Natural Language Processing (NLP) in the Extraction of Clinical Information from Electronic Health Records (EHRs) for Cancer Prognosis

Priyabrata Thatoi[1*], Rohit Choudhary[2], Ashish Shiwlani[3], Hamza Ahmed Qureshi[4], Sooraj Kumar[5]

[1] Oklahoma State University

[2] University of Texas at Dallas

[3] Illinois Institute of technology, Chicago

[4] Mercer University

[5] DePaul University, Chicago

Abstracts: NLP has become an important tool in healthcare, particularly in extracting clinical information from EHRs in order to help enhance cancer prognosis. EHRs store vast amounts of structured and unstructured data, offering tremendous potential for improvement in patient outcomes through the delivery of critical insights into the conditions of patients, their responses to treatment, and possible prognostic outcomes. Nevertheless, meaningful information extracted from these huge amounts of unstructured data, like clinical notes, is still hard to gain. The current review thus shows the developments in NLP techniques that aim to extract and analyze clinical data from EHRs, focusing on cancer prognosis, and also showcases some progress in NLP over the last decade, including various methods like named entity recognition, sentiment analysis, and text classification. Some of the limitations and challenges of current models elaborated on in the paper concern variability in clinical language and high-quality annotated data. Finally, it proposes further improvements and future directions for NLP-based approaches toward more accurate, more individualized cancer prognosis and therefore highlights further research and development as needed in this area of rapid growth.

Keywords: Natural Language Processing, NLP, Electronic Health Records, EHRs, Cancer Prognosis, Clinical Information Extraction, Entity Recognition, NER, Text Classification, Sentiment Analysis, Personalized Medicine

1. INTRODUCTION

NLP, short for natural language processing [1], has revolutionized the field of healthcare; especially in relation to extracting clinical data from Electronic Health Records (EHRs), data sets may play a central role in revolutionizing cancer prognosis by generating information that can substantially contribute to better patient care and improved clinic operations [2]. Health care systems all over the world are using electronic health records (EHR) in an increasing manner [3], wherein a large source of clinical data has been generated with and through analysis, it can provide crucial information related to diagnosis and management for various cancers. The statistics relevant to NLP in EHRs for Cancer Prognosis are shown in Table 1.

Table 1. Key Statistics Relevant to NLP in EHRs for Cancer Prognosis

Statistic	Value
Annual global cancer cases (all types)	19.3 million (2020)
Annual cancer deaths worldwide	10 million (2020)
Percentage of clinical data in EHRs that is unstructured	~80%
Estimated time saved by NLP in data extraction	50-70%
Accuracy of NLP algorithms in extracting clinical data	>90%
Percentage of healthcare facilities using EHRs	89% (U.S. hospitals, 2021)
Growth rate of EHR adoption worldwide (2016-2020)	16% annually

Both structured and unstructured data [4]include patient demographics, clinical notes, diagnostic codes, prescription lists, and test results that make up EHRs. The major problem is the unstructured data, mostly in the form of free-text clinical notes [5], [6], although the structured data is relatively easy to extract and analyze. NLP [7]is very useful here. Enormous amounts of unstructured textual data can be processed and analyzed by NLP techniques which may then be used to extract appropriate information that would be useful in making clinical decisions and improving prognosis for cancer patients.

A patient's status, reaction to treatment, and likely course of events must be assessed sensitively and on time so that a prognosis of cancer can be made. Extracting such data from EHRs using standard techniques is erroneous and labor-intensive. By automating the process of extraction, NLP plays the role of a savior in ensuring that important information is collected accurately and efficiently—thus reducing the chances of overlooking some valuable pieces that would otherwise impact patient care on top of saving time.

The development of new technology in natural language processing (NLP) has resulted in the production of intricate algorithms and models capable of effectively capturing and understanding clinical language. Named entity recognition (NER)[8], [9], [10], sentiment analysis [11], [12]and text classification [13], [14], [15] are among the different techniques used to identify specific clinical entities [16], assess patients' sentiments and classify clinical data. It is crucial for predicting cancer patient prognosis because timely and accurate data acquisition can enhance treatment plans as well as predict patient outcomes.

NLP also enables the discovery of patterns and trends in clinical data which are not visible through manual review. For instance, NLP can find links between certain clinical factors and cancer outcomes by mining large clinical note repositories. This has the potential to contribute to more accurate prognosis models and identification of new prognostic biomarkers. Therefore, NLP is instrumental in advancing both the field of cancer research [17], [18]and improving clinical information extraction processes.

In summary, NLP provides a powerful tool for extracting clinical data from EHRs with important implications for canc er. Natural language processing (NLP) has the potential to enhanceclinical decision making [19]through process ext raction and discovery from unrelated data, personalized medicine, and advance cancer research. However, despite the achievements in the field and its integration with clinical practice, there is still an urgent need for a good evaluati on of the application of NLP in EHRs for cancer diagnosis.

Although many studies demonstrate the potential of NLP, the field is rapidly changing and many methods, tools, and applications remain unexplored or unsuitable and require scrutiny. A comprehensive evaluation will aid in synthesizing current information, pointing out research gaps, and addressing issues including data variability, model correctness, and integration into clinical workflows. A review can also point out excellent practices and direct future studies, ensuring that NLP keeps evolving in ways that will improve patient care and cancer prognosis. This review will help to advance the field and increase the efficacy of NLP-based cancer prognosis by identifying areas for improvement and the current state of the field.

This review answers the following questions:

RQ1: Which NLP techniques are most commonly used for extracting clinical information from EHRs?

RQ2: What are the common approaches for preprocessing structured and unstructured data in EHRs before applying NLP models?

RQ3: Which clinical features extracted from EHRs are most informative for predicting cancer prognosis?

RQ4: How are textual features extracted from unstructured clinical notes utilized in cancer prognosis?

RQ5: Which public EHR datasets are available, and which NLP algorithms have proven to be most effective for extracting clinical information?

RQ6: How can structured and unstructured data in EHRs be combined using NLP to improve cancer prognosis predictions?

A. Topology of Review

This review is structured as follows: Section 1 provides an introduction to the topic, setting the stage for the review. The review process used to compile and evaluate pertinent studies is described in Section 2. Examining clinical data from EHRs and distinguishing between structured and unstructured data, Section 3. Sentiment analysis [20], [21], text classification [22], [23], and Named Entity Recognition (NER) [24], [25] are just a few of the NLP techniques covered in Section 4. The use of NLP in cancer prognosis is covered in Section 5, with particular attention paid to outcome prediction, therapy response tracking, and patient condition analysis. The function of NLP in personalized medicine [26], [27] is highlighted in Section 6, with a focus on thorough patient perspectives, customized treatment regimens, and improved decision-making. The difficulties in this field are covered in Section 7, including issues with privacy and security, data quality [28] and annotation, and variability in clinical language. The review's discussion and conclusion are finally presented in Section 8. The visual representation of topological view is shown in Figure 1.

Figure 1. *Organization of the Review*

LITERATURE REVIEW

According to Hossain et al., the study culled 127 papers for full-text evaluation after screening 261 articles from 11 databases. These studies were divided into seven different NLP-related EHR priority areas, including clinical entity recognition and medical note classification. Among the chosen papers, EHRs were found to be the most often utilized data type, with a notable focus on unstructured datasets. This demonstrates the difficulties and possibilities involved in handling such data. The International categorization of Diseases, Ninth Revision (ICD-9) categorization, clinical note analysis, and named entity recognition (NER) for clinical descriptions—particularly in psychiatric disorders—were among the most common uses of NLP techniques. The review discovered that the ML models used in the research were not sufficiently evaluated, pointing to a gap in the assessment of their dependability and efficacy. Data imbalance was one of the major issues found, and it matters in the context of NLP applications in healthcare [29].

Gholipour et al. provided the systematic review that first found 6708 papers on natural language processing (NLP)-based cancer idea extraction. 2503 articles were left for additional review after duplicates were eliminated, during which time titles and abstracts were examined. After 2436 papers were excluded based on inclusion and exclusion criteria, 67 studies were found to be relevant for full-text review. In the end, 17 publications were chosen for in-depth data extraction. The review highlighted that UMLS and SNOMED-CT were the most commonly used terminologies in the field of NLP for extracting cancer concepts. Rule-based methods were the most frequently used techniques among the reviewed NLP algorithms. The study also noted that many included studies did not report the content coverage of the applied terminological systems, suggesting a need for future research to address this gap [30].

Li et al. revealed the review which categorizes 23 studies into four main subsections based on cancer types: breast cancer, colorectal cancer, lung cancer, and other cancers (including liver, prostate, pancreatic, and brain tumors). The studies are organized in tables that display information such as the year of publication, text source, cancer type, purpose, algorithm used, evaluation metrics, and dataset. This arrangement allows for easy comparison and understanding of the advancements in NLP applications over time. The paper highlights the performance of the best-performing models in the studies reviewed. It also provides explanations for models that were not previously introduced, enhancing the reader's understanding of the methodologies employed. The review covers a range of natural language processing (NLP) methods used with electronic medical records (EMR) and health records (EHR), such as rule-based, machine learning-based, and deep learning-based methods. These methods are essential for obtaining important data that helps oncologists make defensible decisions. The paper identifies current limitations in NLP applications that hinder their effectiveness in clinical practice. It suggests potential future research directions to address these challenges and improve the integration of NLP in oncology. The meticulous process of choosing pertinent studies is described in the methodology section. It includes inclusion and exclusion criteria, such as the particular role that NLP plays in CAD (removal of duplicates and irrelevant papers). This procedure made sure that the review contained only the most relevant papers [31].

Sangariyavanich et al. provided a total of 17 studies that were recommended for inclusion in the review, with the majority (53%) recently published in 2021 and 2022 (-III). Studies analyzed a variety of medical records, including pathology reports (7 studies), electronic medical records (10 studies), and other medical records (12 studies). Studies covered different cancer types, with breast cancer being the most common (13 studies), followed by breast cancer (4 studies), and lung cancer (3 studies). Outcomes of interest included metastasis/distal recurrence (10 studies), local recurrence (2 studies), and single recurrence (6 studies). In terms of model performance, deep learning (DL) outperforms other methods in all metrics. The average AUROC score of the DL model was 0.98, recall was 0.88, precision was 0.79, and F1 score was 0.76. Convolutional Neural Networks (CNN) are the most popular deep learning algorithms. Two studies have been proven to be outliers; one using the right-justified algorithm achieved 0.97-1 yield and 0.99 specificity. The review identified three main text representation techniques: statistical, context-free, and contextual representations (including BERT). The median F1 scores for these representations were 0.71 for rule-based, 0.43 for machine learning, and 0.76 for deep-learning approaches. The review acknowledged limitations, such as the exclusion of studies from gray literature and the lack of meta-analysis due to variable reporting of model performance metrics. It emphasized the need for standardized medical terminology and document templates to improve model performance and reduce uncertainty [32].

Nunez et al. demonstrated the NLP models developed in this study achieved impressive performance metrics. The models were evaluated on a never-before-seen internal holdout set, achieving accuracy, balanced accuracy (BAC), and area under the curve (AUC) scores exceeding 0.800. The best models even achieved an AUC greater than 0.900, indicating a strong ability to distinguish between patients who would survive and those who would not over various time frames (6, 36, and 60 months). The performance of this model is comparable to or better than previous studies on predicting cancer survival. In particular, this model uses more general and readily available data by focusing on first-hand conversational data rather than relying on established data. This study demonstrates the generality of the sample because they studied a large population of patients across cancer sites rather than focusing on a single cancer type. This method allows the findings to be applied generally. The model uses definitions of letters in discussion articles consistent with known mortality risk factors. For example, oncologists' emphasis on hospital care was associated with shorter survival, supporting the model's prediction. While the results are encouraging, the study also acknowledges limitations that could impact accuracy in existing patients, such as lack of external validation and the possibility of changing treatment over time [33]**.**

Zhou et al. observed that the CancerBERT model generally outperforms the other models (CRF and BiLSTM-CRF). It shows a better ability to learn the complexities of the target phenotypes and adapt to changes encountered in the clinical literature. The evaluation was performed by examining corpora from two institutions, the University of Minnesota (UMN) and the Mayo Clinic (MC). The models were trained and tested on this data and registered according to the same instructions. This study evaluates the performance of the model against different ECR groups to show how well the model performs in different locations. The CancerBERT model trained in one institution and

fine-tuned in another was successfully compared with models generated from local data. For example, the micro-F1 score of the transition model was 0.925, while the micro-F1 score of the local working model was 0.932, indicating its stability and adaptability. A total of 200 and 161 medical records were collected from UMN and MC, respectively. The high similarity of the organizational purpose of the two institutions helps the model to be effective. This study suggests plans to collect more data from other hospitals to evaluate the generality of the model, especially regarding the activities and extraction of activities for predicting background diseases. [34].

Huang et al. compared 268 histopathology reports to the gold standard with an accuracy of 61.2% to 99.0%. Accuracy varies with the complexity of the extraction task. Greater than 95% accuracy was assessed for 8 of 11 variables; this showed that the prediction model for these variables was comparable to the gold standard. However, our exchange did not reach the threshold. The researchers applied machine learning (ML) and deep learning (DL)-based policies for each variable. They choose the method that gives the best performance index for each variable. This choice is affected by the different grammar and patterns found in the reports. For some variables, such as lymph nodes, the study used regular expressions to extract numbers from phrases such as “3 good lymph nodes.” This approach works well due to the nature of the report. The rule-based model outperformed the ML/DL model in extracting eight variables, including nuclear location, histological type, and margin criteria. In contrast, the ML/DL model performed well in extracting large tumors. When rule-based and ML/DL approach were combined, the average microscopic accuracy reached 93.3%. This demonstrates the efficiency in extracting medical information from reports. Rule-based methods are increasingly effective in their interpretation, especially due to the grammar of histology data, which helps to obtain similar results. [35].

Fanconi et al. published studies compare various models including language model, language and fused LASSO model, and language and fused BERT model. Results are presented by showing the best performing measures for each type of marker and showing the effectiveness of different methods in predicting ACU risk. The results show that the Structured Health Data (SHD) model outperforms the NLP model. Specifically, the C-statistic of the 1-penalized logistic regression model using SHD was 0.748 (95%-CI: 0.735, 0.762). In comparison, the C-statistic for the same model with language model was 0.730 (95%-CI: 0.717, 0.745) and the C-statistic for the Transformer-based model was 0.702 (95%-CI: 0.688, 0.717). This study used a total of 760 clinical data points obtained from previous studies and required for model training. Linguistic models, especially LASSO and BERT, provide features from clinical data, revealing the potential of inappropriate data in gambling. Preprocessing of clinical data includes removing unique characters, flagging negative content, and filtering most content using the time frequency-extracted data frequency (TF-IDF) algorithm. This meticulous study aims to improve the quality of the data input model. The findings highlight the importance of combining NLP approaches with data modeling to improve risk management for oncology patients. Research highlights that although the SHD model is currently more influential, the NLP approach has great potential to contribute clinical insights. [36].

Laurent et al. proposed model achieved an overall accuracy of 0.88 when applied to the validation of 603 internal documents. The collection includes reports prepared by 49 different electrical engineers, showing a variety of experiences and guidelines. The model was able to classify 76.7% of the reports and performed well in classifying tumor response as growth or failure. When tested on 189 external data from 46 different electrical sources, the standard control accuracy was 0.82. This shows that the validity of the model is beyond the original study area. It has been shown that the pattern is similar to progression and that there is no distribution indicating the activity of different lymph nodes. The reports used in this study have been stripped to protect patient privacy and to ensure compliance with the European General Data Protection Regulation. This includes removing identifying information while preserving the information necessary for analysis. This study focuses on the extraction of the report result required in French publications. This approach allows for a more systematic application of the rules, as results are shown in 98.3% of the analyzed data. Create an oncology dashboard using the vis.js library's timeline tool, which visualizes and organizes data in a dynamic timeline format, improving the usability of the data extracted. [37].

Maghsoudi et al. showed that the study examined how well various demographic groups' Eastern Cooperative Oncology Group (ECOG) performance status reporting was followed. Black patients had an odds ratio of 1.24 (95% CI: 0.74-2.08) with a p-value of 0.40, showing no significant difference in documentation compliance compared to

White patients, according to the odds ratios for various racial and ethnic groups compared to White patients. With a p-value of 0.95 and an odds ratio of 0.98 (95% CI: 0.54-1.78) for Hispanic patients, there did not appear to be a meaningful difference. Asian patients also showed a lack of statistical significance, with an odds ratio of 1.55 (95% CI: 0.74-3.24) and a p-value of 0.24. With a p-value of 0.48 and an odds ratio of 1.18 (95% CI: 0.73-1.91) for all non-White patients, the result that there are no appreciable racial variations in documentation compliance is further supported. The impact of language on documentation compliance was also investigated in this study. Patients who spoke Spanish had an odds ratio of 0.99 (95% CI: 0.58-1.69) and a p-value of 0.99, meaning that there was no discernible variation in compliance between them and English speakers. With a p-value of 0.66 and an odds ratio of 1.10 (95% CI: 0.70-1.74), non-English speakers once more did not demonstrate a significant difference. The lack of significant results was further supported by the odds ratio of 1.07 (95% CI: 0.48-2.37) and p-value of 0.85 for non-Spanish speakers. With a p-value of 0.56, the odds ratio for male patients relative to female patients was 1.12 (95% CI: 0.75-1.68), suggesting that there was no discernible variation in documentation compliance based on gender [38].

METHODOLOGY

The steps that we adopted to conduct this review are as follows:

A. Articles Collection

Several protocols were adhered to for a systematic review of the literature on the use of NLP for extracting clinical information related to determining the prognosis of cancer patients from EHRs. The literature search of peer-reviewed publications was conducted until September 2023. Short papers, reports, editorials, posters, and dissertations were screened out. The investigators followed the recommendations outlined in the PRISMA-P statements. Search terms: Natural Language Processing, NLP, Electronic Health Records, EHRs, cancer prognosis, clinical information extraction, named entity recognition, NER, sentiment analysis, text classification, personalized medicine. The study used databases such as Web of Science, PubMed, Google Scholar, MDPI, Elsevier, and IEEE Xplore. A double-check during the process of searching found a total of 1,320 peer-reviewed publications. Only papers entirely dedicated to NLP techniques for extraction of clinical information, considered specifically relevant for the estimation of cancer prognosis, were included.

i. Search Strategy

Clearly defining inclusion and exclusion criteria is very important in evaluating the validity of the literature review during the selection process. We took the following quality standards, inspired by the relevant research, to guide the selection process. In this work, we included studies focusing on NLP-based extraction of clinical information from EHRs for cancer prognosis. The papers were screened initially by title, then by abstract, and lastly by full text. The following quality criteria were applied that determined the inclusion of the research articles to be used:

a) Focus Area

It focused on research that reviewed the application of NLP techniques to extract clinical information relevant to cancer prognosis, including NER, sentiment analysis, and text classification.

b) Technical Details

Studies that clearly explained NLP techniques, data preprocessing methods, and any specific algorithms used in the analysis.

c) Quantifiable Results

Research that discussed measurable outcomes on accuracy, precision, recall, F1-score, and other relevant metrics concerning the assessment of the effectiveness of NLP in extracting clinical information.

d) Source credibility

To assure a very high level of quality and reliability, this review only referred to peer-reviewed journals and conference papers.

RESEARCH FINDINGS

The results of the research in this review are presented below:

A. Clinical Information from Electronic Health Records

EHRs [39], a comprehensive collection of patient-related information, have become the cornerstone of modern healthcare. Structured and unstructured data are the two main types of medical data managed by EHRs. To use the NLP [39] method to extract useful information, especially on cancer, it is necessary to understand the nature and characteristics of different types of information.

a) *Structured Data*

Electronic health records that contain pre-arranged information in a format that can be easily searched and identified are said to have structured information [40], [41]. Test results, prescriptions, vital signs, diagnostic codes, and demographic information are just a few examples. One type of data storage is a table, where each entry follows a data structure, such as the International Classification of Diseases (ICD) [42]code for the diagnosis. The quality of the design information allows it to be easily used directly in computer models, allowing for rapid access and analysis. The methodology can provide important results, such as tumor size, stage, and biomarker levels, which are essential for developing predictive models in the context of cancer.

b) *Unstructured Data*

In contrast, plain text that does not meet the specified criteria becomes invalid information. This includes patient descriptions, electronic medical records, medical reports, medical records, and discharge summaries. Unstructured data is more difficult to interpret than structured data due to the variety of words, phrases, and concepts. However, it often contains important medical information that was not captured in the design, such as patient descriptions of symptoms, physician observations, and reasons for treatment options. Natural language processing (NLP) is useful in extracting relevant information from raw data to transform irrelevant data into a structure that can be analyzed with other medical information [43], [44].

The tabular representation of types of clinical data in EHRs is shown in Table 2 and Table 3.

Table 2. Types of Clinical Data in EHRs

Data Type	Description	Examples
Structured Data	Data arranged in predefined formats for easy retrieval and analysis	Lab test results, diagnosis codes, medication lists, vital signs, demographic information
Unstructured Data	Free-text data is not organized in a fixed format, requiring advanced techniques for analysis	Patient narratives, clinical notes, radiology reports, pathology reports, discharge summaries

Table 3. Types of Clinical Data in EHRs

Data Type	Example Category	Example Content
Structured Data	Laboratory Test Results	Tumor size: 3 cm, Stage: II, Biomarker levels: HER2 positive
Unstructured Data	Clinical Notes	“The patient experienced significant fatigue and nausea after the last chemotherapy cycle.”

Figure 2 shown below represents the EHR data integration for cancer prognosis of Structured vs. Unstructured Data.

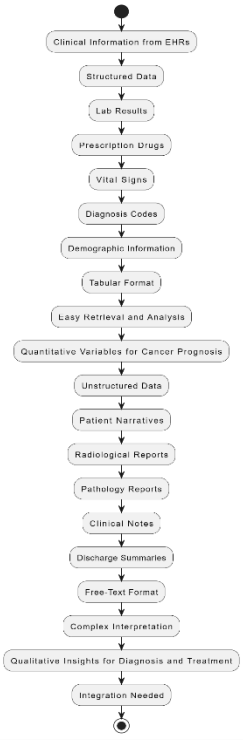


Figure 2. *EHR Data Integration for Cancer Prognosis: Structured vs. Unstructured Data*

B. NLP Techniques

Specifically, the attributes of perceiving and valuing structured and unstructured data are magnificently essential in the application of clinical knowledge [45], [46]provided in EHRs for cancer prediction. Big data complements clinical data by creating context to the data and increase dimension of data that arises from unstructured data while clinical data is structured information used to make decisions. Otherwise, unstructured data of needed clinical records would remain unprocessed, and all necessary data available within the discrete patient record would be used for presenting the client with the probability of developing cancer in the future, as close to the given profile as possible.

a) Named Entity Recognition (NER)

One popular NLP technique is the so called named entity recognition (NER) that aims at finding the spatially delimited text chunks belonging to predefined classes such as disease names, drug names, symptom names, or anatomical terms. NER is useful for identification of particular medical entities from the text of clinical notes in the framework of EHRs. To build the models for the prognosis of cancer, NER [47], [48]helps in the identification of the relevant entities like cancer types, stage, line of treatment and patient history.
In order to ensure the possibility to identify the medical term correctly, NER systems in the healthcare sector often use large and all-embracing medical ontologies and lexica, e.g., the UMLS. To enhance the NER accuracy, there is the use of deep learning as well as machine learning such as Bidirectional Long Short-Term Memory networks (BiLSTM) [49]and Conditional Random Fields (CRF) [50]. The sources of the information used in cancer prognosis [51], [52] are enhanced by NER since it eliminates items irrelevant for prognosis. Enhancing cancer prognosis with Named Entity Recognition is shown in Figure 3.

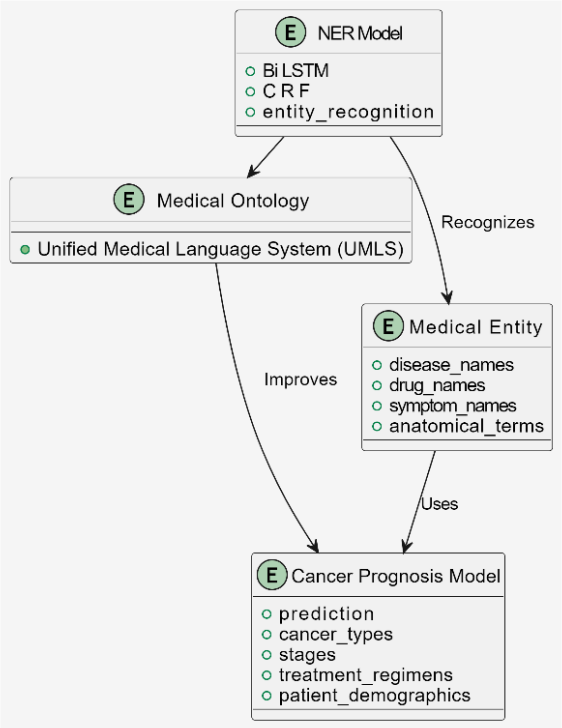


Figure 3. *Enhancing Cancer Prognosis with Named Entity Recognition: Methods, Instruments and Uses*

b) Sentiment Analysis

Sentiment analysis or opinion mining is a subfield of NLP that discovers that feeling, whether favorable, unfavorable, or neutral, conveyed in text. Healthcare application of sentiment analysis [53], [54] can be performed to define the emotional context of clinical notes, patient feedback or physician comments as captured in Electronic Health Records. Cancer prediction by means of SA can provide data about the psychological and the emotional state of a person, which can correlate with the treatment outcomes.

For example positive remarks in the patient's clinical record may indicating a good clinical response to treatment, while negative sentiments may signal issues or symptoms which worsen. There are some deep learning algorithms for sentiment categorization Recurrent Neural Networks (RNN) [55], [56] and for machine learning algorithms [57], [58], Support Vector Machines (SVM) [59], [60]. Other clinical information can be used in combination with sentiment analysis to support healthcare professionals to get a better understanding of the state of a patient. This could bring about unique and more effective treatment procedures or plan for the patient. The visual representation of sentiment analysis in healthcare data is shown in Figure 4.

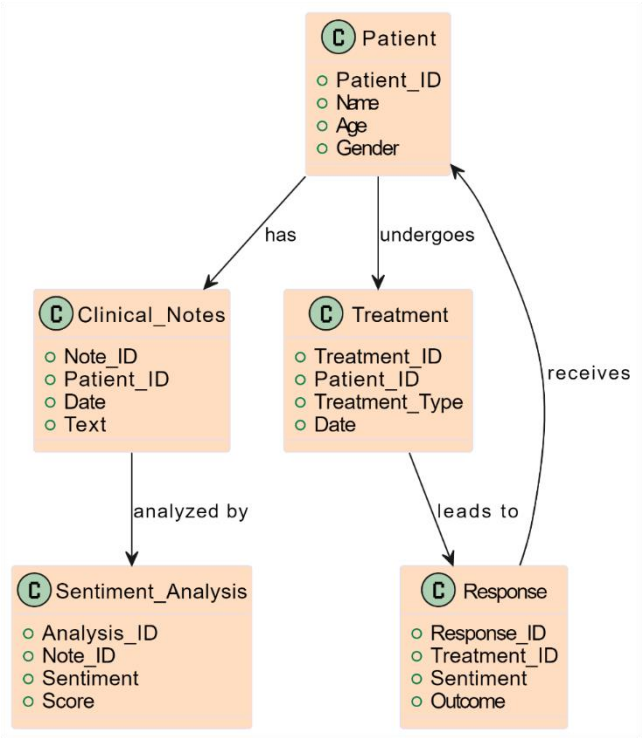
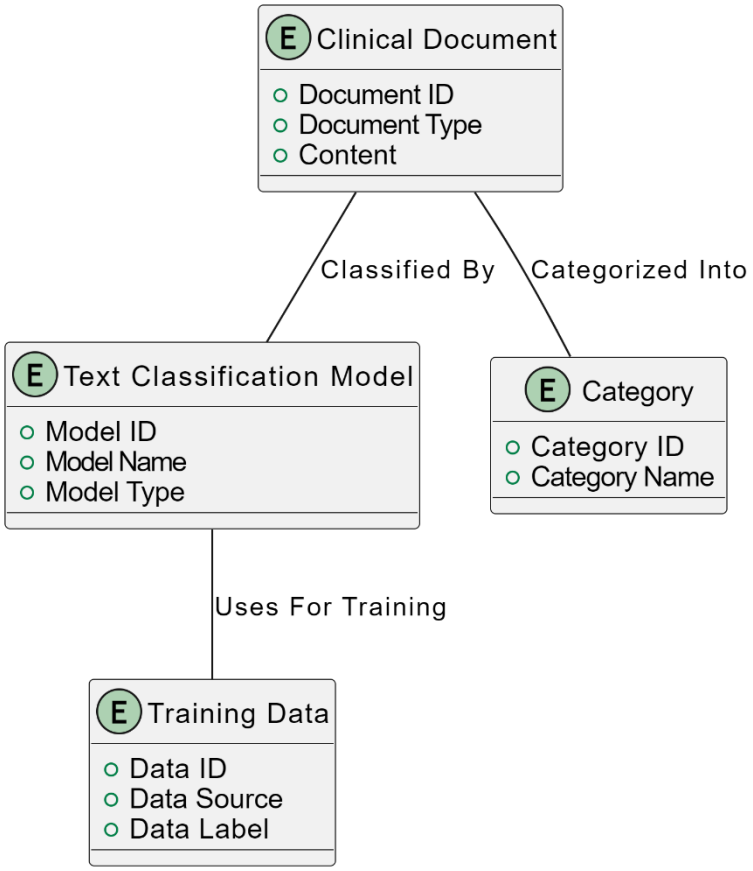


Figure 4. *Entity-Relationship Diagram for Sentiment Analysis in Healthcare Data*

c) Text Classification

Text classification is a procedure of categorizing these texts into pre-specified classes or categories. In EHRs, text classification [61], [62] is applied for other purposes of grouping clinical documents into categories relevant to cancer prediction. Among such records include the discharge summaries, pathology reports and therapy notes. Example of text classification is to identify papers that talk about specific cancer types, therapies, or results. Gathering such an overwhelming amount of unstructured data in EHRs, it has to be sorted and filtered using this method to make the data easier to analyze. The general classification of a text model usually employs deep learning algorithms such as transformers and Convolutional Neural Networks (CNNs) [63], [64]; or supervised learning algorithms such as Naïve Bayes [65], [66] & Decision Trees [67], [68]. Concerning the categorization of fresh, unidentified documents these models are trained on existing labeled data sets so as to identify the connection between the text characteristics and the respective categories. Depiction of text classification in EHRs for Cancer Prognosis is shown in Figure 5.

Figure 5. *ER Diagram for Text Classification in EHRs for Cancer Prognosis*



C. Applications in Cancer Prognosis

a) Patient Condition Analysis

Among the various applications of natural language processing (NLP) in electronic health record (EHRs), perhaps one of the most critical is in the assessment of patient conditions for prognosis of cancer. Imaging descriptions, pathology reports, and clinical notes of the patients are examples of unstructured data that can be present in the EHR. Such data sources are often quite detailed portraying the health status of a patient. Being an unstructured content, the information from these resources has to be converted by NLP algorithms such as text categorization and Named Entity Recognition (NER) to enable clinicians identify relevant clinical entities [69], [70] such as size, stage, and grade of the tumor. With the patient data collected and organized systematically by healthcare professionals, it is impossible not to attain comprehensive knowledge of a patient in formulating individual treatment plans.

Some examples of such comorbidities [71], complications, or other factors that can influence the patient's status and prognosis can also be identified with the assistance of NLP techniques for patient condition evaluation. For instance, other diseases such as diabetes or cardiovascular diseases [72] may influence the way cancer is managed, or the likelihood of the patient to benefit from treatment. Such new risk variables can be identified and measured in EHRs by using natural language processing by clinicians. This enables the advancement of more precise and detailed evaluation of the patient's condition.

b) Treatment Response Monitoring

One of the most significant aspects of the cancer management is the assessment of the treatment outcomes because timely adjustment of the treatment regimen may significantly influence the patients' survival and quality of

life. NLP has significant use in this area since, due to its capabilities, it allows the information related to the treatment in the EHR to be collected and analyzed. Clinical notes typically encompass offerings in relation to dosage of radiation, cycles of chemotherapy [73], surgeries and other related procedures, and outcomes of the procedures. From these remarks, one can use NLP for identifying patterns in the treatments, such as the lessening of the size of the tumor or increase in biomarker's values.

Furthermore, it will be feasible to consider patient-reported outcomes, for instance, side effects or symptoms, documented in EHRs for the analysis using NLP[74], [75]. Healthcare professionals can get understanding of how patient is responding to the therapy in real time by identifying and quantifying these narratives. This makes early detection of side effects or treatment resistance possible and this puts doctors in a position where they can easily change their strategies. Therefore, averting NLP-driven therapy response monitoring may help in promising cancer care models that are personalized and more adaptive.

c) Outcome Prediction

Outcome prediction is perhaps one of the most vital areas where NLP can be used for the assessment of patients with cancer. There are still many broad and historic data about outcomes, diagnosis and treatment, and clinical characteristics of patients in EHRs. Using NLP [76], [77]on these records, researchers and medical professionals can develop quantitative models that predict patient outcomes such as the chances of survival or recurrence of the disease, as well as the chances of success in the particular line of treatment. As will be detailed below, these models may entail a number of different variables such as history of treatment and tumor characteristics and demographic data to generate individualised risk profiles.

D. Personalized Medicine

a) Comprehensive Patient View

One of the nascent forms of patient care known as "personalized medicine" primarily relies on using massive amounts of clinical data stored in EHRs to give a holistic picture of the patient in front of their physicians. This is especially important in the context of cancer prognosis as it most of the time involves inclusion of structured data including images and lab results and unstructured data including patient history, treatment plan and doctor's notes. To this end, this unstructured data must be captured and analyzed and healthcare practitioners must be provided with fast and accurate access to precise patient information and for this natural language processing or NLP is inevitable. By applying NLP to the clinical notes, patient narratives and other free-text entries recorded in the EHRs, healthcare providers can build a richer and layered picture of a patient's condition and come to much better-informed judgments about cancer diagnosis and treatment.

b) Individualized Treatment Plans

The goal of individualized approach is therapeutic plan [78] creation based on individual characteristics of every patient, which is appropriate in oncology. This is so because the disease is very wide and can be manifested in different ways hence when you take two people with similar cancer they can respond differently to the same treatment. NLP makes the identification of such distinctive patient features as genetic mutations, past treatment outcomes, and coexisting diseases, easier owing to the fact that it can pull out the pertinent clinical data from the EHRs. From it, individualized treatment intervention strategies that can help a given patient can be formulated. HCPs can consequently design organ-specific medical care management strategies that are custom designed to each individualized patient's cancer type with the add-on of proteomics and genetics among other advanced technologies that NLP interfaces. That will increase the patient satisfied rate and decrease the unnecessary side effects on patients.

c) *Enhanced Decision-Making*

One of the core components of personalized medicine is clinical decision making and this is benefited by the use of NLP for EHRs. NLP tools provide physicians with means to come up with right decision regarding diagnosis, treatment and prognosis by processing large amounts of clips provider data. Let's take for instance unstructured clinical notes; through the data, the NLP algorithms can identify some specific pattern or even make some inference or prediction that would otherwise be hard to achieve through methods such naked eye analysis. These include informing possible risk factors, the probability of a patient's status, status, and optimal management based on research findings. Further, NLP can help physicians to be aware of the clinical guidelines and new discovery of research findings relevant to the patients hence improving the provision of differentiated care in the management of cancer.

Table 4. Summary of literature review

Author, Year	Target Variable	Input	Architecture	Pre-Processing	Dataset	Outcome	Output	Result
Hossain et al., 2023	NLP Techniques in EHRs for Cancer Prognosis	EHR Data	Various NLP Models	Unstructured Data Handling	127 papers from 11 databases	ICD-9 Classification, NER for Clinical Descriptions	Multi-class	Gap in ML Model Evaluation, Data Imbalance
Gholipour et al., 2023	Extraction of Cancer Concepts	EHR Data	Rule-based NLP	UMLS, SNOMED-CT Terminologies	17 articles selected from 2503	Cancer Concept Extraction using UMLS and SNOMED-CT Terminologies	Rule-based	Frequent Use of Rule-based Methods, Content Coverage Issues
Li et al., 2023	NLP Techniques for Various Cancer Types	EHR and EMR	Rule-based, ML, DL Approaches	Text Categorization by Cancer Type	23 studies categorized into four main cancer types	Information Extraction for Cancer Prognosis	Multi-class	Best Models Noted, Future Research Directions Identified
Sangariyavanich et al., 2023	Cancer Prognosis using Local and Public Data	Clinical Documents	Deep Learning (CNN)	Data from Pathology and Radiology Reports	17 studies with local data and MIMIC-III public data	Recurrence Prediction across Multiple Cancer Types	Multi-class	Deep Learning Outperforms Other Methods, Use of Public Data
Nunez et al., 2023	Cancer Survival Prediction	EHR Data	NLP Models	Data from Initial Consultation	Internal Holdout Set for Validation	Cancer Survival Prediction over Time	Binary	AUC > 0.900, Generalizability of Models
Zhou et al., 2023	Generalizability of NLP Models	Clinical Corpora	CancerBERT, CRF, BiLSTM-CRF	Annotated Clinical Texts	Clinical Corpora from University of Minnesota and Mayo Clinic	Entity Recognition Across Varying Levels of Coverage	Multi-class	CancerBERT Outperforms, Comparable Performance on Transferred Data
Huang et al., 2023	Accuracy in Extracting Clinical Information	Histopathology Reports	Rule-based, ML, DL	Task-Specific Approach	268 Histopathology Reports	Variable Accuracy Depending on Task Complexity	Multi-class	Rule-based Outperforms for Structured Data, ML/DL for Tumor Size
Fanconi et al., 2023	Risk Prediction for Oncology Patients	Clinical Notes	LASSO, BERT Models	Preprocessing Clinical Notes	760 Structured Health Data Points	Risk Prediction Using Unstructured and Structured Data	Binary	SHD Models Outperform NLP Models, Potential for NLP Approaches
Laurent et al., 2023	Tumor Response Classification	Radiology	Rule-based, NLP	De-identified Reports,	Validation Set with	Tumor Progression or	Binary	Accuracy of 0.88, Robust

		Reports		Conclusion Extraction	603 Internal Documents	No Progression Classification		Performance Beyond Training Environment
Maghsoudi et al., 2023	Documentation Compliance Across Demographics	EHR Data	Statistical Analysis	ECOG Performance Status Documentation	Compliance Analysis Across Demographic Groups	Compliance with Documentation Standards	Binary	No Significant Differences in Compliance Across Demographics

Challenges and Limitations

Its application of Natural Language Processing (NLP) at Electronic Health Records (EHR) for extraction of clinical data for cancer prognosis is highly possible. Here are some of the obstacles that need to be overcome to achieve successful enactment and credibility of the field; These difficulties are mostly connected with the specific terminology used at clinics or other research facilities from the point of biology and such problems as the stability and get ability of the information being processed; also, privacy and security concern.

A. Clinical Language Variability

The problem with incorporating NLP to EHR also lies in the fact that there is abundance of clinical terms. There are countless specialties and subspecialties and terms vary from facility to facility, practice to practice, and clinician to clinician. This is one of the important issues in many EHRs to have less time-efficient text mining and extract data from unstructured text as there is no standard format. Moreover, relative to other fields, identifying familiar words, expressions, abbreviations and peculiar styles of clinical notes can inconvenience for the NLP algorithms. The problem is compounded by the fact that many of the expressions can be referred to as multifunctional, which means that their meaning can be different in different cases. It can lead to wrong understanding and wrong conclusions concerning information that is extracted [79].

B. Data Quality and Annotation

There is another problem, and that is of the data quality [80], [81], which is a feature of many electronic health records systems. This can greatly impact the functionalities of NLP systems due to finding information in poorly maintained Electronic Health Records as the collected data is often inaccurate, inconsistent or insufficient. The data cleaning [82], [83] and preparation phase are cardinal to the NLP process since the quality of the input data affects the obtained data quality: In addition, supervised training of NLP models entails labelling of clinical information, which is a tiresome process. Elements such as correct labeling of data require professionals; in some cases, the labeling may differ depending on the annotator's work. This remains a major disadvantage since it may lead to irregularity in the training of the models, thus irregularity in the NLP model forecasts.

C. Privacy and Security

Due to the fact that EHRs contain patients' personal information, the issues of privacy and security are very important. Patients' records are sensitive information that requires protection as practice clinicians apply Natural Language Processing (NLP) to analyze clinical data. This is a special concern for managing unstructured data because in may be challenging to remove or discard personal identifiers when compared to managing structured data. One of the biggest ethical and legal risks is the threat of data leaks or, in other words, the unintentional loss of privacy of the patient. One more challenge that complicates the process of construction and deployment of NLP systems is the legislation problem, for example, HIPAA [84] in the USA.

DISCUSSION

This section discusses the consequences and importance of the use of NLP in the extraction of EHR clinical data related to cancer prognosis. It stresses that NLP is capable of changing healthcare, specifically in oncology, where

promptness and accuracy of the extracted information make a huge difference for the prognosis of the patient. NLP technologies fill the information gap between raw data and actionable intelligence by way of automated extraction of insightful information from huge volumes of unstructured clinical data, hence enabling healthcare providers to make decisions more rapidly and knowledgeably.

Of course, one of the major topics of discussion is the precision and dependability of NLP algorithms during the extraction process. Much as NLP has progressed, several challenges still exist, especially where management of the complexity and variety that comes with clinical language is concerned. The complexity of medical jargon, acronyms, and the contextual knowledge to be used in interpreting precise clinical notes remains extremely high. Some NLP systems have already given great promise when fully trained and optimized. Not to mention, the review ran multiple experiments by which it reached an accuracy of over 90% for particular tasks, despite these obstacles.

Another important issue that was covered by the debate is the integration of unstructured and structured data in the EHR. The review underlines the need for sophisticated NLP methods that would combine effectively these two sources of information so that a more complete picture of the state of the patient could be given. NLP can help in making the prognosis of cancer more individualized and focused by fusing structured data, like laboratory results and vital signs, with less-structured data, including physician notes and discharge summaries.

Finally, it elaborates on personalized medicine—a developing field where the medications are tailored to every patient's unique profile—in this dialogue. Such a review can be performed by NLP, as shown, to project a holistic view of the patient with past data and present information. Such integration will facilitate the construction of individual programs for treatment, taking into consideration factors like genetic data, case history, and current health condition of a patient. This improves outcomes and reduces side effects in the long run.

Other major topics of discussion are concerns about security and privacy. The principal point raised in the review is that there should be very strong regard for adherence to stipulations on data privacy, especially where sensitive health data is being processed. While NLP has many benefits regarding cancer prognosis, this has to be weighed against strict security measures for protecting the privacy of patients and preventing criminal access to EHR data.

Moreover, it also speaks to how much the NLP systems are scalable in medical environments. It discusses the scope for the application of NLP in healthcare settings, from large hospitals to small clinics, and it also speaks to the challenges it poses for such wide diffusion. Because the big task is to be executed, barriers like variation of the EHR systems used, variability in the modes of recording clinically, and the requirement for giant volumes of the training datasets are being spoken of.

The paper underlines continuous developments going on in machine learning and deep learning [85], [86]methodologies, which keep on increasing the potentials of Natural Language Processing in the medical domain. It says information extraction from clinical texts is getting more accurate and efficient due to modern models such as transformers and attention-based architectures. Because of these developments, more complex NLP applications are possible in the healthcare industry, like real-time data analysis and predictive modeling.

Further research in this regard into the explainability and understandability of NLP models, applied in clinical contexts, remains to be done, says the review for the future approaches. Therefore, understanding the process by which the model comes to its conclusion becomes very essential when health providers are increasingly reliant on AI-driven technologies for decision-making. It also demands greater transparency of NLP systems so that they can justify explicitly their predictions and suggestions.

CONCLUSION

This review consolidates the conclusion drawn regarding how critical NLP is in better prognosis by extracting relevant data from EHRs. It also represented that although there were still obstacles in the way, there were huge advantages that were seen in incorporating NLP into clinical workflows. These are more precise diagnosis, treatment planning, and in the end, patient outcomes. Further, it closes with an appeal for further collaboration

between scientists, physicians, and technologists to overcome the remaining barriers that prevent the full integration of NLP into health care. To sum up, this review highlights how NLP can revolutionize the process of extracting clinical data for cancer prognosis from electronic health records. It draws attention to the field's successes as well as its difficulties and suggests a time when NLP will be crucial to personalized medicine and sophisticated healthcare analytics. The application of NLP to ordinary clinical practice has the potential to transform cancer care and improve patient outcomes globally as the discipline develops.

REFERENCES

[1] T. August, L. L. Wang, J. Bragg, M. A. Hearst, A. Head, and K. Lo, "Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing," ACM Transactions on Computer-Human Interaction, vol. 30, no. 5, Sep. 2023, doi: 10.1145/3589955.

[2] A. J. Nashwan and A. A. AbuJaber, "Harnessing the Power of Large Language Models (LLMs) for Electronic Health Records (EHRs) Optimization," Cureus, Jul. 2023, doi: 10.7759/cureus.42634.

[3] M. El Khatib, S. Hamidi, I. Al Ameeri, H. Al Zaabi, and R. Al Marqab, "Digital Disruption and Big Data in Healthcare-Opportunities and Challenges," ClinicoEconomics and Outcomes Research, vol. 14, pp. 563–574, 2022, doi: 10.2147/CEOR.S369553.

[4] I. Li et al., "Neural Natural Language Processing for unstructured data in electronic health records: A review," Comput Sci Rev, vol. 46, p. 100511, Nov. 2022, doi: 10.1016/J.COSREV.2022.100511.

[5] K. R. Siegersma et al., "Development of a Pipeline for Adverse Drug Reaction Identification in Clinical Notes: Word Embedding Models and String Matching," JMIR Med Inform, vol. 10, no. 1, 2022, doi: 10.2196/31063.

[6] J. Sanyal, D. Rubin, and I. Banerjee, "A weakly supervised model for the automated detection of adverse events using clinical notes," J Biomed Inform, vol. 126, 2022, doi: 10.1016/j.jbi.2021.103969.

[7] V. Balasubramanian, S. Vivekanandhan, and V. Mahadevan, "Pandemic tele-smart: a contactless tele-health system for efficient monitoring of remotely located COVID-19 quarantine wards in India using near-field communication and natural language processing system," Med Biol Eng Comput, vol. 60, no. 1, 2022, doi: 10.1007/s11517-021-02456-1.

[8] N. Abdelmageed et al., "BiodivNERE: Gold standard corpora for named entity recognition and relation extraction in the biodiversity domain," Biodivers Data J, vol. 10, 2022, doi: 10.3897/BDJ.10.e89481.

[9] K. Jeon, G. Lee, S. Yang, and H. D. Jeong, "Named entity recognition of building construction defect information from text with linguistic noise," Autom Constr, vol. 143, 2022, doi: 10.1016/j.autcon.2022.104543.

[10] N. Nath, S. H. Lee, and I. Lee, "NEAR: Named entity and attribute recognition of clinical concepts," J Biomed Inform, vol. 130, 2022, doi: 10.1016/j.jbi.2022.104092.

[11] G. D'Aniello, M. Gaeta, and I. La Rocca, "KnowMIS-ABSA: an overview and a reference model for applications of sentiment analysis and aspect-based sentiment analysis," Artif Intell Rev, vol. 55, no. 7, 2022, doi: 10.1007/s10462-021-10134-9.

[12] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," Artif Intell Rev, vol. 55, no. 7, 2022, doi: 10.1007/s10462-022-10144-1.

[13] A. Mohammed and R. Kora, "An effective ensemble deep learning framework for text classification," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 10, 2022, doi: 10.1016/j.jksuci.2021.11.001.

[14] Q. Li et al., "A Survey on Text Classification: From Traditional to Deep Learning," 2022. doi: 10.1145/3495162.

[15] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, " Survey on Text Classification Algorithms: From Text to Predictions," Information (Switzerland), vol. 13, no. 2, 2022, doi: 10.3390/info13020083.

[16] Y. Okamura et al., "Obsessive-Compulsive Disorder with Psychotic Features: Is It a Clinical Entity?," Healthcare (Switzerland), vol. 10, no. 10, 2022, doi: 10.3390/healthcare10101910.

[17] Y. Chen, L. Hao, V. Z. Zou, Z. Hollander, R. T. Ng, and K. V. Isaac, "Automated medical chart review for breast cancer outcomes research: a novel natural language processing extraction system," BMC Med Res Methodol, vol. 22, no. 1, 2022, doi: 10.1186/s12874-022-01583-z.

[18] L. Wang et al., "Assessment of Electronic Health Record for Cancer Research and Patient Care Through a Scoping Review of Cancer Natural Language Processing," JCO Clin Cancer Inform, no. 6, 2022, doi: 10.1200/cci.22.00006.

[19] D. Goodman-Meza et al., "Natural Language Processing and Machine Learning to Identify People Who Inject Drugs in Electronic Health Records," Open Forum Infect Dis, vol. 9, no. 9, 2022, doi: 10.1093/ofid/ofac471.

[20] N. V. Babu and E. G. M. Kanaga, "Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: A Review," 2022. doi: 10.1007/s42979-021-00958-1.

[21] A. Nazir, Y. Rao, L. Wu, and L. Sun, "Issues and Challenges of Aspect-based Sentiment Analysis: A Comprehensive Survey," 2022. doi: 10.1109/TAFFC.2020.2970399.

[22] S. M. Mishal and M. M. Hamad, "Text Classification Using Convolutional Neural Networks," Fusion: Practice and Applications, vol. 7, no. 1, 2022, doi: 10.54216/FPA.070105.

[23] A. Occhipinti, L. Rogers, and C. Angione, "A pipeline and comparative study of 12 machine learning models for text classification," Expert Syst Appl, vol. 201, 2022, doi: 10.1016/j.eswa.2022.117193.

[24] N. Liu, Q. Hu, H. Xu, X. Xu, and M. Chen, “Med-BERT: A Pretraining Framework for Medical Records Named Entity Recognition,” IEEE Trans Industr Inform, vol. 18, no. 8, 2022, doi: 10.1109/TII.2021.3131180.

[25] F. W. Mutinda, K. Liew, S. Yada, S. Wakamiya, and E. Aramaki, “Automatic data extraction to support meta-analysis statistical analysis: a case study on breast cancer,” BMC Med Inform Decis Mak, vol. 22, no. 1, 2022, doi: 10.1186/s12911-022-01897-4.

[26] B. N. Hiremath and M. M. Patil, “Enhancing Optimized Personalized Therapy in Clinical Decision Support System using Natural Language Processing,” Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 6, 2022, doi: 10.1016/j.jksuci.2020.03.006.

[27] C. Ni Ki, A. Hosseinian-Far, A. Daneshkhah, and N. Salari, “Topic modelling in precision medicine with its applications in personalized diabetes management,” Expert Syst, vol. 39, no. 4, 2022, doi: 10.1111/exsy.12774.

[28] M. Nesca, A. Katz, C. K. Leung, and L. M. Lix, “A scoping review of preprocessing methods for unstructured text data to assess data quality,” 2022. doi: 10.23889/ijpds.v7i1.1757.

[29] E. Hossain et al., “Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review,” Mar. 01, 2023, Elsevier Ltd. doi: 10.1016/j.compbiomed.2023.106649.

[30] M. Gholipour, R. Khajouei, P. Amiri, S. Hajesmaeel Gohari, and L. Ahmadian, “Extracting cancer concepts from clinical notes using natural language processing: a systematic review,” BMC Bioinformatics, vol. 24, no. 1, Dec. 2023, doi: 10.1186/s12859-023-05480-0.

[31] C. Li, Y. Zhang, Y. Weng, B. Wang, and Z. Li, “Natural Language Processing Applications for Computer-Aided Diagnosis in Oncology,” Jan. 01, 2023, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/diagnostics13020286.

[32] E. Sangariyavanich et al., “Systematic review of natural language processing for recurrent cancer detection from electronic medical records,” Jan. 01, 2023, Elsevier Ltd. doi: 10.1016/j.imu.2023.101326.

[33] J. J. Nunez, B. Leung, C. Ho, A. T. Bates, and R. T. Ng, “Predicting the Survival of Patients With Cancer From Their Initial Oncology Consultation Document Using Natural Language Processing,” JAMA Netw Open, vol. 6, no. 2, p. e230813, Feb. 2023, doi: 10.1001/jamanetworkopen.2023.0813.

[34] S. Zhou et al., “A cross-institutional evaluation on breast cancer phenotyping NLP algorithms on electronic health records,” Comput Struct Biotechnol J, vol. 22, pp. 32–40, Jan. 2023, doi: 10.1016/j.csbj.2023.08.018.

[35] H. Huang et al., “Natural language processing in urology: Automated extraction of clinical information from histopathology reports of uro-oncology procedures,” Heliyon, vol. 9, no. 4, Apr. 2023, doi: 10.1016/j.heliyon.2023.e14793.

[36] C. Fanconi, M. Van Buchem, and T. Hernandez-Boussard, “Natural Language Processing Methods to Identify Oncology Patients at High Risk for Acute Care with Clinical Notes.”

[37] G. Laurent, F. Craynest, M. Thobois, and N. Hajjaji, “Automatic Classification of Tumor Response From Radiology Reports With Rule-Based Natural Language Processing Integrated Into the Clinical Oncology Workflow,” JCO Clin Cancer Inform, vol. 7, p. 2200139, 2023, doi: 10.1200/CCI.22.

[38] A. Maghsoudi et al., “e13582 Publication Only Application of natural language processing to assess the performance status docu-mentation quality metric in patients with non-small-cell lung cancer,” 2023.

[39] S. Pais, J. Cordeiro, and M. L. Jamil, “NLP-based platform as a service: a brief review,” J Big Data, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00603-5.

[40] G. T. Gobbel et al., “Leveraging structured and unstructured electronic health record data to detect reasons for suboptimal statin therapy use in patients with atherosclerotic cardiovascular disease,” Am J Prev Cardiol, vol. 9, p. 100300, Mar. 2022, doi: 10.1016/J.AJPC.2021.100300.

[41] J. Shi et al., “Identifying Patients Who Meet Criteria for Genetic Testing of Hereditary Cancers Based on Structured and Unstructured Family Health History Data in the Electronic Health Record: Natural Language Processing Approach,” JMIR Med Inform, vol. 10, no. 8, 2022, doi: 10.2196/37842.

[42] Y. Hong and M. L. Zeng, “International Classification of Diseases (ICD),” Knowledge Organization, vol. 49, no. 7, 2022, doi: 10.5771/0943-7444-2022-7-496.

[43] D. Hopkins, D. J. Rickwood, D. J. Hallford, and C. Watsford, “Structured data vs. unstructured data in machine learning prediction models for suicidal behaviors: A systematic review and meta-analysis,” 2022. doi: 10.3389/fdgth.2022.945006.

[44] Y. Zhao and J. Chen, “A Survey on Differential Privacy for Unstructured Data Content,” ACM Comput Surv, vol. 54, no. 10 s, 2022, doi: 10.1145/3490237.

[45] L. E. S. e. Oliveira et al., “SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks,” J Biomed Semantics, vol. 13, no. 1, 2022, doi: 10.1186/s13326-022-00269-1.

[46] H. Wu et al., “A survey on clinical natural language processing in the United Kingdom from 2007 to 2022,” 2022. doi: 10.1038/s41746-022-00730-6.

[47] Z. Zheng, X. Z. Lu, K. Y. Chen, Y. C. Zhou, and J. R. Lin, “Pretrained domain-specific language model for natural language processing tasks in the AEC domain,” Comput Ind, vol. 142, 2022, doi: 10.1016/j.compind.2022.103733.

[48] S. Fu, H. Lyu, Z. Wang, X. Hao, and C. Zhang, “Extracting historical flood locations from news media data by the named entity recognition (NER) model to assess urban flood susceptibility,” J Hydrol (Amst), vol. 612, 2022, doi: 10.1016/j.jhydrol.2022.128312.

[49] Y. Li and L. Wang, “Human Activity Recognition Based on Residual Network and BiLSTM,” Sensors, vol. 22, no. 2, 2022, doi: 10.3390/s22020635.

[50] T. Paul et al., “Utility of Features in a Natural-Language-Processing-Based Clinical De-Identification Model Using Radiology Reports for Advanced NSCLC Patients,” Applied Sciences (Switzerland), vol. 12, no. 19, 2022, doi: 10.3390/app12199976.

[51] N. Mollaei, C. Cepeda, J. Rodrigues, and H. Gamboa, “Biomedical Text Mining: Applicability of Machine Learning-based Natural Language Processing in Medical Database,” 2022. doi: 10.5220/0010819500003123.

[52] P. R. Deshmukh and R. Phalnikar, “Prognostic elements extraction from documents to detect prognostic stage,” Comput Methods Biomech Biomed Engin, vol. 25, no. 4, 2022, doi: 10.1080/10255842.2021.1955359.

[53] S. Hosgurmath, V. Petli, and V. K. Jalihal, “An omicron variant tweeter sentiment analysis using NLP technique,” Global Transitions Proceedings, vol. 3, no. 1, 2022, doi: 10.1016/j.gltp.2022.03.025.

S. Yadav, “Detecting Presence of Mental Illness Using Nlp Sentiment Analysis,” International Research Journal of Modernization in Engineering Technology and Science, no. 06, 2022.

[54] Z. Ebrahimi, M. Loni, M. Daneshtalab, and A. Gharehbaghi, “A review on deep learning methods for ECG arrhythmia classification,” 2020. doi: 10.1016/j.eswax.2020.100033.

[55] A. K. Munnangi, S. UdhayaKumar, V. Ravi, R. Sekaran, and S. Kannan, “Survival study on deep learning techniques for IoT enabled smart healthcare system,” Health Technol (Berl), vol. 13, no. 2, 2023, doi: 10.1007/s12553-023-00736-4.

[56] A. Mavrogiorgou, A. Kiourtis, S. Kleftakis, K. Mavrogiorgos, N. Zafeiropoulos, and D. Kyriazis, “A Catalogue of Machine Learning Algorithms for Healthcare Risk Predictions †,” Sensors, vol. 22, no. 22, 2022, doi: 10.3390/s22228615.

[57] A. R. Ismail, N. Z. Abidin, and M. K. Maen, “Systematic Review on Missing Data Imputation Techniques with Machine Learning Algorithms for Healthcare,” 2022. doi: 10.18196/jrc.v3i2.13133.

[58] P. Khan, P. Ranjan, Y. Singh, and S. Kumar, “Warehouse LSTM-SVM-Based ECG Data Classification With Mitigated Device Heterogeneity,” IEEE Trans Comput Soc Syst, vol. 9, no. 5, 2022, doi: 10.1109/TCSS.2021.3116428.

[59] M. Payal, S. A. Ajagbe, and T. Ananth Kumar, “Support Vector Machines (SVMS) Based Advanced Health Care System Using Machine Learning Techniques,” International Journal of Innovate Research in Computer and Communication Engineering, vol. 10, no. 5, 2022.

[60] V. Dogra et al., “A Complete Process of Text Classification System Using State-of-the-Art NLP Models,” 2022. doi: 10.1155/2022/1883698.

[61] Z. Li et al., “A Unified Understanding of Deep NLP Models for Text Classification,” IEEE Trans Vis Comput Graph, vol. 28, no. 12, 2022, doi: 10.1109/TVCG.2022.3184186.

[62] A. Jamali and M. Mahdianpari, “Swin Transformer and Deep Convolutional Neural Networks for Coastal Wetland Classification Using Sentinel-1, Sentinel-2, and LiDAR Data,” Remote Sens (Basel), vol. 14, no. 2, 2022, doi: 10.3390/rs14020359.

[63] [K. De Angeli et al., “Class imbalance in out-of-distribution datasets: Improving the robustness of the TextCNN for the classification of rare cancer types,” J Biomed Inform, vol. 125, 2022, doi: 10.1016/j.jbi.2021.103957.

[64] S. Chaichulee, C. Promchai, T. Kaewkomon, C. Kongkamol, T. Ingviya, and P. Sangsupawanich, “Multi-label classification of symptom terms from free-text bilingual adverse drug reaction reports using natural language processing,” PLoS One, vol. 17, no. 8 August, 2022, doi: 10.1371/journal.pone.0270595.

[65] V. Gupta, A. Sood, and T. Singh, “Disease Detection Using RASA Chatbot,” in 2022 International Mobile and Embedded Technology Conference, MECON 2022, 2022. doi: 10.1109/MECON53876.2022.9752338.

[66] F. Gomollón et al., “Clinical characteristics and prognostic factors for Crohn’s disease relapses using natural language processing and machine learning: a pilot study,” Eur J Gastroenterol Hepatol, vol. 34, no. 4, 2022, doi: 10.1097/MEG.0000000000002317.

[67] P. C. Nair, D. Gupta, and B. Indira Devi, “Automatic Symptom Extraction from Unstructured Web Data for Designing Healthcare Systems,” in Lecture Notes in Electrical Engineering, 2022. doi: 10.1007/978-981-16-1342-5_46.

[68] Y. An, X. Xia, X. Chen, F. X. Wu, and J. Wang, “Chinese clinical named entity recognition via multi-head self-attention based BiLSTM-CRF,” Artif Intell Med, vol. 127, 2022, doi: 10.1016/j.artmed.2022.102282.

[69] A. Fang et al., “Extracting clinical named entity for pituitary adenomas from Chinese electronic medical records,” BMC Med Inform Decis Mak, vol. 22, no. 1, 2022, doi: 10.1186/s12911-022-01810-z.

[70] P. Rogliani, J. Ora, L. Calzetta, M. G. Matera, and M. Cazzola, “Asthma and comorbidities: recent advances,” 2022. doi: 10.20452/pamw.16250.

[71] C. Hu, X. Zhang, T. Teng, Z. G. Ma, and Q. Z. Tang, “Cellular Senescence in Cardiovascular Diseases: A Systematic Review,” 2022. doi: 10.14336/AD.2021.0927.

[72] M. Ptasiewicz, P. Maksymiuk, and R. Chałas, “Oral hygiene considerations in adult patients with leukemia during a cycle of chemotherapy,” Int J Environ Res Public Health, vol. 19, no. 1, 2022, doi: 10.3390/ijerph19010479.

[73] W. Boughattas, M. Ben Salha, and N. Moella, “Mental training for young athlete: A case of study of NLP practice,” SSM - Mental Health, vol. 2, 2022, doi: 10.1016/j.ssmmh.2022.100076.

[74] Y. Gu et al., “Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing,” ACM Trans Comput Healthc, vol. 3, no. 1, 2022, doi: 10.1145/3458754.

[75] V. Shankar and S. Parsana, “An overview and empirical comparison of natural language processing (NLP) models and an introduction to and empirical application of autoencoder models in marketing,” J Acad Mark Sci, vol. 50, no. 6, 2022, doi: 10.1007/s11747-022-00840-3.

[76] A. Madsen, S. Reddy, and S. Chandar, “Post-hoc Interpretability for Neural NLP: A Survey,” ACM Comput Surv, vol. 55, no. 8, 2022, doi: 10.1145/3546577.

[77] NCT05648175, “Comparing Clinical Decision-making of AI Technology to a Multi-professional Care Team in eCBT for Depression,” https://clinicaltrials.gov/ct2/show/NCT05648175, 2022.

[78] H. MacFarlane, A. C. Salem, L. Chen, M. Asgari, and E. Fombonne, “Combining voice and language features improves automated autism

detection," Autism Research, vol. 15, no. 7, 2022, doi: 10.1002/aur.2733.

[79] B. Shiner et al., "Improvements to PTSD quality metrics with natural language processing," J Eval Clin Pract, vol. 28, no. 4, 2022, doi: 10.1111/jep.13587.

[80] J. H. Bae et al., "Natural Language Processing for Assessing Quality Indicators in Free-Text Colonoscopy and Pathology Reports: Development and Usability Study," JMIR Med Inform, vol. 10, no. 4, 2022, doi: 10.2196/35257.

[81] W. Gu, X. Yang, M. Yang, K. Han, W. Pan, and Z. Zhu, "MarkerGenie: an NLP-enabled text-mining system for biomedical entity relation extraction," Bioinformatics Advances, vol. 2, no. 1, 2022, doi: 10.1093/bioadv/vbac035.

[82] K. Mavrogiorgos, A. Mavrogiorgou, A. Kiourtis, N. Zafeiropoulos, S. Kleftakis, and D. Kyriazis, "Automated Rule-Based Data Cleaning Using NLP," in Conference of Open Innovation Association, FRUCT, 2022. doi: 10.23919/FRUCT56874.2022.9953810.

[83] M. Heath, T. H. Porter, and G. Silvera, "Hospital characteristics associated with HIPAA breaches," Int J Healthc Manag, vol. 15, no. 2, 2022, doi: 10.1080/20479700.2020.1870349.

[84] J. Liu, D. Capurro, A. Nguyen, and K. Verspoor, "'Note Bloat' impacts deep learning-based NLP models for clinical prediction tasks," J Biomed Inform, vol. 133, 2022, doi: 10.1016/j.jbi.2022.104149.

[85] I. Lauriola, A. Lavelli, and F. Aiolli, "An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools," Neurocomputing, vol. 470, 2022, doi: 10.1016/j.neucom.2021.05.103.

DOI: https://doi.org/10.15379/ijmst.v6i2.3784