuracy, precision, recall, and F1-score are calculated to provide a comprehensive evaluation of model performance.

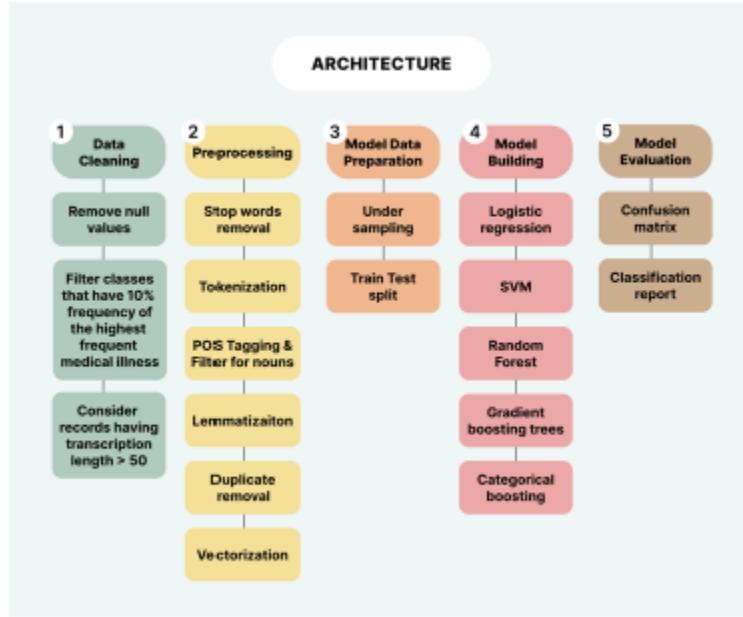


Fig. 1. Workflow of the Model

Figure 1 shows the pipeline in a detailed way.

## IV. TECHNICAL DETAILS

### A. Data Cleaning

In this project, Data cleaning is a crucial step to ensure the accuracy and reliability of the dataset used for analysis. One common challenge involves handling null values. To mitigate this, rows or columns containing null values are identified and subsequently removed from the dataset. After this, additional filtering is performed by excluding classes with records that amount to less than 10% of the maximum number of records in any class. Finally, only records with transcription lengths exceeding 50 characters are retained for further processing. Figure 2 shows a brief look at the dataset's attributes.

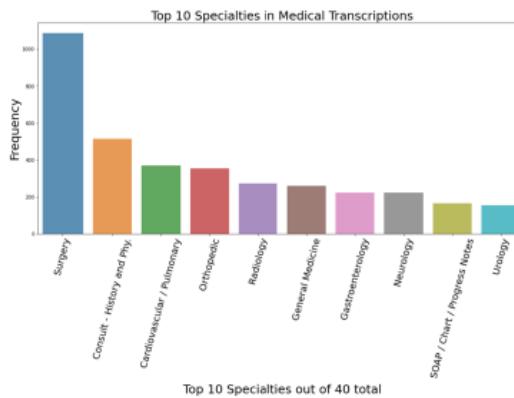


Fig. 2. Ten most frequent medical illness in the dataset.

## B. Preprocessing

### 1. Stop-Word Removal Using NLTK and Domain-Specific Filters

Stop-word removal is a technique used to eliminate common, non-informative words that appear frequently in the text but do not contribute significantly to the understanding or classification task.

- What Are Stop-words?  
Stop-words include words such as "the," "and," "is," "of," and "in," which often occur in text but provide little or no value in distinguishing between different classes. While they are essential for grammatical structure, they do not carry substantial semantic weight in most classification problems.
- Use of NLTK:  
The Natural Language Toolkit (NLTK) provides a pre-defined list of stop-words for multiple languages, including English. These lists are widely used in NLP pipelines to filter out common words efficiently. The implementation involves loading the stop-word list from NLTK and removing these words during preprocessing.
- Domain-Specific Filters:  
In addition to standard stop-words, domain-specific terms that are overly generic or irrelevant to the classification task are also removed. For example, in medical text processing, words like "patient," "report," or "note" may frequently appear across all classes and provide no distinguishing value. Custom stop-word lists are created by analyzing term frequencies in the dataset and excluding terms that occur uniformly across different specialties.
- Benefits of Stop-Word Removal:
  - Reduces the size of the vocabulary, thereby lowering computational complexity.
  - Improves the signal-to-noise ratio by retaining only meaningful and domain-relevant terms.
  - Enhances the focus of machine learning models on features that are critical for classification.

### 2. Tokenization and Part-of-Speech (POS) Tagging to Focus on Nouns

Tokenization and POS tagging are essential for breaking text into manageable components and identifying the grammatical roles of words, respectively.

- Tokenization:
  - Tokenization is the process of splitting the input text into smaller units called tokens, which can be words, phrases, or subwords, depending on the granularity required.
  - Using libraries like NLTK or spaCy, the text is divided into individual words or phrases. For instance, a sentence like "*The patient was diagnosed with pneumonia*" is split into tokens: ["The," "patient," "was," "diagnosed," "with," "pneumonia"].
  - Tokenization enables the system to process textual data in a structured format, making it easier to extract and analyze individual components.
- Part-of-Speech (POS) Tagging:
  - POS tagging assigns grammatical labels (e.g., noun, verb, adjective) to each token in the text. For example, in the sentence above, "patient" is labeled as a noun, "diagnosed" as a verb, and "pneumonia" as a noun.
  - This step is particularly important in medical NLP tasks, as nouns often represent critical medical entities such as conditions, symptoms, or procedures. Focusing on nouns allows the system to capture relevant information that distinguishes one medical specialty from another.
- Focusing on Nouns:
  - After tagging, only nouns are retained for further processing, as they typically carry the most significant information in medical text. For example, nouns like "pneumonia," "surgery," and "cardiology" are directly linked to medical specialties and are more relevant for classification than auxiliary words or verbs.
- Advantages:
  - Reduces dimensionality by filtering out less relevant grammatical constructs.
  - Increases the accuracy of feature extraction by focusing on the most meaningful terms.

- Improves interpretability by highlighting key medical terms that are useful for domain experts.

### 3. Lemmatization for Reducing Text to Root Forms

Lemmatization is a text normalization technique that reduces words to their root or base forms, ensuring consistency in textual representation.

- What Is Lemmatization?
    - Lemmatization transforms inflected or derived words into their base form, known as a "lemma." Unlike stemming, which performs a similar function, lemmatization uses linguistic rules to ensure that the transformed word is valid in the language. For instance, "diagnoses" and "diagnosing" are reduced to "diagnose," and "better" is reduced to "good" (its lemma).
  - Why Lemmatization Is Important:
    - In medical transcription datasets, variations in word forms can increase vocabulary size unnecessarily, leading to redundancy and noise. For example, a dataset may contain terms like "diagnosed," "diagnosis," and "diagnosing," all of which refer to the same concept. Lemmatization ensures that these terms are treated as a single feature, simplifying the dataset and improving model efficiency.
  - How It Works:
    - Lemmatization tools such as NLTK's WordNetLemmatizer or spaCy are employed to analyze the context and part of speech of each word to determine its correct lemma.
    - For example, the sentence "*The patient is being treated for various diagnoses*" is lemmatized to "*The patient be treat for various diagnose.*"
  - Advantages of Lemmatization:
    - Reduces redundancy by consolidating word variants into a single feature.
    - Lowers computational complexity by decreasing vocabulary size.
    - Enhances the performance of machine learning models by providing a more consistent and meaningful representation of text.



Fig. 3. Word Cloud for Medical Transcripts

### C. Feature Selection

Feature selection plays an essential role in the machine learning workflow, aiming to identify and retain the most influential features from a dataset while discarding those that contribute little or no value. This step is particularly important for improving the efficiency and effectiveness of machine learning models. In this project, feature selection helped streamline the dataset by reducing its dimensionality, which not only enhanced computational efficiency but also minimized the impact of irrelevant or redundant features. By focusing on the most significant predictors, feature selection improved the interpretability of the model, making it easier to understand the underlying relationships in the data and the rationale behind predictions.

This project employed forward feature selection as the feature selection strategy due to its systematic and iterative approach. Forward feature selection is a stepwise process that begins with no features and gradually builds an optimal feature set by adding one feature at a time. The method starts by evaluating each feature independently, training a model for each, and selecting the feature that provides the highest improvement in performance metrics. After the first feature is selected, the process continues by testing

combinations of the already-selected features along with one additional feature. At each step, the feature that results in the best overall model performance is added to the selection. This iterative process continues until the model achieves the desired accuracy or until a predefined number of features is selected.

This methodology was especially suited to the project's dataset, which contained numerous variables, some of which might have introduced noise or irrelevant information. By systematically evaluating and selecting features based on their contribution to the prediction task, the project was able to narrow down the dataset to the most impactful predictors, thereby improving both model efficiency and accuracy.

The benefits of forward feature selection were many. It provided a computationally efficient way to focus only on subsets of features rather than testing all possible combinations, which would have been infeasible given the dataset size. Additionally, it offered a transparent approach to determining which features contributed most significantly to the classification task. This was crucial in building a reliable and interpretable system for classifying medical transcription data.

However, forward feature selection is not without its challenges. One key consideration is the risk of overfitting, especially if too many features are selected relative to the size of the dataset. Overfitting occurs when the model performs well on the training data but struggles to generalize to new, unseen data. To address this, cross-validation was incorporated into the feature selection process. Cross-validation ensured that the selected features consistently contributed to improved performance across different subsets of the data, mitigating the risk of overfitting and enhancing the model's ability to generalize.

Forward feature selection proved to be a powerful technique for refining the dataset and isolating the most predictive features. By focusing on features with the greatest impact, the project achieved better model accuracy while maintaining interpretability. The careful balance of computational efficiency, robustness, and validation ensured that the final model was both effective and generalizable, highlighting the importance of feature selection in the machine learning pipeline.

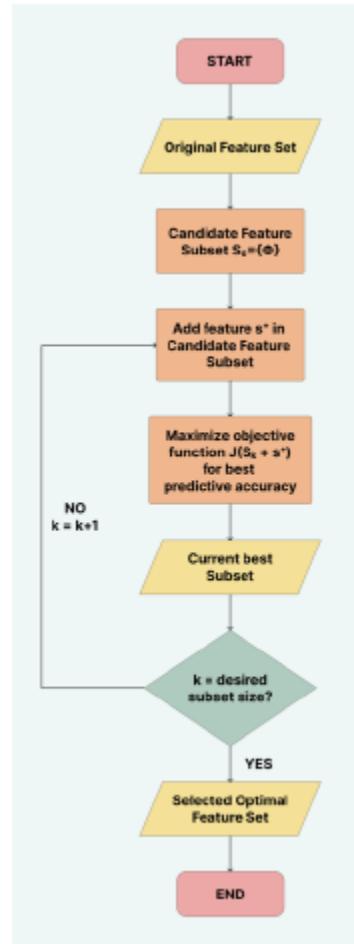


Fig. 4. Flowchart for Sequential Forward Selection (Feature Extraction Algorithm)

#### D. Model Training

Model Training for our project consisted of 3 steps as follows:

- Using classifiers for mapping:

Classifiers take input data (features) and map it to predefined classes (labels). This is done by learning patterns from a labeled dataset during the training phase. Once trained, the classifier can predict the class of new, unseen data. For this project, following classifiers were used:

- Logistic Regression: Logistic regression is a statistical technique used for predicting binary outcomes—determining whether a certain event will happen or not—based on a set of independent variables. It creates a connection between the input features and the probability of the target outcome. Additionally, logistic regression can be adapted to handle multi-class classification problems, making it a flexible option for machine learning tasks.
- Support Vector Machines (SVM): SVM is a supervised learning algorithm frequently utilized for both classification and regression tasks. It operates by identifying a hyperplane in a multi-dimensional feature space that effectively distinguishes data points into separate classes. In situations where the data cannot be linearly separated, SVM uses kernel functions to change the input features into higher-dimensional spaces, allowing it to classify intricate patterns adeptly. This adaptability makes SVM a strong choice for managing non-linear datasets.
- Random Forest: Random Forest is a powerful ensemble learning technique that excels in various tasks, including classification, regression, and feature selection. The algorithm builds multiple decision trees during the training phase and makes predictions by combining their outputs—either through majority voting for classification or averaging for regression. Random Forest is highly resistant to noisy data and outliers, and it performs effectively even with substantial datasets containing numerous input features, making it a dependable and commonly used algorithm.
- Gradient Boosting Trees: Gradient Boosting is a boosting approach that merges several weak models to create a robust predictive model. Gradient Boosting Trees build on this method by incorporating decision trees as the foundational learners. The procedure iteratively trains each subsequent tree to address the residual errors of the previous ones, enabling the algorithm to grasp complex, non-linear relationships between input features and target variables. Gradient Boosting Trees are highly efficient for both classification and regression tasks, yielding strong results across various datasets.
- Categorical Boosting: Categorical Boosting is a distinct variant of Gradient Boosting tailored to manage categorical data effectively. In contrast to traditional boosting methods, it integrates decision trees with categorical splits rather than continuous ones and employs encoding techniques like one-hot encoding or target encoding. This method is particularly advantageous for datasets featuring high-cardinality categorical attributes or imbalanced class distributions, making it an ideal choice for classification tasks that involve categorical inputs.

- Data splitting into training, validation, and test sets.

Data splitting is a crucial step in preparing the dataset for training and evaluating machine learning models. The dataset is divided into three distinct subsets: training, validation, and test sets. This division ensures that the model is trained effectively, tuned appropriately, and evaluated fairly on unseen data.

- Training Set: The training set comprises the majority of the data and is used to train the model by enabling it to learn patterns, relationships, and dependencies within the dataset. In this project, the training data underwent extensive preprocessing, such as stop-word removal, tokenization, and lemmatization, to ensure high-quality input. During training, the machine learning models, including Logistic Regression, SVM, Random Forest, Gradient Boosting Trees, and Categorical Boosting, were exposed to this subset to adjust their internal parameters and minimize errors in predictions.
- Validation Set: The validation set plays a critical role in hyperparameter tuning and model selection. This subset allows us to assess how well the model generalizes to data it has not seen during training, helping us identify issues like overfitting or underfitting. For instance, during this project, the validation set was used to tune hyperparameters such as the learning rate for Gradient Boosting or the maximum depth of trees in Random Forest. The feedback from the validation set guided iterative improvements to the system and informed decisions about feature selection and model adjustments.

- Test Set: Finally, the test set is reserved for evaluating the final performance of the model after training and validation. This set remains untouched during the earlier stages to provide an unbiased measure of how well the model can generalize to new data. In this project, the test set was used to generate performance metrics such as accuracy, precision, recall, and F1-score, providing a comprehensive assessment of the effectiveness of the proposed system.

The splitting process was conducted using an appropriate ratio, ensuring that each subset retained the diversity and representativeness of the original dataset. This careful division was particularly important for the project, as it aimed to build a scalable and reliable classification system for medical transcription tasks, where small errors can have significant implications.

- Addressing imbalance using RandomUnderSampler.

Addressing class imbalance was a pivotal step to ensure that the machine learning models could effectively classify medical transcriptions across all categories. The dataset, sourced from medical transcription records, exhibited significant class imbalance. For instance, certain specialties, such as "Surgery" or "Consultation and History," had a disproportionately higher number of records compared to others, such as "Cardiovascular/Pulmonary." This imbalance posed a challenge because machine learning models tend to favor majority classes, leading to biased predictions and poor performance on underrepresented categories.

To address this challenge, **RandomUnderSampler** was employed as a data preprocessing technique. RandomUnderSampler is a resampling method that tackles class imbalance by reducing the number of samples in the majority class to match the size of the minority class. This results in a balanced dataset where all classes have equal representation during model training.

#### E. Evaluation

- Metrics include accuracy, precision, recall, and F1-score.

## V. EXPERIMENTS

For data processing, we experimented various techniques to minimize the volume of data, facilitating easier feature selection, and developed a method to extract only the nouns from the transcripts using POS Tagging.

We explored different class quantities and documented the outcomes to reach certain conclusions.

In feature selection, we tried various amounts of features, finding that 10 features resulted in underfitting, while 20 features caused overfitting; thus, we decided to use 15 features.

Subsequently, we examined different models such as Logistic Regression, SVM, Random Forest, Gradient Boosting Trees, and Categorical Boosting to determine which model was most effective for this problem statement.

We also calculated Cosine and Jaccard similarity, finding that many of the transcriptions were similar yet categorized under different specialties.

#### Output:

- Binary Classification: Achieved 99% accuracy for "Surgery" and "Consultation and History."
- Three-Class Classification: Performance dropped to 75% due to overlaps in specialty descriptions.
- Insights suggest more granular feature engineering and larger datasets for improvements

## VI. RELATED WORK

The existing literature part from above in the report shows that several studies have explored the integration of Sequential Forward Selection (SFS) and Natural Language Processing (NLP) techniques to enhance clinical text classification in healthcare. These studies underscore the importance of feature selection, which significantly improves model performance by focusing on the most relevant data features while eliminating redundant or irrelevant ones. In particular, SFS has been shown to be highly effective in reducing the dimensionality of clinical datasets, making them more manageable and computationally efficient while maintaining the model's ability to make accurate predictions. This feature selection method is especially valuable in healthcare contexts, such as medical transcription classification, where the dataset often contains numerous features that can overwhelm the learning process.

Recent research has highlighted the use of deep learning techniques, including Convolutional Neural Networks (CNNs), to capture complex patterns in clinical text. CNNs, typically used for image classification, have been successfully adapted for text data due to their ability to learn hierarchical features and relationships in sequences of words. Studies have shown that CNNs can effectively capture local dependencies within clinical narratives, which is crucial for identifying key medical terms and their contexts. However, CNNs require large, annotated datasets to train effectively, which can be a significant limitation in healthcare applications where labeled data is often scarce.

In addition to CNNs, semi-supervised Support Vector Machines (SVMs) have also been widely studied for clinical text classification, particularly when labeled data is limited. Semi-supervised SVMs leverage both labeled and unlabeled data, making them an attractive option for healthcare applications where large amounts of unlabeled data are available. These models have demonstrated the ability to improve classification accuracy by using the unlabeled data to better generalize the decision boundaries between classes. Despite their promising results, semi-supervised SVMs also require substantial datasets to achieve optimal performance, and their complexity can be a challenge in real-world applications.

While CNNs and semi-supervised SVMs have been shown to offer high accuracy in clinical text classification, they often demand extensive computational resources and large labeled datasets. Instead, the project focused on using more accessible and computationally efficient techniques, such as SFS for feature selection and traditional machine learning classifiers like Logistic Regression, SVM, Random Forest, and Gradient Boosting. These models, although simpler, have been effective in handling the medical transcription dataset used in this project, especially given the limitations of available labeled data and computational resources. Nevertheless, the integration of more advanced techniques such as CNNs and semi-supervised SVMs could be explored in future work, particularly as larger datasets become available for training and validation.

## VII. RESULTS

We observed that the dataset used for training and testing the model is relatively small, which poses a significant challenge for making highly accurate predictions. The limited number of records available in the dataset restricts the model's ability to learn diverse patterns and may lead to suboptimal performance, particularly when dealing with more complex or nuanced data. To enhance the model's ability to make precise and reliable predictions, it is clear that a larger, more comprehensive dataset with a greater number of medical records is required. A larger dataset would allow the model to capture more varied and representative examples of each class, thereby improving its generalizability and overall prediction accuracy.

Despite the limitations of the dataset size, the model demonstrated notable success when tasked with classifying two specific medical specialties—‘Surgery’ and ‘Consult and History’—achieving a remarkable accuracy of 99% using Categorical Boosting. This high level of accuracy indicates that the model performed exceptionally well in distinguishing between these two classes, which may be due to their distinct and well-defined characteristics in the dataset. The data for these specialties appeared to have more clearly separable features, making the classification task easier and leading to strong performance.

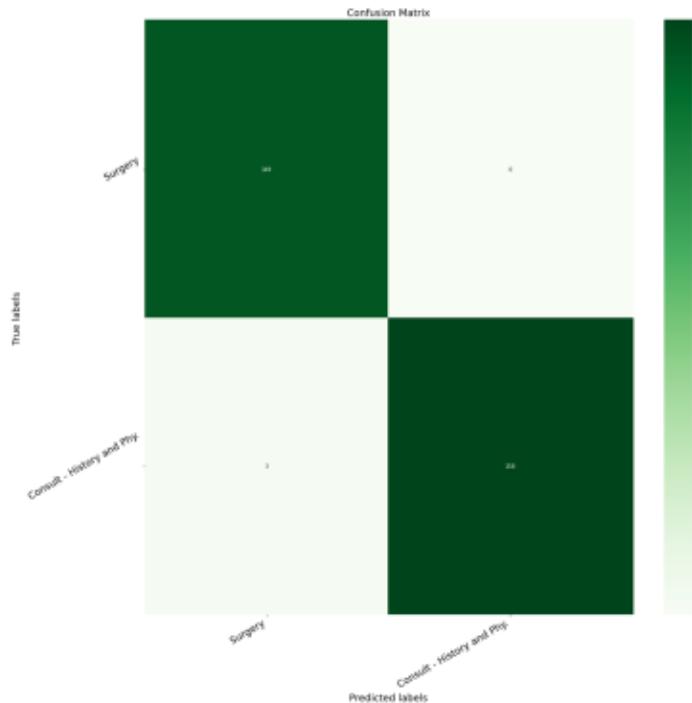


Fig. 5. Confusion Matrix (Heatmap) for 2 Medical Illnesses using Categorical Boosting

	precision	recall	f1-score	support
Surgery	0.98	1.00	0.99	149
Consult - History and Phy.	1.00	0.98	0.99	161
accuracy			0.99	310
macro avg	0.99	0.99	0.99	310
weighted avg	0.99	0.99	0.99	310

Fig. 6. Classification Report for 2 Medical Illnesses using Categorical Boosting

However, when the classification task was expanded to include a third medical specialty—‘Cardiovascular/Pulmonary’—the accuracy dropped to 75%. This reduction in performance can be attributed to several factors, with one of the primary reasons being the overlapping nature of the medical transcripts across different specialties. Many of the records related to ‘Cardiovascular/Pulmonary’ were misclassified as belonging to either the ‘Surgery’ or ‘Consultation and History’ categories. This is because a significant portion of the medical transcripts describing cardiovascular or pulmonary conditions also involved surgical procedures or detailed patient histories, which are common elements in both the ‘Surgery’ and ‘Consultation and History’ classes.

For example, medical records in the ‘Cardiovascular/Pulmonary’ category often describe heart surgeries, such as bypass surgeries or valve replacements, or discuss a patient’s cardiac history, which may overlap with descriptions found in the ‘Surgery’ or ‘Consultation and History’ categories. As a result, the model had difficulty distinguishing between these classes and frequently misclassified such records into one of the two overlapping categories. This issue of misclassification highlights the need for more refined features or additional data to better differentiate between these specialties, especially in cases where the medical transcriptions describe complex or interconnected clinical scenarios.

Overall, while the model performed exceptionally well on simpler, binary classifications, the challenge of handling more complex, multi-class classification tasks revealed limitations due to both the dataset’s size and the inherent overlap in the medical transcriptions. Moving forward, addressing these challenges through techniques such as dataset augmentation, improved feature engineering, and leveraging more advanced classification algorithms could help mitigate these issues and improve the model’s performance across a broader set of classes.

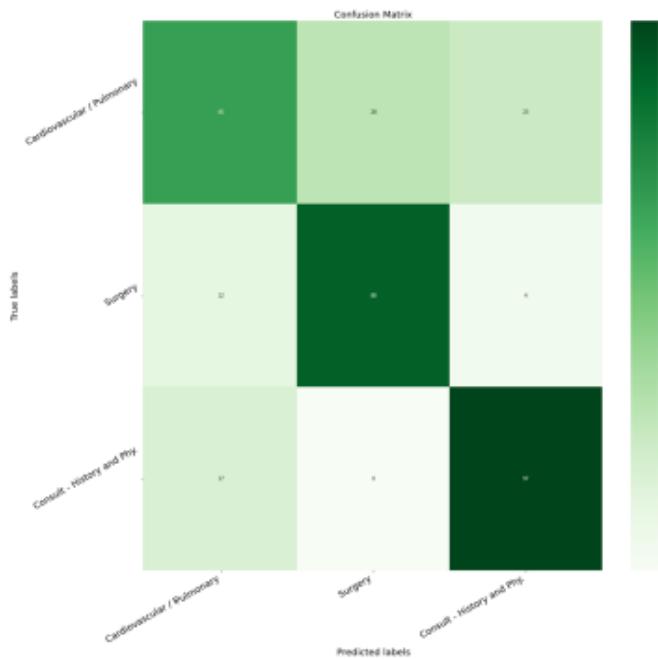


Fig. 7. Confusion Matrix (Heatmap) for 3 Medical Illnesses using Categorical Boosting

	precision	recall	f1-score	support
Cardiovascular / Pulmonary	0.69	0.56	0.62	116
Surgery	0.76	0.85	0.80	104
Consult - History and Phy.	0.78	0.85	0.82	114
accuracy			0.75	334
macro avg	0.74	0.75	0.74	334
weighted avg	0.74	0.75	0.74	334

Fig. 8. Classification Report for 3 Medical Illnesses using Categorical Boosting

### VIII. CONCLUSION

The project highlights the potential of SFS and traditional classifiers in clinical text classification. Despite achieving high accuracy in binary settings, performance suffers with increased complexity due to dataset limitations.

We can utilize our understanding of the subject to organize similar categories, thereby decreasing the total number of categories we need to consider. Developing customized features manually might enhance the performance of the dataset, but these tailored features may not be as effective for other transcription datasets that differ from this one. We have concluded that additional data is necessary to accurately categorize the transcriptions into their respective medical classifications. Future work may involve segmenting the text in a way that facilitates multiclass classification.

Our evaluation indicates that we require more data to effectively categorize transcriptions into medical classifications. The existing dataset is insufficient, which has hindered our ability to achieve the level of accuracy we desire. To remedy this, we will aim to break the transcriptions into smaller, more defined units in the future. This approach will allow us to adopt a multiclass classification method, leading us to classify the medical content in a more detailed and refined manner. We believe this will considerably enhance the precision of our classification outcomes, making our medical transcription system more effective and trustworthy. To put it simply, we need to gather more data and subdivide it into smaller, more targeted segments. This will assist us in more accurately classifying the transcriptions into the appropriate medical categories.

### REFERENCES

- [1] Yao, L., Mao, C. & Luo, Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Med Inform Decis Mak* 19 (Suppl 3), 71 (2019). <https://doi.org/10.1186/s12911-019-0781-4>
- [2] Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, Yang GZ. Deep Learning for Health Informatics. *IEEE J Biomed Health Inform.* 2017 Jan;21(1):4-21. doi: 10.1109/JBHI.2016.2636665. Epub 2016 Dec 29. PMID: 28055930.
- [3] Marcano-Cedeño, Alexis & Quintanilla, Joel & Cortina-Januchs, Guadalupe & Andina, Diego. (2010). Feature selection using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network. *IECON Proceedings (Industrial Electronics Conference)*. 2845 - 2850. 10.1109/IECON.2010.5675075.
- [4] Vijay Garla, Caroline Taylor, Cynthia Brandt, Semi-supervised clinical text classification with Laplacian SVMs: An application to cancer case management, *Journal of Biomedical Informatics*, Volume 46, Issue 5, 2013, Pages 869-875, ISSN 1532-0464, <https://doi.org/10.1016/j.jbi.2013.06.014>.
- [5] Dina Demner-Fushman, Wendy W. Chapman, Clement J. McDonald, What can natural language processing do for clinical decision support?, *Journal of Biomedical Informatics*, Volume 42, Issue 5, 2009, Pages 760-772, ISSN 1532-0464, <https://doi.org/10.1016/j.jbi.2009.08.007>.
- [6] Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc.* 2008 Jan-Feb;15(1):14-24. doi: 10.1197/jamia.M2408. Epub 2007 Oct 18. PMID: 17947624; PMCID: PMC2274873.
- [7] Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Peter Bruza. 2014. Medical Semantic Similarity with a Neural Language Model. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)*. Association for Computing Machinery, New York, NY, USA, 1819–1822. <https://doi.org/10.1145/2661829.2661974>
- [8] Last, Mark & Kandel, Abraham & Maimon, Oded. (2002). Information-theoretic algorithm for feature selection. *Pattern Recognition Letters.* 22. 799-811. 10.1016/S0167-8655(01)00019-8.
- [9] Mineichi Kudo, Jack Sklansky, Comparison of algorithms that select features for pattern classifiers, *Pattern Recognition*, Volume 33, Issue 1, 2000, Pages 25-41, ISSN 0031-3203, [https://doi.org/10.1016/S0031-3203\(99\)00041-2](https://doi.org/10.1016/S0031-3203(99)00041-2).
- [10] D. Xiao and J. Zhang, "Importance Degree of Features and Feature Selection," 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, Tianjin, China, 2009, pp. 197-201, doi: 10.1109/FSKD.2009.625.