Proceedings of the 2nd International Conference on Modern Medicine and Global Health
DOI: 10.54254/2753-8818/46/20240618

Applications of machine learning to electronic health record data in liver-related disease

Jie Luo[1,6,†], Yuqi Sun[2,7,†], Jiachen Liu[3,5,8,†], Yu Zhou[4,9,†]

[1]College of computer science and cyber security (Oxford Brookes College), Chengdu University of Technology. Chengdu,610059, China
[2] Faculty of engineering, University of Sydney, Sydney, NSW2050, Australia
[3]Faculty of Mathematics, University of Waterloo, Waterloo, N2L 3G1, Canada
[4]Collage of information Engineering, Shanghai Maritime University, Shanghai, 200135, China
[5]Corresponding author

[6]KevinLuo021@gmail.com
[7]ysun2877@uni.sydney.edu.au
[8]a326liu@uwaterloo.ca
[9]1293570363@qq.com
[†]All the authors contributed equally to this work and should be considered as co-fist author.

Abstract. Electronic Health Records (EHRs) has gained its increasing significance in modern healthcare as its promising prospects in the application of machine learning. The accumulation of vast clinical data holds potential for repurposing in clinical research such as prediction, diagnosis and prognosis. However, the collection and preparation of EHR data present challenges, primarily due to the inherent incompleteness of the data and the associated privacy and security concerns. To address the issue, text-mining tools based on domain-specific lexicons, data sharing and multiple de-identification methods have been suggested. In terms of methodologies, various machine learning models and algorithms used in EHR data analysis are analyzed, including logistic regression, decision trees, random forests, and natural language processing, each with its unique application scenarios in the healthcare domain. Liver related diseases, including HAV, HBV, HCV and especially Liver Cancer, has affected hundreds of millions of people around the world. The incidence and mortality rates for these diseases are still rising continually. With recent advancements of Machine Learning techniques, such as the attention mechanism and BERT-based embedding, which have shown exceptional results in EHR analysis when applying to liver diseases. While EHRs offer a treasure trove of data for clinical research, the challenges associated with their collection, processing, and analysis cannot be ignored. It underscores the need for robust methodologies and tools to harness the full potential of EHRs while ensuring data integrity and patient privacy. In this paper, we will gather and review the existing application in the realm of liver-related diseases.

Keyword: EHR, machine learning, deep learning, text-mining, embedding, NLP, liver-related disease

1. Introduction

Liver cirrhosis, the end-stage consequence of various chronic liver diseases, is characterized by the accumulation of fibrillar collagen-rich extracellular matrix[1]. This condition signifies a critical juncture wherein liver function becomes severely compromised, leading to a cascade of potential complications. Ascites and bleeding varices stand out as common and life-threatening consequences of cirrhosis.

The liver, a vital organ with multifaceted functions, plays a pivotal role in metabolism, detoxification, and protein synthesis. Its well-being is crucial for overall health, but a range of liver diseases pose significant global health challenges. Concurrently, the integration of machine learning in healthcare offers promising avenues for improving clinical decision-making and early disease detection.

Hepatitis A virus (HAV) globally causes hepatitis through oral-fecal transmission, leading to liver inflammation and damage. Symptoms range from mild to severe, with untreated cases potentially resulting in cirrhosis, cancer, and death.[2]. Chronic HBV infection affects 350 million globally, causing 1 million annual deaths from liver disease. Immune response often leads to cirrhosis, liver failure, or HCC in 40%. Factors like age, gender, immune/viral status, and external influences impact progression.[3]. HCV infection, a global pandemic with 170 million cases, is 5x more common than HIV-1[4]. Mainly transmitted through blood (needles, drug use), it’s widespread, especially in specific populations. Most HCV-infected develop chronic disease, a major reason for liver transplants. [4].

Liver cancer (Hepatocellular Carcinoma or HCC) is a global, aggressive malignancy from liver cells, often linked to pre-existing liver conditions like cirrhosis. It ranks fourth in cancer-related deaths worldwide. By 2025, over 10,000 people are projected to be diagnosed annually [5], with rising incidence and mortality rates, posing significant health and societal challenges.

Fatty liver disease, a metabolic condition with a global prevalence, is marked by the abnormal buildup of fat within liver cells. This condition encompasses two main types: non-alcoholic fatty liver disease (NAFLD) and alcoholic fatty liver disease (AFLD).

With the flourishing development of machine learning, implementations of machine learning models to EHR data have gained increasing significance in modern health care. For their potential to seek latent insights in massive amounts of medical data, machines are enabled to improve clinical decision-making and detect diseases at an early stage. On the flipside, machine learning techniques could also handle unstructured data such as the medical history and the medical prescription of individuals which could also assist the clinical decision-making.

2. EHRs and machine learning

2.1. EHR

EHRs is a digital way to collect and store patients’ longitudinal healthcare data, including Medical History, Medications, Test Results, Allergies, Clinical Notes, etc. The primary purpose of EHRs is to provide more accurate and higher-quality healthcare. These data are maintained by healthcare providers [5], such as hospitals and clinics, and collected during routine delivery of patients’ health care. Nowadays, EHRs are omnipresent in healthcare, virtually replacing traditional health records. Its widespread use led to the accumulation of substantial clinical data, which is considered can be effectively repurposed for clinical research [6].

2.2. Data Preprocessing

Collection and preparation of EHR data are essential steps prior to employing machine learning, given the data’s inherent incompleteness and the significant privacy and security considerations involved. Hence, specific methods are necessary to alleviate their impacts on the machine learning process and potential ethical concern.

Incompleteness stands as a primary challenge to the EHR data collection, arising from a variety of underlying factors. Researchers has discovered several different methods to make improvements. To reduce the amount of unavailable, inaccessible or incomputable data within a data warehouse, Botsis et al. [7] suggest using a text-mining tool, which requires a source- and domain-specific lexicon to extract valuable data. In their research, they also noticed that data fragmentation is very likely to happen on individuals with terminal illnesses. This occurs due to the possibility of these patients being referred to different institutions for more effective treatments [7]. Thus, establishing a health information exchange network across healthcare entities, such as Health Information Exchange (HIE), is recommended to mitigate this problem [7]. Additionally, implementing an industrial standing for EHRs, including using clinical registries with predefined format and defining "standard content" for data, is also recognized as effective ways to enhance data completeness and quality [5], [7].

Automatic de-identification is an important data preparation procedure to solve privacy and security concerns of EHR data before being used in secondary research purposes. It involves the process of eliminating sensitive personal information while retaining the clinical data at the same time. Nowadays, named entity recognition (NER) is considered as the most common method for de-identification [8]. A long time before the adoption of NER, researchers have evaluated different de-identification systems at that time to find the best one. They have found some effective techniques that encompass machine learning methodologies, such as CRF, Decision Trees, Maximum Entropy models, and SVM [9]. These are coupled with dictionaries and occasionally regular expressions. However, these researchers claimed that many systems perform well on the specific documents, but they should be tested on a wider variety of documents to assess their generalizability [9]. Then, a system based on LSTM-CRF model, with both a label prediction bidirectional LSTM layer and a CRF layer, was introduced. This model exhibited a slight performance improvement over CRF-based methods [8]. Still, it requires fine-tuning to achieve best performance. In recent years, Ahmed et al. [10] invented a new model that combines the self-attention mechanism and stacked Recurrent Neural Network. They claimed that this system not only maintains computational efficiency but also surpasses the performance of state-of-the-art models. However, like above, it is hard to say how good is the general performance of this model without testing on multiple heterogeneous documents.

2.3. Algorithm and model

For the analysis of structured data such as body mass index, laboratory results, and heart rate, supervised learning is generally applied which let the machine learn the features from labeled training data to make predictions or decisions. The frequently used models and algorithms in the field of EHR data analysis encompasses [6]:

1) Logistic regression: a classification algorithm that is usually used to predict the probability of a binary output and it is effective in coping with the linearly separable data. Its application scenarios lie in risk assessment and diagnostic assistance.

2) Decision tree: a tree-like algorithm that outputs decisions based on the condition of the branches which is generally used for classification and regression. It is often implemented for diagnostic accuracy improvement.

3) Random forest: an algorithm that combines various decision trees to generate accurate decision while each decision tree is trained on a portion of the original input data and makes its own decision. The final prediction is based on a majority vote (for classification) or an average (for regression) of predictions from all the trees.

4) Support vector machine: a classification algorithm that maps the input data to a hyperplane that distinguishes them linearly. It is commonly applied to enhance clinical calculations and improve surgical outcomes.

In terms of the unstructured data, which is mostly text-based including pathology or radiology inspection reports, admission or discharge details, and progress notes, natural language processing (NLP) is more commonly applied to effectively analyze and extract crucial information from them. Within the NLP algorithm, it first tokenizes the text into small units and then analyzes the structure of the text linguistically by recognizing the grammar regulations after which the machine could parse the text semantically. In the case of EHR data, the NLP algorithm is commonly applied to [8]:

1) Medical Text Classification and prediction: Enabling the machine to categorize clinical notes into different sections or types, such as patient pathology history, diagnosis, and treatment. Besides, segmenting the sentences and phrases semantically from the text and identifying specific medical terminology like drug names, disease names, and patient identifiers.

2) Bidirectional Encoder Representations from Transformers (BERT) -based Embeddings [8]: a powerful pre-trained language model that contextualizes word representations in the text, which capture the meaning of words based on the surrounding words in a sentence or text. It is shown to be effective in capturing the complex semantics and domain-specific language present in medical texts.

3) Information Extraction: Extracting relevant and imperative information such as radiology inspection results, medications, and symptoms from unstructured clinical narratives.

4) Generation: Providing feedback and comment based on the clinical text generation and medical language translation that parses the diagnoses and predictions generated from other supervised machine learning models.

2.4. Validation and Evaluation

Validating machine learning models applied to EHR data is crucial to ensure that the models are accurate, reliable, robust, and generalizable. EHR data is complex and heterogeneous, presenting unique challenges for model development and validation. General performance of a certain model could be measured by its confusion matrix which summarizes the performance of a classification model. It encompasses metrics such as false negatives (FN), true negatives (TN), false positives (FP), and true positives (TP), which can help assess model accuracy and errors. Other specific and more detailed validation methods that could be applied to the machine learning models mentioned previously are listed below:

1) Cross-Validation: Dividing the dataset into several subsets to train and validate the model iteratively, which provides an estimation of model performance with more robustness

2) External Validation: Testing the model on an independent dataset that was not used during training to assess generalization to new, unseen data.

3) Temporal Validation: Evaluating the model's performance over time, using historical data for training and testing on future data points to assess predictive capabilities.

Validation procedures for machine learning models should be carefully designed that is catered to address the specific challenges of healthcare data, including privacy concerns, data quality issues, and potential biases. The purpose is to develop models that are accurate, safe, and ethically sound for use in real-world clinical scenarios.

3. Applications on liver disease management

3.1. Utilizing Supervised Learning on Structured and Unstructured Data

The 11th leading cause of death worldwide is liver disease, particularly cirrhosis, which is a significant health problem. It accounts for 2.1% of global deaths and impacts 5.2016% of disability-adjusted life years in a two-year span. Each year, 320,000 deaths occur due to Chronic Liver Disease (CLD), with a gender distribution of two-thirds male and one-third female. One of the primary challenges is early detection, given its almost imperceptible initial symptoms. Today's healthcare system produces a deluge of patient data, posing challenges for clinicians. However, artificial intelligence, particularly supervised machine learning methods like Support Vector Machines (SVM), Decision Trees (DT), and Random Forests (RF), holds promise in aiding early diagnosis. These methods learn from labeled data and predict outcomes for new, unlabeled data. Applied to liver diseases, they can foresee disease progression and treatment outcomes by analyzing patient records and biomarkers. Such models equip doctors with tools to make precise diagnostic and treatment choices, enhancing patient outcomes and survival.

Incorporating these techniques is crucial for innovating liver disease management and could save numerous lives.

Firstly, machine learning techniques have demonstrated their effectiveness in tracking and predicting liver disease development, in terms of disease progression. Researchers have analyzed different liver disease datasets using various machine learning algorithms to evaluate the analysis performance with different parameters and optimization techniques. This study underscores the significance of machine learning optimization in adjusting hyperparameters to minimize cost functions[11]. Concurrently, the study by Mostafa et al. utilized various machine learning algorithms to analyze patients' clinical data to more accurately predict the progression of the disease[12]. The application of this technique can not only assist doctors in diagnosing diseases earlier but also provide more personalized treatment recommendations for patients. For liver disease diagnosis, a mixed approach using the Adaptive Neuro Fuzzy Inference System (ANFIS) and Particle Swarm Optimization (PSO) has been suggested by some experts. This intelligent diagnostic method combines inference systems and optimization processes, aiming to adjust ANFIS hyperparameters based on the dataset[12].

Next, when it comes to death rates, we have shown that computer programs for learning can guess right about the chance of dying after getting a new liver. Deep learning models used by Li et al. try to predict risk factors after liver transplantation, providing doctors with a more accurate tool to assess the risks and benefits of transplantation. The application of this technology can help doctors better select suitable donors for patients, thereby increasing the success rate of transplantation[13].

Lastly, machine learning techniques have been effectively used to identify valuable clinical features in radiology reports, resulting in more precise forecasts about a patient's future health condition. The study by Liu et al.[14] Used machine learning models and natural language processing methods to examine radiology reports, giving doctors important data about patient outcomes. [14]. Some researchers have proposed a way based on a deep neural network model for predicting Non-Alcoholic Fatty Liver Disease (NAFLD) using multimodal inputs, including metadata and facial images. This method outperforms techniques that use only metadata[15].

3.2. Classification

3.2.1. Medical text classification

The process of categorizing or classifying clinical narratives into predefined classes has gained significant attention due to its potential to enhance healthcare decision-making. Gao et al. [16] have provided a comparative analysis on a bunch of methods of clinical text classification techniques, discussing limitations and breakthroughs related to unstructured narratives and emphasizing the role of self-attention mechanism and long document splitting. Building on this they dive into deep learning methods, particularly CNNs, and hierarchical self-attention network (HiSAN), showcasing their prowess in capturing intricate patterns from clinical texts. Notably, the introduction of BERT [16] has revolutionized medical text classification, as researchers fine-tune this transformer-based model on medical text datasets to achieve state-of-the-art results in multiple clinical classification tasks.

3.2.2. Segmentation

Efficiently segmenting clinical narratives into meaningful sections is pivotal for extracting relevant information and improving data organization. Ganesan and Subotin [17] had an overview on the previous clinical text segmentation methods, highlighting the complexity of unstructured data and the challenges in model generalization. They propose a Logistic regression model that is competent to segment the top-hierarchical sections of clinical narratives and meanwhile could remain its segmentation accuracy to unseen dataset. Recognizing the lack of labeled data for unsupervised segmentation Tepper et al. [18] introduce a heuristics-free machine learning approach for the rapid adaptation to unlabeled data, specifically, they applied maximum entropy model with beam search and Gaussian prior smoothing.

As healthcare data continues to grow, the integration of advanced NLP techniques offers a promising avenue for extracting valuable insights and improving patient care through accurate classification and effective segmentation of clinical narratives. The field remains dynamic, with ongoing research and developments shaping the future of NLP in healthcare.

3.3. BERT-based Embeddings

In NLP, embedding denotes a numerical representation of a word, phrase, or text within a vector space. These representations have found applications in modeling the biomedical text semantics, tracking the patients' trajectory, and a multitude of other assignments [8]. By utilizing a masked language model that can predict words concealed within a contextual order at random, BERT is asserted to possess superior word embedding capabilities compared to their predecessors [8], [19]. However, pretraining is required due to its poor generalizability. Hence, many different versions of pretrained BERTs have emerged. For instance, BioBERT for biochemical domain is pretrained on biomedical research papers, while ClinicalBERT, EhrBERT and MS-BERT for EHR domain are trained on clinical notes [8]. They all have better performance in their respective areas.

BERT has already been widely implemented in solving liver-related problems. Liu et al. [19] utialized BERT-based deep learning to figure out the evidence related to liver cancer diagnosis. This research is based on Chinese Radiology Reports, and Zhang et al. [20] provides a fine-tuning BERT for Chinese clinical documents. Furthermore, Liu et al. [19] introduce a BERT-BiLSTM-CRF model in their research, where this model is composed of the fine-tuned BERT language model for word embedding and BiLSTM-CRF technique for pattern extraction. While extracting features of APHE (hyperintense enhancement in the arterial phase) and PDPH (hypointense in the portal and delayed phases), two important diagnosis evidence for hepatocellular carcinoma, these researchers compared three distinct models: CRF, BiLSTM-CRF, and BERT-BiLSTM-CRF. The result turns out that, no matter in report level or character level, the recognition result of BERT-BiLSTM-CRF outperformed both other two models significantly. Overall, Liu et al. [19] acknowledge the exceptional performance of the BERT-based machine learning model in extracting radiological features from Chinese radiology reports, and demonstrate that both APHE and PDPH stood out as the two most crucial features for diagnosing liver cancer.

3.4. Information Extraction

Information Extraction (IE) involves the automated identification of crucial information within unstructured natural language text [9]. In this section, our primary emphasis will be on presenting an overview of its applications within EHRs.

The focus of the relationship extraction task is to identify and capture specific types of relationships between different entities from the text. In a medical context, a precise understanding of the relationships between various medical entities is essential to a comprehensive understanding of a patient's medical records. The early relationship extraction problem was often viewed as a multi-label classification task, where categories represented labels with a particular type of relationship. However, as time goes by, more complex feature modeling methods are gradually introduced into deep relation extraction systems to improve the modeling effect of entity relationships.

Convolutional neural networks (CNNs) are one of the methods used to solve the problem of relation extraction. In clinical discharge summary, standard CNN combined with word-level features is applied to extract the relationship between entities [21]. Another investigation integrated Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs) to extract biomedical relationships from both linear and dependency graph representations of candidate sentences [26]. This hybrid CNN-RNN model was subjected to benchmarking across various protein-protein and drug-drug interaction corpora, showcasing the collaborative efficacy of CNNs and RNNs in relation extraction tasks [9]. In addition, some studies employ RNNs to label entities of interest one by one and infer relationships between them. In one study, by comparing SVMs, RNNs, and rule induction systems, it was found that in adverse drug event (ADE) detection tasks, the RNN model performed best in relation extraction [22]. Similarly, in the task of intra-sentence relationship extraction, researchers adopt a similar BiLSTM-CRF model [23] and use the Transformer-based method to capture the long-distance dependencies of inter-sentence relationships [24].

3.5. Generation

Research in the terms of Natural Language Generation (NLG) of generating EHR has become increasingly important. [25]. An encoder-decoder model enables the creation of synthetic main complaints based on factors such as age group, gender, and discharge diagnosis in EHRs. This method is effective at creating realistic chief complaint text and also keeps a lot of the original epidemiological data intact. A unique benefit of this approach is that the synthetic complaints don't contain any personally identifiable information (PII), which provides a way to de-identify text in EHRs. When integrated with technologies such as Generative Adversarial Networks (GANs), the method can even be expanded to generate fully synthetic EHRs. This facilitates better sharing of data between healthcare providers and researchers, and also enhances the optimization of healthcare-related machine learning models. Particularly in cases involving diseases such as cirrhosis, hepatitis, liver cancer, and fatty liver disease, this technology may contribute to a more profound understanding and analysis of the prevalence trends and characteristics of these conditions. Consequently, it could provide valuable insights to support clinical decision-making processes[26].

4. Discussion

4.1. Current ML-related limitations

4.1.1. Data limitation

The main part of the EHR data is often unstructured and could be incomplete or inconsistent. It may contain errors due to various reasons such as typos, missing information, not recorded regularly. Apart from that, the EHR data would only be approachable with the permission of both the patients as well as the medical clinics. Besides, the EHR data collected from different hospitals, even different doctors could be varied in structure and content, which may pose greater challenges for the machine to parse the data.

4.1.2. Model performance

These issues mentioned can drastically affect the performance of machine learning models, which includes the failure for models to converge or reduced prediction and classification accuracy. Especially for the unstructured data, its chaotic form and disorder will hamper the deep learning model from extracting the pattern from the training data as the lack of context or clear structure can make it challenging to disambiguate word meanings or infer relationships between words [27] thus the irrelevant noise could further confuse the model and even mislead it to biased state, in addition, many NLP tasks involve understanding long-range dependencies and relationships within text based on the proper sequence of the data, and failure of capturing these long-term dependencies could lead to suboptimal performance.

4.2. Ethical and Privacy Considerations

Despite the implementation of data de-identification and data privacy protection laws in the context of EHR data collection, concerns about the potential breach of personal privacy continue to trouble many patients. First of all, automatic de-identification is process where its performance larges depends on the source of documents and clinical institutions. It can lead to poor performance if the model is not fine-tuned. Also, with the development of re-identification techniques, the security of de-identification can become compromised. Secondly, for some institutions that are lack of data security awareness, unauthorized individuals may access and abuse these data. They are vulnerable to hacking and data breaches if these servers are not well maintained. Even doctors or employees with access to EHR systems might abuse their privileges by selling patients' private information to illegal organizations.

Additionally, the health information exchange network and the secondary utilization of EHRs mentioned earlier aim to enhance data accuracy and utility. However, both approaches involve the sharing of data among institutions, consequently increasing the risk of data exposure. All these privacy and security concerns will create challenges in obtaining consents of patients sharing their EHR data to clinical institutions. Consequently, this will diminish the comprehensiveness and value of EHR data.

4.3. Future Research Trends and Prospects

Machine Learning for Privacy: As privacy concerns about EHR data rise, future research should focus on building improved privacy-preserving machine learning algorithms. Improving data de-identification methods, investigating safe multi-party computation, and creating federated learning systems that enable collaborative analysis without revealing raw data are all part of this effort. Researchers may build comprehensive ethical frameworks and protocols for EHR data collection, sharing, and analysis to address the ethical challenges connected with EHR data utilization. These frameworks should take into account patient consent, data ownership, and data management. Interoperability and Data Standardization: enhancing data quality and accessibility requires enhancing interoperability across different EHR systems and standardized data formats. Future study might concentrate on the development of universal standards for EHR data interchange in order to improve data integration and analysis. Deep Learning Advances: Advances in deep learning approaches, such as self-attention mechanisms and transformer models like BERT, will almost certainly play a large role in extracting important insights from unstructured clinical narratives. Further research should be conducted to determine how these models might be fine-tuned for specific healthcare purposes. Initiatives for Collaborative Research: Collaboration is vital among healthcare organizations, researchers, and regulatory agencies. Collaborative research projects and data-sharing networks can help to expedite development in the area while also addressing data privacy and ethical concerns.

5. Conclusion

In this comprehensive review, the paper delves into the potential of machine learning (ML) techniques, especially in the domain of liver disease, using Electronic Health Records (EHRs). It examines the challenges and opportunities that EHRs present for clinical research, such as data incompleteness and privacy concerns. The study extensively reviews various ML methods for EHR data analysis tailored to liver disease, spotlighting recent advancements like self-attention mechanisms and stacked Recurrent Neural Networks. Furthermore, it underscores the importance of robust methodologies to extract relevant insights for liver disease while maintaining data integrity and patient confidentiality. Although this review advocates the establishment of Health Information Exchange Networks (HIE) and the implementation of industrial standards as potential solutions, it acknowledges that the domain has not yet fully tackled data privacy and ethical challenges. Notably, the BERT-BiLSTM-CRF model, as highlighted, excels in medical diagnostics, suggesting that its application in liver disease diagnostics could be a valuable research focus. While this study marks an essential stepping stone in EHR-based research for liver disease using advanced ML techniques, it also flags the critical need for addressing data privacy and ethics in subsequent studies. Overall, this research offers both theoretical and practical contributions, with a pronounced emphasis on liver disease diagnostics and treatment enhancements using ML algorithms.

Acknowledgement

Jie Luo, Yuqi Sun, Jiachen Liu and Zhou Yu contributed equally to this work and should be considered co-first authors.

References

[1] J. W. Jang, ‘Current status of liver diseases in Korea: Liver cirrhosis’, *Korean J Hepatol*, Dec. 2009, doi: 10.3350/kjhep.2009.15.S6.S40.

[2] O. Gholizadeh *et al.*, ‘Hepatitis A: Viral Structure, Classification, Life Cycle, Clinical Symptoms, Diagnosis Error, and Vaccination’, *Can. J. Infect. Dis. Med. Microbiol.*, vol. 2023, p. e4263309, Jan. 2023, doi: 10.1155/2023/4263309.

[3] T. L. Wright, ‘Introduction to Chronic Hepatitis B Infection’, *Off. J. Am. Coll. Gastroenterol. ACG*, vol. 101, p. S1, Jan. 2006.

[4] G. M. Lauer, ‘HEPATITIS C VIRUS INFECTION’, *N. Engl. J. Med.*, vol. 345, no. 1, Jul. 2001, [Online]. Available: www.nejm.org

[5] M. R. Cowie *et al.*, ‘Electronic health records to facilitate clinical research’, *Clin. Res. Cardiol.*, vol. 106, no. 1, pp. 1–9, Jan. 2017, doi: 10.1007/s00392-016-1025-6.

[6] W.-C. Lin, J. S. Chen, M. F. Chiang, and M. R. Hribar, ‘Applications of Artificial Intelligence to Electronic Health Record Data in Ophthalmology’, *Transl. Vis. Sci. Technol.*, vol. 9, no. 2, p. 13, Feb. 2020, doi: 10.1167/tvst.9.2.13.

[7] T. Botsis, G. Hartvigsen, F. Chen, and C. Weng, ‘Secondary Use of EHR: Data Quality Issues and Informatics Opportunities’, *Summit Transl. Bioinforma.*, vol. 2010, pp. 1–5, Mar. 2010.

[8] I. Li *et al.*, ‘Neural Natural Language Processing for Unstructured Data in Electronic Health Records: a Review’. arXiv, Jul. 06, 2021. Accessed: Aug. 12, 2023. [Online]. Available: http://arxiv.org/abs/2107.02975

[9] S. M. Meystre, F. J. Friedlin, B. R. South, S. Shen, and M. H. Samore, ‘Automatic de-identification of textual documents in the electronic health record: a review of recent research’, *BMC Med. Res. Methodol.*, vol. 10, no. 1, p. 70, Dec. 2010, doi: 10.1186/1471-2288-10-70.

[10] T. Ahmed, M. M. A. Aziz, and N. Mohammed, ‘De-identification of electronic health record using neural network’, *Sci. Rep.*, vol. 10, no. 1, p. 18600, Oct. 2020, doi: 10.1038/s41598-020-75544-1.

[11] S. S. Nigatu, P. C. R. Alla, R. N. Ravikumar, K. Mishra, G. Komala, and G. R. Chami, ‘A Comparitive Study on Liver Disease Prediction using Supervised Learning Algorithms with Hyperparameter Tuning’, in *2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, May 2023, pp. 353–357. doi: 10.1109/InCACCT57535.2023.10141830.

[12] F. B. Mostafa and M. E. Hasan, ‘Machine Learning Approaches for Binary Classification to Discover Liver Diseases using Clinical Data’. arXiv, Jun. 05, 2021. doi: 10.48550/arXiv.2104.12055.

[13] C. Li, X. Jiang, and K. Zhang, ‘A Transformer-Based Deep Learning Approach for Fairly Predicting Post-Liver Transplant Risk Factors’. arXiv, Apr. 05, 2023. doi: 10.48550/arXiv.2304.02780.

[14] H. Liu *et al.*, ‘A Natural Language Processing Pipeline of Chinese Free-text Radiology Reports for Liver Cancer Diagnosis’, *IEEE Access*, vol. 8, pp. 159110–159119, 2020, doi: 10.1109/ACCESS.2020.3020138.

[15] ‘Fatty Liver Disease Prediction Using Supervised Learning | SpringerLink’. https://link.springer.com/chapter/10.1007/978-981-16-3660-8_54 (accessed Aug. 30, 2023).

[16] S. Gao *et al.*, ‘Limitations of Transformers on Clinical Text Classification’, *IEEE J. Biomed. Health Inform.*, vol. 25, no. 9, pp. 3596–3607, Sep. 2021, doi: 10.1109/JBHI.2021.3062322.

[17] K. Ganesan and M. Subotin, ‘A general supervised approach to segmentation of clinical texts’, in *2014 IEEE International Conference on Big Data (Big Data)*, Washington, DC, USA: IEEE, Oct. 2014, pp. 33–40. doi: 10.1109/BigData.2014.7004390.

[18] M. Tepper, D. Capurro, F. Xia, L. Vanderwende, and M. Yetisgen-Yildiz, ‘Statistical Section Segmentation in Free-Text Clinical Records’.

[19] H. Liu *et al.*, ‘Use of BERT (Bidirectional Encoder Representations from Transformers)-Based Deep Learning Method for Extracting Evidences in Chinese Radiology Reports: Development of a Computer-Aided Liver Cancer Diagnosis Framework’, *J. Med. Internet Res.*, vol. 23, no. 1, p. e19689, Jan. 2021, doi: 10.2196/19689.

[20] X. Zhang *et al.*, ‘Extracting comprehensive clinical information for breast cancer using deep learning methods’, *Int. J. Med. Inf.*, vol. 132, p. 103985, Dec. 2019, doi: 10.1016/j.ijmedinf.2019.103985.

[21] S. K. Sahu, A. Anand, K. Oruganty, and M. Gattu, ‘Relation extraction from clinical texts using domain invariant convolutional neural network’. arXiv, Jun. 30, 2016. doi: 10.48550/arXiv.1606.09370.

[22] T. Munkhdalai, F. Liu, and H. Yu, ‘JMIR Public Health and Surveillance - Clinical Relation Extraction Toward Drug Safety Surveillance Using Electronic Health Record Narratives: Classical Learning Versus Deep Learning’, *JMIR Pubulications*, Apr. 25, 2018. https://publichealth.jmir.org/2018/2/e29/ (accessed Aug. 28, 2023).

[23] B. Dandala, V. Joopudi, and M. Devarakonda, ‘Adverse Drug Events Detection in Clinical Notes by Jointly Modeling Entities and Relations Using Neural Networks’, *Drug Saf.*, vol. 42, no. 1, pp. 135–146, Jan. 2019, doi: 10.1007/s40264-018-0764-x.

[24] F. Christopoulou, T. T. Tran, S. K. Sahu, M. Miwa, and S. Ananiadou, ‘Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods’, *J. Am. Med. Inform. Assoc.*, vol. 27, no. 1, pp. 39–46, Jan. 2020, doi: 10.1093/jamia/ocz101.

[25] I. Li *et al.*, ‘Neural Natural Language Processing for Unstructured Data in Electronic Health Records: a Review’. arXiv, Jul. 06, 2021. Accessed: Aug. 13, 2023. [Online]. Available: http://arxiv.org/abs/2107.02975

[26] ‘Natural language generation for electronic health records | npj Digital Medicine’. https://www.nature.com/articles/s41746-018-0070-0 (accessed Aug. 30, 2023).

[27] Z. Ye, P. Liu, J. Fu, and G. Neubig, ‘Towards More Fine-grained and Reliable NLP Performance Prediction’, 2021, doi: 10.48550/ARXIV.2102.05486.