Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Natural Language Processing (NLP) in Electronic Health Records (EHR) of Chronic Diseases in Relation to Healthcare Decision

Yujie Zhou[1, 2] and Songyan Yu[1, 2]

[1]School of Medicine, Southern University of Science and Technology, Shenzhen 518055, China

[2]Contribute equally

Corresponding author: Yujie Zhou (e-mail: 12110522@mail.sustech.edu.cn), Songyan Yu (e-mail: 12112023@mail.sustech.edu.cn).

ABSTRACT Electronic health records (EHR), as automated compilations of healthcare activities and assessments, have garnered attention in healthcare, health prevention, and research. The potential for extracting clinical insights from EHR relies on the advancements in natural language processing (NLP) methods. Against the backdrop of a continuously increasing global incidence of chronic diseases, approaches leveraging machine learning and deep learning to process EHR are enhancing the understanding of patient clinical trajectories and predicting disease risks. This provides an opportunity for early prevention and precision medicine for chronic diseases. This review summarizes machine learning and deep learning models used for natural language processing in EHR, along with their applications in chronic cardiovascular diseases and chronic digestive system diseases.

INDEX TERMS Cardiovascular disease; Chronic disease; Deep learning; Digestive system disease; Natural language processing; Machine learning.

I. INTRODUCTION

Chronic diseases also called noncommunicable diseases (NCDs) have been widely viewed as one of the main challenges of healthcare. According to WHO, chronic diseases kill 41 million people each year, equivalent to 74% of all deaths globally [1]. Although great progress has been made in discovering new treatments and prevention strategies for chronic diseases, the risk from chronic diseases still exists, which even evolves into a phenomenon of rising incidences of chronic diseases [2]. Therefore, totally different and novel approaches beyond clinical medicine itself should be carried out to mitigate the negative impact of chronic disease on the whole of society in the future.

Electronic health records (EHRs) are automated collections of clinical data generated in the healthcare activity and assessment [3]. The analysis based on EHR can improve understanding of the patient's clinical trajectory, provide possibilities for better patient stratification and risk prediction and inform clinical decision-making [4-6]. Considering that the long-term nature of chronic diseases can provide a very large and continuous stream of data, the reprocessing of EHR offers a potential direction for ameliorating chronic diseases, including delaying or preventing their onset.

Nowadays, EHRs have been increasingly popular and significant in clinical diagnosis, management and even research all over the world. For the United States, the adoption of EHRs had increased from 9.4% in 2008 to 96% in 2017 [7]. A similar trend has been seen in the health services of China, Australia and European countries [8, 9]. With the wide adoption and application of EHR, the volume of accumulated data in EHR including their modification has been so enormous that humans can't interpret every detail in the EHR. In addition, the noise, heterogeneity, random errors, incompleteness and even systematic biases aroused by such a large lumber of data lead to difficulty in integrating processing and modeling [10]. As a result, computer-based tools should be developed to mine, process and organize the information behind the data.

The components of EHR can be divided into structured data and unstructured data [11]. The structured EHR data is composed of numerical or categorical sources including laboratory values or prescriptions. On the contrary, clinical documentation or notes and discharge summaries containing free texts comprise the unstructured EHR data. For the structured data, several efforts have been made to increase the availability to it. For instance, a study showed a graphical approach to quantify the number of clinical notes above the structured data for rheumatoid arthritis and Alzheimer

disease [12]. In addition, a great deal of progress in treating the unstructured free texts in the medical area has been realized by machine learning (ML). For example, as an emerging ML-integrated tool, autoML has been applied in EHR analysis [13]. All of these have provided a possibility to handle EHR for clinical tasks, especially in chronic diseases. However, one of the main challenges for medical information studies is the unique language and clinical idioms used by clinicians.

Natural language processing (NLP) is a subfield of artificial intelligence (AI) techniques allowing the interactions between computers and human languages [14]. It has been used for clinical text mining, which bridges the gap between clinical human language and computational systems [15]. NLP consists of tasks that computationally use human languages such as written or spoken language to detect the underlying concepts, which meets the need to the extraction of the wealthy information about patient clinical history generally locked behind EHRs. Although the use of NLP in the clinical domain obtains an increasing uptake with diverse applications, few reviews have mainly focused on the EHR-NLP applications on chronic diseases which are represented by cardiovascular diseases and digestive system diseases. In this review, several basic NLP algorithms and models based on machine learning (ML) and deep learning (DL) and used for EHR processing to extract information about healthcare decisions will be concluded. Subsequently, the detailed cases of two typical chronic diseases, cardiovascular diseases and gastrointestinal diseases, will be discussed. In the end, several challenges and future trends in this area.

II. COMMONLY USED MODELS IN NLP FOR EHR-BASED HEALTHCARE DESITION-MAKING

A. ML MODELS

Machine learning (ML) is a field of computer science that empowers computer systems to progressively enhance performance from data using statistical techniques, without the need for explicit programming [16]. While early NLP methods enabled computers to handle and analyze text data written in human languages, the emergence of ML methods has allowed NLP to extract and measure more complex structures. Therefore, since 2015, there has been a growing preference for machine learning methodologies over rule-based approaches (Fig. 1) [17]. As a paradigm for analyzing EHRs, ML has gained tremendous development and is widely used to analyze EHR data in the last few years. Generally, machine learning methods can be categorized into supervised learning, semi-supervised, and unsupervised learning. Supervised learning utilizes labeled data to train algorithms for classification or accurate outcome prediction. Semi-supervised learning is a learning problem that involves a small number of labeled examples and a large number of unlabeled examples. In contrast, unsupervised learning utilizes unlabeled data for training by uncovering underlying structures. According to the quantity of published articles, it seems that algorithms pertaining to supervised learning have garnered greater favor. Nevertheless, a recent novel approach tackling NLP problems by combining unsupervised, supervised, and rule-based learning has demonstrated its potential for clinical NLP tasks. Within the research context, this method exhibited notable interpretability as well as superior performance [18].

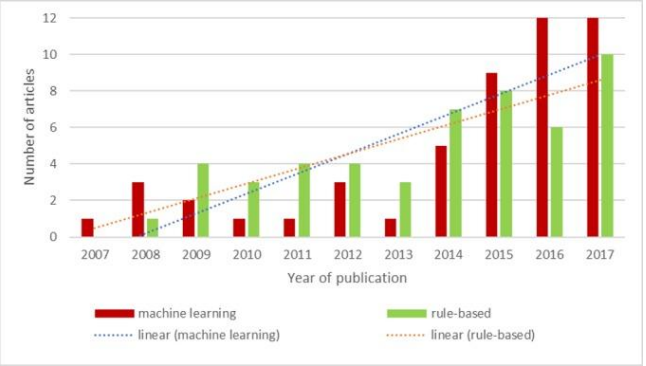

FIGURE 1. Natural language processing rule-based methods versus machine learning for chronic diseases.

1) SUPERVISED LEARNING

The most widely used supervised learning algorithms for dealing with EHR is support vector machine (SVM), followed by Naïve Bayes (NB) [19]. The prevalence of these two algorithms could be attributed to their popularity, relative simplicity, and lower demand for training data [19]. Moreover, logistic regression (LR), eXtreme Gradient Boosting (XGBoost), AdaBoost, random forest (RF), linear regression, gradient boosting (GB), and decision tree (DT) models have also been employed in EHR [20].

SVM has powerful capabilities in linear or nonlinear classification, regression, and even outlier detection tasks. It can be used for text classification, handwriting identification gene classification and so on. Clinical narratives exhibit characteristics of high-dimensional feature spaces, a few irrelevant features, and sparse instance vectors. SVM is considered to have good performance in solving these problems [21]. Therefore, the use of SVM is considered effective and has gained wide recognition. Several studies also have indicated that SVM outperforms Naïve Bayes, Bayesian networks, decision trees, and rule-based systems in text classification. Additionally, SVM active learning techniques have shown the potential to reduce the required sample size [22-24].

NB is a popular supervised machine learning algorithm used for classification tasks such as text classification. It is based on Bayes' Theorem with an independent assumption among predictors. NB doesn't require the inference of a dependency network. Moreover, they are convenient to apply when dealing with high-dimensional features. NB demonstrates the superiority of its algorithm when dealing

with large-scale data and effectively reduces the likelihood of overfitting [25]. Even in the presence of missing values, NB can learn and demonstrate less reliance on missing data imputation [26, 27]. Currently, NB has been employed to predict heart diseases in medical data, classify smoking status, search EMR records to identify multiple sclerosis and categorize obesity and cancer based on EMR records [10, 28-32].

Regression methods have been extensively employed for computational tasks over an extended period. They possess the advantage of being straightforward and convenient for model construction or adjustment, adjusting their parameters to maximize the conditional likelihood of the data. Further, regression models do not require a lot of effort in building or tuning, and the feature statistics derived from these regression models can be easily interpreted for meaningful insights.

2) UNSUPERVISED LEARNING

Due to the necessity for manual labeling in supervised learning, which demands high human resources, sample quantity, and quality, unsupervised learning, in contrast, offers an annotation-free approach that alleviates the burdensome labeling task, thereby enhancing the scalability of research and alleviating the tedious labeling task. Common unsupervised learning tasks include clustering and density estimation. Luo et al. reported that they established a model for clinical narrative texts using unsupervised learning. By introducing a new architecture called subgraph augmented non-negative tensor factorization (SANTF), they classified lymphoma patients into three subtypes, achieving an accuracy of 75% [33].

B. DL MODELS

The primary task of NLP is to provide an in-depth representation of the text or the language, including feature learning [34]. Conventional methods usually start with time-consuming handcrafting of features via careful human analysis of a specific application, and are followed by the development of algorithms to extract and utilize instances of those features. However, it has been proved that simple representations of the language or the text coupled with large amounts of data might work as well or better than more complex representations based on the instance of the bag-of-words (BoW) model [35]. Meanwhile, deep supervised feature learning methods are highly data-driven and can be used in more general efforts aimed at providing a robust data representation [36]. This is because deep learning (DL) can learn the features from unlabeled data to provide a low-dimensional representation of a high-dimensional data space. Then, DL takes the advantage in processing the vast amounts of unlabeled data in NLP. As a result, deep learning has become the precursor in NLP applications. Nowadays, the clinical NLP has also been revolutionarily reshaped by DL architectures. To be more specific, the novel DL models have been applied in various clinical NLP tasks, including classification, prediction, word embedding, medical text summarization, language modeling, ICD-9 classification, clinical notes analysis, mental health issue identification and medical dialogue analysis [20]. Meanwhile, the frequently used deep learning architectures to analyze EHR are introduced as follows.

1) CONVOLUTIONAL NEURAL NETWORKS

As a subclass of feed-forward neural networks, convolutional neural networks (CNNs) are inspired by the human visual cortex and based on the underlying mathematical operation, convolution, to measure the interoperability of its input functions [37]. In the cases of NLP, the inputs of sentences or documents are represented as matrices. In the training phase, most of the CNN structures learn word or sentence representations in which each matrix row is associated with a language element like a word or a character [38].

In the clinical field, CNN has been widely utilized in NLP tasks. For instance, Li et al. designed a deep learning system (DeepLabeler) to automatically classify international classification of diseases (ICD-9), which consists of CNN and the document-to-vector (D2V) technique to search and encode local and global features (Fig. 2). Specifically, this model achieved its target through feature extraction and multi-label classification. In addition, this architecture effectively extracted global and local features from the medical information mart for intensive care (MIMIC) dataset without any useful information loss. In addition, the multi-label classification step would utilize a fully connected neural network (FCNN) to anticipate the likelihood of each ICD-9 code [39].

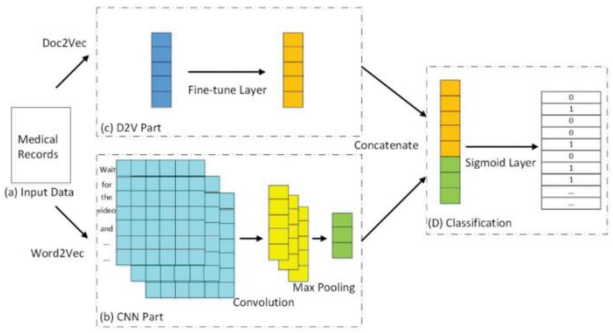

FIGURE 2. The architecture of the DeepLabeler based on CNN structure.

Also, several NLP solutions associated with CNN are not widely available. The clinical name entity recognition (NER) is one of examples. An NER model has been developed to extract several medical entities like drug names, route of administration, frequency, dosage, strength, form and duration from a large number of medical records. The core architecture of this proposed NER model is based on a CNN network in which the token representations are hashed

Bloom embeddings of specific word prefixes, suffixes and lemmatizations. For this model, the data from the United Kingdom Secondary Care Mental Health Record (CRIS) has evaluated its transferability [40].

2) RECURRENT NEURAL NETWORK AND LONG SHORT-TERM MEMORY

A recurrent neural network (RNN) is a sequence of feed-forward neural networks (FNNs) with the output of each FNN corresponding to the input of the next one. In all, layers in an RNN can be categorized into input, hidden, and output layers [34]. At each time step, predictions are made and parameters of the current hidden layer are used as input to the next time step. In particular, hidden layers in RNNs can carry information from the past, which is useful in language modeling like identifying the meaning behind a pronoun to work as the memory. Especially, the long short-term memory network (LSTM) is one of the most widely used classes of RNNs, aiming at capturing the long-time dependencies between inputs from different time steps [41].

The structures of RNN and LSTM have been utilized in the clinical text classification. A study showed a hybrid model of gated attention incorporated bidirectional long short-term memory (ABLSTM) and attention-based bidirectional LSTM to classify the clinical document [42]. To finish this NLP task, RNN was first added in this study to model time-sensitive sequences. Then, LSTM in this study served as "gates" to regulate or control the data flow to RNN. Furthermore, due to the disadvantages of "black box" methods when dealing with medical multi-class classification, the researcher introduced a bidirectional LSTM framework containing an attention layer to weigh the words in a phrase automatically based on the perceived relevance (Fig. 3).

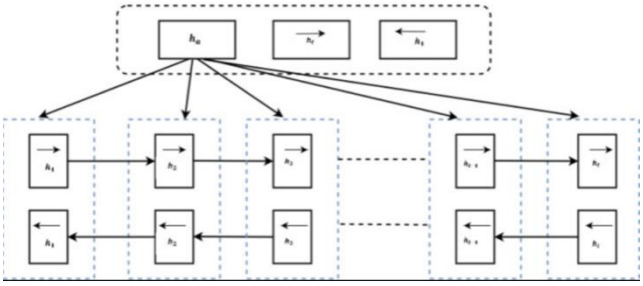

FIGURE 3. The architecture of the hybrid model of ABLSTM.

3) TRANSFORMER

In all, the transformer architecture utilizes a self-attention mechanism to capture long-range relationships in the input and to process input sequences with various lengths simultaneously. Vaswani et al. first proposed the transformer architecture in 2017 [43]. Since then, transformer-based architectures have created novel models for various NLP tasks, especially for the application of bidirectional encoder representation from transformer (BERT) on clinical records [44]. For example, a Multitask-Clinical BERT (MT-Clinical BERT) model can realize multitask learning on eight different information retrieval tasks, including clinical text embedding learning, entity retrieval, and the recognition of personal health indicators (PHI) (Fig. 4). At the same time, these embeddings serve as inputs to these prediction functions [45]. Interestingly, several transformer-based models have been trained on actual clinical data. MS-BERT created by Costa et al. can be the most representative one. To be more specific, this model has been trained on more than 70,000 medical notes from multiple sclerosis (MS) patients after de-identified processing. According to an evaluation of the expanded disability status scale (EDSS) from a study, MS-BERT performed better compared with models based on CNN [20].

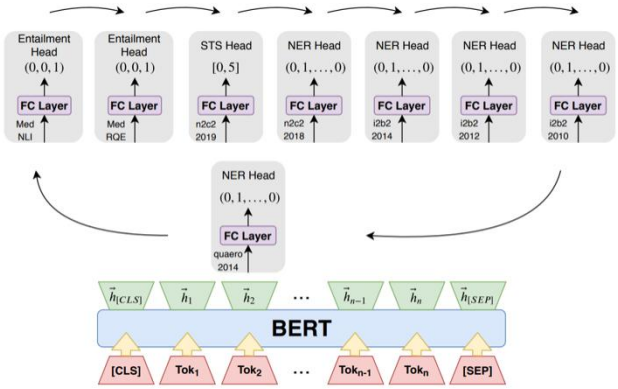

FIGURE 4. The architecture of the eight-headed MT-Clinical BERT.

III. APPLICATION ON THE NLP OF CHRONIC DISEASE'S EHR

A. CARDIOVASCULAR DISEASES

Chronic cardiovascular diseases include chronic heart failure, coronary heart disease, hypertension, chronic infectious endocarditis, chronic pericarditis, and so on. Cardiovascular diseases make a substantial contribution to the incidence and mortality rates among both men and women worldwide, affecting not only high-income countries but also middle and low-income nations. Globally, deaths caused by coronary heart disease account for 40% of the total mortality rate [46]. Presently, much of the focus is on utilizing NLP to the risk of heart diseases. As one of the major risk factors for individuals over 65 years old, developing NLP in EHR for analyzing patient data related to cardiovascular system diseases will be conducive to clinical and translational research. It assists clinical professionals in effectively extracting clinically significant information for guiding clinical decisions. Brodnick et al. developed a machine learning-based, stakeholder-informed, automated, NLP system to assess the quality of heart failure (HF) inpatient care. Through training and testing the congestive heart failure information extraction framework (CHIEF), it reliably captured clinical data, reducing or eliminating costs associated with human review of HF patient records [47].

Karystianis et al. evaluated the identification of cardiac risk factors in clinical notes of diabetic patients. The study demonstrated that the system, applied to 514 unseen assessments, showed relatively good outcomes in identifying coronary heart disease family history, medications, and some related disease factors (such as hypertension, diabetes, and hyperlipidemia), but faced challenges in identifying specific indicators of coronary heart disease [48]. Kim et al. developed a left ventricular ejection fraction (LVEF) extraction module aimed at identifying LVEF information from various types of clinical notes and using this information for heart failure quality measurements. Additionally, the authors indicated that when clinical data reports are highly structured, less training data is required. However, when reports are less structured or have rich vocabularies, combining the predictions of the existing LVEF extraction module can improve information extraction [49]. Similarly, Byrd et al. developed an NLP program to identify information and symptoms of HF in early-admission patients. They utilized clinical notes from EHR for early-pattern analysis of HF and to provide clinical decision support [50]. Other heart diseases have also received significant attention apart from heart failure. Yang et al. developed an information extraction system based on machine learning, rule-based methods, and dictionary-based keyword spotting to automatically identify coronary heart disease risk factors in medical records, addressing the inherent complexity of various risk factors within clinical contexts. The system exhibited good performance in the test data of the 2014 i2b2/UTHealth NLP Challenge [51]. Echocardiography is one of the most common diagnostic techniques in cardiology. Due to the lack of structure in most echocardiographic reports, analyzing them has been challenging in terms of scope of data retrieval, automation, and accuracy in earlier years. In recent years, Nath et al. proposed an NLP-based system, EchoInfer, capable of transforming heterogeneous echocardiography reports into structured data format on a large scale. It demonstrated high sensitivity and accuracy in extracting and identifying key indicators from echocardiography reports, offering potential clinical applications [52].

Peripheral artery disease (PAD) and coronary artery disease (CAD) are chronic diseases with high incidence and mortality rates, associated with increased risks of myocardial infarction and stroke [53, 54]. Several studies have reported their contributions to the automatic identification and extraction of clinical free-text and report data related to PAD and CAD. These contributions encompass utilizing new text analysis techniques to quantify adverse events associated with the only FDA-approved drug, cilostazol, used in clinical settings for treating PAD, to evaluate drug effectiveness and safety [55]. Additionally, automatic identification of PAD cases from clinical narrative notes [56], followed by the same team's expansion of the PAD identification algorithm to develop a subtyping NLP algorithm for identifying complications of late-stage peripheral arterial disease—severe limb ischemia cases from clinical notes [14]. Furthermore, the development of models based on Naïve Bayes, Maximum Entropy, and SVM to automatically predict CAD development from clinical free-text [57], along with mining CAD risk factor-related data from unstructured clinical narratives for assessment [58].

Hypertension is also a common chronic disease within the circulatory system, with a relatively high incidence rate. Among individuals aged 60 and above, the prevalence of hypertension exceeds 60% [59]. Many efforts have been made in recent years towards the assessment of hypertension. It has been reported that in the classification of hypertension, billing codes or blood pressure readings have shown good performance in hypertension classification. Even simple combinations of input categories can enhance performance. Sophisticated algorithms can achieve highly accurate classifications [60]. However, consensus has not been reached on the definition of hypertension monitoring based on EHRs. Applying different criteria to define hypertension using EHR data has a large effect on hypertension prevalence estimates. Hohman et al. recently proposed an electronic phenotype for hypertension monitoring in EHRs. Their work emphasizes the substantial impact of different analytical decisions on defining the numerator and denominator in EHR-based estimates of chronic disease prevalence and control, contributing to standardization efforts [61]. Additionally, Martin et al. recently utilized EHR data to develop a hypertension identification algorithm based on the Gradient Boosting algorithm XGBoost [62].

B. DIGESTIVE SYSTEM DISEASES

Digestive system diseases are highly prevalent worldwide, cause considerable distress, and can be fatal. To be more specific, the impacts of digestive system diseases in 2019 could be equivalent to the one healthy year loss of 88.99 million people [63]. Digestive system diseases include organic and functional diseases of the esophagus, stomach, intestine, liver, biliary, pancreas and other organs, which are very common in clinics. As a significant part of digestive system diseases, chronic digestive diseases such as cirrhosis and other chronic liver diseases constituted the highest proportion of categorized digestive disease disability-adjusted life-year (DALY) burdens globally [64]. Despite the high prevalence of these diseases, a large number of chronic digestive system diseases are underdiagnosed and lack systematic screening protocols. For instance, although hepatic steatosis presents in approximately 25% of the US population, research on the natural history with the use of EHR data can be difficult due to the requirement of long-term follow-up and the lack of gold standard for diagnosis [65, 66]. In recent years, a vast number of studies have proposed various NLP methods to extract information relevant to medical decisions from the EHR of chronic digestive system diseases.

Firstly, for chronic gastric diseases such as gastritis and gastric cancer, they are usually screened by esophagogastroduodenoscopies (EGDs) [67]. However, on the one hand, the unstructured format of the reports and usage of endoscopic abbreviations made it hard to extract information about diagnosis and specific phenotypes of gastric disease. On the other hand, the endoscopic procedures discord with the pathologic reports in EHRs. Then, a group of researchers developed an effective NLP-based pipeline including text preprocessing, concept mapping, concept extraction and summarizing to extract gastritis-associated information on the presence and anatomical extent or degree of the lesion and gastric-cancer-associated information on the presence, anatomic location, size of the lesion and cancer classification from the EGDs and pathologic reports in EHRs. Furthermore, the feasibility of this NLP-based algorithm built on Python 3.7 and the regular expression package 're' has been verified by 1000 EHRs over 10 years [68]. Similarly, Soroush et al. designed an NLP system based on regular expressions and MetaMapLite to extract endoscopy-related quality metrics such as dysplasia and intestinal metaplasia in Barrett's Esophagus (BE) patients' EHRs for BE-related treatment [69].

In addition, various NLP-based information extraction methods are also widely used for EHRs of patients with gut-related chronic diseases. For inflammatory bowel disease, extraintestinal manifestations (EIMs) are important symptomatic components of it. EIMs can impact various organs to cause inflammation like iritis and are closely associated with the disease course, clinical outcomes, the medication need and increased relapse rate of IBD [70, 71]. However, due to the unconformity in the description of EIM occurrence along with unreliable and inaccurate diagnostic codes, it is hard to understand and extract EIM-related information from EHRs [72]. A study showed a new NLP pipeline to realize automatic detection of EIMs and inference of EIM activities from the EHRs to better diagnose IBD. In this research, Ryan et al. conducted an NLP flow including document preparation, identification of EIM keywords and concepts, tokenization of EIM description window and status concept identification, negation detection, EIM document section identification, and document-level EIM status determination to predict EIM status. Especially, the keys to implementing this NLP approach comprise the use of Natural Language Tool Kit functions from Python modules, the use of SecTag to localize the EIM-associated document and document section prioritization [73]. As another significant characteristic of IBD, Crohn's disease (CD) has also been involved in the application of clinical NLP. A group developed an initial NLP model to describe the clinical characteristics of patients with CD and generate a predictive model for the CD relapses, which identified information from patients' EHRs and utilized ML algorithms such as logistic regression, decision tree and random forest [74]. Meanwhile, several researchers have focused on the celiac disease. Chen et al. designed a NLP system in 2016 to improve the identification of celiac disease patients based on the pathology reports in EHRs, which was achieved with the help of n-gram feature extraction by Java-based Weka and classification model from Bayes, function-based, lazy model and tree classifiers [75]. Furthermore, colorectal cancer (CRC) is a kind of malignant tumor to cause the third rank globally in incidence [76]. A large number of NLP models have been applied to treat CRC-related EHRs. For instance, the combination method of SVM and feature selection has been designed to extract the hidden information indicating the CRC-related complication, anastomosis leakage [77]. Also, the relationship between time course and symptoms of CRC and the onset of CRC based on family history have been realized by improved NLP-based methods to extract information from EHRs [78, 79].

Although some criteria such as mild elevations in aspartate aminotransferase (AST) and alanine aminotransferase (ALT) levels can determine the occurrence of chronic liver disease, the information hidden in the unstructured text of the EHR needs to be further explored to quickly identify liver-related chronic diseases [80]. Then, several NLP methods can solve this. For instance, researchers have utilized an NLP algorithm with a Linguamatics literature text mining tool to identify undiagnosed hepatic steatosis with EHRs [65]. Also, a commercial model, CLiX clinical NLP engine, had been improved to better reflect progressive risks for non-alcoholic fatty liver disease with the process of unstructured data in EHRs [81].

IV. DISCUSSION

NLP has found extensive applications in handling clinical notes of various chronic diseases. Advances in machine learning and deep learning models can facilitate health learning tasks in Electronic Health Records (EHRs), enhancing understanding of patient clinical trajectories and predicting chronic disease risks, thus driving forward intelligent healthcare. In this review, we primarily summarize the algorithmic foundation of NLP for EHRs and its specific applications in chronic diseases of the circulatory and digestive systems.

While machine learning occupies a considerable portion of chronic disease EHRs, deep learning algorithms exhibit a rapid growth trend and demonstrate significant potential. However, the development of deep learning still faces challenges. Compared to Machine Learning models commonly used in health records, Deep Learning models have room for improvement in data availability, complexities of specific domain text data, and interpretability. Furthermore, DL-based algorithms require substantial data for superior performance over other algorithms, imposing significant demands on financial support and workstation capacity.

Although numerous studies indicate the use of machine learning to automatically detect and predict patient safety

events, many of these algorithms lack external validation or prospective testing. Thus, further research is needed to enhance the performance of these automated systems.

Issues related to the accuracy of medical text spelling and abbreviations increase the time and difficulty required for model training. Due to the specificity and linguistic diversity of medical terminologies, some doctors use Latin abbreviations to specify drug frequency (e.g., "BD" spelled as "bis die") or use conventional abbreviations to spell diseases (e.g., "cancer" spelled as "CA"). This makes it challenging for computers to correctly identify these complex abbreviation patterns.

Data scarcity remains a significant challenge in medical NLP research. A large amount of data is a prerequisite for model accuracy. However, due to various issues concerning patient privacy, ethics, and other considerations associated with EHRs, acquiring substantial data can be challenging, with healthcare systems hesitating to provide patient data. One potential solution is utilizing synthetic data [82]. However, the practicality of using machine-generated data for machine training needs careful consideration.

In summary, NLP has made significant strides in EHRs. The advent of deep learning-based algorithms will expedite its development. Future research should focus on enhancing accuracy and practical clinical translation.

REFERENCES

[1] WHO. "SDG Target 3.4 Non-communicable diseases and mental health." https://www.who.int/data/gho/data/themes/topics/sdg-target-3_4-noncommunicable-diseases-and-mental-health. (accessed.

[2] S. Mendis, S. Davis, and B. Norrving, "Organizational update: the world health organization global status report on noncommunicable diseases 2014; one more landmark step in the combat against stroke and vascular disease," (in eng), *Stroke,* vol. 46, no. 5, pp. e121-e122, 2015, doi: 10.1161/STROKEAHA.115.008097.

[3] M. R. Cowie *et al.*, "Electronic health records to facilitate clinical research," *Clinical Research in Cardiology,* vol. 106, pp. 1-9, 2017.

[4] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records," (in eng), *Scientific Reports,* vol. 6, p. 26094, 2016, doi: 10.1038/srep26094.

[5] C. Ye *et al.*, "Prediction of Incident Hypertension Within the Next Year: Prospective Study Using Statewide Electronic Health Records and Machine Learning," (in eng), *Journal of Medical Internet Research,* vol. 20, no. 1, p. e22, 2018, doi: 10.2196/jmir.9268.

[6] B. A. Goldstein, A. M. Navar, M. J. Pencina, and J. P. A. Ioannidis, "Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review," (in eng), *J Am Med Inform Assoc,* vol. 24, no. 1, pp. 198-208, 2017, doi: 10.1093/jamia/ocw042.

[7] J. Liang *et al.*, "Adoption of electronic health records (EHRs) in China during the past 10 years: consecutive survey data analysis and comparison of Sino-American challenges and experiences," *Journal of medical Internet research,* vol. 23, no. 2, p. e24813, 2021.

[8] A. J. Hodgkins, J. Mullan, D. J. Mayne, C. S. Boyages, and A. Bonney, "Australian general practitioners' attitudes to the extraction of research data from electronic health records," *Australian journal of general practice,* vol. 49, no. 3, pp. 145-150, 2020.

[9] K. A. Cairns *et al.*, "Building on antimicrobial stewardship programs through integration with electronic medical records: the Australian experience," *Infectious diseases and therapy,* vol. 10, pp. 61-73, 2021.

[10] K. Jensen *et al.*, "Analysis of free text in electronic health records for identification of cancer patient trajectories," (in eng), *Scientific Reports,* vol. 7, p. 46226, 2017, doi: 10.1038/srep46226.

[11] H. Consultant, "Why unstructured data holds the key to intelligent healthcare systems [Internet]. Atlanta (GA): HIT Consultant; 2015. cited at 2019 Jan 15," ed, 2015.

[12] 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 performance," (in eng), *J Am Med Inform Assoc,* vol. 23, no. e1, pp. e20-e27, 2016, doi: 10.1093/jamia/ocv130.

[13] J. Waring, C. Lindvall, and R. Umeton, "Automated machine learning: Review of the state-of-the-art and opportunities for healthcare," *Artificial intelligence in medicine,* vol. 104, p. 101822, 2020.

[14] N. Afzal *et al.*, "Natural language processing of clinical notes for identification of critical limb ischemia," (in eng), *Int J Med Inform,* vol. 111, pp. 83-89, 2018, doi: 10.1016/j.ijmedinf.2017.12.024.

[15] Y. Juhn and H. Liu, "Artificial intelligence approaches using natural language processing to advance EHR-based clinical research," (in eng), *J Allergy Clin Immunol,* vol. 145, no. 2, pp. 463-469, 2020, doi: 10.1016/j.jaci.2019.12.897.

[16] J. B. Edgcomb and B. Zima, "Machine Learning, Natural Language Processing, and the Electronic Health Record: Innovations in Mental Health Services Research," (in eng), *Psychiatr Serv,* vol. 70, no. 4, pp. 346-349, 2019, doi: 10.1176/appi.ps.201800401.

[17] S. Sheikhalishahi, R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi, and V. Osmani, "Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review," *JMIR Med Inform,* vol. 7, no. 2, p. e12239, Apr 27 2019, doi: 10.2196/12239.

[18] G. T. Berge, O.-C. Granmo, T. O. Tveit, A. L. Ruthjersen, and J. Sharma, "Combining unsupervised, supervised and rule-based learning: the case of detecting patient allergies in electronic health records," (in eng), *BMC Med Inform Decis Mak,* vol. 23, no. 1, p. 188, 2023, doi: 10.1186/s12911-023-02271-8.

[19] S. Sheikhalishahi, R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi, and V. Osmani, "Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review," (in eng), *JMIR Med Inform,* vol. 7, no. 2, p. e12239, 2019, doi: 10.2196/12239.

[20] E. Hossain *et al.*, "Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review," (in eng), *Comput Biol Med,* vol. 155, p. 106649, 2023, doi: 10.1016/j.compbiomed.2023.106649.

[21] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," in *Machine Learning: ECML-98*, Berlin, Heidelberg, C. Nédellec and C. Rouveirol, Eds., 1998// 1998: Springer Berlin Heidelberg, pp. 137-142.

[22] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," in *Proceedings of the seventh international conference on Information and knowledge management*, 1998, pp. 148-155.

[23] R. J. Carroll, A. E. Eyler, and J. C. Denny, "Naïve Electronic Health Record phenotype identification for Rheumatoid arthritis," (in eng), *AMIA Annu Symp Proc,* vol. 2011, pp. 189-196, 2011. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/22195070.

[24] W.-Q. Wei, C. Tao, G. Jiang, and C. G. Chute, "A high throughput semantic concept frequency based approach for patient identification: a case study using type 2 diabetes

mellitus clinical notes," (in eng), *AMIA Annu Symp Proc,* vol. 2010, pp. 857-861, 2010. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/21347100.

[25] J. Huang, J. J. Lu, and C. X. Ling, "Comparing naive bayes, decision trees, and SVM with AUC and, accuracy," (in English), *Third Ieee International Conference on Data Mining, Proceedings,* pp. 553-556, 2003. [Online]. Available: <Go to ISI>://WOS:000188999400080.

[26] Y. Luo, P. Szolovits, A. S. Dighe, and J. M. Baron, "Using Machine Learning to Predict Laboratory Test Results," (in English), *Am J Clin Pathol,* vol. 145, no. 6, pp. 778-788, Jun 2016, doi: 10.1093/Ajcp/Aqw064.

[27] Y. Luo, P. Szolovits, A. S. Dighe, and J. M. Baron, "3D-MICE: integration of cross-sectional and longitudinal imputation for multi-analyte longitudinal clinical data," (in English), *J Am Med Inform Assn,* vol. 25, no. 6, pp. 645-653, Jun 2018, doi: 10.1093/jamia/ocx133.

[28] M. Torii *et al.*, "Risk factor detection for heart disease by applying text analytics in electronic medical records," (in eng), *J Biomed Inform,* vol. 58 Suppl, no. Suppl, pp. S164-S170, 2015, doi: 10.1016/j.jbi.2015.08.011.

[29] J. Jonnagaddala, S.-T. Liaw, P. Ray, M. Kumar, H.-J. Dai, and C.-Y. Hsu, "Identification and Progression of Heart Disease Risk Factors in Diabetic Patients from Longitudinal Electronic Health Records," (in eng), *Biomed Res Int,* vol. 2015, p. 636371, 2015, doi: 10.1155/2015/636371.

[30] H. S. Chase, L. R. Mitrani, G. G. Lu, and D. J. Fulgieri, "Early recognition of multiple sclerosis using natural language processing of the electronic health record," (in eng), *BMC Med Inform Decis Mak,* vol. 17, no. 1, p. 24, 2017, doi: 10.1186/s12911-017-0418-4.

[31] R. L. Figueroa and C. A. Flores, "Extracting Information from Electronic Medical Records to Identify the Obesity Status of a Patient Based on Comorbidities and Bodyweight Measures," (in eng), *J Med Syst,* vol. 40, no. 8, p. 191, 2016, doi: 10.1007/s10916-016-0548-8.

[32] S. N. Kasthurirathne *et al.*, "Toward better public health reporting using existing off the shelf approaches: The value of medical dictionaries in automated cancer detection using plaintext medical data," (in eng), *J Biomed Inform,* vol. 69, pp. 160-176, 2017, doi: 10.1016/j.jbi.2017.04.008.

[33] Y. Luo, Y. Xin, E. Hochberg, R. Joshi, O. Uzuner, and P. Szolovits, "Subgraph augmented non-negative tensor factorization (SANTF) for modeling clinical narrative text," (in eng), *J Am Med Inform Assoc,* vol. 22, no. 5, pp. 1009-1019, 2015, doi: 10.1093/jamia/ocv016.

[34] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, "Natural language processing advancements by deep learning: A survey," *arXiv preprint arXiv:2003.01200,* 2020.

[35] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," *arXiv preprint cs/0205070,* 2002.

[36] S. Sun and M. Iyyer, "Revisiting simple neural probabilistic language models," *arXiv preprint arXiv:2104.03474,* 2021.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems,* vol. 25, 2012.

[38] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882,* 2014.

[39] M. Li *et al.*, "Automated ICD-9 coding via a deep learning approach," *IEEE/ACM transactions on computational biology and bioinformatics,* vol. 16, no. 4, pp. 1193-1202, 2018.

[40] A. Kormilitzin, N. Vaci, Q. Liu, and A. Nevado-Holgado, "Med7: A transferable clinical natural language processing model for electronic health records," *Artificial Intelligence in Medicine,* vol. 118, p. 102086, 2021.

[41] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation,* vol. 9, no. 8, pp. 1735-1780, 1997.

[42] X. Li, M. Cui, J. Li, R. Bai, Z. Lu, and U. Aickelin, "A hybrid medical text classification framework: Integrating attentive rule construction and neural network," *Neurocomputing,* vol. 443, pp. 345-355, 2021.

[43] A. Vaswani *et al.*, "Attention Is All You Need," (in English), *Adv Neur In,* vol. 30, 2017. [Online]. Available: <Go to ISI>://WOS:000452649406008.

[44] D. Kici, G. Malik, M. Cevik, D. Parikh, and A. Basar, "A BERT-based transfer learning approach to text classification on software requirements specifications," in *Canadian Conference on AI*, 2021, vol. 1, p. 042077.

[45] A. Mulyar and B. T. McInnes, "MT-Clinical BERT: Scaling Clinical Information Extraction with Multitask Learning," p. arXiv:2004.10220doi: 10.48550/arXiv.2004.10220.

[46] S. A. Phillips and M. Guazzi, "The vasculature in cardiovascular diseases: will the vasculature tell us what the future holds?," (in eng), *Prog Cardiovasc Dis,* vol. 57, no. 5, pp. 407-408, 2015, doi: 10.1016/j.pcad.2014.12.004.

[47] J. H. Garvin *et al.*, "Automating Quality Measures for Heart Failure Using Natural Language Processing: A Descriptive Study in the Department of Veterans Affairs," (in eng), *JMIR Med Inform,* vol. 6, no. 1, p. e5, 2018, doi: 10.2196/medinform.9150.

[48] G. Karystianis, A. Dehghan, A. Kovacevic, J. A. Keane, and G. Nenadic, "Using local lexicalized rules to identify heart disease risk factors in clinical notes," (in eng), *J Biomed Inform,* vol. 58 Suppl, no. Suppl, pp. S183-S188, 2015, doi: 10.1016/j.jbi.2015.06.013.

[49] Y. Kim *et al.*, "Extraction of left ventricular ejection fraction information from various types of clinical reports," (in eng), *J Biomed Inform,* vol. 67, pp. 42-48, 2017, doi: 10.1016/j.jbi.2017.01.017.

[50] R. J. Byrd, S. R. Steinhubl, J. Sun, S. Ebadollahi, and W. F. Stewart, "Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records," (in eng), *Int J Med Inform,* vol. 83, no. 12, pp. 983-992, 2014, doi: 10.1016/j.ijmedinf.2012.12.005.

[51] H. Yang and J. M. Garibaldi, "A hybrid model for automatic identification of risk factors for heart disease," (in eng), *J Biomed Inform,* vol. 58 Suppl, no. Suppl, pp. S171-S182, 2015, doi: 10.1016/j.jbi.2015.09.006.

[52] C. Nath, M. S. Albaghdadi, and S. R. Jonnalagadda, "A Natural Language Processing Tool for Large-Scale Data Extraction from Echocardiography Reports," (in eng), *PLoS One,* vol. 11, no. 4, p. e0153749, 2016, doi: 10.1371/journal.pone.0153749.

[53] A. T. Hirsch *et al.*, "ACC/AHA 2005 guidelines for the management of patients with peripheral arterial disease (lower extremity, renal, mesenteric, and abdominal aortic): executive summary a collaborative report from the American Association for Vascular Surgery/Society for Vascular Surgery, Society for Cardiovascular Angiography and Interventions, Society for Vascular Medicine and Biology, Society of Interventional Radiology, and the ACC/AHA Task Force on Practice Guidelines (Writing Committee to Develop Guidelines for the Management of Patients With Peripheral Arterial Disease) endorsed by the American Association of Cardiovascular and Pulmonary Rehabilitation; National Heart, Lung, and Blood Institute; Society for Vascular Nursing; TransAtlantic Inter-Society Consensus; and Vascular Disease Foundation," (in eng), *J Am Coll Cardiol,* vol. 47, no. 6, pp. 1239-1312, 2006. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/16545667.

[54] A. K. Malakar, D. Choudhury, B. Halder, P. Paul, A. Uddin, and S. Chakraborty, "A review on coronary artery disease, its risk factors, and therapeutics," (in eng), *J Cell Physiol,* vol. 234, no. 10, pp. 16812-16823, 2019, doi: 10.1002/jcp.28350.

[55] N. J. Leeper, A. Bauer-Mehren, S. V. Iyer, P. Lependu, C. Olson, and N. H. Shah, "Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes," (in eng), *PLoS One,* vol. 8, no. 5, p. e63499, 2013, doi: 10.1371/journal.pone.0063499.

[56] N. Afzal *et al.*, "Mining peripheral arterial disease cases from narrative clinical notes using natural language processing," (in

eng), *J Vasc Surg,* vol. 65, no. 6, pp. 1753-1761, 2017, doi: 10.1016/j.jvs.2016.11.031.
[57] K. Buchan, M. Filannino, and Ö. Uzuner, "Automatic prediction of coronary artery disease from clinical narratives," (in eng), *J Biomed Inform,* vol. 72, pp. 23-32, 2017, doi: 10.1016/j.jbi.2017.06.019.
[58] J. Jonnagaddala, S.-T. Liaw, P. Ray, M. Kumar, N.-W. Chang, and H.-J. Dai, "Coronary artery disease risk assessment from unstructured electronic health records using text mining," (in eng), *J Biomed Inform,* vol. 58 Suppl, no. Suppl, pp. S203-S210, 2015, doi: 10.1016/j.jbi.2015.08.003.
[59] C. K. Chow *et al.*, "Prevalence, awareness, treatment, and control of hypertension in rural and urban communities in high-, middle-, and low-income countries," (in eng), *JAMA,* vol. 310, no. 9, pp. 959-968, 2013, doi: 10.1001/jama.2013.184182.
[60] P. L. Teixeira *et al.*, "Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals," (in eng), *J Am Med Inform Assoc,* vol. 24, no. 1, pp. 162-171, 2017, doi: 10.1093/jamia/ocw071.
[61] K. H. Hohman *et al.*, "Development of a Hypertension Electronic Phenotype for Chronic Disease Surveillance in Electronic Health Records: Key Analytic Decisions and Their Effects," (in eng), *Prev Chronic Dis,* vol. 20, p. E80, 2023, doi: 10.5888/pcd20.230026.
[62] E. A. Martin, A. G. D'Souza, S. Lee, C. Doktorchik, C. A. Eastwood, and H. Quan, "Hypertension identification using inpatient clinical notes from electronic medical records: an explainable, data-driven algorithm study," (in eng), *CMAJ Open,* vol. 11, no. 1, pp. E131-E139, 2023, doi: 10.9778/cmajo.20210170.
[63] R. Wang, Z. Li, S. Liu, and D. Zhang, "Global, regional, and national burden of 10 digestive diseases in 204 countries and territories from 1990 to 2019," (in eng), *Front Public Health,* vol. 11, p. 1061453, 2023, doi: 10.3389/fpubh.2023.1061453.
[64] A. F. Peery *et al.*, "Burden and Cost of Gastrointestinal, Liver, and Pancreatic Diseases in the United States: Update 2021," (in eng), *Gastroenterology,* vol. 162, no. 2, pp. 621-644, 2022, doi: 10.1053/j.gastro.2021.10.017.
[65] C. V. Schneider *et al.*, "Large-scale identification of undiagnosed hepatic steatosis using natural language processing," (in eng), *EClinicalMedicine,* vol. 62, p. 102149, 2023, doi: 10.1016/j.eclinm.2023.102149.
[66] T. Arshad, P. Golabi, L. Henry, and Z. M. Younossi, "Epidemiology of Non-alcoholic Fatty Liver Disease in North America," (in eng), *Curr Pharm Des,* vol. 26, no. 10, pp. 993-997, 2020, doi: 10.2174/1381612826666200303114934.
[67] N. H. I. Service, "National health screening statistical yearbook," *Seoul: National Health Insurance Service,* 2014.
[68] G. Song *et al.*, "Natural Language Processing for Information Extraction of Gastric Diseases and Its Application in Large-Scale Clinical Research," (in eng), *Journal of Clinical Medicine,* vol. 11, no. 11, 2022, doi: 10.3390/jcm11112967.
[69] A. Soroush *et al.*, "Natural Language Processing Can Automate Extraction of Barrett's Esophagus Endoscopy Quality Metrics," (in eng), *medRxiv,* 2023, doi: 10.1101/2023.07.11.23292529.
[70] M. Harbord *et al.*, "The First European Evidence-based Consensus on Extra-intestinal Manifestations in Inflammatory Bowel Disease," (in eng), *J Crohns Colitis,* vol. 10, no. 3, pp. 239-254, 2016, doi: 10.1093/ecco-jcc/jjv213.
[71] S. Jansson, M. Malham, A. Paerregaard, C. Jakobsen, and V. Wewer, "Extraintestinal Manifestations Are Associated With Disease Severity in Pediatric Onset Inflammatory Bowel Disease," (in eng), *J Pediatr Gastroenterol Nutr,* vol. 71, no. 1, pp. 40-45, 2020, doi: 10.1097/MPG.0000000000002707.
[72] A. N. Ananthakrishnan *et al.*, "Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach," (in eng), *Inflamm Bowel Dis,* vol. 19, no. 7, pp. 1411-1420, 2013, doi: 10.1097/MIB.0b013e31828133fd.
[73] R. W. Stidham *et al.*, "Identifying the Presence, Activity, and Status of Extraintestinal Manifestations of Inflammatory Bowel Disease Using Natural Language Processing of Clinical Notes," (in eng), *Inflamm Bowel Dis,* vol. 29, no. 4, pp. 503-510, 2023, doi: 10.1093/ibd/izac109.
[74] F. Gomollón *et al.*, "Clinical characteristics and prognostic factors for Crohn's disease relapses using natural language processing and machine learning: a pilot study," (in eng), *Eur J Gastroenterol Hepatol,* vol. 34, no. 4, pp. 389-397, 2022, doi: 10.1097/MEG.0000000000002317.
[75] W. Chen, Y. Huang, B. Boyle, and S. Lin, "The utility of including pathology reports in improving the computational identification of patients," (in eng), *J Pathol Inform,* vol. 7, p. 46, 2016. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/27994938.
[76] Y. Zhou, S. Yu, and W. Zhang, "NOD-like Receptor Signaling Pathway in Gastrointestinal Inflammatory Diseases and Cancers," *International Journal of Molecular Sciences,* vol. 24, no. 19, p. 14511, 2023. [Online]. Available: https://www.mdpi.com/1422-0067/24/19/14511.
[77] C. Soguero-Ruiz *et al.*, "Support Vector Feature Selection for Early Detection of Anastomosis Leakage From Bag-of-Words in Electronic Health Records," (in eng), *IEEE J Biomed Health Inform,* vol. 20, no. 5, pp. 1404-1415, 2016, doi: 10.1109/JBHI.2014.2361688.
[78] J. C. Denny *et al.*, "Extracting timing and status descriptors for colonoscopy testing from electronic medical records," (in eng), *J Am Med Inform Assoc,* vol. 17, no. 4, pp. 383-388, 2010, doi: 10.1136/jamia.2010.004804.
[79] D. L. Mowery *et al.*, "Determining Onset for Familial Breast and Colorectal Cancer from Family History Comments in the Electronic Health Record," (in eng), *AMIA Jt Summits Transl Sci Proc,* vol. 2019, pp. 173-181, 2019. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31258969.
[80] D. Amarapurkar *et al.*, "Prevalence of non-alcoholic fatty liver disease: population based study," (in eng), *Ann Hepatol,* vol. 6, no. 3, pp. 161-163, 2007. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/17786142.
[81] T. T. Van Vleck *et al.*, "Augmented intelligence with natural language processing applied to electronic health records for identifying patients with non-alcoholic fatty liver disease at risk for disease progression," (in eng), *International Journal of Medical Informatics,* vol. 129, pp. 334-341, 2019, doi: 10.1016/j.ijmedinf.2019.06.028.
[82] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. Schuller, "Survey of Deep Representation Learning for Speech Emotion Recognition," (in English), *Ieee T Affect Comput,* vol. 14, no. 2, pp. 1634-1654, Apr-Jun 2023, doi: 10.1109/Taffc.2021.3114365.