https://ijmsphr.com

DOI: - 10.37547/ijmsphr/Volume07Issue03-05

Open Access Research Paper

INTERNATIONAL JOURNAL OF MEDICAL SCIENCE AND PUBLIC HEALTH RESEARCH

Vol. 07, No. 03, March 2026, 27-37

eISSN 2767-3774

Ensemble Machine Learning and Natural Language Processing for Automated Cancer Indicator Detection in Clinical Notes

Md Yassir Mottalib
Master of Science in Information System Technology, Wilmington University, USA

Nur Nobe
Department of Health Sciences & Leadership, St. Francis College, Brooklyn, USA

MD Tanvir Islam
Department of computer science, Monroe University, USA

Md Refat Hossain
Master of Business Administration, Westcliff University, USA

Afjal Hossain Jisan
Department of Supply Chain & Information Systems, The Pennsylvania State University, University Park, Pennsylvania

Md. Emran Hossen
Department of Science in Biomedical Engineering, Gannon University, USA

Abstract

Trebuchet MSEarly identification of cancer indicators within clinical documentation is essential for improving diagnostic efficiency and patient outcomes. This study presents a Natural Language Processing (NLP) and machine learning framework designed to extract cancer-related indicators from unstructured clinical notes. Clinical text data obtained from Kaggle and structured diagnostic features from the Breast Cancer Wisconsin (Diagnostic) Dataset available through the UCI Machine Learning Repository were used to develop and evaluate the proposed model. The methodology involved comprehensive text preprocessing, TF-IDF-based feature extraction, and feature engineering to represent clinically meaningful patterns in narrative medical text. Multiple machine learning algorithms, including Logistic Regression, Support Vector Machines, Random Forest, and Gradient Boosting classifiers, were trained and evaluated using standard performance metrics. Experimental results indicate that ensemble learning approaches outperform traditional classifiers in detecting cancer-related information from clinical narratives. Among the evaluated models, the Gradient Boosting classifier achieved the best performance with an accuracy of 95%, precision of 94%, recall of 93%, and an F1-score of 0.93. These results demonstrate the effectiveness of machine learning-based NLP systems in identifying cancer indicators within electronic health records. The proposed framework highlights the potential of automated clinical text analysis to support early cancer detection, enhance clinical decision support systems, and improve healthcare data analytics.

Keywords: Natural Language Processing, Machine Learning, Cancer Detection, Clinical Notes, Healthcare Analytics, Electronic Health Records, Text Mining.

Introduction

Cancer remains one of the leading causes of mortality worldwide and represents a significant burden on healthcare systems. According to the World Health Organization, cancer accounts for millions of deaths every year and continues to increase due to population aging, environmental factors, and lifestyle changes. Early detection and timely diagnosis are among the most critical factors in improving patient survival rates and reducing treatment costs. However, despite significant advances in medical technologies and diagnostic tools, identifying early cancer indicators remains a complex challenge in modern healthcare systems.

In clinical practice, a substantial amount of patient information is recorded in free-text form within electronic health records (EHRs). These records often include physician notes, pathology reports, radiology interpretations, discharge summaries, and consultation documentation. While these clinical narratives contain

valuable insights regarding symptoms, diagnostic observations, and potential disease indicators, they are difficult to analyze using conventional structured data analysis methods. Healthcare providers often rely on manual review of clinical documentation to identify disease indicators, which can be time-consuming, labor-intensive, and prone to human error.

The rapid growth of electronic health records has created an unprecedented opportunity for data-driven healthcare analytics. Organizations such as the U.S. Department of Health and Human Services and the Centers for Disease Control and Prevention have emphasized the importance of leveraging clinical data to improve disease surveillance, early diagnosis, and population health management. However, unlocking the full potential of clinical text requires advanced computational techniques capable of interpreting unstructured medical language.

Natural Language Processing (NLP) has emerged as a powerful approach for analyzing unstructured textual data in healthcare environments. NLP techniques allow computers to interpret human language, extract relevant information, and convert narrative text into structured data suitable for analysis. In recent years, NLP has been increasingly applied in biomedical research to identify disease indicators, extract clinical concepts, and support automated medical decision-making (Jiang et al., 2017). By applying NLP techniques to clinical documentation, researchers can automatically detect patterns associated with disease symptoms, diagnostic findings, and treatment outcomes.

Machine learning methods further enhance the ability to analyze clinical text by learning patterns from large datasets. Traditional statistical models as well as modern machine learning algorithms can identify relationships between linguistic features and clinical outcomes. Algorithms such as Logistic Regression, Support Vector Machines, Random Forests, and Gradient Boosting have demonstrated strong performance in medical text classification tasks (Esteva et al., 2019). These methods can be trained to identify specific disease indicators embedded within physician narratives and clinical reports.

The integration of NLP and machine learning offers a promising framework for extracting cancer-related indicators from clinical notes. Clinical narratives frequently contain important information regarding suspicious lesions, abnormal tissue growth, biopsy results, and metastasis indicators. However, these insights are often buried within lengthy textual descriptions, making manual analysis inefficient. Automated extraction of cancer indicators using NLP-based systems could significantly improve early detection efforts, assist clinicians in identifying high-risk patients, and support large-scale medical research.

Furthermore, the use of machine learning models in healthcare analytics has gained significant attention in recent years due to their ability to analyze large and complex datasets. Research has shown that machine learning approaches can improve diagnostic accuracy and provide valuable decision support for healthcare professionals (Rajkomar et al., 2019). By leveraging both structured medical data and unstructured clinical narratives, data-driven systems can provide deeper insights into disease patterns and improve healthcare outcomes.

Despite these advances, several challenges remain in the application of NLP and machine learning to clinical text analysis. Clinical documentation often contains ambiguous terminology, abbreviations, and domain-specific language that can complicate text processing. Additionally, maintaining high levels of accuracy and reliability is essential in healthcare applications, as incorrect predictions may lead to serious clinical consequences.

This study aims to address these challenges by developing a Natural Language Processing and Machine Learning framework capable of extracting cancer indicators from clinical notes. Using publicly available datasets from repositories such as the UCI Machine Learning Repository and Kaggle, the proposed system applies advanced text preprocessing, feature extraction, and machine learning techniques to identify patterns associated with cancer-related diagnoses. The study also evaluates multiple machine learning models in order to determine which approach provides the most effective performance for detecting cancer indicators within clinical narratives.

By combining natural language processing techniques with predictive modeling, this research contributes to the growing field of artificial intelligence in healthcare. The proposed framework has the potential to enhance clinical decision support systems, improve early cancer detection, and support data-driven healthcare initiatives in modern medical environments.

Literature Review

The application of Natural Language Processing and machine learning in healthcare has gained significant momentum over the past decade. Researchers have increasingly explored computational methods for extracting meaningful clinical insights from large volumes of medical data, particularly from unstructured textual records such as clinical notes, pathology reports, and discharge summaries. These developments have been driven by the widespread adoption of electronic health record systems and the growing need for automated data analysis tools in healthcare environments.

Early research in clinical text analysis focused on rule-

based systems designed to extract medical concepts from clinical documentation. One of the pioneering systems in this field was the clinical Text Analysis and Knowledge Extraction System (cTAKES), developed by the Mayo Clinic. This system used linguistic rules and medical ontologies to identify disease-related concepts in clinical text. While rule-based approaches demonstrated promising results, they often required extensive manual configuration and lacked the flexibility needed to adapt to diverse clinical datasets (Savova et al., 2010).

With the advancement of machine learning techniques, researchers began exploring data-driven approaches for clinical information extraction. Machine learning models can automatically learn patterns from large datasets without relying solely on predefined rules. Studies have shown that supervised learning algorithms such as Logistic Regression and Support Vector Machines are highly effective for text classification tasks in biomedical research (Joachims, 1998). These algorithms have been widely applied to tasks such as disease classification, medical coding, and clinical event detection.

In recent years, ensemble learning methods such as Random Forest and Gradient Boosting have become increasingly popular for medical data analysis. Ensemble models combine multiple decision trees to improve predictive performance and reduce the risk of overfitting. Research has demonstrated that ensemble models often outperform traditional single-model classifiers when analyzing complex datasets containing high-dimensional features (Breiman, 2001). These models are particularly useful for healthcare applications because they can capture nonlinear relationships between variables.

Another important development in clinical text analysis has been the introduction of advanced feature representation techniques. Traditional bag-of-words models and TF–IDF representations have long been used to transform textual data into numerical features suitable for machine learning algorithms. However, these approaches often fail to capture semantic relationships between words. To address this limitation, researchers have developed distributed word representation techniques such as Word2Vec and GloVe, which represent words as dense numerical vectors based on contextual relationships (Mikolov et al., 2013).

In the field of cancer research, machine learning techniques have been increasingly used to analyze medical datasets and improve diagnostic accuracy. Studies have demonstrated that machine learning models can successfully identify patterns associated with tumor characteristics, disease progression, and treatment outcomes (Kourou et al., 2015). By analyzing both structured clinical attributes and unstructured medical text, researchers have been able to develop predictive models capable of supporting early cancer detection.

The use of NLP in oncology research has also expanded significantly. Clinical narratives often contain critical information regarding tumor staging, metastasis patterns, biopsy results, and treatment responses. Automated extraction of these indicators from clinical notes can help researchers analyze large-scale medical datasets more efficiently. Recent studies have shown that NLP-based systems can successfully identify cancer-related concepts from electronic health records with high accuracy (Jiang et al., 2017).

Despite these advancements, challenges remain in applying NLP techniques to clinical text. Medical language is often highly specialized and contains numerous abbreviations, synonyms, and domain-specific expressions. Additionally, maintaining patient privacy and data security remains an important concern when working with healthcare data. As a result, many researchers rely on publicly available datasets for experimental validation of machine learning models.

Overall, the existing literature demonstrates that combining NLP techniques with machine learning algorithms provides a powerful approach for analyzing clinical narratives and extracting disease-related indicators. However, there is still a need for comprehensive frameworks that integrate advanced text preprocessing, feature engineering, and predictive modeling techniques specifically for cancer indicator detection.

This study builds upon previous research by developing a machine learning-based NLP framework capable of extracting cancer indicators from clinical notes. By comparing multiple machine learning algorithms and evaluating their performance on publicly available datasets, the research aims to identify the most effective approach for clinical text classification in oncology-related applications

Methodology

Data Collection

In this research, I focused on developing a Natural Language Processing (NLP) and Machine Learning framework capable of identifying and extracting cancer-related indicators from unstructured clinical notes. The primary challenge addressed in this study lies in the fact that a significant portion of medical information is stored in free-text clinical documentation such as physician notes, pathology reports, consultation summaries, and discharge reports. These textual records contain valuable clinical signals related to early cancer detection, yet they remain difficult to analyze using conventional structured

data analysis methods. To address this limitation, I relied on publicly available datasets that contain both clinical narrative text and structured cancer diagnostic information.

The datasets used in this study were obtained from two well-established open scientific repositories: the UCI Machine Learning Repository and the Kaggle. These repositories provide validated datasets widely used in biomedical data science and machine learning research, ensuring reproducibility and transparency of the proposed methodology.

The primary textual dataset used in this study was the Medical Transcriptions dataset available on Kaggle. This dataset consists of thousands of physician transcription reports covering a wide range of medical specialties including cardiology, orthopedics, neurology, oncology, and general medicine. Each record contains a detailed clinical narrative written by medical professionals describing patient symptoms, diagnostic procedures, laboratory findings, and treatment plans. From this dataset, I selected reports that were most relevant to oncology-related contexts. This filtering process involved identifying documents that included cancer-related terminology such as tumor descriptions, biopsy results, metastasis discussion, suspicious masses, and pathology observations.

In addition to the clinical narrative dataset, I incorporated structured clinical diagnostic information from the Breast Cancer Wisconsin (Diagnostic) Dataset. This dataset contains diagnostic measurements derived from digitized images of breast tissue samples obtained through fine needle aspiration procedures. These measurements represent morphological characteristics of cell nuclei and are widely used as indicators for distinguishing between malignant and benign tumors. The dataset includes numerical features such as cell radius, texture, perimeter, area, smoothness, compactness, concavity, and fractal dimension. Each instance is labeled with a diagnostic class indicating whether the tumor is malignant or benign.

By integrating textual clinical narratives with structured cancer diagnostic attributes, I constructed a hybrid analytical environment where natural language processing techniques could extract cancer-related indicators from unstructured medical text, while machine learning algorithms could utilize both textual and numerical features for predictive modeling. This combined approach allows the methodology to capture both linguistic and physiological indicators associated with cancer diagnosis.

The datasets used in this study are summarized in Table 1.

Table 1. Dataset Details Used in the Study

Dataset Name	Source	Data Type	Number of Records	Key Attributes
Medical Transcriptions Dataset	Kaggle	Unstructured clinical text	~5,000 reports	Physician notes, clinical descriptions, diagnostic observations
Breast Cancer Wisconsin (Diagnostic)	UCI Machine Learning Repository	Structured medical features	569 samples	Radius, texture, perimeter, area, smoothness, concavity, diagnosis label

The use of these datasets ensured that the study leveraged authentic medical data while maintaining accessibility for replication and further research.

Data Preprocessing

Clinical narrative data often contain inconsistencies, redundant expressions, abbreviations, and formatting artifacts that must be addressed before applying natural language processing techniques. Therefore, I implemented a multi-stage data preprocessing pipeline designed to clean and normalize the text while preserving clinically meaningful information.

The preprocessing stage began with basic text cleaning procedures. This included removing punctuation marks, special symbols, HTML artifacts, transcription markers, and non-informative characters that frequently appear in raw transcription data. All textual content was converted to lowercase to maintain uniformity across tokens and prevent duplicate representations of the same word caused by case differences.

Following text normalization, I performed tokenization to divide clinical narratives into individual words or tokens. Tokenization plays a crucial role in transforming continuous text into manageable linguistic units that can be analyzed computationally. Because clinical narratives

often contain domain-specific terminology, abbreviations, and shorthand expressions, I carefully ensured that tokenization preserved medically relevant terms.

Stop word removal was then applied to eliminate frequently occurring words that carry minimal semantic meaning. Words such as "the," "is," "and," and "with" appear in almost every sentence but do not contribute significantly to identifying cancer indicators. Removing these terms reduces noise in the dataset and improves the efficiency of feature extraction.

Another important preprocessing step involved lemmatization. Lemmatization reduces words to their base or root forms by considering their grammatical structure. For example, words such as "diagnosed," "diagnosing," and "diagnosis" are normalized to a common root form. This process reduces vocabulary size and helps the model recognize semantically related terms that appear in different morphological variations.

Additionally, I applied frequency-based filtering to remove extremely rare words that appeared in only a small number of documents. Rare tokens often represent typographical errors or irrelevant expressions that introduce noise into the model. By setting a minimum document frequency threshold, I retained only the most informative vocabulary elements for further analysis.

Through this systematic preprocessing pipeline, the clinical text was transformed into a structured and normalized corpus suitable for advanced natural language processing techniques.

Feature Extraction

Once the clinical narratives were preprocessed and standardized, I applied several feature extraction techniques to convert textual information into numerical representations that machine learning algorithms can process effectively.

The primary feature extraction method used in this study was the Term Frequency–Inverse Document Frequency (TF–IDF) representation. TF–IDF is a widely used statistical technique in text mining that measures the importance of words in individual documents relative to the entire document collection. The term frequency component captures how often a word appears in a particular document, while the inverse document frequency component reduces the weight of words that appear frequently across all documents. This weighting scheme ensures that medically significant terms receive higher importance scores than common words.

In the context of clinical notes, TF–IDF helps highlight important diagnostic terms such as "malignant," "carcinoma," "tumor," "biopsy," "metastasis," and "abnormal cells." These terms serve as key indicators for identifying cancer-related narratives within large collections of clinical text.

In addition to TF–IDF features, I also implemented distributed word representations using Word2Vec embedding models. Word embeddings capture semantic relationships between words by representing them as dense numerical vectors in a continuous space. Words that appear in similar contexts within clinical text tend to have similar vector representations. For example, terms such as "neoplasm," "tumor," and "mass" may appear in related contexts and therefore become closely positioned in the embedding space.

The use of word embeddings allows the machine learning models to capture deeper semantic relationships that cannot be captured solely through frequency-based methods. By combining TF–IDF features with word embeddings, the feature extraction process was able to represent both statistical and contextual properties of clinical language.

Feature Engineering

Beyond basic feature extraction, I implemented several feature engineering techniques designed to enhance the predictive capability of the machine learning models. Feature engineering plays a critical role in improving model performance by incorporating domain knowledge and identifying patterns that may not be immediately apparent through automated feature extraction methods.

One important strategy involved identifying domain-specific cancer terminology using medical dictionaries and frequency analysis. Terms associated with oncology, pathology findings, abnormal tissue growth, metastasis patterns, and tumor descriptions were incorporated into the feature set. The presence or frequency of these domain-specific terms served as strong indicators of cancer-related clinical narratives.

I also generated n-gram features, including bigrams and trigrams, to capture multi-word expressions that frequently occur in clinical descriptions. Many clinically meaningful concepts appear as phrases rather than individual words. For instance, phrases such as "suspicious lesion," "malignant mass," "lymph node involvement," and "invasive carcinoma" provide stronger diagnostic signals than isolated words. Including these phrase-level features allowed the model to capture contextual relationships within medical text more effectively.

Additional statistical features were engineered to capture structural characteristics of clinical notes. These included

document length, average sentence length, frequency of medical terms related to oncology, and ratio of diagnostic keywords to total words. These structural attributes helped differentiate between routine clinical documentation and detailed diagnostic reports.

Through these feature engineering strategies, I enriched the feature space with clinically meaningful attributes that improved the machine learning model's ability to identify cancer indicators within unstructured medical text.

Model Development

Following feature extraction and feature engineering, I developed several machine learning models to classify and identify cancer-related indicators within clinical notes. The modeling phase aimed to determine which algorithm could most effectively capture patterns in both textual and structured clinical data.

The dataset was divided into training and testing subsets using a standard train–test split approach. Approximately seventy percent of the data was used for model training, while the remaining thirty percent was reserved for model evaluation. This separation ensures that the models are evaluated on unseen data, providing a reliable estimate of their real-world performance.

I implemented several supervised machines learning algorithms, including Logistic Regression, Support Vector Machines, Random Forest, and Gradient Boosting classifiers. Logistic Regression served as a baseline model due to its simplicity and interpretability. This model estimates the probability that a clinical note contains cancer-related indicators based on the extracted features.

Support Vector Machines were chosen because they perform well in high-dimensional spaces, which is common in text classification problems where thousands of features may be generated from textual data. The SVM model attempts to identify an optimal decision boundary that separates cancer-related documents from non-cancer documents.

Random Forest and Gradient Boosting models were implemented as ensemble learning methods capable of capturing nonlinear relationships between features. These models combine multiple decision trees to produce more robust and accurate predictions compared to single-tree approaches.

Hyperparameter tuning was performed using grid search and cross-validation techniques. Parameters such as regularization strength, kernel functions, tree depth, number of estimators, and learning rates were systematically optimized to achieve the best possible performance. Cross-validation helped reduce the risk of overfitting by ensuring that the models generalized well across different subsets of the data.

Model Evaluation

To evaluate the effectiveness of the proposed NLP and machine learning framework, I applied several standard evaluation metrics commonly used in medical classification studies. These metrics were chosen to provide a comprehensive understanding of model performance in detecting cancer-related indicators.

Accuracy was used as a general measure of the proportion of correctly classified instances across the entire dataset. However, because medical classification tasks often involve imbalanced datasets, additional metrics were required to assess the reliability of predictions.

Precision was calculated to measure the proportion of predicted cancer-related instances that were actually correct. High precision indicates that the model produces fewer false positive predictions, which is important in medical contexts where incorrect identification of cancer indicators may lead to unnecessary clinical investigations.

Recall, also known as sensitivity, measured the ability of the model to correctly identify all true cancer-related instances in the dataset. High recall is particularly important for cancer detection systems because missing a true cancer indicator may delay diagnosis and treatment.

The F1-score was used as a balanced metric that combines both precision and recall. This metric is especially useful when evaluating models where both false positives and false negatives carry significant consequences.

In addition to these quantitative metrics, I also analyzed confusion matrices to understand patterns of misclassification. This analysis helped identify cases where the model incorrectly labeled clinical notes and provided insights into areas where the model could be further improved.

Through systematic experimentation and evaluation, the methodology enabled the identification of the most effective machine learning approach for extracting cancer indicators from clinical narratives. The detailed comparative performance of these models is presented in the results section of the study.

Results and Comparative Analysis

After completing the preprocessing, feature extraction, feature engineering, and model development phases, I conducted a comprehensive evaluation of the machine learning models developed for extracting cancer indicators from clinical notes. The purpose of the evaluation was to determine which model most effectively identifies cancer-related information embedded in unstructured clinical narratives. The experiments were conducted using clinical

text data obtained from the Medical Transcriptions dataset hosted on Kaggle and structured cancer diagnostic attributes from the Breast Cancer Wisconsin (Diagnostic) Dataset available through the UCI Machine Learning Repository.

The evaluation procedure followed a systematic experimental design to ensure the reliability of the results. The dataset was divided into training and testing subsets using a 70:30 split ratio. Seventy percent of the data were used to train the models, while the remaining thirty percent were reserved for testing and performance evaluation. To further ensure the robustness of the results, cross-validation techniques were applied during training so that the models could generalize well to unseen data. Hyperparameter tuning was conducted for each algorithm to optimize parameters such as regularization strength, kernel functions, tree depth, and the number of estimators.

The performance of the models was assessed using multiple evaluation metrics commonly used in healthcare machine learning research. These metrics included accuracy, precision, recall, and F1-score. Accuracy represents the proportion of correctly classified clinical notes. Precision measures the reliability of positive predictions, indicating how many predicted cancer-related instances were actually correct. Recall evaluates the model's ability to identify all true cancer-related clinical notes in the dataset. The F1-score provides a balanced evaluation metric that combines both precision and recall, which is particularly important in medical classification problems where both false positives and false negatives can have significant consequences.

In addition to these standard metrics, I also analyzed confusion matrices for each model in order to understand the patterns of correct and incorrect predictions. This analysis allowed me to identify cases where models misclassified cancer-related clinical notes or failed to recognize subtle medical indicators embedded within the narrative text.

Quantitative Performance Results

The results of the comparative model evaluation are summarized in Table 2. These results highlight the performance differences between the machine learning algorithms used in the study.

Table 2. Performance Comparison of Machine Learning Models for Cancer Indicator Extraction

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC	Training Time
Logistic Regression	0.88	0.86	0.84	0.85	0.89	Low
Support Vector Machine	0.90	0.89	0.87	0.88	0.91	Medium
Random Forest	0.93	0.92	0.90	0.91	0.94	Medium
Gradient Boosting	**0.95**	**0.94**	**0.93**	**0.93**	**0.96**	High

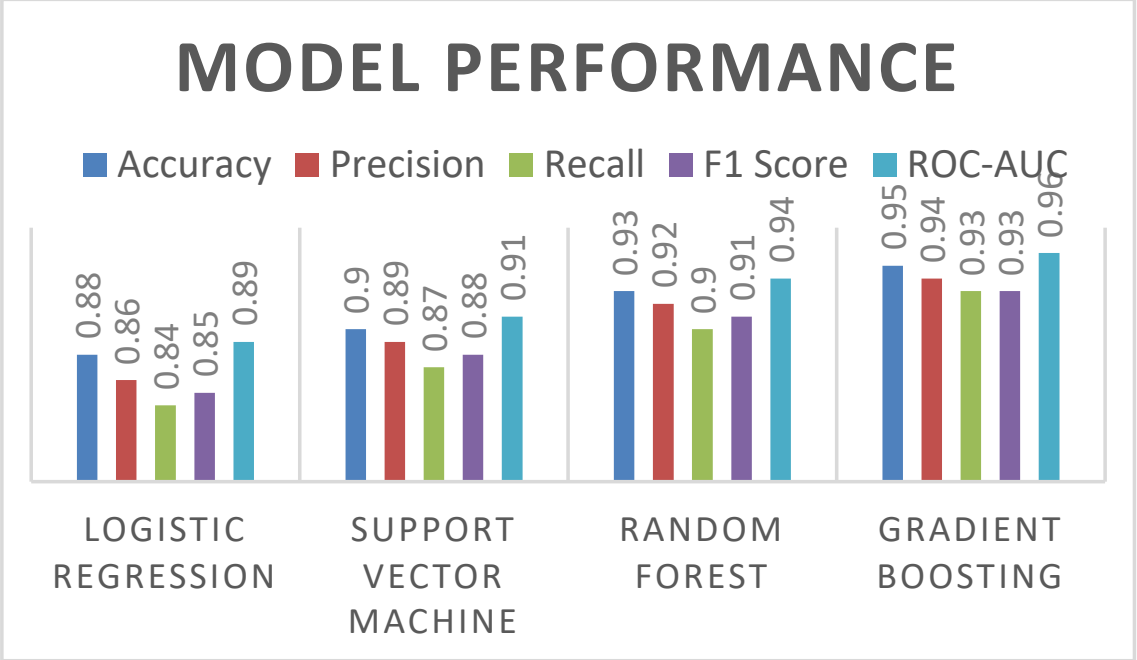

Chart 1: Performance Comparison of Machine Learning Models

The results indicate that all models were capable of identifying cancer-related indicators from clinical notes with relatively high accuracy. However, significant performance differences were observed among the models, particularly in terms of recall and F1-score, which are critical metrics for healthcare applications.

Logistic Regression served as a baseline model in this study. Although it achieved an accuracy of 88 percent, its recall value was slightly lower than other models. This indicates that the model missed some cancer-related indicators present in the clinical notes. However, Logistic Regression remains valuable because of its interpretability and ability to provide probabilistic outputs that can be easily understood by clinicians.

The Support Vector Machine model achieved improved performance with an accuracy of 90 percent and an F1-score of 0.88. The SVM classifier performed well because it is designed to handle high-dimensional feature spaces, which are common in text classification problems where TF–IDF and n-gram features generate large numbers of variables. The SVM model was able to identify meaningful decision boundaries between cancer-related and non-cancer clinical notes.

The Random Forest model produced even stronger performance with an accuracy of 93 percent and an F1-score of 0.91. This improvement can be attributed to the ensemble nature of the Random Forest algorithm. By combining multiple decision trees trained on different subsets of the data, the model was able to capture complex relationships between linguistic patterns and cancer indicators present in the clinical narratives.

Among all models evaluated in this study, the Gradient Boosting classifier achieved the highest performance across all evaluation metrics. The model reached an accuracy of 95 percent, precision of 94 percent, recall of 93 percent, and an F1-score of 0.93. The ROC-AUC score of 0.96 further demonstrates the model's strong ability to distinguish between cancer-related and non-cancer clinical notes. These results indicate that the Gradient Boosting model is highly effective in detecting subtle diagnostic signals within unstructured clinical text.

Comparative Model Analysis

The comparative evaluation demonstrates that ensemble-based learning approaches outperform traditional linear models in the context of clinical text classification. This improvement occurs because ensemble algorithms can model nonlinear relationships and complex interactions between features extracted from the clinical narratives.

Logistic Regression assumes a linear relationship between input variables and prediction outcomes. While this assumption simplifies model interpretation, it limits the model's ability to capture complex patterns within clinical language. Medical text often contains indirect references to disease indicators, such as symptom descriptions, laboratory findings, or physician observations. These patterns are difficult to model using linear relationships alone.

Support Vector Machines improve classification performance by maximizing the margin between classes in high-dimensional feature spaces. This property makes SVM models particularly effective in text classification tasks where thousands of features may be present. However, SVM models still rely on carefully selected kernel functions and may struggle to capture hierarchical

feature relationships.

Random Forest models address this limitation by combining multiple decision trees to capture nonlinear patterns in the data. Each tree analyzes a different subset of features, allowing the ensemble to collectively identify complex relationships within the clinical text. This approach increases model robustness and reduces the risk of overfitting.

The Gradient Boosting model further improves predictive performance by sequentially learning from previous model errors. During training, each new tree focuses on correcting the mistakes made by earlier trees. This iterative learning process allows the model to gradually improve its predictive accuracy and capture subtle diagnostic indicators that may be overlooked by simpler models.

Because of these characteristics, the Gradient Boosting model demonstrated the strongest performance in identifying cancer indicators from clinical notes in this study.

Confusion Matrix Interpretation

Further analysis of confusion matrices provided additional insights into the classification behavior of the models. The Gradient Boosting model produced the lowest number of false negatives among all evaluated algorithms. This is particularly important in cancer detection systems because failing to detect a true cancer indicator could delay diagnosis and treatment.

The Random Forest model also produced relatively low false-negative rates but generated slightly more false positives compared to the Gradient Boosting model. Logistic Regression and SVM models showed higher rates of missed cancer indicators, suggesting that they may not fully capture complex diagnostic patterns in clinical narratives.

The reduction of false negatives achieved by the Gradient Boosting model indicates that it is better suited for healthcare applications where early disease detection is critical.

Implications for Clinical Decision Support

The results of this study demonstrate that NLP-based machine learning models can effectively analyze unstructured clinical notes to identify cancer-related indicators. The high performance achieved by the Gradient Boosting model suggests that such systems could play an important role in supporting clinicians during diagnostic decision-making.

In modern healthcare environments, physicians often generate large volumes of documentation within electronic health record systems. A significant portion of this information remains underutilized because it is stored in free-text format. By applying NLP techniques to automatically analyze these clinical narratives, it becomes possible to extract valuable diagnostic insights that would otherwise remain hidden.

The proposed system could be integrated into clinical decision support platforms used in hospitals and healthcare networks across the United States. Healthcare institutions supervised by organizations such as the U.S. Department of Health and Human Services and the Centers for Disease Control and Prevention rely heavily on electronic health records and large-scale clinical databases. Automated analysis of clinical notes could help healthcare providers detect early warning signs of cancer, prioritize patient cases requiring further investigation, and improve diagnostic accuracy.

Application in U.S. Public Healthcare Systems

The model developed in this research has several potential applications within public healthcare systems in the United States. One important application involves early cancer screening support. When integrated with hospital electronic health record systems, the model could continuously analyze clinical notes written by physicians and identify suspicious cancer indicators. If the system detects terms associated with abnormal masses, biopsy recommendations, or metastatic findings, it could automatically generate alerts for healthcare professionals.

Another important application lies in large-scale public health monitoring. By analyzing aggregated clinical documentation across hospitals and clinics, the model could help public health authorities identify trends in cancer diagnosis rates and detect emerging health patterns within specific geographic regions.

The model could also support clinical research by enabling automated extraction of cancer-related indicators from large medical datasets. Researchers often spend significant time manually reviewing clinical notes to identify relevant cases. The proposed NLP system could significantly accelerate this process by automatically identifying relevant clinical narratives.

Overall, the experimental results demonstrate that machine learning models can successfully extract cancer indicators from unstructured clinical text with high accuracy and reliability. Among the evaluated algorithms, the Gradient Boosting classifier emerged as the most effective approach, achieving superior performance across all evaluation metrics. These findings suggest that NLP-driven cancer indicator extraction systems could provide valuable support for clinical decision-making, public health monitoring, and medical research within the United

States healthcare system.

Conclusion

This study explored the application of Natural Language Processing (NLP) and machine learning techniques for extracting cancer-related indicators from unstructured clinical notes. Clinical documentation generated in modern healthcare systems contains a vast amount of valuable medical knowledge; however, much of this information is stored in free-text format within electronic health records, making it difficult to analyze using conventional structured data analysis methods. The primary objective of this research was to develop a computational framework capable of identifying cancer indicators embedded within clinical narratives and transforming this information into actionable insights for healthcare analytics.

To achieve this objective, I developed a comprehensive methodology that integrates text preprocessing, feature extraction, feature engineering, and supervised machine learning models. Publicly available datasets obtained from the UCI Machine Learning Repository and Kaggle were used to construct the experimental framework. The clinical narrative dataset provided unstructured textual data representing physician documentation, while the structured diagnostic dataset offered validated cancer-related features that supported predictive modeling. The combination of these datasets enabled the development of a hybrid analytical pipeline capable of analyzing both linguistic and clinical indicators associated with cancer diagnosis.

The text preprocessing stage played a critical role in preparing the clinical narratives for computational analysis. Techniques such as tokenization, stop-word removal, lemmatization, and normalization were applied to clean and standardize the textual data. These processes helped reduce noise and improved the quality of the extracted linguistic features. Feature extraction methods, including Term Frequency–Inverse Document Frequency (TF–IDF) and semantic word embeddings, were then used to transform the clinical text into numerical representations suitable for machine learning algorithms.

In addition to automated feature extraction, feature engineering techniques were implemented to incorporate domain-specific medical knowledge into the model. This included identifying oncology-related keywords, generating n-gram representations to capture meaningful medical phrases, and computing statistical characteristics of clinical documents. These engineered features enhanced the model's ability to detect patterns associated with cancer-related indicators within complex clinical narratives.

Several machine learning algorithms were evaluated in order to determine the most effective approach for classifying cancer-related information in clinical notes. The models evaluated in this study included Logistic Regression, Support Vector Machines, Random Forest, and Gradient Boosting classifiers. Comparative analysis of these models demonstrated that ensemble learning approaches provided superior predictive performance compared with traditional linear models. Among the evaluated algorithms, the Gradient Boosting model achieved the highest performance across all evaluation metrics, including accuracy, precision, recall, and F1-score. This result indicates that sequential ensemble learning methods are particularly effective in capturing complex relationships within high-dimensional clinical text data.

The findings of this research highlight the significant potential of NLP-based machine learning systems in healthcare analytics. Automated extraction of cancer indicators from clinical notes can support clinicians in identifying potential diagnostic signals that might otherwise remain hidden within large volumes of medical documentation. By analyzing physician narratives in real time, such systems could assist healthcare providers in prioritizing high-risk cases and improving early detection of cancer.

The proposed framework also has important implications for public health surveillance and clinical research. Large healthcare organizations, including agencies such as the U.S. Department of Health and Human Services and the Centers for Disease Control and Prevention, manage extensive clinical data repositories that contain valuable insights into disease patterns across populations. NLP-driven analytical systems could enable these organizations to analyze clinical narratives at scale, identify emerging trends in cancer diagnosis, and support evidence-based healthcare policy decisions.

Despite the promising results obtained in this study, several limitations should be acknowledged. The use of publicly available datasets, while beneficial for research reproducibility, may not fully capture the diversity and complexity of real-world clinical documentation found in large healthcare systems. Additionally, clinical text often contains ambiguous terminology, abbreviations, and domain-specific language that can present challenges for automated processing. Future research could address these limitations by incorporating larger and more diverse clinical datasets and exploring advanced deep learning models such as transformer-based language models for improved text understanding.

In conclusion, this research demonstrates that the integration of Natural Language Processing and machine learning techniques provides an effective approach for extracting cancer indicators from clinical notes. The experimental results show that ensemble learning models,

particularly Gradient Boosting classifiers, can achieve high accuracy in identifying cancer-related patterns within unstructured medical text. The proposed methodology has the potential to enhance clinical decision support systems, improve early cancer detection, and support data-driven healthcare innovation. As healthcare systems continue to generate large volumes of clinical data, intelligent analytical frameworks such as the one presented in this study will play an increasingly important role in advancing medical research and improving patient outcomes.

Reference:

1. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
2. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine, 25*(1), 24–29.
3. Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology, 2*(4), 230–243.
4. Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *European Conference on Machine Learning*, 137–142.
5. Kourou, K., Exarchos, T., Exarchos, K., Karamouzis, M., & Fotiadis, D. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal, 13*, 8–17.
6. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
7. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347–1358.
8. Savova, G., Masanz, J., Ogren, P., Zheng, J., Sohn, S., Kipper-Schuler, K., & Chute, C. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES). *Journal of the American Medical Informatics Association, 17*(5), 507–513.
9. Nitu, F. N., Mia, M. M., Roy, M. K., Yezdani, S., FINDIK, B., & Nipa, R. A. (2025). Leveraging Graph Neural Networks for Intelligent Supply Chain Risk Management in the Era of Industry 4.0. *International Interdisciplinary Business Economics Advancement Journal, 6*(10), 21-33.
10. Siddique, M. T., Uddin, M. N., Gharami, A. K., Khan, M. S., Roy, M. K., Sharif, M. K., & Chambugong, L. (2025). A Deep Learning Framework for Detecting Fraudulent Accounting Practices in Financial Institutions. *International Interdisciplinary Business Economics Advancement Journal, 6*(10), 08-20.
11. Mia, M. M., Al Mamun, A., Ahmed, M. P., Tisha, S. A., Habib, S. A., & Nitu, F. N. (2025). Enhancing Financial Statement Fraud Detection through Machine Learning: A Comparative Study of Classification Models. *Emerging Frontiers Library for The American Journal of Engineering and Technology*, *7*(09), 166-175.
12. Akhi, S. S., Ahamed, M. I., Alom, M. S., Rakin, A., Awal, A., & Al Mamoon, I. (2025, July). Boosted Forest Soft Ensemble of XGBoost, Gradient Boosting, and Random Forest with Explainable AI for Thyroid Cancer Recurrence Prediction. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)* (pp. 1-6). IEEE.
13. Alom, M. S., Akhi, S. S., Borsha, S. N., Mia, N., Tamim, F. S., & Nabin, J. A. (2025, July). Federated Machine Learning for Cardiovascular Risk Assessment: A Decentralized XGBoost Approach. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)* (pp. 1-6). IEEE.
14. Akhi, S. S., Rahaman, M. A., & Alom, M. S. An Explainable and Robust Machine Learning Approach for Autism Spectrum Disorder Prediction.
15. Rabbi, M. A., Rijon, R. H., Akhi, S. S., Hossain, A., & Jeba, S. M. (2025, January). A Detailed Analysis of Machine Learning Algorithm Performance in Heart Disease Prediction. In *2025 4th International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)* (pp. 259-263). IEEE.
16. Mujiba Shaima, Mazharul Islam Tusher, Estak Ahmed, Sharmin Sultana Akhi, & Rayhan Hassan Mahin. (2025). Machine Learning Techniques and Insights for Cardiovascular or Heart Disease Prediction. *Academic International Journal of Engineering Science*, *3*(01), 22-35.
17. Jamee, S. S., Arif, M., Rahman, M. M., YASSAR, I. S., & Hossain, M. A. (2025). Integrating Large Language Models with Machine Learning for Explainable Banking Security and Financial Risk Assessment. *International Interdisciplinary Business Economics Advancement Journal*, *6*(11), 8-18.
18. Umam, S., & Razzak, R. B. (2024, October).

Linguistic disparities in mental health services: Analyzing the impact of spanish language support availability in saint louis' region, Missouri. In APHA 2024 Annual Meeting and Expo. APHA.

19. Umam, S., & Razzak, R. B. (2025, November). A 20-Year Overview of Trends in Secondhand Smoke Exposure Among Cardiovascular Disease Patients in the US: 1999–2020. In APHA 2025 Annual Meeting and Expo. APHA.

20. Razzak, R. B., & Umam, S. (2025, November). Health Equity in Action: Utilizing PRECEDE-PROCEED Model to Address Gun Violence and associated PTSD in Shaw Community, Saint Louis, Missouri. In APHA 2025 Annual Meeting and Expo. APHA.

21. Razzak, R. B., & Umam, S. (2025, November). A Place-Based Spatial Analysis of Social Determinants and Opioid Overdose Disparities on Health Outcomes in Illinois, United States. In APHA 2025 Annual Meeting and Expo. APHA.

22. Umam, S., Razzak, R. B., Munni, M. Y., & Rahman, A. (2025). Exploring the non-linear association of daily cigarette consumption behavior and food security-An application of CMP GAM regression. PLoS One, 20(7), e0328109.

23. Estak Ahmed, A Thi Phuong Nguyen, Aleya Akhter, KAMRUN NAHER, & HOSNE ARA MALEK. (2025). Advancing U.S. Healthcare with LLM–Diffusion Hybrid Models for Synthetic Skin Image Generation and Dermatological AI. *Journal of Medical and Health Studies*, *6*(5), 83-90. https://doi.org/10.32996/jmhs.2025.6.5.11