REVIEW ON MACHINE LEARNING

BHUMIKA S K ,ANKITHA B, BHAGYASHREE R P, BHARATH J

STUDENT, INFORMATION SCIENCE AND ENGINEERING, ALVA'S INSTITUTE OF ENGINEERING AND TECHNOLOGY, KARNATAKA , INDIA

ABSTRACT

Numerous real-world data sources are made available by the extensive usage of electronic health record (EHR) systems in the medical field, opening up new directions for clinical research. Due to the fact that clinical narratives in electronic health records include a significant quantity of important clinical information, natural language processing (NLP) techniques have been employed as an artificial intelligence strategy to extract information from them. However, much clinical facts are still concealed in a clinical narrative structure in free-form texts like electronic health records. Consequently, to fully utilize EHR data and automatically transform clinical narrative text into structured clinical data, biomedical NLP algorithms must be used. Biomedical NLP applications might thus be utilized to guide clinical judgments, recognize health issues, and successfully prevent or delay the onset of a disease. This review analyzes the possibilities, difficulties, and uses of biomedical natural language processing (NLP) techniques and examines the literature that is currently available on the secondary use of electronic health record data for clinical research on chronic diseases. We provide an overview of machine learning and deep learning techniques used to process EHRs and enhance the comprehension of the patient's clinical records and the prediction of chronic disease risk. These techniques offer a great opportunity to extract previously undiscovered clinical information. We also review some of the biomedical NLP systems and methods used over EHRs. Additionally, based on EHR data relevant to chronic diseases, this research describes the application of Deep Learning and Machine Learning algorithms in biomedical NLP applications. In conclusion, this evaluation showcases the future trends and challenges in the biomedical NLP.

INDEX TERMS: *machine learning, natural language processing (NLP), clinical data, deep learning, artificial intelligence (AI), and electronic health records (EHR).*

INTRODUCTION

Digital data processing is greatly impacted by machine learning and natural language processing (NLP) approaches. Since more and more research is being done using digital data, it is critical to utilize data's worth in these various fields.

Applications for information extraction from clinical texts include data mining, identification of research subjects, automatic terminology management, de- identification of the clinical text, analysis of the medication used to treat the illness and its side effects, and prediction of the onset and progression of various chronic diseases. While NLP-based machine learning approaches perform better in the biomedical and healthcare fields, greater expertise

Shen Yin, the associate editor in charge of organizing the manuscript's assessment and approving its publishing, is necessary in the narrative clinical text analysis [1]. Consequently, in order to create new avenues for study in this area, a thorough examination of the issues and difficulties associated with obtaining information from clinical texts is required [2]. Natural language processing, bioinformatics, medical informatics, and computer linguistics are all included in the field of biomedical NLP research [1]. One important aim of Natural Language Processing (NLP) is to extract useful information from freely available clinical texts hidden in unstructured data. This can help with research, administrative reporting, and decision making. Biomedical

NLP applications in EHRs have a significant impact on a number of health care and biomedical research sectors. Medical language processing was made possible using NLP in the healthcare domain. The majority of biomedical data often exist in an unstructured format, which is the outcome of voice recognition software, direct entry, or dictated transcriptions applications. Consequently, because the summarization and decision- support tasks cannot be completed using the input data in its narrative form, data pre- processing is necessary before information can be extracted. Tokenization, part-of- speech tagging, spell checking, sentence splitting, Word Sense Disambiguation (WSD), and some type of parsing are examples of preprocessing. Situation- dependent characteristics such as negation, timing, and event subject identification are important in causing incorrect interpretation of the information gathered.

Numerous methods exist for extracting information, including rule-based methods, pattern matching methods, machine learning methods, and statistical methods. Subsequently, the information that has been extracted can be utilized for the analysis of the clinical text, as well as for enhancing the EHR and decision support systems and connecting concepts to standard terminology. The approaches and research on how to apply natural language processing (NLP) to texts and literature related to biomedical and electronic medical records are included in the field of biomedical NLP. When compared to traditional machine learning (ML) techniques, deep learning techniques have recently outperformed them in a variety of general natural language processing (NLP) tasks, including language modeling, named entity recognition, sentiment analysis, (Part of Speech) tagging, and paraphrase detection. Due to the regular usage of Clinical reports typically face different challenges than general-domain text because of acronyms and non-standard clinical terminology used by healthcare professionals, the disorganized structure of the document, and the requirement for complete de-identification and anonymization to protect patient data privacy. In the end, tackling and resolving these issues might encourage additional study and advancement for a range of biomedical applications, including public health management, pharmacovigilance, drugs, clinical decision support, patient engagement support, identification of patient cohorts, and clinical text summarization.

MOTIVATION

Traditionally, clinical professionals have extracted clinical information manually from narrative clinical texts. This has led to a number of problems, including high costs and scalability challenges. Since clinical notes are more than structured data, these problems have primarily affected chronic diseases. For instance, Wei et al. [4] graphically quantifies the number of clinical notes relative to structured data for chronic diseases like rheumatoid arthritis, Parkinson's disease, and Alzheimer's disease. NLP techniques play a significant role in addressing and resolving a number of the difficulties associated with different therapeutic activities, such as the automatic extraction of pertinent clinical data that might, for example, delay or prevent the start of disease. This study's primary goal is to present a thorough review of NLP in the therapeutic area, covering techniques, this study's primary goal is to present a thorough overview of natural language processing (NLP) in the clinical setting, including system architecture, applications, techniques, and the difficulties that clinical NLP approaches encounter while trying to comprehend clinical narratives. These are described below.We have identified biological NLP and NLP in general along with its technology and approaches. Next, we discussed the biomedical NLP application areas for machine learning and deep learning. We have given a summary of the overall architecture and most widely used biomedical natural language processing systems. We have discovered that NLP applications are used in clinical notes to recognise chronic illnesses and comprehend the difficulties they are now encountering. After that, we looked at a study of the literature on the application of different NLP approaches to narrative clinical notes on chronic illnesses. This included an examination of the challenges that NLP methodology encountered when it came to clinical narrative understanding. In summary, we wrap up this review article by outlining the difficulties that are currently being encountered and the unresolved problems related to processing clinical and biological text, and by offering the NLP field sufficient and opportunities to extract new methodologies.

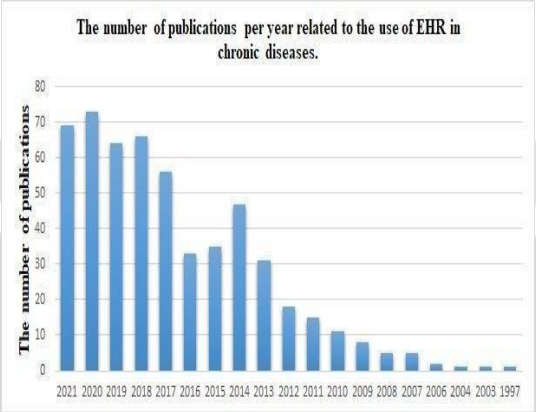
CRITERIA FOR SEARCH AND SELECTION

Using Google Scholar, PubMed, and the Web of Science, we looked up prior research published between 2009 and 2021. The terms electronic health records, electronic medical records, EHR, and EMR were used in all searches, along with the terms machine learning or the name of a specific machine learning method related to chronic illnesses. The number of publications pertaining to the yearly application of machine learning to EHRs is displayed in Figure 1a. The amount of papers published each year about the use of EHR in chronic illnesses is depicted in Figure 1b. Using machine learning and deep learning approaches, we propose to summarise the most notable and important publications and investigations that focus on EHR in this review paper. We begin with an overview of NLP as a whole, with its techniques, technologies, and possible jobs/usecases in the biomedical and healthcare domains in Sections II and III, and then machine learning application areas in the biomedical natural language processing in Section IV. In Section V, we then give a summary of NLP systems and system design. In Section VI, we next go over a review of recent research on the use of NLP to treat chronic illnesses. In Section VII, we then examine the unanswered questions and difficulties that still exist in the field of biomedical NLP. In conclusion, the review paper's conclusion is demonstrated in Section VIII by listing the challenges and unresolved difficulties that exist today.

BASICS AND BACKGROUND

NATURALLANGUAGEPROCESSING OVERVIEW:

NLP is a field that integrates linguistics, computer science, and artificial intelligence (AI) with the goal of processing and interpreting human language for various purposes.



(For instance, translating languages and automatically responding to inquiries). Because of the many difficulties in describing, comprehending, and applying linguistic, social, global, or visual information, NLP is commonly regarded as an AI-complete problem. NLP typically calls for analysing an input text at several levels, including discourse processing, tokenization, morphological analysis, syntactic analysis, and semantic analysis [5]. Natural Language Processing (NLP) is a subfield of artificial intelligence that specialises in perception of human-generated textual or spoken input. Natural Language Query (NLQ), Natural Language Generation (NLG), and Natural Language Understanding (NLU) are some of the subdisciplines of the technology [5].

NATURALLANGUAGEPROCESSING

BIOMEDICINEANDHEALTHCARE

A MEDICAL Natural language processing faces many difficulties when it comes to general language, but a few crucial problems are particularly pertinent to the biomedical and healthcare fields.

Publications, electronic health records, and the Internet are just a few of the many electronic sources of information pertaining to the healthcare industry. Consequently, there are numerous criticisms of biomedical information, the majority of which are textual, concerning the management and application of this data, which is essential for the advancement of health research, 140630 quality enhancement, and cost containment. Because it is necessary to transform narrative clinical texts into structured data that can be employed in computer systems, natural language processing (NLP) is significant [6]. Throughout the past ten years, hospitals have significantly increased their use of electronic health record systems, thanks to by offering $30 billion in incentives to medical practices and hospitals to implement EHR systems; this is made possible in part by the 2009 Health Information Technology for Economic and Clinical Health (HITECH) Act [3]. The most recent report from the Office of the National Coordinator for Health Information Technology (ONC) [7] shows that 84% of hospitals use the basic EHR system, a 9-fold growth since 2008. Furthermore, from 42% to 87% more ce- based physicians are using and adopting certified and basic electronic health records. EHR systems retain data about every patient they come into contact with, including their demographics, diagnosis, lab results, medications, radiological imaging, clinical notes, etc.In general, hospital and outpatient care settings use Electronic Health Records (EHR) systems.has significantly increased [7]. By decreasing errors, increasing efficiency, and enhancing treatment quality, the implementation of EHR in clinics and hospitals can improve patient care while also giving researchers access to a wealth of data [8]. EHR systems are generally divided into three categories based on their functionality: basic EHRs without clinical notes, basic EHRs with clinical notes, and complete systems [7]. Despite lacking more sophisticated features, even entry-level EHR systems may give a wide range of patient information, including medical history, diseases, and

prescription use. Since the hospital's administrative tasks were the primary purpose for which the EHR was created, numerous classification schemes and regulated terminology exist for the purpose of collecting patient medical data and occurrences. The
International Statistical Classification of Diseases and Related Conditions Table 1 provides codes for Diagnostic codes from Health Problems (ICD), procedure codes from Current Procedural Terminology
(CPT), laboratory note codes from Logical Observation Identifiers Names and Codes (LOINC), and drug codes from RxNorm are all included in this list.Organisations may have different codes, and partial mappings may be maintained by the Systemized Nomenclature of Medicine-Clinical Terms (SNOMED CT) and the United Medical Language System (UMLS). Research is still being done in the area of organising and analysing data across organisations and terminologies due to the availability of several classification schemas.EHR systems store a variety of patient data, such as demographics, lab results, physical examination findings, diagnostics, sensor measurements, prescribed or managed medications, and clinical notes. A distinct culture is in order to manage the intricacy of EHR data and its various data kinds, such as the following:

(i) Quantitativenmeasurements such as the body mass index.

(ii) The time and date Things like the admittance time or date of birth.

(iii) Categorical values, such as race, or regulated vocabulary terms, such as CPT procedures or ICD-10 (previously ICD-9) diagnosis codes. (iv) Free-Text Natural Language, such as summaries of discharge or progress reports. It is also possible to arrange certain kinds of data chronologically.

(v) Sequential Data derived from multimodal patient history or vital perioperative signs.
While other biomedical data—like genetic information or medical images—are present and have been the subject of significant recent studies In this review work, we focus on the main categories of data seen in the majority of contemporary EHR systems [9] [11]. Given the rising prevalence of chronic diseases worldwide, new approaches are required in this field to support and progress evidential medicine. There is a strong and beneficial influence of the EHRs are used as a secondary tool for analysing clinical data for translational and biological applications.
Although its primary purpose is to improve operational healthcare performance, a number of research studies have found a secondary use of EHRs in bioinformatics and healthcare applications [12], [13]. Specifically, biomedical tasks including extracting medical concepts [2], [14], modelling patient trajectories [15], detecting diseases [16], [17], supporting clinical choices [18], etc., used patient- relevant data housed in EHR systems. A deeper and better understanding of clinical patient trajectories—which track a patient's status from one health state to another while receiving a diagnosis of a specific clinical condition and predicting their risk of developing chronic diseases—is made possible by processing EHRs using machine learning and deep learning techniques. This presents a rare opportunity to learn previously undiscovered Clinical details. But a lot of clinical history is still hidden under clinical narratives in free- form publications. Therefore, the development of natural language processing (NLP) techniques to automatically transform clinical material from its narrative form into an organised form that may guide therapeutic decisions and perhaps delay or prevent the start of diseases is necessary to fully harness the potential of EHR data [19].Due to its high dimensionality, noise, heterogeneity, sparse design, incompleteness, random errors, and systemic biases, EHR processing and modelling present significant challenges. Furthermore, a great deal of data regarding a patient's health history is typically kept in free-text clinical narratives [20], since written records of clinical occurrences continue to be the most popular and descriptive manner of doing so. It is imperative that NLP approaches be developed and incorporated into machine learning algorithms in order to automatically convert clinical free-text into a structured data format. Since much potentially useful clinical information for pharmaco epidemiological research is contained in unstructured free-text documents, natural language processing (NLP) has been used for a wider range of applications in the clinical domain. These applications include the detection of medical concepts from nursing documentation [21], discharge summaries [22], and radiology reports [23]. Typical stroke codes are used in routine health records, such as Scottish Morbidity Records (SMR01). Unstructured text Reports from the Computerised
Radiological Information System (CRIS) may be able to fill in this blank. By adding data from CRIS reports to SMR01, the number of stroke-type-specific diagnoses will grow. Additionally, the correctness of the data will be evaluated. precision of this approach. It hasn't been common practice, therefore, to use NLP-based frameworks to narrative clinical texts in order to control administrative or decision-support systems in clinical duties and activities.

NLP DUTY IN THE MEDICAL SURGENCE AREAS

The clinical NLP does the following tasks: Word Sense Disambiguation (WSD) is the process of automatically giving a word in question in a particular context its precise meaning (sense). One crucial requirement for biomedical NLP activities is the capacity to effectively grasp ambiguous words within a certain context. Every ambiguous term has a list of all potential meanings (senses) according to the medical word sense disambiguation. The terminology used in clinical notes is often unclear. The acronym PCA has several different meanings, which include prostate cancer, patient-controlled anaesthesia, and principal component analysis. Since WSD is a necessary step for the examination of clinical notes [28], it is a crucial issue in the medical field [24], [25], [26], and [27]. Name Entity Recognition (NER) is an Information Extraction (IE) subtask. Converting unstructured text into computer-readable structured data is one of the most crucial challenges in biological natural language processing [29]. Finding expressions in clinical notes that indicate named entities (diseases, drugs, and lab tests) is known as named entity recognition, or NER. Numerous methods, including dictionary- based, rule-based, statistical, deep learning, and hybrid approaches, can be applied in
NER [30, 31].

Adverse Drug Events (ADEs): Medical research and hospital care both gain from the identification of adverse drug events (ADEs) and details on medications found in clinical notes. ADEs are defined as illnesses brought on by medical errors involving medication, including overdoses, allergic responses, adverse drug reactions, and prescription errors [32].
EHRs provide a plethora of unstructured data, including discharge summaries, procedural notes, medical histories, and test results, that contain rich information about ADEs [33] [35]. Finding and detecting ADE-related information in narrative clinical notes is an extremely difficult and time-consuming operation. For the purpose of automatically processing narrative EHRs and identifying medications, ADEs, and their interactions, an NLP system is therefore required [36]. Information extraction (IE) is a crucial biomedical natural language processing (NLP) job that makes it easier to use EHRs for clinical decision support, quality by autonomously extracting and encoding therapeutic concepts from narrative notes, translation, or improvement research. Within the general realm, automatic idea, entity, and event extraction from free text, along with their relationships and related characteristics, is usually acknowledged as a specialised topic in empirical natural language processing (NLP).
[34, 37] Finding and recognising semantic links between clinical concepts in clinical notes is the main goal of relation extraction (RE), a significant subtask of information extraction (IE) [38], [39]. For instance, the lab test MRI suggests two diseases: a C5-6 disc herniation and cord compression. In this clinical report, an MRI revealed a C5-6 disc herniation with cord compression. Previous studies have mentioned a wide range of associations, including disease- attribute pair extraction [40],
[41],identification of temporal relationships [42], detection of adverse drug events [43], [44], etc. The clinical NLP area has recently introduced a number of shared tasks pertaining to connection extraction from clinical notes, including the 2018 National NLP Challenge, the Semantic Evaluation (SemEval) Challenge, and Integrating Biology and the Bedside (i2b2) Challenges [45, 46].

APPLICATION OF MACHINE

LEARNING AND DEEP LEARNING TECHNIQUES IN THE BIOMEDICAL NLP DOMAIN

Early EHR analyses relied on more conventional statistical methods that were easier to understand [79]. In recent times, machine learning methods including Cox Proportional Hazard Model [82], Support Vector Machines (SVM) [81], and Logistic Regression [80] have been utilised. and Random Forest [83] have been used to mine trustworthy prediction patterns from EHR data. When applying the statistical model to the analysis of EHR data, there are important considerations [84] [86]. By using modelling tools that may be utilised to analyse and extract complicated interactions between nonlinear variables, such as mixed and multimodal data gathered at random times [86], from each patient's complete medical history, such problems can be resolved. The most widely used machine learning technique in medical reports for the identification of diabetes EHR progress notes, the classification of breast radiology reports based on BI-RADS [90], and the prediction of heart disease [88,
[89] is the support vector machine algorithm. Naïve Bayesian machine learning is the second most prevalent method. From free text, numerous clinical occurrences can be identified. Recurrent Neural Network (RNN) architectures are used to apply EHR notes, including illnesses, drugs, tests, and adverse medicine effects [117], and patient data de-identification from electronic health records [118]. Bidirectional RNNs/LSTMs have been suc cessfully applied to several biomedical NLP tasks such as building models for the prediction of the missing punctuation in medical reports [119], the identi cation of biomedical events [120], the modeling of relational and contextual sim ilarities between the named entities in biomedical articles to understand important information to provide appropri ate treatment suggestions [121], the extraction of clinical concepts from EHR reports [122], and

the recognition of named entities in clinical texts [123]. Several recent studies create models with the embedded graph data for Using a bidirectional LSTM transducer, adverse medication reaction detection in social media data was achieved [124]. When used in conjunction with CNNs, RNNs are employed to create recognition models for disease name learning with term- and character-level embedding data [49]. In this part, we present an overview of the state-of-the-art in biomedical applications as a result of the deep learning techniques' recent and rapid development in the context of electronic health records (EHRs). According to current research's logical classification, Table 3 lists the most recent biomedical models that use deep learning techniques along with their primary application, subtask definitions, and kind of input data.

BIOMEDICAL NLP SYSTEMS

In this part, we provide an overview of NLP systems and their architecture. Overview of the Biomedical Natural Language Processing System

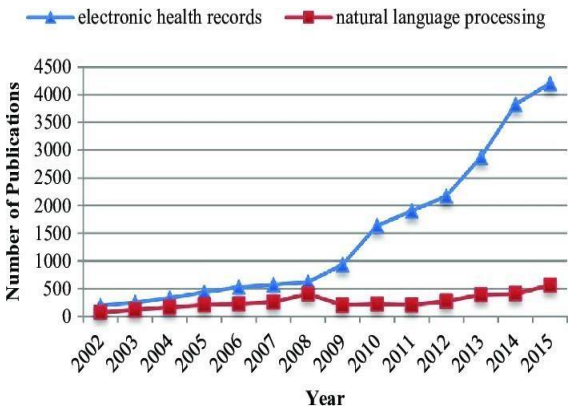
Figure 8a provides an illustration of NLP and its various components based on Friedman and Elhadad's explanation [6]. As seen in Figure 8a, the trained corpora, domain model, domain knowledge, and linguistic knowledge are on the left, and techniques, tools, systems, and applications are on the right. So, NLP aspects can be separated into two categories. An overview of the general architecture of the NLP system is given in Figure 8b. The background knowledge that corresponds to the left part of the gure and the framework that includes NLP tools and modules that correspond to the right part of the gure are the two main components of the NLP system. The two main elements Below is an illustration of biological NLP systems and their functions, showing how NLP tools are integrated into a pipeline built on top of a specific framework. The framework is a software platform that facilitates the control and management of pipeline components such as loading, unloading, and handling.
Its components can be merged, amalgamated, or utilised as plug-ins inside the system. The architecture of biomedical NLP systems consists of two levels: low- level and high-level processors. Low-level processors perform basic NLP tasks such chunking noun phrases, segment tagging, sentence boundary recognition, and part-of- speech tagging. High-level processors carry out semantic level processing, which includes named entities recognition (e.g., disease/disorder, sign/symptoms, medications), relationship identification, and timeline extraction.

LITERATURE REVIEW AND RELATED WORKS

We examine a few publications and studies that make up our literature review throughout this section, along with a list of all relevant works for machine learning applications in biomedical natural language processing, with a focus on chronic disorders. Classification of Diseases: Approximately 106 studies have been analysed, with the majority of them being associated with 43 distinct chronic diseases. Clarifying the use of NLP and the clinical notes associated with it for certain illnesses was one of the goals. Thus, the 43 specific chronic disorders were then divided into ten disease categories using the International Classification of disorders.

DISEASES OF THE CIRCULATORY SYSTEM

Heart-related Conditions Although heart disease is one of the leading causes of death, there has lately been progress in its prediction and prevention. The first step in anticipating and preventing heart disease is identifying risk factors. Numerous investigations have been suggested to Figure shows the annual number of publications on chronic diseases associated to EHRs.



Identify risk factors for heart disease, but no one has made an effort to discover all of the risk variables. The National Centre for Computer Science for Integrating Biology and Beside (i2b2) published a challenge for biomedical natural language processing in 2014. It involved a track (track 2) for identifying heart disease risk factors in clinical documents over time. This track's objective was to categorise data on cardiovascular risks, as as well as to keep an eye on the historical medical records' quality. Sorting tags and traits related to the onset and progression of the illness, risk factors, and drugs in the patient's medical history was crucial. The number of papers pertaining to circulatory system illnesses is compiled in Table 6.

B) Coronary artery disease and its peripheral forms

The chronic illness known as peripheral arterial disease (PAD) affects millions of individuals globally. The NLP algorithm should be used to automatically determine the PAD status using predetermined criteria in clinical reports rather than manually reviewing charts, which is a labor-intensive and time-consuming method when assessing PAD status from clinical notes. NLP has been utilised by numerous researchers to diagnose peripheral artery disease.

Hypertension

High blood pressure High blood pressure (HBP) disorders and hypertension (HTN) are two of the major health issues. The number of persons with hypertension is predicted to rise by 60 percent by 2025. HTN is one of the main risk factors for kidney and cardiovascular disorders. Any patient knowledge pertaining to hypertension is significant when it comes to developing predictive prevention and monitoring models, as well as cohort discovery. The majority of this crucial medical information is usually dispersed across several EHR systems in the form of unstructured clinical records. Finding patient-relevant information in unstructured clinical notes typically requires a significant investment of time and money.

OPEN ISSUES AND CHALLENGES

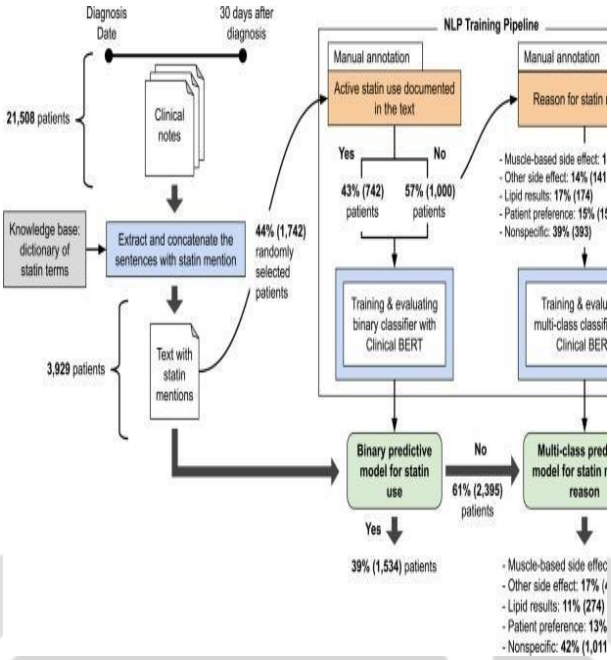


Fig: Summary of the studies that use NLP methods for peripheral and coronary arterial disease.

Extensive techniques and approaches are needed to enhance and broaden the current evidence-based therapies that can lessen the severity of chronic illnesses. A viable route is the secondary use of EHRs for patient data processing, supporting medical research, and improving clinical decision making. Techniques based on EHR processing and modelling improve patient stratification and risk prediction by enhancing our understanding of patient clinical trajectories. Machine learning, and particularly deep learning, can be used to process EHRs and effectively extract unknown clinical knowledge. The vast, continuous stream of data provided by the longitudinal structure of chronic diseases can be used to spot helpful clinical trends and guide treatment decisions in a way that prevents or delays the beginning of the disease.

Due of the several disciplines involved in the professional creation of clinical reports and advancements in NLP studies in 140644 Comparatively speaking, the biomedical field is slow and behind the advancement of general NLP. The main obstacles to the advancement of biomedical natural language processing (NLP) are the following: difficult access to shared data; insufficient annotated datasets for training and benchmarking; insufficient annotation agreements and standards; difficult reproducibility; limited partnerships; and a lack of user-centered development and scalability. The shared tasks for the i2b2/VA Challenge address these issues by giving participants access to annotated datasets that may include solutions. The practice of using clinical notes for the diagnosis of chronic diseases has been beset with a number of problems and obstacles by the development of biomedical NLP. It is noteworthy that these issues, as stated in [53], [106], still exist today:

1) Domain knowledge: The most crucial prerequisite for an NLP is having sufficient domain knowledge. researcher working on the creation of biomedical record processing systems and procedures. The key reason domain knowledge is important is that the system's output can be used in the healthcare industry. For the intended biomedical application, the system must therefore always have sufficient recall, accuracy, and F-measurement along with the requisite performance modification. It's interesting to note that NLP approaches can be used to extract domain knowledge from free text. For instance, the NLP strategy for the automated capture of domain knowledge connected to ontology follows a two-phase methodology: in the first phase, terms are extracted from the language representations of concepts, and in the second phase, semantic relations are extracted.

CONCLUSION

In this review paper, we have covered an overview of natural language processing (NLP) in general as well as NLP in biomedicine and healthcare, along with its methodologies, technologies, possible applications, and tasks in these fields. Next, we discussed the applications of deep learning and machine learning in
biomedical natural language processing. An overview of the most widely used biological NLP systems and their general design is given below. The literature study of several NLP strategies used to narrative clinical notes on chronic diseases was then covered. This included an appraisal of the challenges that NLP methodology faced in clinical narrative understanding. In summary, we wrap up this review article by outlining the difficulties that are currently being encountered and the unresolved problems related to processing biomedical and clinical material.

Essential challenges including domain knowledge, the confidentiality of clinical texts, acronyms, different formats, expressiveness, intra- and interoperability, and information interpretation have all been covered in this review study.
Understanding the intricacy of clinical text processing and the range of viable options is made possible by these conversations. The creation of methodologies for processing the different format of clinical texts is a significant area of research related to the knowledge of the difficulties associated in processing the text. For NLP researchers, each format presents a unique issue that can be investigated using both conventional and hybrid techniques. Our analysis has demonstrated the need for biomedical natural language processing (NLP) methods to be updated and modified to focus more on concept interpretation rather than just extracting clinical terms. This involves integrating clinical data, domain expertise, and general knowledge in the reasoning process, in addition to comprehending the relationships between concepts. extracting knowledge about chronic illnesses from clinical accounts that are not structured. The narrative component remains, despite the development of new standards and improved EHR coding with clinical terminology standards; hence, biomedical natural language processing techniques are crucial for clinical research informatics.

Numerous models and approaches have also been widely applied to biomedical literature; all of these natural language processing (NLP) techniques are significant and can be used to mine electronic health records (EHRs) in an efficient manner to assist critical clinical research operations.
New deep learning techniques have significantly advanced a variety of jobs and will be used more frequently to analyse large amounts of data from electronic health records in an effective and efficient manner, thereby promoting quality improvement, clinical research in general, disease management, and other related areas.

REFERENCES

S.A.HasanandO.Farri, Clinicalnaturallanguageprocessingwithdee p learning, in Data Science for Healthcare. Springer, 2019, pp. 147171.

G. K. Savova, K. C. Kipper-Schuler, J.

F. Hurdle, and S. M. Meystre, Extracting information from textual documents in the electronic health record: A review of recent research, Yearbook Med. Informat., vol. 17, no. 1, pp. 128144, 2008.

D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, What can natural language processing do for clinical decision support? J. Biomed. Inform., vol. 42, no. 5, pp. 760772, 2009.

W.-Q. Wei, P. L. Teixeira, H. Mo, R. M.

Cronin, J. L. Warner, and J. C. Denny, Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping perfor mance, J. Amer. Med. Inform. Assoc., vol.
23, no. e1, pp. e20 e27, Apr. 2016.

W. contributors. (2020). Natural Language Processing Wikipedia, the Free Encyclopedia. Accessed: Oct. 4, 2020. [Online]. Available:

https://en.wikipedia.org/w/index.php?title=

Natural_language_processing

C. Friedman, T. C. Rind esch, and M. Corn, Natural language process ing: State of the art and prospects for signi cant progress, a workshop sponsored by the National Library of Medicine, J. Biomed.
Inform., vol. 46, no. 5, pp. 765773, 2013.

J. Henry, Y. Pylypchuk, T. Searcy, and V. Patel, Adoption of electronic health record systems among US non-federal acute care hospitals: 2008 2015, ONCData
Brief, vol. 35, pp. 19, May 2016.

L. A. Knake, M. Ahuja, E. L.

McDonald, K. K. Ryckman, N. Weathers, T. Burstain, J. M. Dagle, J. C. Murray, and

P. Nadkarni, Quality of EHRdata extractions for studies of preterm birth in a tertiary care center: Guidelines for obtaining reliable data, BMC Pediatrics, vol. 16, no. 1, p. 59, Dec. 2016.

D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.- Z. Yang, Deeplearning for health informatics, IEEE J. Biomed. Health Inform., vol. 21, no. 1, pp. 421, Jan. 2017.

D.Shen, G. Wu,andH.Suk, Deeplearning in medical image analysis, Annu. Rev. Biomed. Eng., vol. 19, pp.

221248, Jun. 2017.

C.Angermueller,T.Pärnamaa,L.Parts,andO. Stegle, Deeplearningfor computational biology, Mol. Syst. Biol., vol. 12, no. 7, p. 878, Jul. 2016. [

T. Botsis, G. Hartvigsen, F. Chen, and C. Weng, Secondary use of EHR:Data quality issues and informatics opportunities, Summit Transl. Bioinf., vol. 2010, p. 1, Oct. 2010

P. B. Jensen, L. J. Jensen, and S. Brunak, Mining electronic health records: Towards better research applications and clinical care, Nature Rev. Genet., vol. 13, no. 6, pp. 395405, 2012.

M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu, Astudyofmachine-learning-based approaches to extract clinical entities and their assertions from discharge summaries, J. Amer. Med. Inform. Assoc., vol. 18, no. 5, pp. 601606, Apr. 2011.

S. Ebadollahi, J. Sun, D. Gotz, J. Hu, D. Sow, and C. Neti, Predicting patients trajectory of physiological data using temporal trends in similar patients: A system for near-term prognostics, in Proc.
AMIA Annu. Symp., 2010, p. 192. [16]D.ZhaoandC.Weng,

CombiningPubMedknowledgeandEHRdat ato develop a weighted Bayesian network for pancreatic cancer prediction, J. Biomed. Informat., vol. 44, no. 5, pp. 859868, Oct. 2011.

P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, and D. S. Lee, Using meth ods from the data-mining and machine-learning literature for disease classi cation and prediction: A case study examining classi cation of heart failure subtypes, J. Clin. Epidemiol., vol. 66, no. 4, pp. 398407, 2013.

G. J. Kuperman, A. Bobb, T. H. Payne,

A. J. Avery, T. K. Gandhi, G. Burns, D. C. Classen, and D. W. Bates, Medication- related clini cal decision support in computerized provider order entry systems: A review, J. Amer. Med. Inform. Assoc., vol. 14, no. 1, pp. 2940, 2007.

S. Sheikhalishahi, R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi, and V. Osmani, Naturallanguage processing of clinical notes on chronic dis eases: Systematic review, JMIR Med. Informat., vol. 7, no. 2, Apr. 2019, Art. no. e12239.

K. Jensen, C. Soguero-Ruiz, K. Oyvind Mikalsen, R.-O. Lindsetmo, I.
Kouskoumvekaki, M. Girolami, S. Olav Skrovseth, and K. M. Augestad, Analysis of free text in electronic health records for identi cation of cancer patient trajectories, Sci. Rep., vol. 7, no. 1, p. 46226, May 2017.

L. L. Popejoy, M. A. Khalilia, M. Popescu, C. Galambos, V. Lyons, M.
Rantz, L. Hicks, and F. Stetzer, Quantifying care coordination using natural language processing and domain-speci c ontology, J. Amer. Med. Inform. Assoc., vol. 22, no. e1, pp. e93 e103, Apr. 2015.

H. Yang, I. Spasic, J. A. Keane, and G. Nenadic, A text mining approach to the prediction of disease status from clinical discharge summaries, J. Amer. Med. Inform. Assoc., vol. 16, no. 4, pp. 596600, Jul. 2009.

R. W. V. Flynn, T. M. Macdonald, N. Schembri, G. D. Murray, and A. S. F. Doney, Automated data capture from free- text radiology reports to enhance accuracy of hospital inpatient stroke codes, Pharmacoepi demiol. Drug Saf., vol. 19, no. 8, pp. 843847, Aug. 2010.

Y.Chen,H.Cao,Q.Mei,K.Zheng,andH.Xu, Applyingactivelearning to supervised word sense disambiguation in MEDLINE, J. Amer. Med. Inform. Assoc., vol. 20, no. 5, pp. 10011006, Sep. 2013.

H. Liu, A multi-aspect comparison study of supervised word sense dis ambiguation, J. Amer. Med. Inform. Assoc., vol. 11, no. 4, pp. 320331, Apr. 2004.

M. J. Schuemie, J. A. Kors, and B. Mons, Word sense disambiguation in the biomedical domain: An overview, J. Comput. Biol., vol. 12, no. 5, pp. 554565, Jun. 2005.

H. Xu, M. Markatou, R. Dimova, H. Liu, and C. Friedman, Machine learning and word sense disambiguation in the biomedical domain: Design and evaluation issues, BMC Bioinf., vol. 7, no. 1, pp. 116, Dec. 2006.

Q.DongandY.Wang,

Enhancingmedical word sense inventories using word sense induction: A preliminary study, in Heterogeneous Data Management, Polystores, and Analytics for Healthcare. Springer, 2020, pp. 151167.

W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, Data processing and text mining technologies on electronic medical records: A review, J. Healthcare Eng., vol. 2018, Apr. 2018, Art. no. 4302425.

M. Allahyari, S. Pouriyeh, M. Asse , S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, A brief survey of text mining: Classi cation, clustering and extraction techniques, 2017, arXiv:1707.02919. [Online]. Available: http://arxiv.org/abs/1707.02919

A. niegula, A. Poniszewska-Mara da, and L. Chomtek, Towards the named entity recognition methods in biomedical eld, in Proc. Int. Conf. Current Trends Theory Pract. Inform. Springer, 2020, pp. 375387