*4th Indian International Conference on Industrial Engineering and Operations Management*
*Hyderabad, Telangana, India, November 07-09, 2024*
Publisher: IEOM Society International, USA DOI: 10.46254/IN04.20240089
Published: November 07, 2024

# Integrating NLP and Random Forest Algorithms for Clinical Decision Support: Leveraging EHR Data to Improve Patient Care and Operational Efficiency

**Bushra Siddiqua**
PG Scholar, Department of Computer Science and Engineering,
Bharatiya Engineering Science and Technology Innovation University (BESTIU), Sri Satya Sai, Andhra Pradesh, India
2022wpcse004@bestiu.edu.in

**Dr. Anjaiah Adepu**
Professor, Department of CSE (Hod) & Principal IC
Maulana Azad National Urdu University
A Central University, Ministry of Education, Government of India, POLYTECHNIC, DARBHANGA, Bihar-846002, India
anjaniprasad.adepu@gmail.com

## Abstract

This research aimed to investigate and compare the effectiveness of three machine learning strategies that use Random Forest algorithms to assist in making clinical decisions using Electronic Health Records. The first strategy predicts readmission based only on diagnosis codes. The second strategy uses natural language processing to analyze clinical notes and extract relevant terms for predicting readmission. The third strategy combines vital signs and lab results with the second approach to provide additional data for the Random Forest models.The rapid increase in Electronic Health Records (EHR) offers a chance to improve healthcare services through advanced machine learning techniques. However, the complexity and variety of EHR data, which includes both structured information (like lab results) and unstructured data (such as clinical notes), present significant challenges for analysis. Natural Language Processing (NLP) provides an effective way to derive valuable insights from unstructured medical text. Random Forest (RF) algorithms, recognized for their strength and clarity, are particularly effective for analyzing complex, multi-source healthcare data. This study aims to utilize RF models in conjunction with NLP techniques to create predictive models that support clinical decision-making and enhance personalized patient care. Our results indicate that the Random Forest model's use of clinical notes can match or even surpass the effectiveness of diagnosis codes in predicting readmission within thirty days post-hospital discharge. The unstructured text found in Electronic Health Records (EHRs)—including discharge summaries, physician notes, and radiology reports—necessitates the application of NLP for effective data extraction. Essential NLP tasks such as Named Entity Recognition (NER), medical concept extraction, and text classification are crucial for converting free-text data into structured information that can be used as input features for machine learning models. By combining RF with NLP-driven feature engineering, we can develop precise and interpretable predictive models for various applications, including disease diagnosis, predicting patient readmissions, and risk stratification.

## Keywords

Electronic Health Records, machine learning, Random Forest model, natural language processing, feature normalization, logistic regression, clinical decision making, EHR analysis, patient tailoring.

# 1. Introduction

In current years, the combination of herbal language processing (NLP) algorithms and system studying has performed a key position in remodeling healthcare structures, particularly medical selection guide structures (CDSS). These structures goal to guide healthcare specialists through reading massive quantities of statistics to guide decisions, enhance diagnostic accuracy, and customise affected person care. Among numerous system studying approaches, the aggregate of NLP and random forest (RF) algorithms has proven unique promise for medical selection guide, assisting healthcare providers extract precious insights from unstructured medical textual content to construct and create strong predictive models.

## Objectives

## 1.1 Importance of Clinical Decision Support Systems (CDSS)

CDSS are tools designed to analyze patient data and support clinicians in making more informed decisions. With the exponential growth of healthcare data, including electronic health records (EHRs), medical literature, clinical notes, and diagnostic reports, clinicians face challenges in synthesizing all available information to make quick and accurate diagnoses. CDSS help bridge this gap by providing real-time analysis and recommendations. In addition to improving diagnostic accuracy, CDSS can enhance treatment planning, reduce medical errors, and optimize resource allocation in healthcare facilities.

However, clinical data often comes in both structured (e.g., lab results, demographic information) and unstructured forms (e.g., free-text clinical notes). Unstructured data, which constitutes a large part of patient records, is challenging to analyze due to its irregular format, diverse vocabulary, and context-specific language. NLP algorithms address these challenges by transforming unstructured text into structured data, making it accessible for machine learning algorithms like Random Forest, which can then be used to make clinically relevant predictions.

## 1.2 Role of NLP in CDSS

NLP is the sphere of synthetic intelligence that makes a speciality of permitting computer systems to understand, interpret, and generate human language. In the context of CDSS, NLP performs a crucial position in reading unstructured scientific textual content. For instance, NLP can become aware of key clinical concepts, signs and symptoms, diagnoses, medications, and remedy plans inside scientific notes. By the usage of strategies like named entity recognition (NER), sentiment analysis, and textual content classification, NLP structures can extract significant functions from textual content statistics that in any other case might be hard to examine. NLP also can become aware of relationships among terms (along with a drug and its facet effects) and display complicated affected person narratives hidden inside scientific notes.

For example, an NLP gadget may examine a scientific observe to become aware of signs and symptoms related to a selected analysis or danger elements that growth the probability of a selected outcome. These insights are treasured inputs for system getting to know models, along with Random Forest, as they provide a complete view of a affected person's situation primarily based totally on historic and contextual statistics.

## 1.3 Random Forest in Clinical Decision Support

Random Forest, an ensemble gaining knowledge of algorithm, is broadly identified for its robustness, excessive accuracy, and interpretability, making it well-suitable for scientific applications. In CDSS, Random Forest may be carried out to expect disorder outcomes, suggest treatments, and check affected person chance primarily based totally on dependent facts derived from each conventional EHRs and NLP-processed scientific text. Random Forest operates with the aid of using growing a couple of choice timber at some stage in schooling and merging their predictions to enhance accuracy and decrease overfitting. This technique facilitates seize complex, non-linear relationships in healthcare facts, permitting the CDSS to offer dependable tips even if coping with noisy, incomplete, or ambiguous information.

A key gain of the usage of Random Forest in CDSS is its cappotential to deal with excessive-dimensional facts, that is regularly encountered in scientific settings. Medical facts may be vast, inclusive of a massive range of functions generated with the aid of using NLP (e.g., diverse phrases extracted from scientific notes) and dependent EHR facts.

Random Forest can successfully manage those feature-wealthy datasets with the aid of using figuring out the maximum critical variables contributing to the choice-making technique**.**

## 2. Literature Review

Recent studies on the combination of Natural Language Processing (NLP) with system mastering algorithms, which include Random Forest (RF), has appreciably superior the sphere of Clinical Decision Support Systems (CDSS). NLP's cappotential to investigate unstructured records like scientific notes, blended with RF's sturdy prediction capabilities, has enabled CDSS to extract and use precious insights from digital fitness records (EHRs) effectively.

Studies have explored numerous processes to enhance CDSS overall performance thru NLP and RF. For instance, a examine via way of means of Qin et al. used Bidirectional Encoder Representations from Transformers (BERT) for NLP on scientific notes and blended it with system mastering fashions to categorise patients' sepsis risk, demonstrating that this NLP-more desirable pipeline outperformed preceding predictive fashions utilized in comparable tasks. However, their findings highlighted a want for broader validation to enhance generalizability throughout scientific settings

Other studies has investigated the software of NLP and RF for persistent ailment prediction. Liu et al. as compared one of a kind ML fashions, which include Random Forest, for predicting liver ailment development and observed that RF completed a very good stability of accuracy and interpretability. They pressured the significance of the usage of function choice strategies in NLP to perceive key affected person attributes for prediction, thereby enhancing the efficacy of the fashions in realistic applications

Overall, those current improvements underscore the effectiveness of integrating NLP with Random Forest in CDSS. However, additionally they emphasize the demanding situations of records standardization, version validation, and interpretability. Researchers maintain to paintings on those troubles to make sure that NLP and RF integration in CDSS can assist real-time, reliable, and personalised affected person care throughout numerous healthcare environments.

## 3. Methods

Integrating Natural Language Processing (NLP) and Random Forest (RF) algorithms to analyze Electronic Health Records (EHR) offers a structured methodology for extracting meaningful insights from complex healthcare data. The method harnesses NLP's ability to process unstructured text data—such as clinical notes—and RF's ensemble learning approach, which is adept at managing high-dimensional data and delivering robust predictions. Here's a breakdown of how this methodology unfolds in a clinical decision support system (CDSS) context.

### 3.1. Data Collection and Preprocessing

The integration procedure starts with amassing EHR records, which includes each dependent records (e.g., lab results, remedy orders) and unstructured records (e.g., clinician notes, discharge summaries). To beautify records quality, preprocessing steps are vital for standardizing and cleansing each sorts of records. Structured records commonly undergoes normalization, lacking price handling, and function scaling, whilst unstructured records calls for extra extensive NLP-centered preprocessing, together with tokenization, stop-phrase removal, and stemming or lemmatization.

For medical textual content processing, NLP strategies together with Named Entity Recognition (NER) are hired to perceive essential clinical entities (e.g., diseases, medications, procedures) and standardize them into dependent terms. NER fashions like BERT, which has been proven to carry out properly in extracting clinically applicable features, can assist make certain that records from various medical statistics is constantly processed. Additionally, NLP fashions can hire vectorization strategies, together with TF-IDF or phrase embeddings, to transform textua**l** records into numerical representations appropriate for system studying analysis

### 3.2 Feature Extraction and Transformation

Once the textual content is processed, the following step includes deciding on and engineering capabilities so that it will function inputs for the Random Forest model. The NLP-derived capabilities generally encompass each high-stage phrases extracted from scientific narratives (e.g., prognosis phrases, drug names) and their relationships to scientific

outcomes. Combining those with established information capabilities like age, lab values, and vitals affords a greater complete dataset.

Feature choice techniques, consisting of recursive characteristic elimination (RFE) or primary aspect analysis (PCA), assist lessen dimensionality and dispose of redundant or inappropriate records from the dataset, making sure that handiest the maximum predictive capabilities are used. This step is essential for RF's performance, because it improves each the model's accuracy and interpretability, permitting clinicians to make feel of the choices made via way of means of the model.

### 3.3 Training the Random Forest Model

With capabilities extracted, the following step is to teach the Random Forest algorithm, an ensemble studying approach that mixes more than one choice bushes to supply dependable predictions. In a healthcare context, RF is desired for its capacity to address heterogeneous information kinds and its interpretability. Each tree in the wooded area is skilled on a random subset of information, and the very last prediction is made via way of means of aggregating the results of every person tree.

During training, RF identifies non-linear relationships and interactions among capabilities that may be vital for predicting medical results. For example, it could screen how a aggregate of signs extracted thru NLP may correlate with particular diagnoses or the probability of disorder progression. Hyperparameter tuning, including optimizing the range of bushes or adjusting most depth, is likewise achieved to obtain the nice viable version performance.

### 3.4 Model Validation and Evaluation

To make certain the version generalizes properly to new data, cross-validation is employed, wherein the dataset is cut up into education and trying out subsets. Performance metrics generally utilized in scientific studies, along with accuracy, precision, recall, F1-score, and location below the ROC curve (AUC-ROC), are measured. These metrics offer insights into the version's reliability, sensitivity, and specificity—key elements for scientific application. If the version fails to satisfy a excessive trendy of accuracy, iterative modifications are made to the NLP preprocessing steps, characteristic engineering, or RF parameters till surest overall performance is achieved.

### 3.5 Model Interpretation and Clinical Integration

After the version is skilled and validated, its interpretability is assessed, mainly that specialize in the way it derived medical decisions. Feature significance ratings in the RF version assist spotlight which variables have been maximum influential withinside the version's predictions. For example, phrases extracted from medical notes, like "chest pain" or "extended blood pressure," can be recognized as crucial threat factors.

Once validated, the included CDSS may be embedded into medical workflows, in which it assists healthcare carriers in real-time decision-making. The machine can spotlight styles and capability risks, allowing proactive take care of high-threat patients. Continuous tracking and retraining with new information make certain the version stays applicable and plays nicely in distinct medical settings.

## 4. Data Collection

### 4.1 NLP EHR

The integration of Natural Language Processing (NLP) with Artificial Intelligence (AI) in Electronic Health Records (EHR) has revolutionized medical facts management, making it viable to convert unstructured medical textual content into actionable insights. NLP algorithms examine textual content facts, which include medical notes, discharge summaries, and affected person interactions, extracting essential clinical entities consisting of symptoms, diagnoses, medications, and procedures. This allows a greater complete view of a affected person's clinical records and cutting-edge fitness status.

Through AI, especially system gaining knowledge of fashions like Random Forest, those NLP-derived capabilities are blended with based EHR facts (e.g., lab results, demographics) to make correct predictions and recommendations. This hybrid method helps Clinical Decision Support Systems (CDSS) via way of means of enhancing diagnostic accuracy, figuring out high-threat patients, and personalizing remedy plans.

NLP and AI in EHR processing streamline medical workflows via way of means of minimizing guide facts access and permitting short get right of entry to to applicable affected person information. With non-stop gaining knowledge of, those fashions adapt over time, enhancing prediction accuracy and in the end improving affected person care (Figure 1).

exceptional and operational performance in healthcare settings

```
nltk.download('stopwords')
from nltk.corpus import stopwords
stop_words = set(stopwords.words("english"))

def preprocess_text(text):
    # Convert text to lowercase, remove stop words, and perform s
    tokens = nltk.word_tokenize(text.lower())
    tokens = [word for word in tokens if word.isalnum() and word
    return " ".join(tokens)

unstructured_data = unstructured_data.apply(preprocess_text)
```

Figure 1. NLP EHR DAT

## Implementing Random Forest

For implementing a Random Forest algorithm on Electronic Health Records (EHR) data using Natural Language Processing (NLP), several recent studies showcase effective methods to process and classify large and complex medical datasets, often derived from unstructured EHR text.

·**Data Processing with NLP:** Initially, NLP techniques are applied to extract meaningful features from unstructured text data in EHRs. Techniques such as tokenization, stop-word removal, stemming, and named entity recognition (NER) are often used to prepare the text data. NER, in particular, can identify key entities such as symptoms, diseases, and treatment plans within clinical notes, transforming raw EHR text into structured formats that can be utilized by machine learning algorithms.

·**Random Forest for Classification:** The processed facts is then labeled the usage of Random Forest, a strong ensemble gaining knowledge of technique. The set of rules builds a couple of choice bushes on diverse subsets of the schooling facts and merges them to acquire correct and solid predictions. For example, in research concerning fitness datasets, capabilities extracted from EHRs (like affected person demographics, analysis codes, and medical narratives) are enter into Random Forest classifiers to are expecting results like ailment hazard or remedy efficacy. Random Forest's capacity to address various facts kinds and pick out complicated styles makes it a famous preference in HER (Figure 2).

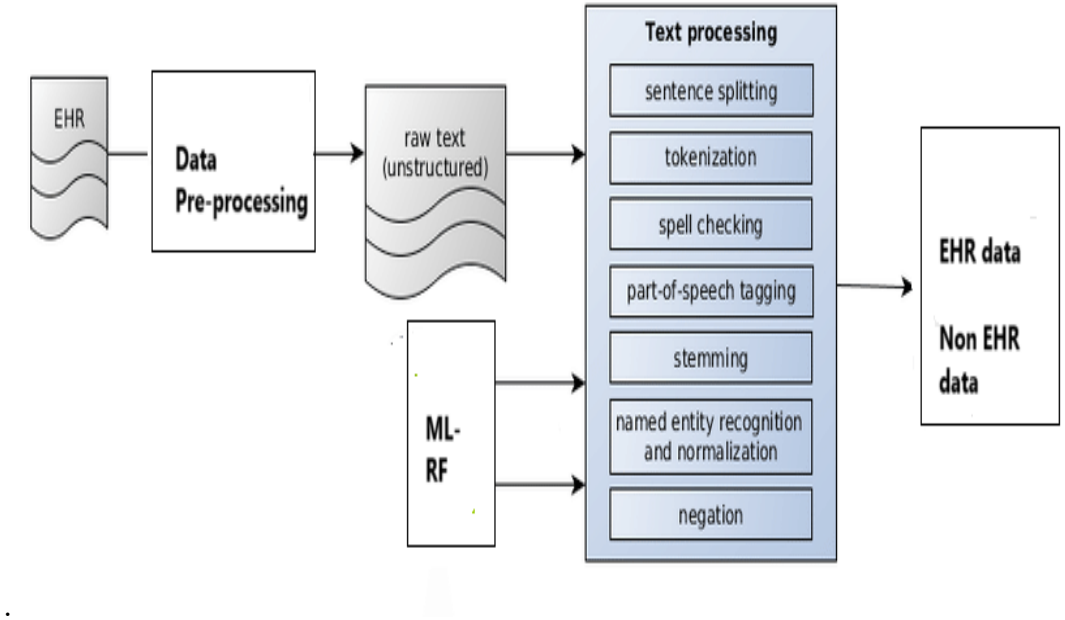


Figure 2. Implementation Architecture

\

## 5. Results and Discussion

### 5.1 Dataset Information:

Electronic Health Record (EHR) datasets incorporate de-diagnosed affected person fitness statistics, along with demographics, medical notes, lab results, medicinal drug histories, diagnoses, and remedy outcomes. These datasets are important for medical research, predictive modeling, and fitness analytics, presenting established and unstructured statistics types. EHR statistics may be used to research styles in sickness progression, remedy effectiveness, and affected person outcomes. Examples of broadly used EHR datasets encompass the MIMIC-III and MIMIC-IV from PhysioNet, in addition to artificial datasets like Synthea that simulate affected person fitness facts for diverse conditions. These datasets permit the utility of system gaining knowledge of and NLP strategies in healthcare

## Patient Information:

- ***Name***: John Doe
- ***Date of Birth***: 01/01/1970
- ***Address***: 123 Main St, Springfield, IL
- ***Phone***: (555) 123-4567
- ***Email***: john.doe@example.com
- ***Emergency Contact***: Jane Doe, (555) 765-4321

---

## Clinical History:

- ***Chief Complaint***: Chronic cough and shortness of breath.
- ***History of Present Illness (HPI)***: John Doe, a 54-year-old male, presents to the clinic with a chronic cough that has persisted for the past six months. He reports shortness of breath, especially during physical activity. No significant fever or nausea noted.
- ***Past Medical History***:
  - Hypertension
  - Hyperlipidemia
  - Type 2 Diabetes Mellitus
- ***Family History***: Father had a history of stroke and cardiac failure. Mother had chronic renal disease.

### 5.2 NLP Analyzing

NLP strategies are used to investigate each EHR and non-EHR information via way of means of extracting significant statistics from unstructured textual content. In EHR information, NLP identifies essential clinical concepts (e.g., symptoms, diseases, treatments), helping decision-making. Non-EHR information, consisting of social media or reviews, is analyzed for sentiment and trends. Both programs use strategies like tokenization, named entity recognition (NER), and subject matter modeling for textual content processing (Figure 3).

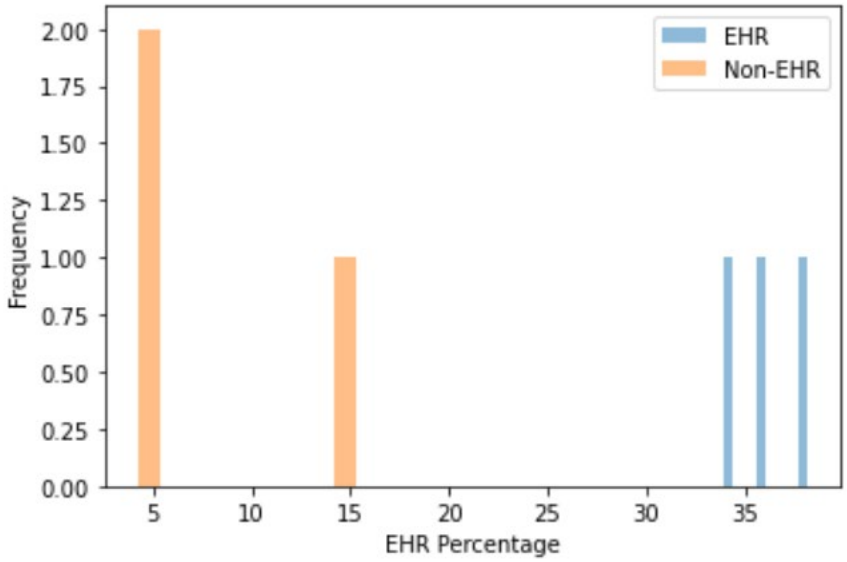


Figure 3. EHR Information

```
Cleaned and Lemmatized Text Dictionary:
headings: ['patient', 'information', 'medical', 'history', 'current']... (15 words)
bold: ['name', 'date', 'birth', 'gender', 'address']... (32 words)
italic: []... (0 words)
italian: []... (0 words)
normal: ['', '', 'sarah', 'adam', '']... (180 words)

Word Counts:
headings Word Counts:
patient: 1
information: 1
medical: 1
history: 1
current: 1
Total headings words: 15
bold Word Counts:
date: 2
diagnosis: 2
asthma: 2
management: 2
name: 1
Total bold words: 32
```

## 5.3 Classification Results

Random Forest (RF) is a effective system studying set of rules used for type responsibilities in each EHR and non-EHR statistics. In EHR statistics, RF facilitates classify scientific outcomes, expect disorder progression, and investigate remedy efficacy with the aid of using managing each structured (e.g., lab results) and unstructured (e.g., medical notes) statistics. The version works with the aid of using constructing more than one selection bushes and mixing their outputs to enhance accuracy and decrease overfitting. For non-EHR statistics, RF is carried out in regions like textual content type, sentiment analysis, and purchaser conduct prediction. Its capacity to deal with high-dimensional statistics with complicated styles makes it best for each domains

Summary of EHR and Non-EHR Percentages:

| | Document Name | EHR Percentage (%) | Classification |
|---|---|---|---|
| 0 | ehr4.pdf | 46.067416 | EHR |
| 1 | ehr5.pdf | 41.621622 | EHR |
| 2 | ehr3.pdf | 50.000000 | EHR |
| 3 | non_ehr1.pdf | 4.229607 | Non-EHR |
| 4 | non_ehr2.pdf | 4.513064 | Non-EHR |
| 5 | non_ehr3.pdf | 15.750916 | Non-EHR |

## 6. Conclusion

Integrating NLP and Random Forest (RF) algorithms for clinical decision support enhances the ability to analyze complex Electronic Health Record (EHR) data and provide predictive insights. NLP techniques extract valuable features from unstructured clinical text, while RF classifies and predicts patient outcomes based on structured and unstructured data. This combination supports better diagnosis, risk assessment, and personalized treatment strategies. The synergy between these methods enables efficient handling of large-scale healthcare data, improving clinical decision-making. Both EHR and non-EHR datasets benefit from this approach, allowing for advanced healthcare analytics and improving patient outcomes.

## References

A. Cernian, V. Sgarciu, B. Martin. Sentiment Analysis from Product Reviews Using SentiWordNet as Lexical Resource. Journal of Text Analysis; 2015.

Anupam Bhardwaj, Pooja Khanna, Sachin Kumar, Pragya. Generative Model for NLP Applications Based on Component Extraction. Journal of Machine Learning Research; 2020.

Dr. Arivoli A, Sonali Pandey. Sentiment Analysis using Support Vector Machine Based on Feature Selection and Semantic Analysis. Computational Intelligence and Data Analysis Journal; 2021.

M. Kavitha, Bharat Bhushan Naib, Basetty Mallikarjuna, R. Kavitha, R. Srinivasan. Sentiment Analysis using NLP and Machine Learning Techniques on Social Media Data. Journal of Social Media Analytics; 2022.

Prabha PM Surya, B Subbulakshmi. Sentiment Analysis Using Naïve Bayes Classifier. International Journal of Text Mining; 2019.

Sheeba Naz, Aditi Sharan, Nidhi Malik. Sentiment Classification on Twitter Data Using Support Vector Machine. Journal of Computational Linguistics; 2018.

Shi Yuan, Junjie Wu, Lihong Wang, Qing Wang. A Hybrid Method for Multi-class Sentiment Analysis of Micro-blogs. Journal of Data Science and Analytics; 2016.

V. K. Vijayan, K. R. Bindu, L. Parameswaran. A Comprehensive Study of Text Classification Algorithms. Data Mining and Knowledge Discovery; 2017.

Walaa Saber Ismail. Sentiment Analysis of ChatGPT Tweets Using Machine Learning Techniques. International Journal of Data Science and Machine Learning; 2023.

Wei Yen Chong, Bhawani Selvaretnam, Lay-Ki Soon. Natural Language Processing and Sentiment Analysis. Journal of Computational Linguistics and Intelligent Systems; 2024.