

**DEVELOPING NOVEL NLP-ENABLED TIMELY AND ACCURATE  
DECISION MAKING FOR PRECISION MEDICINE**

A Dissertation  
Presented to  
The Academic Faculty

by

Yuanda Zhu

In Partial Fulfillment  
of the Requirements for the Degree  
Doctor of Philosophy in the  
School of Electrical and Computer Engineering

Georgia Institute of Technology  
December 2023

**COPYRIGHT © 2023 BY YUANDA ZHU**

**DEVELOPING NOVEL NLP-ENABLED TIMELY AND ACCURATE  
DECISION MAKING FOR PRECISION MEDICINE**

Approved by:

Dr. May D. Wang, Advisor  
Depart of Biomedical Engineering  
*Georgia Institute of Technology*

Dr. Blake Anderson  
School of Medicine  
*Emory University*

Dr. David V. Anderson  
School of Electrical and Computer  
Engineering  
*Georgia Institute of Technology*

Dr. David Wright  
School of Medicine  
*Emory University*

Dr. Faramarz Fekri  
School of Electrical and Computer  
Engineering  
*Georgia Institute of Technology*

Date Approved: November 8th, 2023

Dedicated to my parents and my girlfriend

For all their love, support and encouragement

## ACKNOWLEDGEMENTS

There are no better words to express my deepest gratitude and respect for my thesis and research advisor, Professor May D. Wang. She inspired me to become an independent researcher, a critical thinker and a team player. She guided me through the challenges in academic research and career growth. She also demonstrated the success of a talented yet hard-working scientist. Additionally, I would like to thank my dissertation committee members: Dr. David Anderson, Dr. Faramarz Fekri, Dr. Blake Anderson and Dr. David Wright. Particularly, Dr. David Anderson gave me guidance on my initial stage in graduate research; Dr. Fekri asked a genuine yet fundamental question, which inspired me to change PhD career. Dr. Blake Anderson has not only provided research data for NLP-related projects, but provided constructive feedback and guidance that contributed significantly to my dissertation as well.

I would like to thank the collaborators and colleagues for providing their technical expertise and unlimited support, which inspired me and assisted me through the hardest time: Dr. Ryan Hoffman, Dr. Li Tong, Hang Wu, Ying Sha, Dr. Felipe Giuste, Dr. Monica Isgut, Wenqi Shi, Anirudh Choudhary, Benoit Marteau, Yishan Zhong, Andrew Hornback, and all other current and former BioMIBLab members.

Finally, I would like to thank my parents, Jianping Zhu and Rongling Chen, for their unconditional support and endless encouragement. I would like to thank my girlfriend Lining Jiang for her patience and trust. It was their love that raised me up so that I could overcome all the challenges.

# TABLE OF CONTENTS

<b>ACKNOWLEDGEMENTS</b>	<b>iv</b>
<b>LIST OF TABLES</b>	<b>vii</b>
<b>LIST OF FIGURES</b>	<b>ix</b>
<b>LIST OF SYMBOLS AND ABBREVIATIONS</b>	<b>xi</b>
<b>SUMMARY</b>	<b>xiii</b>
<b>CHAPTER 1. Introduction</b>	<b>1</b>
<b>1.1 Precision Medicine</b>	<b>1</b>
<b>1.2 Natural Language Processing Techniques and Impact</b>	<b>3</b>
<b>CHAPTER 2. Prevention in Precision Medicine: Proposing Novel NLP Framework on Electronic Health Records for Intelligent Mortality Reporting</b>	<b>5</b>
<b>2.1 Prevention in Precision Medicine</b>	<b>5</b>
<b>2.2 Death Reporting in Precision Medicine</b>	<b>6</b>
<b>2.3 Related Works in Intelligent Death Reporting</b>	<b>11</b>
<b>2.4 Causal Sequence of Death</b>	<b>12</b>
2.4.1 Dataset	13
2.4.2 Generating Causal Sequences via Translation	14
<b>2.5 NLP Approaches in Generating Causal Sequence of Death</b>	<b>15</b>
2.5.1 Neural Machine Translation: Encoder and Decoder	15
2.5.2 Decoding and Translation	19
2.5.3 Evaluation	20
<b>2.6 Experiments</b>	<b>24</b>
2.6.1 OpenNMT	25
2.6.2 Optional Preprocessing: Validity Check	26
2.6.3 XLM: Pretraining	26
<b>2.7 Results</b>	<b>27</b>
2.7.1 Attention Comparison	27
2.7.2 Validity Check, Domain Knowledge Constraint and ICD-10 Input Sequence	28
2.7.3 Attention Visualization: A Case Study	30
2.7.4 XLM	31
<b>2.8 Discussion</b>	<b>32</b>
<b>2.9 Conclusion</b>	<b>34</b>
<b>CHAPTER 3. Screening in Precision Medicine: Proposing Novel NLP Frameworks on Unstructured Free-Text Notes for Disease Identification</b>	<b>36</b>
<b>3.1 Using Natural Language Processing on Free-Text Clinical Notes to Identify Patients with Long-Term COVID Effects</b>	<b>36</b>
3.1.1 Introduction: Long COVID	36
3.1.2 Related Works	39

3.1.3	Methods: Novel NLP Frameworks	41
3.1.4	Results and Discussion	46
3.1.5	Conclusion	51
<b>3.2</b>	<b>Use of natural language processing of patient-initiated Electronic Health Messages to Identify patients with COVID-19 Infection</b>	<b>52</b>
3.2.1	Introduction: Surge in Patient Messages	52
3.2.2	Methods: Novel NLP Approach for Automatic Triage	54
3.2.3	Results	58
3.2.4	Discussion	61
3.2.5	Limitations	63
3.2.6	Conclusions	64
<b>CHAPTER 4. Diagnosis in Precision Medicine: Developing Novel NLP Frameworks on Time-Series EEG Signals for Automated Seizure Detection</b>		<b>65</b>
<b>4.1</b>	<b>Introduction</b>	<b>65</b>
<b>4.2</b>	<b>Background and Related Works</b>	<b>67</b>
4.2.1	Seizure Basics	67
4.2.2	How Clinicians Detect Seizure?	69
4.2.3	Automated Seizure Detection Using Machine Learning	71
<b>4.3</b>	<b>Dataset: TUH EEG Seizure Corpus (TUSZ)</b>	<b>73</b>
<b>4.4</b>	<b>Case Study 1: Adopting Regularization of Deep Neural Networks for EEG Seizure Detection to Mitigate Overfitting</b>	<b>75</b>
4.4.1	Methods	76
4.4.2	Results	82
4.4.3	Discussion	86
4.4.4	Conclusion	91
<b>4.5</b>	<b>Case Study 2: Mitigating Patient-to-Patient Variation in EEG Seizure Detection using Meta Transfer Learning</b>	<b>93</b>
4.5.1	Method	93
4.5.2	Experiments and Results	97
4.5.3	Discussion and Conclusion	99
<b>4.6</b>	<b>Case Study 3: Developing Novel CNN + Transformer Framework for Automated Seizure Detection on Multi-Channel EEGs</b>	<b>101</b>
4.6.1	Methods	102
4.6.2	Experiments and Results	106
4.6.3	Discussion	108
4.6.4	Conclusion	110
<b>CHAPTER 5. Discussion and Conclusion</b>		<b>111</b>
<b>REFERENCES</b>		<b>116</b>

## LIST OF TABLES

Table 1	An Example of 1-Gram Precision and 2-Gram Precision in BLEU Score.	21
Table 2	Our Modified BLEU Score for Different Candidate Sequences.	23
Table 3	Average BLEU and Accuracy Scores and Standard Deviation in Parentheses Across Five Folds.	27
Table 4	Average BLEU Scores and Standard Deviation in Parentheses for Five Experiments.	28
Table 5	The Mean Value of Five-fold Cross Validation Results of Three Different Pretrained BERT Models. The highest recall score is highlighted in bold and underlined.	46
Table 6	Top and Bottom 10 Words Across All Clinical Notes of the Last Fold of Testing Data.	49
Table 7	Demographics for patient sample whose EHR messages were used to train the model.	59
Table 8	Sensitivity (recall), specificity, precision, and F1-score of eCOV.	60
Table 9	Differences between COVID-19 positive patients who received antiviral treatment and those who did not. Reported as mean	61

(standard deviation). Statistical significance calculated using paired T-Test ( $\alpha = 0.05$ ).

Table 10	Number of sessions and duration of top three seizure types and non-seizure type in TUSZ dataset.	75
Table 11	Results for 4 second segments using hand-crafted features and traditional machine learning models, as well as a CNN/LSTM model.	83
Table 12	Results for 4 second segments using a CNN/LSTM model with adversarial multitask learning.	84
Table 13	Results for 4 second segments using a CNN/LSTM model with adversarial multitask learning, with randomly rearranged, randomly rescaled channels each minibatch.	85
Table 14	Results For Different Models. Macro-F1 Score and AUROC Score Are Evaluated On Four Classes On the Predefined Test Set.	98
Table 15	Model architecture and output shape.	105
Table 16	Classification report of the proposed approach.	107
Table 17	Performance comparison of our proposed approach against the replicated state-of-the-art deep learning approaches from the literature. All approaches are evaluated on the same TUSZ dataset for binary seizure detection task. MSA refers to multi-head self	108

## LIST OF FIGURES

Figure 1	Overall Structure of this paper. The encoder-decoder model is the main framework for generating sequences of causes of death. Validity check is an optional preprocessing step and domain knowledge constraint is an optional step in decoding.	10
Figure 2	Sample data from the Michigan data set. The casual sequence of death in ICD-10 for this decedent is I500→R688 (Heart failure leading to Other general symptoms and signs), outlined in green. This decedent had a total of 30 ICD-9 diagnostic codes assigned during the last visit to hospital, outlined in blue.	13
Figure 3	Neural machine translation consists of an encoder (stacked recurrent networks in blue) and a decoder (stacked recurrent networks in red). The symbol <i>&lt;eos&gt;</i> is a special token referring to the end of a sentence. Adapted from Luong and Manning, 2016.	15
Figure 4	Attention visualization and explanation. In part (a), the attention score matrix is visualized. From top to bottom are the underlying cause of death and immediately causes of death. Darker blue color indicates higher attention scores (the input code is more related with the output code). In part (b), we provide human-readable description to all ICD codes. The identified causes of death for this decedents are cardiovascular diseases.	30
Figure 5	Long COVID, impact and common symptoms.	37
Figure 6	Overall diagram of this work. The three key components are patient identification, clinical notes data acquisition and NLP prediction.	42
Figure 7	Box plot of long COVID classification results across five-fold using different models.	48
Figure 8	Interpretation visualization of a given clinical note in the testing set. The interpretation is for the ClinicalBERT model trained on training and validation data.	50
Figure 9	Flowchart depicting methodology of model training and cohort study analysis.	55
Figure 10	Electrode locations for a standard 10-20 system with a 22-channel TCP montage.	68
Figure 11	A spike and slow wave pattern in a focal non-specific seizure.	69

Figure 12	How clinician detect specific seizure type. The non-seizure type and top three most popular seizure types are included.	71
Figure 13	Hierarchical structure of the TUSZ dataset.	74
Figure 14	Architecture of our CNN2D/LSTM. We ran multiple architectures but found that the following had consistently strong results. We found in practice that our networks were most likely to converge if we also applied a learning rate decay of 0.9 each epoch. We used an RELU activation function except for the final layer, where we used softmax. The network trains in parallel another session identification dense layer (not shown), which connects to the feature extraction layers of the network and which is prevented from updating weights to the network.	78
Figure 15	Adversarial multi-task learning approach for seizure detection. We used two neural networks with two task, weights that shared and updated between the networks, and “frozen” layers forced to remain the same after each minibatch update to create a multi-task adversarial network. A) The first network attempts to update the feature extraction for seizure detection and against session identification without updating the weights for the final session identification layer. B) The second network updates session identification weights by predicting positively for sessions, without changing the other weights.	81
Figure 16	Training/validation F1 score for a vanilla CNN/LSTM model. Session identification accuracy increases while valid seizure F1 fails to increase.	83
Figure 17	Training/validation F1 score for a CNN/LSTM model with randomly rearranged channels on each minibatch and random rescaling on each minibatch of 2 with a seizure weight of 25 and patient weight of $-1$ .	86
Figure 18	Meta Transfer Learning for EEG Seizure Detection.	96
Figure 19	Overall flowchart diagram for our proposed approach for EEG seizure detection.	102
Figure 20	Using three convolutional layers, the CNN module effectively extracts both spatial and temporal features from the raw EEGs.	104

## LIST OF SYMBOLS AND ABBREVIATIONS

NLP	Natural Language Processing
EHRs	Electronic health records
BERT	Bidirectional Encoder Representations from Transformers
LSTM	Long Short-Term Memory
EEG	Electroencephalogram
PASC	Post-acute sequelae of SARS-CoV-2 infection
ICD	International Classification of Diseases
NICE	National Institute for Health and Care Excellence
RNN	Recurrent Neural Networks
XAI	Explainable Artificial Intelligence
AWS	Amazon Web Services
ANOVA	Analysis of variance
Tukey's HSD	Tukey's Honestly significant difference
CONSORT	Consolidated Standards of Reporting Trials
CCI	Charlson Comorbidity Index
BMI	Body Mass Index
TP	True Positive
FN	False Negative
TN	True Negative
FP	False Positive
SD	Standard deviation
TCP	Temporal Central Parasagittal

FNSZ	Focal non-specific seizure
GNSZ	General non-specific seizures
CPSZ	Complex partial seizures
CNN	Convolutional Neural Networks
TCN	Temporal Convolutional Networks
STFT	Short-time Fourier transform
TUH	Temple University Hospital
TUSZ	TUH EEG Seizure Corpus
SNR	Signal-to-noise
FFT	Fast Fourier Transform
MAML	Model-Agnostic Meta-Learning
MUPS	Meta Update Strategy
BCI	Brain-Computer Interface
SGD	Stochastic Gradient Descent
ELU	Exponential Linear Unit
NCHS	National Center for Health Statistics
CDC	Centers for Disease Control and Prevention
FHIR	Fast Healthcare Interoperability Resources
ACME	Automatic Classification of Medical Entry
NMT	Neural machine translation
BLEU	BiLingual Evaluation Understudy
CMS	Centers for Medicare & Medicaid Services
XLM	Cross-lingual language model
MLM	Masked language modeling

## SUMMARY

Precision medicine uses big data capturing “individual differences in patients’ genes, environments, and lifestyles focuses” to make prevention strategies, screening, diagnosis of diseases and treatment therapies. In recent years, the rapid development of Natural Language Processing (NLP) uses Artificial Intelligence and Machine Learning to make sense of large volumes of language or time-series data. In this dissertation, I have not only investigated how NLP can extract precise information from electronic health records (EHRs) and free-text clinical notes, but also examined how NLP models such as transformers can be used for analyzing human physiological Electroencephalography (EEG) data for accurate clinical decision making. Specifically, on EHRs, I have shown how NLP is able to decipher and sequentially report death events for individual patients; on free-text clinical notes, I have demonstrated how the pretrained transformer-based NLP model BERT can precisely discern long COVID-19 patients from the rest; and on time-series EEG data, I have illustrated how Long Short-Term Memory (LSTM) and transformer architectures, when combining with convolutional neural networks (CNNs), can extract both spatial and temporal features for accurate seizure detection. My PhD research aims to provide a foundation for advancing NLP for personalized patient care in modern Precision Medicine.

# CHAPTER 1. INTRODUCTION

## 1.1 Precision Medicine

Although the concept of “precision medicine” was officially proposed by the US National Research Council in 2021 [1], its application can be traced back to 1901, when human blood types were used to guide blood transfusions [2]. Between 1990 and 2003, the launch of Human Genome Project significantly expanded clinical domain knowledge on human genotypes [3], leading to the great potential that improves the quality of patient care using genome sequencing [4]. In 2015, the US President Obama illustrated a vision for the nation-wide initiative for precision medicine in his State of Union address [2], [4]. Ever since, more research projects in informed decision making have been conducted [5].

So what is the definition of precision medicine? Initially, clinicians and researchers aimed to develop prevention and treatment strategies that take individual patient variability into consideration [2], especially with genome-driven clinical decision making [4]. Nowadays, the definition of precision medicine is more general, as people prioritize to utilize statistical and AI approaches on clinical data to make informed decision making in healthcare [5]. Hence, it is essential to make data-driven, patient-centered computational tools to tackle clinical components, including prevention, screening, diagnosis and treatment.

During the past few years, machine learning and deep learning techniques have been widely used on different medical data modalities for different clinical tasks. For example, medical image processing is a prominent domain for informed decision making.

Novel explainable AI with convolutional neural networks (CNNs) were proposed as innovative approaches for COVID-19 patient diagnosis on X-ray/ CT images [6]–[9]. Weakly-supervised learning and data augmentation techniques, such as generative adversarial networks (GANs)[10], were designed to automatically identify heart transplant rejection on histopathological whole slide images [11]–[15].

Electronic health records (EHRs) are another critical data modality in healthcare. Novel encoder-decoder frameworks with attention module were proposed to generate the causal sequence of death, using discharge records and death certificate data [16]. This is an important step to enable intelligent death reporting, reducing the workload of physicians and improving the accuracy in death report. In addition, EHRs data are also utilized to assist patient diagnosis. Autoencoder models along with gradient boosting tree models were proposed to assess risks and predict adverse patient outcomes using ICU EHRs [17], [18]. Meanwhile, various machine learning techniques were tested on EHRs to assess the risks of COVID-19 patients [19]. Lastly, data-driven and rule-based approach also facilitated potential improvements in patient treatment using structured EHRs. Specifically, hierarchical clustering analysis were applied to 428 adolescent idiopathic scoliosis (AIS) patient EHRs to explore patient sub-phenotypes [20]. Separately, rule-based approaches were implemented and validated against clinical decision workflows for automated treatment recommendation for cerebral palsy patients [21].

In additional to structured EHRs, the unstructured, free-text clinical notes are an essential components of clinical data in clinical practice. A Novel NLP pipeline that includes fine tuning pretrained BERT models is proven successful in identifying patients with long COVID using about 800 outpatient clinical notes [22]. Likewise, a similar NLP

pipeline demonstrates its high accuracy in detecting patients with potential COVID symptoms from over 14,000 patient-initiated messages [23]. Additionally, retrieval-augmented generation with ChatGPT was developed [24] to feed large language models (LLMs) with AIS domain knowledge as chatbots for shared decision making prior to surgery, enhancing patients education and satisfaction.

Lastly, time-series physiological signals are also widely used in clinical decision making. Adversarial multi-task learning [25], meta learning [26] and novel CNN/ LSTM / Transformer models [27] were proposed to improve accuracy in automated EEG seizure detection.

These efforts show a promising progress and future in precision medicine.

## **1.2 Natural Language Processing Techniques and Impact**

Natural Language Processing (NLP) is a group of techniques that aim to learn the temporal or sequential information from the input sequence data, especially the unstructured free-text data [28]. Traditional NLP approaches utilized token frequency to interpret the semantics of texts. Recurrent neural networks (RNN) and its variants [29], long short-term memory (LSTM, proposed in 1997) [30] and Gated Recurrent Units (GRUs, proposed in 2014) [31], were designed to handle sequential information. They show promising results in several NLP tasks. LSTM and GRU can handle limitations of vanilla RNNs, such as the vanishing gradient problem. In 2014, the attention module [32] was proposed to improve the performance of LSTM models, especially in neural machine translation tasks. The attention module was designed to search and align input and output sequences, significantly improving model performance on handling long sequences. In

2017, the self-attention module and the transformer model [33] were proposed, which significantly improves the performance in various NLP tasks. Notably, Bidirectional Encoder Representations from Transformers (BERT) models [34] and other large language models (LLMs), such as Generative Pre-trained Transformer (GPT) models [35], are built upon the decoder-based transformer models.

The breakthrough in NLP models facilitates better informed decision making in healthcare. Specifically, the attention module in RNN variants enables more accurate sequence search and alignment, leading to more accurate sequence generation tasks in the medical domain. Meanwhile, transformer models can capture long-term dependencies more accurately, which make them possible to handle long-term time-series data. Additionally, pretrained BERT models can more accurately capture complex linguistics within the context and recognize phrases in the clinical domain, promoting their usage on clinical notes and messages.

In this work, I would like to promote accurate and timely decision making by designing novel NLP frameworks for precision medicine. In specific aim 1, I will focus on the prevention domain of precision medicine, where I summarized the key requirements and fundamental challenges in prevention. I used a case study to demonstrate the novel NLP pipeline in generating the causal sequence of death, as intelligent death reporting is a key component in public health for disease prevention. In specific aim 2, I will show screening as an effective tool in proactive patient identification, and prove the concept using two case studies in COVID/ long COVID patient identification. In specific aim 3, I showcase three projects in EEG seizure detection and discuss the challenges and future directions in this field.

# **CHAPTER 2. PREVENTION IN PRECISION MEDICINE: PROPOSING NOVEL NLP FRAMEWORK ON ELECTRONIC HEALTH RECORDS FOR INTELLIGENT MORTALITY REPORTING**

## **2.1 Prevention in Precision Medicine**

There exist three key requirements for prevention in precision medicine [36]–[40]: to understand the burden, to identify the causes of the diseases/ symptoms, and to implement the prevention strategies. The burden of a disease is typically involved with its impact/ prevalence, severe consequences and costs associated with healthcare needed. For example, cancer is a one of the leading causes of death in the US, with 1.6 million new cases and 602,000 people passing away due to cancer each year [37], [40]. Meanwhile, cancer treatment cost about 190 billion US dollars in 2015. Research stated that if properly designed and implemented, public health strategies may prevent 40-50% cancer cases [40].

However, there are fundamental challenges in to estimating the impact of individual genomic and environmental factors when determining the true causes of a disease. From the genomic perspective, more research is needed to verify the association between human genome and a specific disease [39]. Meanwhile, it is neither cost effective nor widely available to apply genomics on disease prevention [37], [38]. One the other hand, it is often difficult to quantify the environmental factors that cause a certain disease [40].

Additionally, it is also challenging to customize and implement prevention strategies for individual patients. Existing preventative strategies are often implemented on

the population-level, targeting for high-risk subgroups of patients [40]. Also, certain strategies, for example, chemoprevention, are ineffective [37], as healthcare providers may not start monitoring conditions of the patients with molecular assessment until signs or symptoms of cancer develops [38].

As assessing the genomic and environmental factors for disease prevention is challenging, an alternative strategy is to predict the progression of severe conditions that lead to death. Additionally, I would like to identify the causal relationship between diseases, and identify the causes of death for each individual patient. In this section, I would like to present the work on using neural machine translation to proposing causal sequence of death to facilitate intelligent death reporting. This work was published as a journal paper [16].

## **2.2 Death Reporting in Precision Medicine**

The United States witnesses over 2.7 million fatalities annually, with nearly 57 million deaths reported worldwide each year [41]. As of March 23rd, 2022, the coronavirus pandemic has claimed close to 6.1 million lives from 472 million confirmed cases globally<sup>1</sup>. Despite being ranked the third leading cause of death [42], [43], the specifics of COVID-19 related complications and mortalities are yet to be thoroughly scrutinized [44]–[47]. Hence, precise death reporting is critical for public health agencies such as the U.S. National Center for Health Statistics (NCHS) and the Centers for Disease Control and Prevention (CDC) to devise effective public health strategies.

---

<sup>1</sup> <https://covid19.who.int/>

The death reporting protocol in the U.S. necessitates the documentation of two types of causes on death certificates: a single medical condition known as the underlying cause of death and a series of ordered medical conditions leading to the death, known as the "causal sequence of death" in our context. This causal sequence of death forms the crux of the NCHS's Multiple Causes of Death data, an indispensable resource in public health.

A causal sequence typically consists of an underlying cause and immediate causes of death, which often stem from the underlying cause. For instance, a causal sequence could start with "chronic obstructive pulmonary disease, unspecified (ICD10: J44.9)" and proceed to "other disorders of lung (ICD-10: J98.4)." Here, ICD-10 denotes the "10th revision of the International Statistical Classification of Diseases and Related Health Problems," a commonly used coding system in death reporting<sup>2</sup>.

Establishing the causal sequences of death is often challenging, even for experienced physicians, as the process requires careful reasoning using medical experience and medical domain knowledge. Additionally, the limited availability of electronic health records in cases of sudden death further complicates this process.

Comprehensive and accurate reporting of condition sequences leading to death is invaluable for tracking disease prevalence, creating public health interventions, and monitoring intervention effectiveness. This can potentially improve the quality of clinical care and patient well-being. For healthcare professionals, frequently reported sequences aid in classifying disease conditions and uncovering hitherto unobserved causal

---

<sup>2</sup> <https://www.cdc.gov/nchs/icd/icd10cm.htm>

relationships. For patients, these sequences can serve as a warning for early preventative measures.

In an effort to facilitate accurate and timely reporting and minimize subjectivity by reporting physicians, we have developed a decision support system using deep learning approaches [16]. This system learns the correlation between death and available clinical codes and generates the causal sequence of death based on the decedent's disease histories. We summarized three challenges we faced and the solutions we proposed.

The first challenge stemmed from the differing coding systems for clinical conditions. Since January 1999, causes of death in the U.S. have been documented using ICD-10 codes [48]. However, healthcare institutions were still recording patients' health data using ICD-9 codes until October 2015 [49]. The ICD-10 codes are considerably different from the ICD-9 codes in terms of structure and quantity, with the former having almost five times the diagnosis codes of the latter<sup>3</sup>.

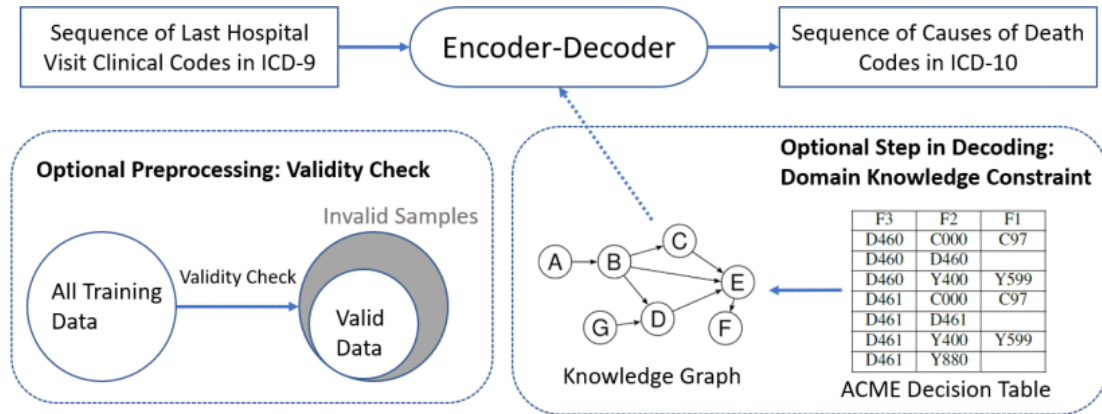
To resolve this discrepancy, we employed natural language translation. Our model accepts diagnosis codes from the last hospital discharge record of the deceased and generates the corresponding causes of death. Analogous to translating English sentences into French, our system can generate a concise ICD-10 code causal sequence from priority-based ICD-9 code discharge records. Autoregressive [32], [50]–[52] and autoencoder models [34], [53], [54] within the Natural Language Processing (NLP) field provide the basis for this translation.

---

<sup>3</sup> [https://www.cdc.gov/nchs/icd/icd10cm\\_pcs\\_faq.htm](https://www.cdc.gov/nchs/icd/icd10cm_pcs_faq.htm)

The second challenge was the necessity for medical domain knowledge. Deep learning models, despite being data-driven, may generate confusing or medically inconsistent sequences, thereby undermining physician trust. To address this, we incorporated expert-curated medical domain knowledge into our deep learning framework. This included the use of expert-curated rules, essentially pairs of causal relationships between clinical condition codes, which were used to guide the model's generation of output sequences, ensuring they align with established medical understanding.

The final hurdle was data interoperability in death reporting. Currently, the U.S. National Center for Health Statistics (NCHS) coordinates with 57 reporting jurisdictions across the United States to collect mortality data [55]. These jurisdictions are governed by different regulations and local laws, making data exchange a challenge. To streamline the process and prepare the data for future large-scale analytics, we utilized Fast Healthcare Interoperability Resources (FHIR) [56] for standardized mortality data reporting. We developed an Android-based FHIR application [57] to access electronic health records data. In addition, we have developed an Android version that can pre-populate sections of the death certificate, extract essential information about the decedent's health history, and interface with the deep learning models to provide clinical decision support.



**Figure 1. Overall Structure of this paper. The encoder-decoder model is the main framework for generating sequences of causes of death. Validity check is an optional preprocessing step and domain knowledge constraint is an optional step in decoding.**

In this study, we used encoder-decoder models with attention mechanism to predict the sequence of causes of death from the decedent's last hospital discharge record. We used attention score visualization to identify death-related conditions, while non-relevant symptoms were filtered out. We also demonstrated the feasibility of using encoder-decoder models with ICD-10 input data by mapping ICD-9 codes to ICD-10 codes. We incorporated expert domain knowledge from an ACME (Automatic Classification of Medical Entry) decision table to ensure that model predictions adhered to known relationships. The overall structure is depicted in Figure 1.

To summarize, our work contributes in the following ways:

- This study is the first to develop encoder-decoder models for predicting causal sequences of death based on death reports and decedents' last hospital visit records.

- It is the first work to use attention visualization to identify death-related conditions from available health records, thereby enhancing model interpretability and aiding physicians in predicting causes of death.
- This work introduces the use of the modified BLEU (BiLingual Evaluation Understudy) score, a standard measure in NLP sequence-to-sequence translation tasks, to evaluate the performance of deep learning prediction of the causal sequence of death.
- The study enhances data interoperability by implementing a user-friendly, FHIR-based application, making it easier for medical professionals to leverage AI solutions in their practice.

### **2.3 Related Works in Intelligent Death Reporting**

The recent surge in research addressing intelligent death reporting testifies to its significance. For instance, Jiang et al. [58] utilized topic modeling on U.S. mortality data, encompassing multiple causes of death, collated by the NCHS between 1999 and 2014. The study successfully categorized co-existing diseases based on their correlations and charted the temporal evolution of these disease groups. However, due to the inherent uncertainty of unsupervised learning, the study fell short in establishing the ideal number of topic groups, thereby limiting its potential application in clinical practice.

Concurrently, Wu and Wang [59] employed a convolutional neural network (CNN) integrated with a dynamic computation graph to infer the underlying cause of death using the same NCHS mortality data. Leveraging a list of relevant medical conditions, the CNN model achieved a noteworthy accuracy of 75% in predicting the single underlying cause of

death. This promising result underscores the potential of applying advanced machine learning methods to death reporting data.

Hoffman et al. [60], in their analysis, highlighted the poor quality of death reporting data, indicating a 20.1% discrepancy in cause-of-death determinations. They recommended implementing validity checks on death reporting data to eliminate invalid causal pairs of death codes. However, the limitation of their work lies in the absence of validation for subsequent tasks, such as the prediction of the single underlying cause of death, which would have showcased the practical value of such validity checks.

In a recent study published in the *Journal of Biomedical and Health Informatics* [61], a novel method was presented for the automatic extraction of the primary cause of death from verbal autopsy questionnaires. This approach utilized a recurrent neural network (RNN) with an attention mechanism to learn the textual representation from the free-text questionnaire data, offering visual attention scores to enhance interpretation of the outcome. This demonstrates the potential of advanced NLP techniques in deciphering unstructured data and contributing to improved death reporting.

Meanwhile, RNN models have also been applied to mortality prediction. Yu et al. [62] proposed a multi-task RNN model with attention mechanisms that successfully predicts patients' hospital mortality, exhibiting higher sensitivity than the Simplified Acute Physiology Score (SAPS-II). The auxiliary task in this multi-task RNN model entailed the reconstruction of patients' physiological time-series data, thus shedding light on the potential benefits of multi-task learning in mortality prediction tasks.

## **2.4 Causal Sequence of Death**

### 2.4.1 Dataset

CCU_USED	CCU_DAYS	AGEGRP1	AGEGRP2	DX1GRP1	DX1GRP2	DXILEAD	CODE800	RELCAUSE1	RELCAUSE2	RELCAUSE3	RELCAUSE4	RELCAUSE5							
0	0	11	5	17170	39	9999	I500	11R688	21I500										
RELCAUSE6	RELCAUSE7	RELCAUSE8	RELCAUSE9	RELCAUSE10	RELCAUSE11	RELCAUSE12	RELCAUSE13	RELCAUSE14	RELCAUSE15	RELCAUSE16									
							NA												
RELCAUSE17	DX_1	DX_2	DX_3	DX_4	DX_5	DX_6	DX_7	DX_8	DX_9	DX_10	DX_11	DX_12	DX_13	DX_14	DX_15	DX_16	DX_17	DX_18	DX_19
	96500	51881	570	5849	2764	V854	5070	34830	5990	0389	99594	E8502	9654	E8504	27801	28860	4280	41401	2449
DX_20	DX_21	DX_22	DX_23	DX_24	DX_25	DX_26	DX_27	DX_28	DX_29	DX_30	DX_31	DX_32	DX_33	DX_34	DX_35	DX_36	DX_37	DX_38	DX_39
496	3051	2724	7245	45981	32723	5859	40390	4169	78791	71590	<NA>	<NA>	<NA>	NA	<NA>	NA	NA	NA	NA
DX_40	DX_41	DX_42	DX_43	DX_44	DX_45														
<NA>	<NA>	<NA>	<NA>	<NA>	<NA>														

Diagnostic codes

Causes of death (Entity axis)

**Figure 2. Sample data from the Michigan data set. The casual sequence of death in ICD-10 for this decedent is I500→R688 (Heart failure leading to Other general symptoms and signs), outlined in green. This decedent had a total of 30 ICD-9 diagnostic codes assigned during the last visit to hospital, outlined in blue.**

This study utilizes discharge records from the final hospital visits of 181,137 decedents, as gathered from the Michigan Vital Statistics Data. Collected by the CDC and its collaborators prior to 2017, this dataset comprises crucial demographic details, diagnostic codes, and procedural codes. It is pertinent to note that the dataset lacks temporal information, featuring only details from the decedents' last hospital visits (as shown in Figure 2). Each decedent's record comprises one line of vital information from the final hospital visit, encompassing up to 45 clinical diagnosis codes, one underlying cause of death, and up to 17 related causes of death. On average, each decedent's record includes 18.84 diagnosis codes and 2.25 causes of death, inclusive of the underlying cause. Conforming to the ICD-9-CM Official Guidelines for Coding and Reporting<sup>4</sup>, diagnosis codes are presented in a priority-based sequence of ICD-9 codes, while causes of death are denoted in ICD-10 codes. We also accessed the NCHS Mortality Multiple Cause Files database<sup>5</sup>, acquired ten years' death codes data (from 2009 to 2018), and computed the

<sup>4</sup> [https://www.cdc.gov/nchs/data/icd/icd9cm\\_guidelines\\_2011.pdf](https://www.cdc.gov/nchs/data/icd/icd9cm_guidelines_2011.pdf)

<sup>5</sup> [https://www.cdc.gov/nchs/data\\_access/vitalstatsonline.htm#Mortalit\\_Multiple](https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm#Mortalit_Multiple)

average length of sequence of death codes as 2.95 codes among over 26 million decedent subjects. Please note that the discharge codes from the last hospital admission could potentially include those from previous admission.

ACME (Automatic Classification of Medical Entry), an ontology comprising medically valid causal relationships between ICD-10 codes, has been developed and propagated by an international team of medical experts [63]. The ACME decision table, which encompasses 95,321 lines of causal relationships, was utilized to constrain medical domain knowledge [60]. To be more specific, if a rule consists of two codes, it can be represented as  $F2 \rightarrow F3$  (cause of death code F2 leading to cause of death code F3); if a rule consists of three codes, it can be interpreted as  $(F1:F2) \rightarrow F3$  (all codes within the subset of F1 and F2 are causes of death that can lead of the cause of death code F3).

#### 2.4.2 *Generating Causal Sequences via Translation*

The task of generating causal sequences can be defined as follows:

**Definition:** The objective of causal sequence generation is to construct an ordered list of clinical codes  $y = y_1, \dots, y_n$  denoting the sequence of conditions leading to death, given a deceased patient's medical history  $x = x_1, \dots, x_m$  represented as a series of clinical codes.

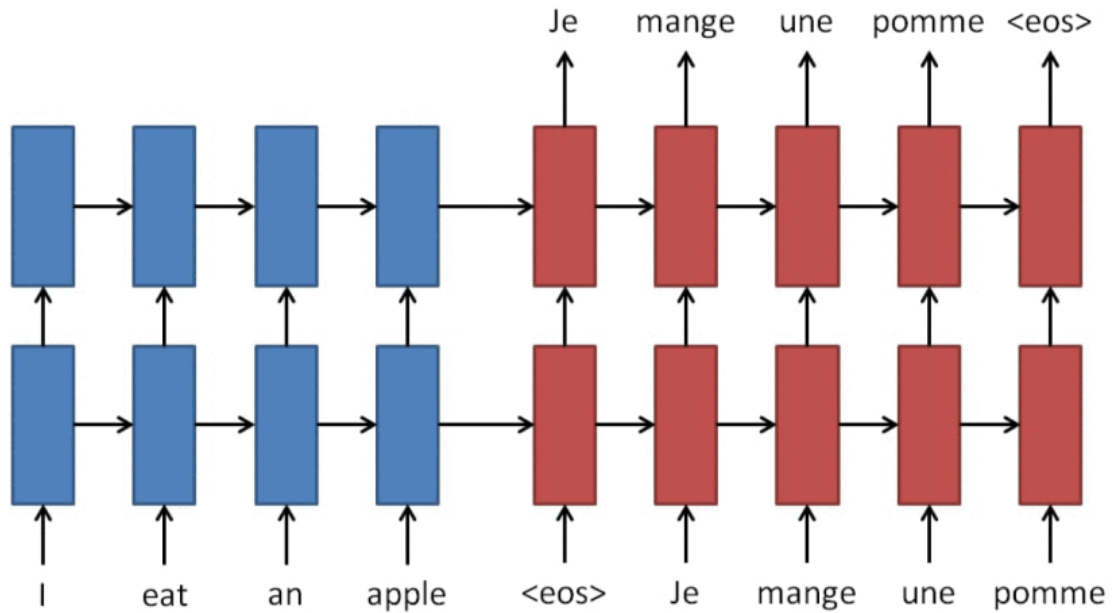
The ultimate goal is to produce a causal sequence of death, delineated as an ordered series of ICD-10 codes representing the causes of death. The input consists of a sequence of ICD-9 diagnosis codes. To construct the output sequence in one domain based on the

input sequence from a separate domain, we resort to the state-of-the-art algorithms of neural machine translation.

The input and output sequence data are partitioned into training, validation, and testing sets in a 7:1:2 ratio. A five-fold cross-validation was applied, and comparable results were obtained through a ten-fold cross-validation, which employed an 8:1:1 split. More comprehensive results can be found in the supplementary file.

## 2.5 NLP Approaches in Generating Causal Sequence of Death

### 2.5.1 Neural Machine Translation: Encoder and Decoder



**Figure 3.** Neural machine translation consists of an encoder (stacked recurrent networks in blue) and a decoder (stacked recurrent networks in red). The symbol *<eos>* is a special token referring to the end of a sentence. Adapted from Luong and Manning, 2016.

The objective of translation is to identify a target sentence  $y = y_1, \dots, y_n$  that maximizes the conditional probability  $p(y|x)$ , given a source sentence  $x = x_1, \dots, x_m$ .

Neural machine translation (NMT) aims to maximize this conditional probability of source-target sentence pairs by fitting a parameterized model using a parallel training corpus. An NMT system essentially comprises two components (as illustrated in Figure 3):

- An encoder, which translates the input sequence  $x$  into a representation  $s$ .
- A decoder, which generates the output sequence  $y$ .

The conditional probability applied by the decoder is expressed as follows:

$$\log p(y|x) = \sum_{t=1}^n \log p(y_t | y_1, y_2, \dots, y_{t-1}, s) \quad (1)$$

The probability of the next generated word,  $y_i$ , is jointly determined by the learned representation vector  $s$  and all previously generated words  $y_1, \dots, y_{t-1}$ .

#### 2.5.1.1 LSTM Encoder - LSTM Decoder

The LSTM Encoder-Decoder framework [64], [65] commences with the encoder reading and encoding an input sequence of embedded vectors  $x$ . Subsequently, the encoder generates a hidden state  $h_t$  at each time step  $t$ , derived from the current input  $x_t$  and the preceding hidden state  $h_{t-1}$ :

$$h_t = f(x_t, h_{t-1}) \quad (2)$$

The source input representation vector  $s$  is given as:

$$s = q(h_1, \dots, h_m) \quad (3)$$

Here,  $f$  and  $q$  denote nonlinear functions. For the fundamental recurrent neural network (RNN)/LSTM model, the conditional probability of the output sequence  $y$  at time  $t$  can be expressed as:

$$p(y_t | y_1, y_2, \dots, y_{t-1}, s) = g(y_{t-1}, h_t, s) \quad (4)$$

Here,  $g$  represents a nonlinear function, potentially multi-layered.

Generic RNN or LSTM encoder-decoder frameworks process sentences word by word using fixed length vectors, which often fail to preserve long-term dependencies. To address this, Bahdanau et al. [32] proposed a soft alignment (soft attention) within a bi-directional RNN model. This allows the model to search for a subset of input words or encoded representation vectors while generating each target word. The soft attention score is computed as:

$$score(s_t, h_i) = v_a^T \tanh(W_a s_{t-1} + U_a h_i) \quad (5)$$

In this case,  $st = f(s_{t-1}, y_{t-1}, c_t)$  is the hidden state of the output word  $y_t$  at position  $t$ , and the context vector  $c_t$  is the weighted sum of the hidden states of the input sequence.  $W_a, U_a, v_a$  are trainable matrices.

Luong et al. [51] introduced global attention, which predicts the position of alignment for the current word prior to computing the context vector using a window

centered around that source position. The general attention score, a subset of the global attention mechanism, is calculated as follows:

$$score(s_t, h_i) = s_t^T W_a h_i \quad (6)$$

Here,  $W_a$  is a learnable weight matrix in the attention layer.

While both global attention [51] and soft attention [32] share a common premise, they differ significantly in their implementation. Global attention utilizes hidden states from the top LSTM layers of both encoders and decoders, whereas soft attention employs the concatenation of forward and backward hidden states in the bi-directional RNN encoder.

In summary, the LSTM encoder-decoder model is relatively straightforward and can be applied to most sequence-to-sequence tasks. However, its performance is somewhat limited, particularly with long sentences.

#### 2.5.1.2 Bidirectional RNN Encoder - LSTM Decoder

A key limitation of the conventional encoder-decoder model lies in its confinement of source sentences to fixed-length vectors, which poses significant challenges for the translation of long sentences [66]. To address this, Bahdanau introduced a bidirectional RNN [32] with soft alignment, enabling the model to jointly learn alignment and translation. Although the bi-directional RNN encoder model improves the learning of word embeddings, it falls short in efficiency compared to the LSTM encoder-decoder framework, and its accuracy is lower than that of transformer models.

### 2.5.1.3 Transformer Model

Despite their advantages, RNN-based encoder-decoder models still struggle with long sentences. Vaswani et al. [33] proposed a solution in the form of the transformer framework, equipped with a multi-head self-attention module that enables parallel encoding of words within the same sentence. A typical transformer comprises a stack of encoders and an equal number of stacked decoders. The input, once embedded, is fed into the bottom-most encoder, with the output from the top-most encoder relayed to all decoders. The output from the uppermost decoder is subsequently passed to a linear layer and a softmax layer, resulting in a predicted sentence. Each encoder is composed of a multi-head self-attention layer and a feed-forward layer, while the decoder incorporates an additional multi-head attention layer that processes both the output from the encoder stack and the output from the preceding attention layer.

Though the transformer model is more time-consuming to train than RNN-based encoder-decoder frameworks, it achieves markedly superior results on NLP tasks [33]. BERT (Bidirectional Encoder Representations from Transformers) [34], a transformer encoder model, has been pre-trained on extensive datasets, including the 800-million-word BooksCorpus and the 2.5-billion-word English Wikipedia. This pre-trained BERT model can be further fine-tuned to enhance performance across multiple NLP tasks.

### 2.5.2 *Decoding and Translation*

A simple approach to decoding is to predict a single word with the highest score based on previous steps. Though efficient and straightforward, this method is susceptible to errors as a single misprediction can corrupt all subsequent predictions. Therefore, a more

reliable strategy, known as "beam search" [67], is adopted. During each decoding step, the decoder generates multiple candidates based on previous outputs, assigning each candidate a non-zero probability value. Beam search maintains the top  $k$  candidates at each step, keeps track of all candidate output paths, and chooses the path with the highest cumulative probability when the sequence ends. Here,  $k$  represents the beam size. A larger  $k$  enhances the robustness of the decoding process but can increase memory requirements and computational time.

Incorporation of medical domain knowledge as constraints during translation is another crucial aspect of our method. The ACME decision table delineates all "feasible" pairwise causal relationships between ICD diagnosis codes [60], [63]. Leveraging this decision table, we construct a domain knowledge graph on all diagnosis codes from the Michigan data before training. Nodes represent diagnosis codes, and directed paths are added between nodes only if a corresponding causal relationship exists in the ACME decision table. During decoding, the networks consult the knowledge graph, including only "feasible" codes in the top  $k$  hypotheses.

### 2.5.3 Evaluation

For quantitative evaluation, we assess how well our proposed causal chain  $\hat{Y} = \{\hat{Y}_1, \dots, \hat{Y}_{M_1}\}$  aligns with physicians' decisions, represented as  $Y = \{Y_1, \dots, Y_{M_2}\}$ . Here,  $Y_i$  refers to the individual codes in the causal sequence of death, while  $M_1$  and  $M_2$  denote the lengths of the chains. Perfect alignment implies  $M_1 = M_2$ , and  $\hat{Y}_i = Y_i$  for  $i = 1, \dots, M_1$ . As this is seldom the case, we calculate a weighted average precision of our alignment in subsequences of varying lengths, i.e., the BLEU score (BiLingual Evaluation Understudy)

[68]. Following natural language processing conventions, we refer to a subsequence of length  $i$  as an " $i$ -gram". BLEU scores range from 0 to 1 (or from 0 to 100 if multiplied by 100), with higher scores indicating better alignment with the clinical decisions made by physicians.

Consider a simple example to illustrate the computation of the BLEU score. In our proposed candidate sequence, the underlying cause of death, Asphyxia and Hypoxemia (R909), leads to Pneumonia, Unspecified Organism (J189), which in turn results in Respiratory Failure, Unspecified (J969).

$$\text{Proposed Sequence: } \hat{Y} = R909 \rightarrow J189 \rightarrow J969$$

The reference sequence ascertained by the physician comprises Asphyxia and Hypoxemia (R909), Pneumonia, Unspecified Organism (J189), and then Acute Respiratory Failure (J960).

$$\text{Reference Sequence: } Y = R909 \rightarrow J189 \rightarrow J960$$

**Table 1. An Example of 1-Gram Precision and 2-Gram Precision in BLEU Score.**

Grams	From Candidate Sequence	Appear in Reference Sequence	Precision
1-gram	(R909), (J189), (J969)	(R909), (J189)	2/3
2-gram	(R909, J189), (J189, J969)	(R909, J189)	1/2

In Table 1, we compute the precision for the listed 1-grams and 2-grams from both  $\hat{Y}$  and  $Y$ . The concept of precision mirrors its definition in a classification setting: it measures the accuracy of predictions in the candidate sequence  $\hat{Y}$  against the reference sequence  $Y$ . The geometric average of these precision metrics constitutes our BLEU score, which is approximately 0.47.

Although BLEU scores typically compute a geometric average for up to 4-gram precision in natural language settings, we only consider up to 2-gram precision and apply clipping to each precision due to the short average length of the causal sequence of death (2.25 codes) in our Michigan dataset. Including 3-gram precision would result in a significant evaluation bias. We also apply a brevity penalty to counteract overly short sentences. Following [68], the modified  $i$ -gram precision is defined as:

$$p_i = \frac{\forall i - \text{grams in } \hat{Y} \text{ that appear in } Y}{\forall i - \text{grams in } \hat{Y}} \quad (7)$$

The brevity penalty BP is defined as:

$$BP = \begin{cases} 1, & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right), & \text{if } c \leq r \end{cases} \quad (8)$$

Here,  $c$  is the length of the candidate sequence (the number of words/codes in the proposed candidate sequence) and  $r$  is the length of the reference sequence (the number of words/codes in the reference sequence).

The BLEU score is thus defined as:

$$BLUE = BP \cdot \exp \left( \sum_{i=1}^N w_i \log p_i \right) \quad (9)$$

In this formula,  $exp$  is the natural exponential function and  $log$  is the natural logarithm function. The weight  $w_i$  is set to  $1/i$  and  $N$  is set to 2.

**Table 2. Our Modified BLEU Score for Different Candidate Sequences.**

	Sequence	BLEU
Reference	I251 → I38 → I429 → I469	
Candidate 1	I429 → I38 → I469 → I251	0.0
Candidate 2	I38 → I429 → I251 → I469	57.7
Candidate 3	I429 → I469 → I251 → I38	81.6
Candidate 4	I38 → I429 → I469 → I251	81.6
Candidate 5	I251 → I38 → I429 → I469	100.0

In the context of clinical interpretation, our modified BLEU score gauges how well our proposed sub-sequences of causal conditions align with the physicians’ decisions. 1-gram precision underscores individual condition code matching, while 2-gram precision assesses the causal relationships between neighboring condition codes. This allows physicians and data-driven algorithms to verify the generated causal relationships for consistency with medical domain knowledge and the ACME decision table.

Table 2 exemplifies different candidate sequences that exhibit perfect 1-gram precision but vary in 2-gram precision. The reference sequence ranges from the underlying cause of death to the immediate cause of death and includes: I251 (Atherosclerotic heart disease of native coronary artery), I38 (Endocarditis, valve unspecified), I429 (Cardiomyopathy, unspecified), and I469 (Cardiac arrest, cause unspecified). The 2-gram precision in our modified BLEU score favors candidate sequences containing a higher number of feasible condition codes with pairwise causal relationships.

In addition to our modified BLEU score, we utilize three other evaluation metrics: accuracy for exact sequence prediction, accuracy for correct individual code prediction (without considering sequence order), and accuracy for correct prediction of the underlying cause of death.

## **2.6 Experiments**

Utilizing the OpenNMT package [69], we trained LSTM encoder-decoder models, bi-directional RNN (BRNN) encoder-decoder models with diverse attention mechanisms, and evaluated the transformer model equipped with a multi-head self-attention module on the Michigan dataset. All models underwent evaluation using the BLEU score and three accuracy metrics.

To broaden the scope of this study, we examined the applicability of encoder-decoder frameworks to current EHRs data, coded in ICD-10. Since the Michigan dataset's input sequence uses ICD-9 codes, we mapped these to ICD-10 codes using the General Equivalence Mappings provided by the Centers for Medicare & Medicaid Services

(CMS)<sup>6</sup>. Specifically, we conducted four experiments on ICD-9 input codes—combining with or without validity checks and knowledge constraints—and one experiment on ICD-10 input codes without these checks or constraints.

Alongside OpenNMT, we integrated the state-of-the-art pretraining model known as the cross-lingual language model (XLM) [54]. Lastly, we visualized attention scores and mapped relationships between source and output sequences.

### 2.6.1 *OpenNMT*

OpenNMT transforms the training, validation, and vocabulary data into PyTorch files for preprocessing. Given the Michigan dataset's relatively small sample size compared to similar NLP tasks, our models feature fewer parameters but maintain an architecture similar to state-of-the-art models. We employed a 2-layer LSTM model during training, with 500 hidden units per layer for the LSTM encoder-decoder framework, in contrast to the 4-layer LSTM model with 1000 units used by Luong et al. [51]. For the bidirectional RNN encoder, we implemented a 2-layer bidirectional LSTM with 500 and 250 hidden units. The transformer model includes six stacked layers, with 2,048 hidden units in feed-forward layers and eight heads in multi-head attention layers.

Training and evaluation of the models utilized an Nvidia GPU Tesla K80. Training an LSTM or bidirectional RNN model for 10,000 steps took around one hour, while training a transformer model took approximately six hours. However, translating all 36,000 testing data points using any of these models took under five minutes.

---

<sup>6</sup> <https://www.nber.org/research/data/icd-9-cm-and-icd-10-cm-and-icd-10-pcs-crosswalk-or-general-equivalence-mappings>

### 2.6.2 *Optional Preprocessing: Validity Check*

In pursuit of improved prediction performance, we introduced an additional preprocessing step, the validity check. For the training and validation data, we employed the algorithm from [60] to remove sentence pairs containing an "invalid" causal relationship between diagnosis codes in the target sentence. This reduced the number of sentences in the training set from 136,753 to 107,711 and in the validation set from 34,385 to 27,009. Following this, we continued to train and translate using the same encoder-decoder models.

### 2.6.3 *XLM: Pretraining*

XLM [54] incorporates BERT's (Bidirectional Encoder Representations from Transformers) [34] masked language modeling (MLM) with the transformer model to enhance translation performance. Preprocessing included tokenizing and applying fastBPE (byte pair encoding) [70] to both monolingual and parallel data. MLM is integral to monolingual language model pretraining. Training comprises three primary steps: denoising auto-encoder, parallel data training, and online back-translation.

Due to our dataset's limited size, we combined all training, validation, and testing data into two corpora for monolingual pre-training. We used MLM perplexities for validation during pre-training. We trained the cross-lingual model with parallel validation data and conducted predictions on parallel test data. Our transformer framework was set with a 512 embedding size and 4 attention heads. The encoder-decoder stacking size varied from 6 layers to 1 layer. We set the dropout rate to 0.1, attention dropout to 0.1, batch size

to 32, and sequence length to 128. We employed GELU for activation and opted for Adam as the optimizer.

## 2.7 Results

### 2.7.1 Attention Comparison

**Table 3. Average BLEU and Accuracy Scores and Standard Deviation in Parentheses Across Five Folds.**

Model	Attention	BLEU	Entire Sequence Accuracy	Individual Codes Accuracy	Underlying COD Accuracy
LSTM	No Attention	17.09 (0.75)	16.55 (0.69)	81.30 (0.39)	54.97 (0.63)
	Soft Attention	17.55 (0.66)	16.87 (0.68)	81.59 (0.41)	55.54 (0.57)
	General Attention	17.62 (1.03)	16.76 (0.96)	81.41 (0.40)	55.15 (0.93)
BRNN	No Attention	17.76 (0.60)	16.66 (0.84)	80.89 (0.39)	55.29 (0.60)
	Soft Attention	<b>17.87</b> (0.74)	16.43 (0.57)	80.77 (1.04)	<b>55.64</b> (1.49)
	General Attention	17.61 (0.74)	<b>16.91</b> (0.69)	<b>81.68</b> (0.28)	55.49 (0.52)
Transformer	Self Attention	17.77 (0.55)	16.47 (0.59)	79.64 (0.50)	54.91 (0.67)

As illustrated in Table 3, the bi-directional RNN (BRNN) encoder-decoder model with soft attention achieves the peak BLEU score, followed by the transformer model and the BRNN with no attention. In comparing various attention mechanisms, the LSTM model with soft or general attention outperforms its counterpart with no attention in terms of BLEU scores. However, BRNN models with diverse attention mechanisms show similar BLEU scores. When juxtaposing LSTM models against BRNN models, LSTM with no attention or with soft attention yields lower BLEU scores compared to its BRNN equivalents. Yet, the LSTM model with general attention exhibits a BLEU score

comparable to the BRNN model with general attention. The highest performing model achieved a BLEU score of 17.87, surpassing the state-of-the-art performance in the natural language domain (English-Czech translation task with a BLEU score of 17.7 and a similar vocabulary size of approximately 10,000) [52].

Furthermore, we also report results for the other three evaluation metrics. The BRNN model with general attention records the highest accuracy for generating both the entire sequence and individual codes correctly. Meanwhile, the BRNN model with soft attention shows the most accurate predictions for the underlying cause of death.

It's noteworthy that all these models, despite the different attention mechanisms, display closely similar performance (less than a 5% difference). Notably, BRNN models utilizing either soft or general attention outperform all other frameworks.

### 2.7.2 Validity Check, Domain Knowledge Constraint and ICD-10 Input Sequence

**Table 4. Average BLEU Scores and Standard Deviation in Parentheses for Five Experiments.**

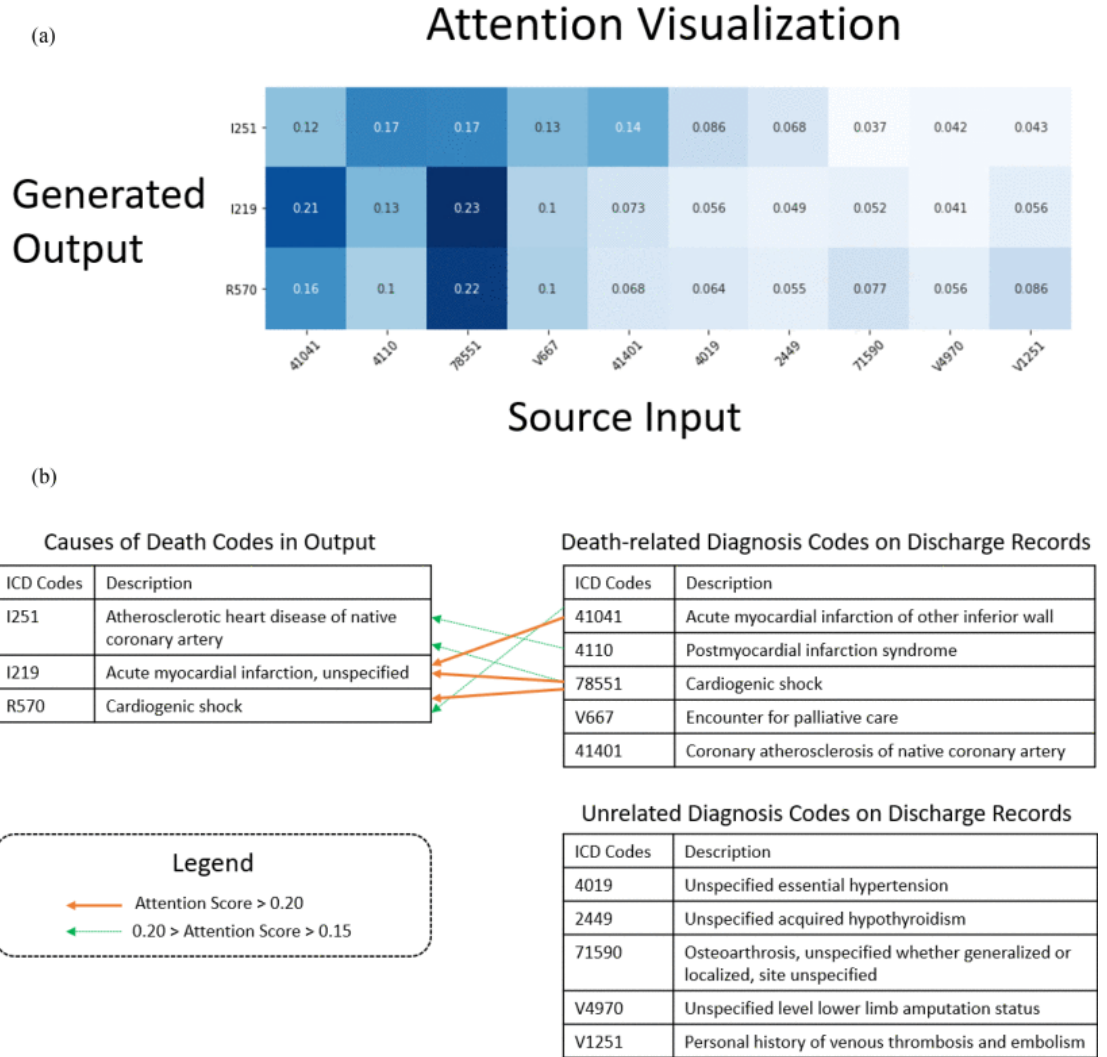
Experiment	Input Data	Validity Check	Knowledge Constraint	LSTM	BRNN	Transformer
1	ICD-9	Not checked	Non-constrained	17.62 (1.03)	17.61 (0.74)	<b>17.77</b> (0.55)
2	ICD-9	Checked	Non-constrained	17.85 (1.18)	<b>18.26</b> (1.10)	15.32 (0.37)
3	ICD-9	Not checked	Constrained	12.61 (6.33)	12.46 (6.26)	14.76 (0.51)
4	ICD-9	Checked	Constrained	12.95 (6.50)	13.16 (6.61)	14.99 (0.42)
5	ICD-10	Not checked	Non-constrained	17.86 (0.50)	17.89 (1.06)	16.31 (0.61)

Table 4 presents the average BLEU scores and their standard deviations (in parentheses) for each encoder-decoder framework across five folds. In Experiment 1 (lacking a validity check in training/validation data and a knowledge constraint in decoding), the transformer model achieves the top BLEU score. Comparing Experiments 1 and 2, we find that adding a validity check in the preprocessing step for training and validation data increases the average BLEU score for LSTM and BRNN models, yet it detracts from the transformer model's performance. This reveals that the validity check has a varied impact on the average performance of different models.

Notably, in Experiments 3 and 4, LSTM and BRNN models witness a marked drop in average BLEU scores, accompanied by a significant surge in their standard deviations. In contrast, the knowledge constraint lowers the average BLEU score for transformer models. Consequently, we demonstrate that the encoder-decoder frameworks can effectively learn the causal relationship between diagnosis codes to such an extent that it's not essential to learn and incorporate the medical domain knowledge constraint from the ACME decision table during the decoding process.

An intriguing comparison lies between Experiments 1 and 5. After post-mapping the input ICD-9 codes into ICD-10 codes, LSTM, BRNN, and transformer models yield average BLEU scores similar to those in Experiment 1. These results carry significant implications: 1) The encoder-decoder frameworks hold promise and stability in generating the causal sequence of death, regardless of whether the input and output data share the same or different coding systems. 2) In situations with no or limited access to the latest EHRs data, we can leverage pre-2015 data to train the models and generate the causal sequence of death.

### 2.7.3 Attention Visualization: A Case Study



**Figure 4. Attention visualization and explanation. In part (a), the attention score matrix is visualized. From top to bottom are the underlying cause of death and immediately causes of death. Darker blue color indicates higher attention scores (the input code is more related with the output code). In part (b), we provide human-readable description to all ICD codes. The identified causes of death for this decedents are cardiovascular diseases.**

In an effort to better comprehend the causal relationship between clinical conditions as reflected in discharge records, we visualized the attention scores produced by the bi-directional RNN model. For instance, ten ICD-9 diagnosis codes from the decedent’s discharge record were used in this visualization. Remarkably, the generated causal

sequence of death was identical to the ground truth as annotated by physicians. We plotted the attention scores for all diagnosis codes in the input sequence (x-axis) against the cause of death codes in the output sequence (y-axis). As depicted in part (a) of Figure 4, a darker shade of blue represents a higher attention score, signifying a strong association between the input diagnosis code and the code in the causal sequence of death. By setting an empirical threshold of 0.1, we identified five diagnosis codes as conditions directly related to death (shown in part (b) of Figure 4). Notably, four out of these five codes corresponded to severe cardiac conditions, aligning perfectly with the identified causes of death. The remaining five diagnosis codes were deemed unrelated to the death due to their lower attention scores.

This attention score visualization not only enhances the model's interpretability by unveiling the relationship between diagnosis codes and causes of death, but also aids researchers and clinicians in identifying death-related conditions from available symptoms on discharge records.

#### 2.7.4 *XLM*

Unexpectedly, the state-of-the-art XLM algorithm significantly underperforms compared to the other encoder-decoder frameworks. Despite exploring various combinations of hyper-parameters, all the BLEU scores remained below 1.

The core algorithm underpinning BERT and XLM, known as masked language modeling, appears incompatible with our dataset. The concept behind masked language modeling involves randomly obscuring a few words in the sentence (either in the source or target sentence) during training, with the intention to recover these masked words based on

their surrounding context. However, given that our target sentence averages only 2.25 words, masking even a single word can pose an extreme challenge to its recovery. Further complicating matters, over 31% of our target sentences comprise only a single word. In such cases, masking the lone word renders its recovery practically impossible.

## **2.8 Discussion**

As noted by [52], larger vocabulary sizes typically enable models to deliver higher BLEU scores. Their hybrid NMT model, utilizing a vocabulary size of 10,000, achieved a BLEU score of 17.7 on an English-Czech translation task. In comparison, our source set vocabulary size is 7616, while our target set is 2649. Despite this, our results surpass those of state-of-the-art outcomes in natural language processing with a similar vocabulary size. Even when comparing with other neural machine translation models [32], [70] possessing larger vocabulary sizes (excluding English-French translations), our outcomes are comparable. A potential extension of this study could involve applying causal inference algorithms [18], [71] to causes of death codes and assessing the average treatment effect.

Wu et al. [59] attained a higher accuracy (75%) in predicting underlying causes of death. However, our model outperforms theirs in generating individual codes, achieving an accuracy of 81%. Blanco's recent work in the Journal of Biomedical and Health Informatics [61] used a similar RNN model to predict a single cause of death code from verbal autopsy questionnaire data, resulting in accuracies of 45.6% and 53.3% for adult and children groups respectively. This is akin to our accuracy for predicting the underlying cause of death. We argue that our models are competent in generating the majority of individual causes of death codes while encompassing the underlying cause of death.

Furthermore, we incorporate medical domain knowledge as a constraint in generating the output sequence. While this constraint negatively impacts the encoder-decoder models, we demonstrate that these frameworks can nonetheless learn the causal relationship between diagnosis codes from the data. We also show that the validity check, a potential pivotal step in the pipeline, could subtly enhance results.

Nevertheless, our work is not without limitations. Firstly, the constraint of medical domain knowledge negatively affects the generation of the causal sequence of death. As the causal relationship derived from the ACME decision table was solely applied during the beam search process in decoding, the constraint failed to positively influence model performance. Furthermore, neither the alignment nor attention mechanisms—core components of the encoder-decoder framework—made use of this constraint. Additionally, despite the demonstrated efficacy of XLM in natural language translation, it failed in our task. This could be due to the inadequacy of masked language modeling in handling extremely short sentences (averaging 2.25 words per sentence).

A potential remedy could be the application of more recent models and pretrained embeddings. Notably, Med-BERT [72] is a pretrained embedding of the BERT model on diagnosis codes from structured electronic health records of over 28 million patients. Med-BERT, pretrained on in-hospital length of stay (LOS) prediction tasks and fine-tuned with disease prediction tasks, could potentially improve performance in generating the causal sequence of death.

Finally, a persisting challenge is the one-word target sentence. Such cases are rare in natural language processing tasks; however, one-word target sentences constitute

31.77% of our training data, 31.68% of validation data, and 31.27% of testing data. These samples significantly attenuate the perceived efficacy of neural machine translation models.

## **2.9 Conclusion**

In this study, we are the first to effectively employ neural machine translation frameworks for predicting the causal sequence of death, bolstering the timeliness, accuracy, and comprehensiveness of death reporting. Our model's performance is gauged using three different accuracy measures, with an impressive achievement of 81.68% accuracy in generating individual codes within the output sequence. Additionally, we leverage attention score visualizations to elucidate the causal relationship between diagnosis codes derived from discharge records. In particular, we manage to identify death-related conditions from available symptoms by mapping all diagnosis codes in the input sequence against all cause of death codes in the output sequence. Finally, we illustrate the utility of a FHIR-based mobile app to retrieve, amend, and upload cause of death data, thereby enhancing clinical integration.

Looking forward, several potential avenues for research present themselves. Firstly, clinical domain knowledge constraints could be configured to interact with attention scores to influence the causal relationship during the model learning phase. Secondly, newer models or pretrained embeddings like Med-BERT could be employed to enhance outcomes. Thirdly, given that our dataset does not contain temporal diagnosis codes, future studies might source data with time-stamped information for more precise analysis. Lastly, our dataset, gathered before 2017, does not encompass COVID-related deaths. Future

collaborations could expand to include discharge records and death certificate records collated during and after the pandemic, thereby allowing our approach to be tested in identifying COVID-related severe symptoms and causes of death.

# **CHAPTER 3. SCREENING IN PRECISION MEDICINE: PROPOSING NOVEL NLP FRAMEWORKS ON UNSTRUCTURED FREE-TEXT NOTES FOR DISEASE IDENTIFICATION**

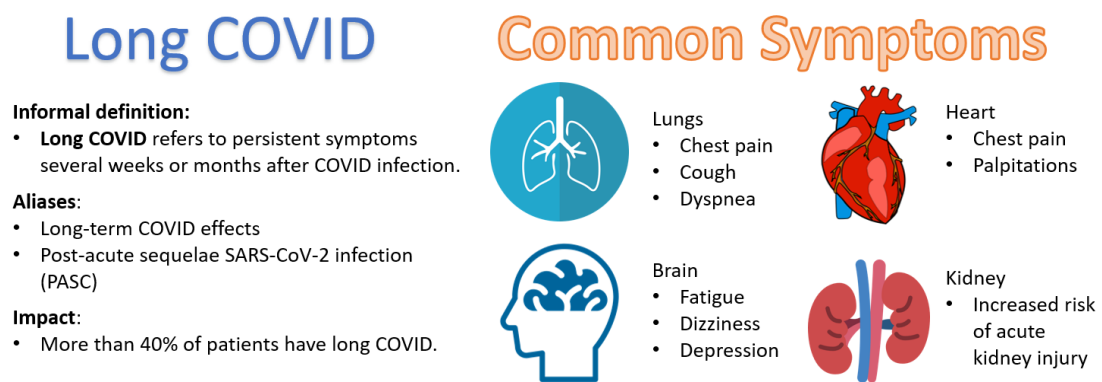
In this chapter, I will present several case studies where natural language processing pipelines are applied on free-text clinical notes and patient messages to identify patients at risk of long COVID and COVID symptoms. I show that NLP pipelines are effective in patient and disease identification, indicating a promising direction for precision medicine.

## **3.1 Using Natural Language Processing on Free-Text Clinical Notes to Identify Patients with Long-Term COVID Effects**

### *3.1.1 Introduction: Long COVID*

Recent studies reveal that approximately 40% to 87% of patients who survive the acute phase of SARS-CoV-2 infection continue to endure one or more long-term conditions [73]–[75]. These conditions, often referred to as "long COVID," comprise a spectrum of symptoms including fatigue, headache, attention disorder, hair loss, and dyspnea [73]. Beyond these symptoms, patients can also experience enduring cardiovascular effects [76] and mental health disorders [77]. Indeed, a recent study shows that the health status of patients, a full 12 months post-COVID infection, can still be significantly impaired relative to the healthy population [77]. Figure 5 summarizes the common symptoms of long COVID.

Though "long COVID" is the most common term, others such as "long-term COVID effects" and "post-acute sequelae of SARS-CoV-2 infection (PASC)" are also utilized to describe the enduring symptoms post-COVID infection [78]. Despite significant efforts devoted to understanding, diagnosing, treating, and preventing COVID-19 [6], [79], [80], no concrete definitions, clinical biomarkers, pathophysiology, or disease courses have yet been identified for long-term COVID effects.



**Figure 5. Long COVID, impact and common symptoms.**

The current clinical practice for detecting long COVID primarily relies on the manual review of patient-reported symptoms on a case-by-case basis. While certain laboratory, radiographic, and physiological tests can assist in diagnosis by excluding other causes, no consensus exists regarding objective test results that would aid in diagnosis. Population-based analysis remains limited due to inconsistent use of specific International Classification of Diseases 10th Revision (ICD-10) codes for long COVID. Manual chart review can help identify patients with persistent symptoms, but this process is labor-intensive and requires clinical expertise. An automatic data-driven approach using non-text data also presents challenges due to inconsistent use of objective tests and ICD-10 codes.

In response to these challenges, this work explores long-term COVID effects utilizing free-text clinical notes and Natural Language Processing (NLP) techniques. A baseline population was generated using results from long-term symptom surveys conducted by Emory Healthcare staff. This enabled clinicians at Emory Clinic to extract outpatient clinical notes from patients who tested positive for COVID and reported persistent symptoms. These notes were subsequently reviewed by clinicians and labeled in a binary fashion for the presence of documented long COVID symptoms. We apply state-of-the-art NLP models to identify long COVID based on these clinical notes and discern potential computational phenotypes using attention interpretation.

Our research represents a significant advancement in the field by providing an objective data-driven solution to identify patients with long COVID. By classifying at the clinical note level, rather than the patient level, we capture the complex reality that an individual patient may display varying signs of long COVID symptoms across multiple visits. This enables clinicians to respond swiftly after high-risk visits, implementing pre-emptive actions to enhance the quality of patient care.

We summarize our key contributions as twofold. 1) This research represents the first known application of NLP techniques on free-text clinical notes to identify patients exhibiting long-term COVID effects. Our novel pipeline achieves a robust recall (sensitivity) score of 0.881 at the note-level classification. 2) We extend our work further by interpreting the classification outcomes and identifying potential patient clinical phenotypes. These data-driven insights can help clinicians better understand and trust our model results, contributing to improved clinical practice.

### 3.1.2 *Related Works*

#### 3.1.2.1 Clinical Diagnosis of Long COVID

"Long COVID" refers to a series of symptoms that persist weeks or even months following initial SARS-CoV-2 infection [74]. The vast majority of patients with long COVID are PCR-negative, suggesting microbiological recovery [81]. The time discrepancy between microbiological and clinical recovery forms the basis of the definition for Post-COVID syndrome [82]. Depending on symptom duration, long COVID can be classified into two stages; post-acute COVID, characterized by symptoms lasting 3 - 12 weeks, and chronic COVID, where symptoms persist beyond 12 weeks [83].

The management of long COVID-19 in outpatient settings often involves numerous medical specialties. These visits typically occur at least one month post-SARS-CoV-2 infection diagnosis, focusing on managing long COVID-19 symptoms through a holistic longitudinal approach. This approach combines primary care follow-up, multidisciplinary rehabilitation services, and the empowerment of affected patient groups with self-care strategies [84]. The broad clinical spectrum of SARS-CoV-2 stems from a lack of pre-existing immunity, leading to an exponential increase in global infection rates [85].

As recommended by the National Institute for Health and Care Excellence (NICE), three consecutive visits form the cornerstone of long COVID management [86], [87]. The first visit, four weeks post-SARS-CoV-2 infection diagnosis, focuses on understanding the patient's health history and assessing any potential underlying causes of prolonged symptoms. The second visit, scheduled after eight weeks, involves evaluating test results, formulating differential diagnoses, and employing diagnostic algorithms to identify

potential symptom causes. Finally, the third visit, 12 weeks post-diagnosis, re-evaluates long-term symptoms and their potential causes [86], [87].

### 3.1.2.2 NLP and Clinical Notes

A critical component of NLP involves learning word embeddings, or real-valued vectors representing single words based on their contextual use [88]. The superior performance of real-valued vectors compared to one-hot vectors can be attributed to two reasons: they reflect each word's semantic meaning through the calculation of distances between the target word and surrounding words, and they are not constrained by vocabulary size, thus offering computational and memory efficiency.

Several state-of-the-art word embedding algorithms in NLP include Word2Vec [89], FastText [90], [91], ELMo [92], and BERT [34]. BERT, along with its two specialized extensions, BioBERT [93] and ClinicalBERT [94], have shown promise in handling various clinical tasks using unstructured clinical notes and other electronic health record (EHR) forms. These tasks include COVID-19 diagnosis [95], [96], acute symptom analysis [97], mortality prediction [98]–[100], and hospital readmission prediction [93].

To our knowledge, no existing work has applied NLP algorithms to detect the presence of long-term COVID effects or provide insights into associated symptoms.

### 3.1.2.3 Common Symptoms of Long COVID

The impact of COVID-19 extends beyond the acute phase of infection, with emerging evidence pointing towards long-lasting health implications that demand the attention of health systems and policymakers worldwide.

In the acute phase of COVID-19, most patients experience mild respiratory infection with symptoms such as odynophagia, anosmia, ageusia, muscle aches, diarrhea, chest pain, headaches, fever, cough, and dyspnea [85]. However, those who do not fully recover and progress to long COVID present with symptoms like fatigue, breathlessness, chest pain, insomnia, dizziness, heart palpitations, joint pain, and brain fog [101].

Emerging research has identified two predominant symptom clusters in individuals with long COVID [102]. The first cluster includes fatigue, headache, and upper respiratory complaints (shortness of breath, sore throat, persistent cough, and loss of smell). The second cluster includes these symptoms along with multi-system complaints, such as ongoing fever and gastroenterological symptoms.

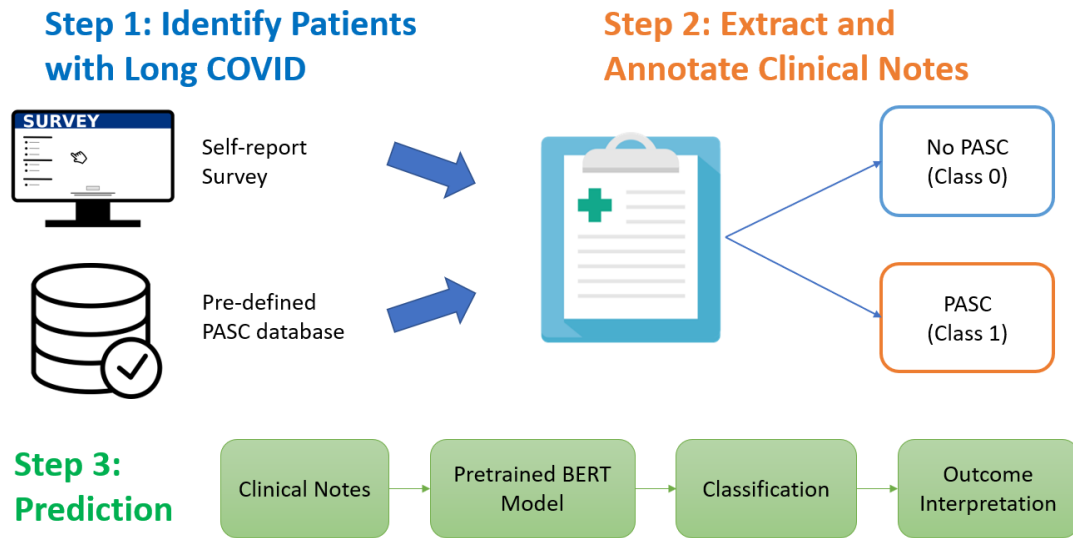
### *3.1.3 Methods: Novel NLP Frameworks*

#### 3.1.3.1 Data Acquisition and Annotation

The methodology employed in this study, as shown in Figure 6, comprises three key steps. Firstly, we identified patients exhibiting long-term COVID symptoms. Following this, we collected and annotated relevant clinical notes. Finally, we utilized advanced BERT models for prediction and analysis.

The data collection process was carried out by clinicians at the Emory Clinic who used two workflows to compile the clinical notes data and assign annotations.

**Workflow 1 for Data Acquisition.** Patients were given surveys inquiring if they experienced persistent symptoms subsequent to their confirmed COVID diagnosis. For patients who reported continued symptoms, clinical notes were obtained from outpatient



**Figure 6. Overall diagram of this work. The three key components are patient identification, clinical notes data acquisition and NLP prediction.**

encounters with healthcare providers (physician/NP/PA) between 30 days and 365 days post-diagnosis. These clinical notes were categorized as follows:

- Clinical notes without mentioned symptoms were labeled as “no PASC” or class 0 (negative).
- Clinical notes with referenced symptoms were included in the dataset and labeled as “PASC” or class 1.

**Workflow 2 for Data Acquisition.** In order to diversify and augment the sample size, we included an equal number of notes from patients who never contracted COVID and those who attended a dedicated PASC clinic. These notes were obtained in a time-matched fashion.

- If the patient never experienced COVID or long COVID (or the clinical note was recorded before their first COVID diagnosis), we included the note and labeled it as “No COVID” or class 0 in the current analysis.
- We randomly selected samples of clinical notes from a dedicated long-COVID clinic, led by a pulmonologist [103]. If symptoms were reported in the clinical note, the note was included in the dataset and labeled as “PASC” or class 1.

### 3.1.3.2 Implementation of NLP BERT Models

While Recurrent Neural Networks (RNN) [51], [65] used to represent state-of-the-art technology in NLP, particularly for sequence-to-sequence generation tasks, their performance is limited on long sequences [66]. The Transformer model, with its multi-head attention modules [33], was introduced to enable the parallel encoding and alignment of words within a sentence. This facilitates substantial improvements in semantic accuracy, and performance in downstream tasks.

A Transformer model is usually composed of an equal number of encoders and decoders. The input sequence embeddings are fed into the bottom encoder module, and the resultant representation is passed onto all decoder modules. The top decoder module passes its learned representation to a linear layer to generate an output sequence. The encoder module is composed of a multi-head self-attention layer and a feed-forward layer, whereas the decoder module has an additional attention layer which processes the output representation learned from the last encoder module and the representation learned from the previous attention layer.

The core component of the Transformer model is the self-attention layer, where the self-attention score is computed as a weighted sum of the Value matrix  $V$ , the weights being determined by the dot-product of matrices Query and Key ( $Q$  and  $K$ ):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (10)$$

The matrices Query, Key, and Value are generated via linear transformations  $Q = XW^Q$ ,  $K = XW^K$ ,  $V = XW^V$ , where  $X$  is the latent representation from the previous module,  $d_k$  is the embedding dimension, and  $W^Q$ ,  $W^K$ , and  $W^V$  are learnable weight matrices.

Despite being more computationally intensive than RNN-based models, Transformer models can yield superior results. BERT [34] is a Transformer encoder model that is pretrained on large English datasets, employing masked language processing for the pretraining stage. Here, around 20% of words are randomly obscured during training; by learning to recover these masked words, BERT is able to learn semantic representation. Fine-tuning pretrained BERT models can enhance performance on various downstream NLP tasks, such as sequence classification.

### 3.1.3.3 Interpretation of Outcomes

AI models often fall short when it comes to outcome interpretation. However, the advent of Explainable Artificial Intelligence (XAI) has opened up new avenues,

particularly in biomedical and clinical applications [7]. Various XAI solutions, such as perturbation-based [104], [105], activation-based [106], gradient-based [107], and mixed approaches [108], [109], have seen wide application. In this study, we used the Transformer-Interpret Python package<sup>7</sup>, which is based on Integrated Gradients [110] and its variation, Layer Integrated Gradients. These algorithms enhance the interpretability of AI models by offering insights into the decision-making process.

#### 3.1.3.4 Implementation of NLP Classification Pipeline

Clinical notes were extracted and stored on Amazon Web Services (AWS) S3 Bucket, with Amazon SageMaker providing a Python interface to process the clinical notes data and run the deep learning models.

The dataset included 719 annotated clinical notes: 542 were labeled as “no PASC” (class 0) and 177 as “PASC” (class 1). Our task was therefore defined as a binary classification problem, predicting whether a clinical note indicates PASC.

We utilized a five-fold cross validation and divided the training, validation, and testing data into a 3:1:1 ratio. The pretrained tokenizer from Hugging Face was employed, truncating the word embeddings at a maximum of 512 tokens. For sequence classification, we fine-tuned the pretrained ClinicalBERT model [94], the base BERT model (BERT-base) [34], and the PubMedBERT model [111] from Hugging Face. Model training was conducted over 50 epochs, which was more than adequate to stabilize the validation loss curve. Evaluation metrics included accuracy, F1 score, precision, and recall.

---

<sup>7</sup> <https://github.com/cdpierse/transformers-interpret>

For statistical analysis, we followed similar steps as those in [17], conducting an analysis of variance (ANOVA) [112]. The null hypothesis posits that the average classification performance using different models is identical. Consequently, the alternative hypothesis would suggest that one model outperforms the others. Alongside variable-level statistical analysis, we performed a pairwise comparison using the Tukey’s honestly significant difference (HSD) test [113], rejecting the null hypothesis if  $p < 0.05$ .

### 3.1.4 Results and Discussion

**Table 5. The Mean Value of Five-fold Cross Validation Results of Three Different Pretrained BERT Models. The highest recall score is highlighted in bold and underlined.**

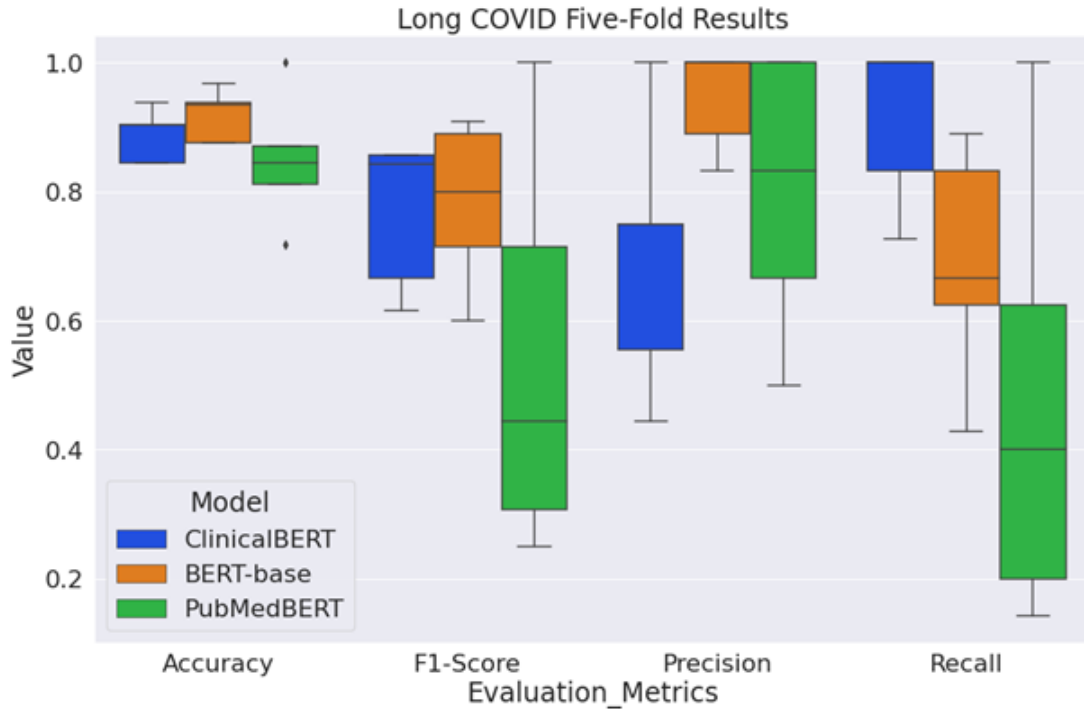
Model	ClinicalBERT	BERT-base	PubMedBERT
Accuracy	0.922	0.910	0.921
F1 Score	0.847	0.815	0.841
Precision	0.821	0.823	0.843
Recall	<b><u>0.881</u></b>	0.808	0.850

#### 3.1.4.1 Classification

Table 5 presents the average value of each evaluation metric for each pretrained model throughout a five-fold cross validation process. These results are also visualized using a box plot in Figure 7.

Here, we would like to discuss about the classification performance for each individual evaluation metric.

- The ClinicalBERT model achieves the highest mean accuracy score; however, the differences are not statistically significant when compared to BERT-base or PubMedBERT ( $p > 0.05$ ). Given the imbalance in our dataset (with a negative to positive ratio of approximately 3:1), the accuracy score may not provide a comprehensive representation of model performance.
- The ClinicalBERT and PubMedBERT models yield comparable F1 scores, suggesting a similar performance in terms of binary classification, provided precision and recall scores are valued equally. Statistical analysis did not reveal any significant difference between these models ( $p > 0.05$ ).
- Although the PubMedBERT model registered the highest precision score, suggesting it correctly identifies the highest fraction of positive samples, it did not significantly outperform the ClinicalBERT or BERT-base models ( $p > 0.05$ ).
- The ClinicalBERT model exhibited a higher recall score than the other two models. Nonetheless, the superiority of the ClinicalBERT model was not statistically significant when compared to BERT-base or PubMedBERT ( $p > 0.05$ ). The recall score, or sensitivity, holds particular importance in this context as it reveals the model's ability to correctly recognize the highest number of clinical notes with long COVID symptoms.



**Figure 7. Box plot of long COVID classification results across five-fold using different models.**

Given the clinical context, recall can be considered the most critical evaluation metric. Our aim is to accurately identify all patients potentially developing long-term COVID symptoms, enabling early intervention and treatment. Therefore, despite the statistical insignificance, we advocate for the ClinicalBERT model due to its superior recall score.

#### 3.1.4.2 Outcome Interpretation

Following the training of the ClinicalBERT model using the training and validation data, we applied the model to the test dataset. To facilitate interpretation, we selected a representative clinical note from the testing data and visualized the token-level attention

information. As depicted in Figure 8, this example is correctly classified as positive for PASC, both in the true and predicted labels.

In this decision-making process, the ClinicalBERT model assigns different values to each word or token. A greener shade (indicating larger positive value) suggests that the token contributes more positively towards the PASC prediction. We consolidated token-

**Table 6. Top and Bottom 10 Words Across All Clinical Notes of the Last Fold of Testing Data.**

Top 110 Words		Bottom 10 Words	
Words	Scores	Words	Scores
Covid	7.89	He	-0.19
Denies	1.08	Measurements	-0.15
and	0.74	fog	-0.14
illness	0.67	granddaughter	-0.13
clinic	0.66	post	-0.12
stairs	0.56	interim	-0.12
system	0.54	had	-0.12
cardiovascula	0.52	resent	-0.11
hospitalization	0.50	carrying	-0.11

Legend: ■ Negative □ Neutral ■ Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	LABEL_1 (1.00)	LABEL_1	10.68	<p>[CLS] 63 ##old , span ##ish speaking only , originally from me ##ico , who comes with her daughter for her 1st pulmonary office visit . she was recently 2 months ago with a diagnosis of co ##vid pneumonia at es ##j ##h . she hasn ##o significant past history e ##ce ##pt h ##yper ##tens ##ion . she was discharge w / o requiring any o ##2 . she feels she is still not quite at her base ##line with some dry to productive cough , e ##pect ##ora ##ting clear secret ##ions , and short of breath during long walks . she hasn ' t been very active since she got ill with co ##rona virus . she states ##lo ##os ##ing app ##roi ##mate ##ly 10 ##l ##bs ##the past 2 months . her appetite is good , denies any fever or chill ##s , no chest pain or fluid retention . she got her modern ##a vaccine in march and planning to get her boost ##er net week . denies any history of tobacco abuse but was e ##pose ##s a lot to wood fire cooking in her native me ##ico . p ##a / la ##t c ##x ##r order today showed improved on both lobes per my interpretation . done today showed no des ##at ##duration 92 - 96 % walk 300 ##meter ##s . good effort . 10 / 22 / 21 / f ##v ##c 84 % 1 . 51 ( 62 % pre ##d ) 1 . 27 ( 68 % pre ##d ) 65 % 18 . 7 ( 47 % pre ##d ) o ##bs ##ttruction . moderate restrict ##ive defect per my interpretation . of systems : no fever , no chill ##s , no fatigue , some weight loss app ##ro 10 ##l ##bs . : no visual disturbance ##s . / nose / mouth / throat : no nasal congestion , no r ##hino ##rr ##hea , no sore throat . : no chest pain , no p ##al ##pit ##ations , no peripheral ed ##ema , no s ##ync ##ope . : + some d ##ys ##p ##nea w e ##ert ##ion , dry to productive cough , clear s ##put ##ums ##put ##um production , no hem ##op ##ty ##sis , no w ##hee ##zing . : no nausea , no vomit ##ing , no heart ##burn , no di ##ar ##hea , n [SEP]</p>

LABEL\_1  
1

"63old , Spanish speaking only , originally from Meico , who comes with her daughter for her 1st pulmonary office visit . She was recently 2 months ago with a diagnosis of COVID pneumonia at ESJH . She hasno significant past history except hypertension . She was discharge w/o requiring any O2 . She feels she is still not quite at her baseline with some dry to productive cough , e pectorating clear secretions , and short of breath during long walks . She hasn't been very active since she got ill with coron a virus . She statesloosing approximately 10lbs the past 2 months . Her appetite is good , denies any fever or chills , no chest pain or fluid retention . She got her Moderna vaccine in March and planning to get her booster net week . denies any history of tobacco abuse but was eposes a lot to wood fire cooking in her native Meico . PA/LAT CXR order today showed improved on both lo bes per my interpretation . done today showed no desaturation 92-96% walk 300meters . Good effort . 10/22/21/FVC 84% 1 . 51 ( 6 2% pred ) 1 . 27 ( 68% pred ) 65% 18 . 7 ( 47% pred ) obstruction . Moderate restrictive defect per my interpretation . of Syst ems: No fever , No chills , No fatigue , some weight loss appro 10lbs . : No visual disturbances . /Nose/Mouth/Throat: No nasal congestion , No rhinorrhea , No sore throat . : No chest pain , No palpitations , No peripheral edema , No syncope . : + some d yspnea w eertion , dry to productive cough , clear sputumsputum production , No hemoptysis , No wheezing . : No nausea , No vom iting , No heartburn , No diarrhea , N"

**Figure 8. Interpretation visualization of a given clinical note in the testing set. The interpretation is for the ClinicalBERT model trained on training and validation data.**

level attention scores into word-level attention scores for each clinical note and subsequently aggregated these scores across all clinical notes in the final fold of testing data. The top and bottom 10 words, along with their corresponding attention scores, are outlined in Table 6. This table elucidates which words significantly influence the model's decision-making process.

While such interpretations may not directly serve as clinical biomarkers, they provide data-driven insights that can help clinicians deepen their understanding of the long-

term effects of COVID. Furthermore, these findings could potentially contribute to the development of new diagnostic criteria or the refinement of current practices.

### *3.1.5 Conclusion*

In this study, we proposed a data-driven methodology to identify patients who display persistent symptoms following their initial COVID-19 infection. Utilizing clinical notes collected from outpatient encounters at the Emory Clinic, we applied state-of-the-art NLP algorithms to predict note-level outcomes. We further dissected the classification results, presenting a ranked list of words that potentially serve as phenotypic indicators. The overarching objective of our work is to provide clinicians with a tool that streamlines the process of identifying patients with long COVID symptoms, thereby enhancing patient care.

Despite these contributions, our study has two primary limitations. Firstly, the existing BERT models are restricted to text inputs of no more than 512 tokens. Consequently, we were required to truncate our clinical notes, inevitably leading to loss of potentially significant information. Secondly, the interpretability of both token-level and word-level attention scores remains limited. Many of the top-ranking words do not directly correlate with symptoms, making it challenging for clinicians to utilize this phenotypic information to accurately diagnose patients who may develop persistent symptoms.

Despite these limitations, future research directions are as follows. 1) Given that our dataset is relatively small, comprising only 719 annotated clinical notes, further expansion of the dataset will be a priority to enhance model robustness and credibility of outcomes. To achieve this, we aim to annotate a larger number of clinical notes to provide

more substantial training and validation data for our NLP models. 2) The restriction to only the first 512 tokens of the clinical notes inevitably result in loss of key information. To mitigate this, future efforts will focus on segmenting clinical notes by their source and subdividing each note into multiple sections. We aim to leverage weakly-supervised or multiple-instance learning on each section of the clinical note, taking into account the fact that we only have note-level, and not section-level annotations. By harnessing the information from all sentences of clinical notes, we expect to enhance classification performance. 3) Additionally, we plan to investigate named entity recognition algorithms and adapt open-source models to our specific dataset. This approach will facilitate the extraction of patterns in medical terminology (pertaining to diseases, symptoms, etc.) that are unique to long-term COVID-19 effects. These patterns could potentially serve as biomarkers, offering more precise and valuable insights to clinicians in their daily practice.

In conclusion, this research lays the groundwork for improved identification and understanding of long COVID symptoms, paving the way for optimized clinical interventions and ultimately, improved patient outcomes.

### **3.2 Use of natural language processing of patient-initiated Electronic Health Messages to Identify patients with COVID-19 Infection**

#### *3.2.1 Introduction: Surge in Patient Messages*

The global health landscape has been dramatically reshaped with the advent and swift propagation of the novel coronavirus, SARS-CoV-2, in 2019 and the subsequent emergence of its variants. These have collectively posed an unprecedented challenge to healthcare systems across the globe. One noteworthy trend in this new reality has been a

surge in telemedicine usage and increased dependence on electronic health record (EHR) communications, with patient-initiated EHR message rates soaring by more than 200% [114]. Although electronic messaging holds considerable potential to enhance communication between clinicians and patients, a surge in message volumes can hamper productivity, impose stress on clinicians already grappling with burnout, and potentially lead to patient morbidity due to response delays [115].

Rapid at-home testing for SARS-CoV-2 has become a viable strategy to receive near-instant results, improve testing accessibility, and enable swift implementation of appropriate isolation measures without exposing healthcare professionals and other patients to risk [116]. A significant number of patients convey positive SARS-CoV-2 test outcomes to clinicians via patient-initiated EHR messages. However, response delays resulting from the burgeoning burden of EHR messaging may lead to missed treatment opportunities [117]–[119]. Oral antiviral therapies, which hold the potential to reduce hospitalization rates, long-term complications, and mortality, have only been researched and authorized for use within a narrow window of five days post symptom onset [116], [120]. Furthermore, there exist limited instances of using natural language processing (NLP) for the analysis of patient messages in a clinical context [121], [122].

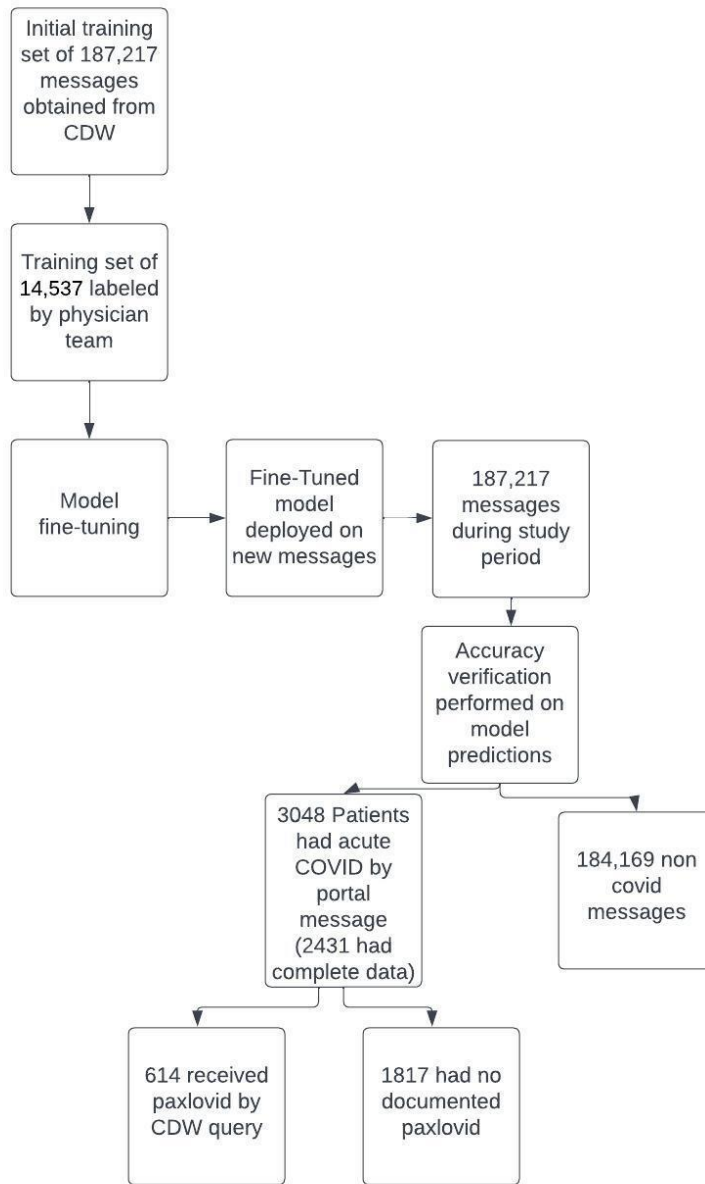
In response to this challenge, we developed an artificial intelligence solution aimed at alleviating the EHR message burden by swiftly identifying patient-composed messages that report positive SARS-CoV-2 tests. The goal is to facilitate the prompt administration of oral antivirals. Through a retrospective cohort analysis, we investigated any potential correlation between the time lapse from when a patient dispatches a message reporting a

positive test to when their message is initially opened by a clinical team member, and whether the patient subsequently received antiviral treatment.

### *3.2.2 Methods: Novel NLP Approach for Automatic Triage*

Our research, involving a retrospective analysis of data from five hospitals in Atlanta, received the requisite approval from Emory University's Institutional Review Board (STUDY1408). We acquired data related to internally administered SARS-CoV-2 PCR and rapid antigen tests, alongside flow sheet-recorded external tests, from Emory's clinical data warehouse through a designated honest broker informatician. As shown in Figure 9, we analyzed a total of 187,217 messages sent by adult patients to their respective healthcare teams from March 30, 2022, to September 1, 2022. For the purpose of training the model, a randomly selected sample of 14,537 messages was categorized by study clinicians into three distinct groups: (1) messages referring to a confirmed COVID-19 infection, (2) messages discussing COVID-19 without mention of a positive test, and (3) messages unrelated to COVID-19. To classify these messages, we utilized transfer learning with a Bidirectional Encoding Representations of Transformers (BERT) Natural Language Processing (NLP) model. Our model, named eCOV, was trained and validated on the 14,537 clinician-categorized messages, maintaining a train/validation/testing ratio of 6:2:2. We then applied a five-fold cross-validation to evaluate performance across all samples in the testing set.

#### 3.2.2.1 Development and evaluation of NLP model



**Figure 9. Flowchart depicting methodology of model training and cohort study analysis.**

Several variations of the BERT model were implemented and assessed, including the base BERT model [34], Bio\_ClinicalBERT [94], and distilBERT [123]. The Bio\_ClinicalBERT model, initiated from the BioBERT model and pre-trained on MIMIC-

III clinical notes, demonstrated superior performance over general-purpose embeddings on several clinical NLP tasks such as the MedNLI natural language inference task and named entity recognition tasks. DistilBERT, a more lightweight transformer model, distilled from the base BERT model, maintained 95% performance in language understanding tasks, despite having 40% fewer parameters and 60% faster computation. All three models presented similar classification performance. However, because the language used in patient messages greatly differs from that of clinical notes, Bio ClinicalBERT did not provide superior classification performance over distilBERT for patient authored EHR text. Therefore, owing to its computational efficiency and speed, we selected the distilBERT model as optimal for this specific application.

To assess the robustness of the distilBERT model, we conducted experiments evaluating its classification performance with limited message samples. We split the dataset into a training and validation set of 10,000 samples and a separate hold-out testing set of 4,537 samples. We randomly extracted 1,000 to 10,000 samples from the training and validation set, maintaining a 6:2 training/validation ratio. Each model was then trained and evaluated on the hold-out testing set. We found that a larger number of training/validation samples resulted in improved results. However, even with just 1,000 samples, the model achieved notable performance. This finding is significant as it suggests that even with a limited number of labeled messages, clinicians could achieve reasonable results by using the same NLP approach in a new hospital setting or for a new clinical task.

All data were gathered and reported in accordance with the CONSORT (Consolidated Standards of Reporting Trials) guidelines for cohort studies. After clinical care for the episode under investigation had concluded, we utilized SQL queries to extract

EHR data. The process of matching chronic conditions, demographics, medications, and other factors was accomplished by correlating unique identifiers and ensured these were active during the clinical period of interest by employing date-matching.

#### 3.2.2.2 Deployment of NLP model

After training, the eCOV three-label classifier was prospectively applied to 2,907 new messages to assess its accuracy. We included only the first message per patient reporting positive results and disregarded any subsequent messages indicating positive outcomes within the study timeframe.

A subset of messages, characterized by patients self-reporting positive outcomes, was reviewed by clinicians to ascertain if the message fell within the window of efficacy for oral antivirals<sup>8</sup>. Patients without documented prescriptions for nirmatrelvir-ritonavir or molnupiravir were categorized as "untreated", while those with a prescription for either of the mentioned antivirals within five days of the message creation date were categorized as "treated". We implemented propensity score matching based on parameters such as age, gender, race, Charlson Comorbidity Index (CCI), and Body Mass Index (BMI), yielding a similar distribution to previous analyses of monoclonal antibody recipients, with 269 patients per category [124].

#### 3.2.2.3 Evaluation of model accuracy

We compared clinician-assigned labels with model-assigned labels to compute the class-specific and the weighted/unweighted average of sensitivity (recall), specificity,

---

<sup>8</sup> <https://www.fda.gov/media/158165/download>

precision, and F1-score. We calculated True Positive (TP), False Negative (FN), True Negative (TN), and False Positive (FP) for each class using a one-vs-rest approach. Class-specific sensitivity (recall), defined as  $TP/(TP+FN)$ , gauged the proportion of patients with COVID-19 correctly identified by the model. Class-specific specificity, calculated as  $TN/(TN+FP)$ , measured the proportion of patients without COVID-19 accurately labeled as COVID-negative by the model. Class-specific precision, or  $TP/(TP+FP)$ , indicated the proportion of patients identified as positive by the model who were indeed part of the target class. Class-specific F1-score, given as  $2TP/(2TP+FP+FN)$ , is a measure combining precision and recall, thereby providing a balanced summary metric of NLP model performance [125]. The unweighted average (macro average) of each performance metric is the arithmetic mean of performance metrics across all classes, disregarding the class sample size. The weighted average takes into account the number of samples (support) in each class for computation.

We determined the time from message creation to the first staff interaction—defined as the moment when a clinical staff member first viewed the message—using clinical data warehouse timestamps. A paired t-test was used to compare time to first message interaction between untreated and treated patients. Figure 9 presents a flowchart illustrating the study methodology, encompassing NLP model training and evaluation as well as cohort study analyses.

### 3.2.3 Results

The analysis incorporated messages from 10,172 patients, presenting an average age of 58 years (SD = 16.65); 6509 (63.99%) of these were female and 3663 (36.01%) were male; 2544 (25.01%) identified as African American or Black, 1508 (14.83%) as Asian, 20 (0.20%) as American Indian or Alaskan Native, 91 (0.89%) as Mixed Race, 28 (0.28%) as Native Hawaiian or Other Pacific Islander, 5980 (58.79%) as White, and 1

**Table 7. Demographics for patient sample whose EHR messages were used to train the model.**

<b>Demographics</b>	<b>N (%)</b>
<b>Total</b>	10172
<b>Age</b>	
18-30 years	615 (6.05%)
31-40 years	1327 (13.05%)
41-50 years	1544 (15.18%)
50-60 years	1865 (18.33%)
61-70 years	2162 (21.25%)
71-80 years	1946 (19.13%)
81-90 years	713 (7.01%)
<b>Sex</b>	
Male	3663 (36.01%)
Female	6509 (63.99%)
<b>Race</b>	
African American or Black	2544 (25.01%)
Asian	1508 (14.83%)
American Indian or Alaskan Native	20 (0.20%)
Mixed Race	91 (0.89%)
Native Hawaiian or Other Pacific Islander	28 (0.28%)
Other	1 (0.01%)
White	5980 (58.79%)

(0.01%) as Other. For comprehensive demographic information, see Table 7. In total, 3048 (1.7%) incoming patient-generated messages pertained to self-reported positive SARS-CoV-2 tests. Among these, 2982 (97.8%) were not otherwise documented in structured EHR data. These 2982 patients, who conveyed COVID-19 diagnoses not recorded in structured EHR data, were examined in our retrospective cohort analysis. The eCOV model demonstrated a macro F1 score of 94%, with individual class F1 scores of 87% for "COVID-other", 95% for "COVID-positive", and 100% for "non-COVID". The respective sensitivities were 85%, 96%, and 100% (Table 8).

Both the treated and untreated groups had comparable qualifications for nirmatrelvir-ritonavir/molnupiravir treatment, with no significant differences noted in age,

**Table 8. Sensitivity (recall), specificity, precision, and F1-score of eCOV.**

	<b>N (%)</b>	<b>Sensitivity</b>	<b>Specificity</b>	<b>Precision</b>	<b>F1-Score</b>
<b>COVID-other</b>	148	0.85	1.00	0.90	0.87
<b>COVID-positive</b>	307	0.96	0.99	0.93	0.95
<b>Non-COVID</b>	2452	1.00	0.99	1.00	1.00
<b>Unweighted Average</b>	N/A	0.94	0.99	0.94	0.94
<b>Weighted Average</b>	N/A	0.99	0.99	0.99	0.99

BMI, and CCI (Table 9). The mean response time was significantly quicker for the treated group (mean = 364.10, SD = 784.47) in comparison to the untreated group (mean = 490.38, SD = 1132.14);  $p = 0.034$ .

### 3.2.4 Discussion

While EHR messaging facilitates patient self-reporting of COVID-19 test results, high message volumes and significant demands on clinician time, coupled with the lack of integration into structured EHR data elements, present significant barriers to timely treatment. Our model, eCOV, offers a novel approach. It is the first clinically-deployable NLP model designed to triage real-time incoming messages of self-reported positive

**Table 9. Differences between COVID-19 positive patients who received antiviral treatment and those who did not. Reported as mean (standard deviation). Statistical significance calculated using paired T-Test (alpha = 0.05).**

	<b>Did not receive antiviral treatment</b>	<b>Received antiviral treatment</b>	<b>P-value*</b>
<b>Age</b>	54.48 (15.90)	56.10 (15.37)	0.955
<b>BMI</b>	28.79 (6.67)	29.59 (6.77)	0.931
<b>CCI</b>	2.71 (3.06)	3.05 (2.94)	0.506
<b>Minutes to First Interaction</b>	490.38 (1132.14)	364.10 (784.47)	0.034

COVID-19 cases. In our study, eCOV successfully identified acute COVID-19 cases from patient-initiated EHR messages with 94% sensitivity.

In large healthcare systems that handle thousands of daily messages, a technological solution is the only sustainable approach for triaging clinically urgent communications. The timing of a physician's response—whether on the same day or several days later—may be the deciding factor in the timely administration of oral antiviral treatment, potentially reducing hospitalization rates or mortality [126]. This remains crucially important, especially as healthcare capacities are strained by concurrent epidemics of COVID-19, influenza, and respiratory syncytial virus. This method also improves accuracy in reporting positive cases; our cohort revealed that only 2.2% of identified cases were documented in the structured data elements typically accessed via traditional EHR database queries.

The ease and accessibility of at-home SARS-CoV-2 tests make them an appealing choice for patients. Those who receive positive COVID-19 results at home and are experiencing severe symptoms can then send an EHR message to their primary care physician for advice on next steps. In some instances, remote treatment can be adequately administered by their primary care physician, reducing hospitalization rates, infection risk for healthcare workers, and the overall strain on the healthcare system. This, however, necessitates prompt initiation of treatment, which is hindered by the high volume of EHR tasks, including patient-initiated messages. Patients whose messages garner slower response times are less likely to receive antiviral prescriptions within the 5-day treatment initiation window. The eCOV model swiftly and accurately identifies and triages patient-initiated messages reporting positive COVID-19 tests.

By accurately classifying patient messages and accelerating the initiation of treatment, NLP integrated into the EHR can potentially enhance clinical outcomes while simultaneously alleviating the burden on healthcare systems. Further analysis and longitudinal studies following real-time integration are required to assess the true clinical impact of this approach.

### *3.2.5 Limitations*

Our study identified a correlation between extended response times and a lack of antiviral prescriptions. While propensity score matching was applied to establish similar classes, factors such as age and pre-existing medical conditions might independently influence the decision to prescribe antivirals, regardless of response time. Furthermore, the scope of our study might not have captured patients who received other forms of treatments, such as monoclonal antibody therapies or intravenous remdesivir, within the treatment window. Numerous factors, including patient preference, vaccination status, insurance coverage, and medication costs, can impact treatment decisions, yet were not considered in our analysis. Therefore, more extensive research is necessary to establish a causal link between response time and treatment.

Additional limitations of this study encompass the lack of visual validation of test results and our inability to systematically verify prescriptions from other facilities or adherence to treatment. Nonetheless, we believe that underreporting rates are significantly higher than instances of false positive reporting [116], [127]. It is plausible that some misclassifications might have arisen due to prescriptions of antivirals from external sources not detected by our database. However, based on an assessment of 50 patients' external

pharmacy records in the EHR, we found a 90% concordance between internal and external prescriptions, suggesting that the probability of such misclassification is likely low. Furthermore, due to our inability to systematically evaluate the onset of symptoms, a subset of patients identified as candidates for antiviral treatment might have exceeded the treatment window by the time when they sent their messages.

### *3.2.6 Conclusions*

Our findings reveal that approximately 98% of patients who reported their positive at-home COVID-19 test results were not otherwise recorded as SARS-CoV-2 positive in the EHR. This indicates numerous opportunities for patients, whose messages might otherwise be overlooked due to the high volume of EHR communications, to receive oral antivirals within the 5-day treatment window. eCOV, an NLP model, demonstrates high accuracy in classifying messages and holds the potential for deployment as an automated triage tool, which can expedite the identification of suitable candidates for treatment.

## **CHAPTER 4. DIAGNOSIS IN PRECISION MEDICINE: DEVELOPING NOVEL NLP FRAMEWORKS ON TIME-SERIES EEG SIGNALS FOR AUTOMATED SEIZURE DETECTION**

### **4.1 Introduction**

Epilepsy is a pervasive neurological disorder characterized by recurrent seizures, impacting over 3 million individuals in the United States alone, with a global reach of approximately 50 million [71], [72]. The core characteristic of these seizures is prolonged aberrant electrical brain activity that disrupts regular cerebral function, and consequently, the individual's normal functioning [128], [129]. Seizures may be triggered by various factors including concussions, fever, photic stimulation, among other elusive factors [129]. The complexity of epilepsy and its prognosis is further demonstrated by recent studies linking Sudden Unexpected Death in Epilepsy (SUDEP) with other top diagnostic codes derived from insurance datasets [18]. This stark reality underscores the critical importance of early and accurate seizure detection. Timely intervention, supported by effective detection, can significantly mitigate the risk of mortality [130].

The multifarious causes of epilepsy often complicate its treatment, necessitating a systematic diagnostic approach. The first line of diagnostic evaluation typically involves an electroencephalogram (EEG), which offers a graphical representation of the brain's electrical activity via electrodes placed on the scalp [128]. The EEG can reveal crucial information, such as the specific location of a seizure, the patterns of its progression and

resolution during an episode, the trajectory of seizure propagation across brain regions, and the unique patterns before (pre-ictal) and during (ictal) a seizure [128].

Seizure detection is an essential clinical task that necessitates accurate algorithms capable of correctly identifying seizures. Hours of EEG data collected from epileptic patients requires clinical annotation to identify optimal segments for human analysis, informing future treatment [128]. Automated seizure detection can enable timely clinical intervention, potentially preventing sudden unexpected death during episodes [131], thereby adding significant value. Consequently, there is considerable interest in developing effective detection algorithms, and a range of human-engineered signal markers have been proposed, including entropy [132], coherence [133], and frequency and wavelet-based features [134].

Seizure detection, nevertheless, still presents significant challenges. Seizures exhibit considerable variability across patients, with possible indications of seizure activity ranging from focal spikes in patient-specific channels to multiple spikes dispersed across all channels [135]. Despite the training clinicians receive to identify these markers, the agreement between raters is approximately 46% on EEG interpretation, showcasing sensitivities ranging from 0.6 to 0.7 and an incidence of 4 to 8 false alarms daily [136]. Therefore, only the combined use of clinical observation and EEG readings can accurately distinguish between a true seizure and a seizure-like state [135].

Additional challenges are due to class imbalance and the unique seizure phenotype exhibited by individual patients [137]. The rarity of events could increase false alarm rates, and in clinical settings, alarms with high false rates may be dismissed as unreliable.

Moreover, for intra-patient seizure detection, the variability of seizure phenotypes suggests that crucial biomarkers for one patient may not exist in the training set, potentially diminishing the algorithm's sensitivity.

In this work, I present three case studies that apply novel deep learning frameworks to improve the accuracy of automatic seizure detection. I show that deep learning techniques, especially those derived from natural language processing, have shown promising results for seizure detection.

## **4.2 Background and Related Works**

### *4.2.1 Seizure Basics*

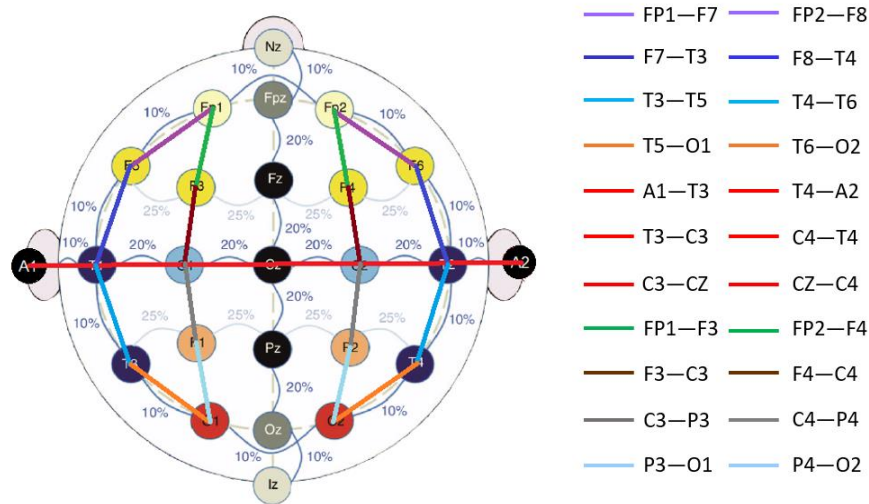
Seizure is a sudden, uncontrolled electrical disturbance in the brain. Seizure impacts behavior, movements, or feelings, and in levels of consciousness <sup>9</sup>. Seizure may cause serious or even deadly complications, such as fall accident, drowning and car accident. To describe seizures, physicians generally consider three important areas: the onset, which refers to the beginning of the seizure, patient's level of awareness during a seizure, and whether any movements happen during a seizure <sup>10</sup>. Seizure can be divided into three different categories: focal onset, generalized onset, and unknow onset. Epileptiform activities starting in one region of the scalp and on one side of the brain are considered as focal seizures. If the patient's awareness is impaired during the seizure, the seizure type is called focal impaired awareness seizure (previously named complex partial seizures) [15].

---

<sup>9</sup> <https://www.mayoclinic.org/diseases-conditions/seizure/symptoms-causes/syc-20365711>

<sup>10</sup> <https://www.epilepsy.com/learn/types-seizures>

Generalized onset seizures typically impact larger area of the scalp, effecting both sides of the brain.



**Figure 10. Electrode locations for a standard 10-20 system with a 22-channel TCP montage.**

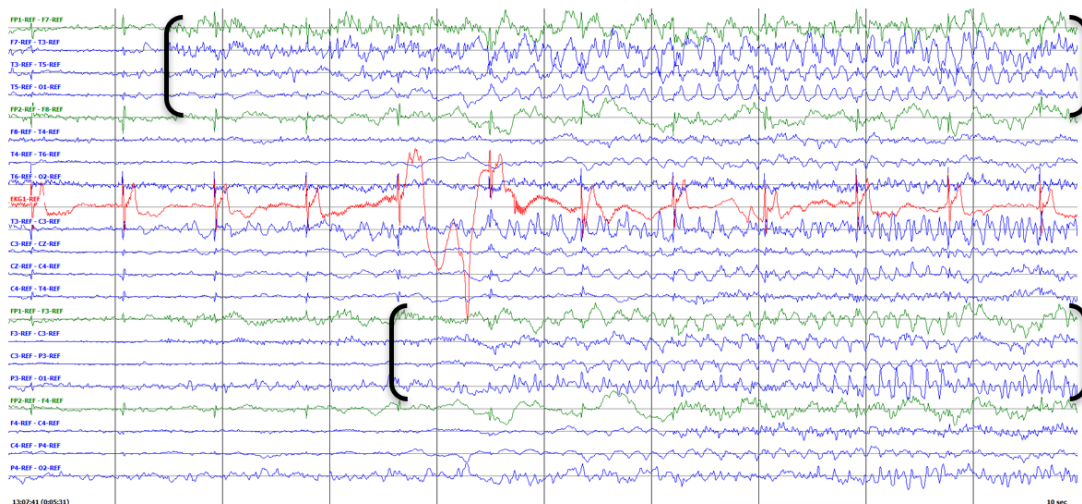
Electroencephalogram (EEG) is a non-invasive method to diagnose seizure in clinical practice. EEGs show electrical activity through a series of electrodes placed on the scalp [128] and can reveal the specific location of a seizure, spreading pattern of the seizure, and the unique morphological patterns before and during the seizure [25]. Common steps in EEG clinical examination includes electrode placement, data acquisition, collecting EEG recordings and generating clinical reports. The most widely used arrangement of scalp electrodes is the standard 10/20 system. The electrodes are symmetrically placed on both left side and right side of the scalp, measuring the electrical potential on scalp of pre-frontal, frontal, parietal, temporal and occipital lobes. A unipolar montage is the difference between the electrical potential recorded at an electrode and a reference node. In clinical practice, however, the neurologists prefer to use bipolar

Temporal Central Parasagittal (TCP) montage, as bipolar TCP montage can help remove signal noise and improve spatial information interpretation of the EEG signal [138]. Figure 10<sup>11</sup> shows a 22-channel TCP montage generated from a standard 10/20 electrode system.

#### 4.2.2 How Clinicians Detect Seizure?

Reliable seizure detection and diagnosis in clinical settings often leverage video EEG tests or monitoring [139]. This technique concurrently captures the patient's physical behavior via video and their brain's electrical activity through scalp electrodes. This bimodal recording provides a holistic view of the patient's physiological and neurological states during a seizure.

Video EEG allows clinicians to establish a correlation between a seizure or event and abnormal electrical brain activity, identify distinctive EEG features, and confirm the seizure type. Clinicians assess key EEG signal characteristics such as evolution, spike and



**Figure 11. A spike and slow wave pattern in a focal non-specific seizure.**

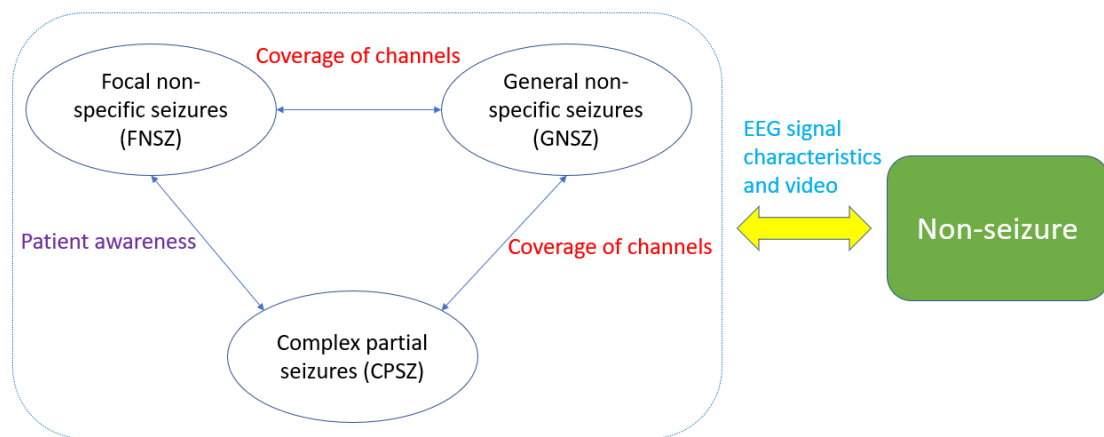
<sup>11</sup> [https://www.isip.piconepress.com/publications/reports/2020/tuh\\_eeg/electrodes/](https://www.isip.piconepress.com/publications/reports/2020/tuh_eeg/electrodes/)

wave morphology, rhythmicity, synchrony, and frequency. Each of these elements provides crucial insights into the seizure type, informing a more targeted treatment plan. By diligently analyzing these parameters, clinicians can achieve a more precise diagnosis and design personalized treatment strategies for individuals afflicted with seizures.

Focal non-specific seizure (FNSZ) is one common category of focal seizures. It covers a broad range of seizure etiologies; consequently, there is significant variation in length, focality and appearance. In a focal non-specific seizure, morphology is the primary indicator of a seizure event. As shown in Figure 11, the seizure shows a spike and slow wave around 5 Hz. To be more specific, spikes are abrupt changes in amplitude that occur in no more than 0.1 second. Slow waves are typically greater than 2 Hz but less than 10 Hz, and may be in the order of alpha, theta, or delta waves. On some occasions, morphology may not be distinct enough to indicate the presence of a focal non-specific seizure. One key indicator would be an increase in amplitude with the decrease in frequency during the seizure event. Another key indicator is postictal slowing, characterized by a slow wave between 1 Hz and 3 Hz just after a seizure.

General non-specific seizures (GNSZ) are very similar to the focal non-specific seizures. They follow the same morphology, frequency, and evolution features, but covering a greater number of channels and effecting larger area of the skull. One major indicator is synchrony, as the complexes of each channel shall have temporal overlapping within 0.2 second.

Complex partial seizures (CPSZ) are focal onset seizures that patient remains impaired awareness when seizure occurs. As complex partial seizures have the same EEG characteristics, including morphology, evolution and postictal slowing, the only way to differentiate CPSZ from FNSZ is to watch the video monitoring tapes or read the clinical notes. Figure 12 summarizes how clinicians detect the top three most common seizures from non-seizure using EEGs and video recordings.



**Figure 12. How clinician detect specific seizure type. The non-seizure type and top three most popular seizure types are included.**

#### 4.2.3 Automated Seizure Detection Using Machine Learning

Seizure patterns displayed in EEGs present significant inter-patient variability, ranging from focal spikes in specific patient channels to generalized spikes across all channels [135]. The task of identifying these markers is labor-intensive and subjective, often resulting in low inter-rater agreement with scores varying from 0.46 to 0.87 [140], [141].

To mitigate subjectivity and reduce manual labor, attention has gradually shifted towards automated annotation systems. These systems utilize hand-engineered features

such as coherence [133], entropy [132], and a spectrum of other statistical and spectral features [142]. These features, often labeled as 'shallow features' before 2014, typically fall into the time domain, frequency domain, and wavelet domain [143].

Deep learning systems for EEG data analysis have recently garnered interest, promising automatic extraction of relevant features from EEG waveforms and potentially boosting seizure detection's efficiency and accuracy. CNNs [144]–[147], Recurrent Neural Networks (RNNs) [25], [148]–[150], Temporal Convolutional Networks (TCNs) [151], [152], and Transformer models [153], [154] are the four primary categories of deep neural networks employed in this endeavor.

CNN-based models, predominantly utilized in image processing, excel at learning local and spatial information but often struggle to capture long-term temporal dependencies [144], [145]. In contrast, RNNs and their derivative, Long Short-Term Memory (LSTM) models, are well-suited for sequence modeling tasks due to their proficiency in capturing temporal sequence information. However, they often face challenges in extracting multi-channel spatial information inherent in EEGs. To address this, CNNs and RNNs are often employed in combination [25], [150], [155], although RNNs still remain susceptible to issues like exploding gradients and have difficulty capturing long-term temporal dependencies.

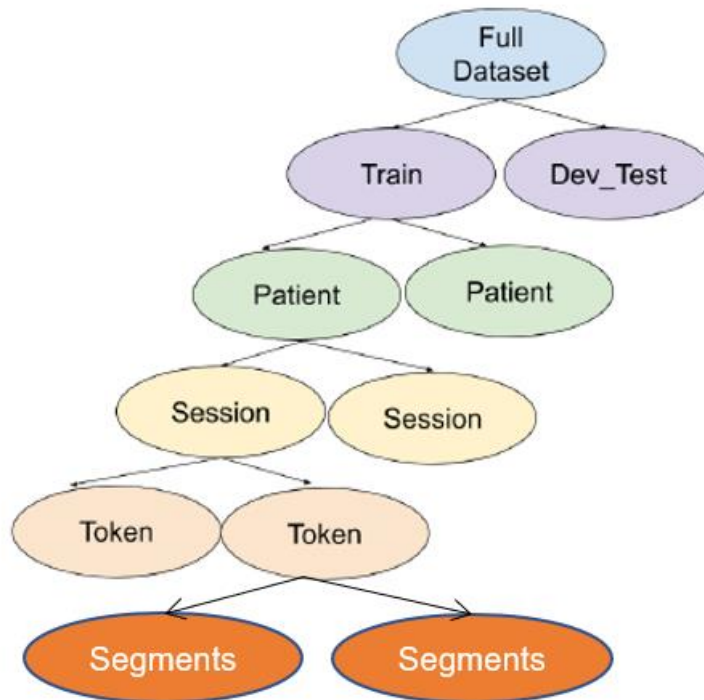
More recent methodologies, such as Temporal Convolutional Networks (TCNs) [151], [152] and Transformer models [33], have gained traction in sequence modeling and text analysis. While they effectively learn long-term temporal information, they, too, can struggle to capture spatial domain information intrinsic to multi-channel EEGs. To

surmount this challenge, Sun et al. [153] proposed a deep learning pipeline that employs a 2D CNN model to generate feature vectors from multi-channel intracranial EEGs (iEEGs) and a transformer encoder module to capture both temporal information within individual channels and attention across different channels. Similarly, Li et al. [156] used a short-time Fourier transform (STFT) to extract frequency domain features from EEGs before deploying CNNs and transformer models for seizure prediction.

While each of these deep learning models demonstrates distinct strengths, further research is required to optimally integrate these models and thereby enhance seizure detection accuracy.

### **4.3 Dataset: TUH EEG Seizure Corpus (TUSZ)**

The Temple University Hospital EEG Corpus is the world's largest public dataset of EEG recordings [157]. The subset, named TUH EEG Seizure Corpus (TUSZ), is the largest public dataset for EEG seizure detection. There are 592 patients with 1185 collective EEG sessions in the predefined training set and 50 patients with 238 EEG sessions in the predefined testing set. The TUSZ uses electrodes arranged over the scalp with a standard 10-20 format. This means that 22 channels are common to all EEG records within the dataset. EEGs are either sampled at 512 Hz, 400 Hz, 256 Hz, or 250 Hz and therefore must be resampled to a common frequency before use. As shown in Figure 13, this dataset uses a hierarchical system with patients arranged between a train and dev/test set, recording sessions for each patient, and smaller token files within each session. Each EEG token file is associated with a time-stamped annotation file that labels the starting time, ending time and seizure type for each seizure event.



**Figure 13. Hierarchical structure of the TUSZ dataset.**

TUSZ suffers from significant class imbalance. This dataset is annotated using 8 separate seizure classes for each seizure event. Many of the classes are not well represented, with generalized non-specific seizures (GNSZ), focal non-specific seizures (FNSZ), complex partial seizures (CPSZ) being the most common classes of seizures in the dataset; For binary classification of EEG seizure detection tasks in this dissertation, all seizure types are included in the analysis as the seizure class. For multi-class classification tasks, only the top three most common seizure types, as well as the non-seizure type, are taken into consideration. Detailed seizure sessions and seizure duration for each seizure type is listed in Table 10.

**Table 10. Number of sessions and duration of top three seizure types and non-seizure type in TUSZ dataset.**

<b>Seizure Type</b>	<b>FNSZ</b>	<b>GNSZ</b>	<b>CPSZ</b>	<b>Non-seizure</b>
Sessions	814	256	79	4,408
Duration (sec)	54,093	28,456	11,180	1,746,772

Another major challenge is the lower signal-to-noise ratio. TUSZ dataset is derived from hospital recording session with far less preprocessing compared to other datasets; the recordings can have significant noise. Furthermore, EEGs are not necessarily taken in standardized environments but instead taken throughout several departments in the hospital. Other authors have suggested that models trained and evaluated on this dataset will have lower metrics than other cleaner and easier datasets [158].

In summary, the TUSZ suffers from multiple challenges which better reflect real world data: 1) more patients would increase patient-to-patient variation; 2) lower signal-to-noise (SNR) ratio that directly reduces the quality of EEG signals; 3) far less preprocessing and 4) significant class imbalance.

#### **4.4 Case Study 1: Adopting Regularization of Deep Neural Networks for EEG Seizure Detection to Mitigate Overfitting**

In this case study, we aim to compare various models and propose a novel deep learning framework for a generalizable intra-patient seizure detection. We will use automatically extracted features tailored primarily for seizure detection to create a session identification model, with its performance serving as an overfitting proxy. Preliminary

findings suggest that generalizing to patients outside of the training set presents significant challenges. We aim to address this by deploying unique strategies to reduce overfitting and specifically target this unwanted bias. To our knowledge, this constitutes the first use of random rescale, random rearrange, and adversarial learning for seizure detection in an intra-patient population.

#### 4.4.1 *Methods*

##### 4.4.1.1 Preprocessing

In EEG recordings, voltage signifies a potential difference between points, allowing various baseline references to define distinct "montages" of EEGs. The TUSZ dataset comprises data from several montage systems, including a linked-ear reference and an average reference system. We selected the average reference system due to its extensive representation in the dataset, surpassing other available reference sets. To ensure consistency, we resampled all records to 250 Hz. As per the Nyquist Theorem, this implies that we would be constrained to frequencies below 125 Hz. The frequency of the EEG signal typically falls between 1 Hz and 50 Hz, with 60 Hz line noise and low frequency muscle and heart artifacts (EMG and ECG) interfering near 1 Hz. Consequently, we employed a fifth order Butterworth bandpass filter to eliminate noise below 1 Hz and above 50 Hz.

Subsequently, we partitioned each file into 4-second segments of non-overlapping data from the record, assigning seizure or non-seizure classifications based on the annotations. To prevent possible information leakage or signal ambiguity, we discarded segments immediately preceding or following a signal.

Given the absence of a predefined validation set in the dataset, we determined our training and validation sets from the existing TUSZ training set, apportioned in an 80:20 ratio, respectively. We utilized the predefined TUSZ test set as the test set for both the traditional machine learning and the deep learning approaches.

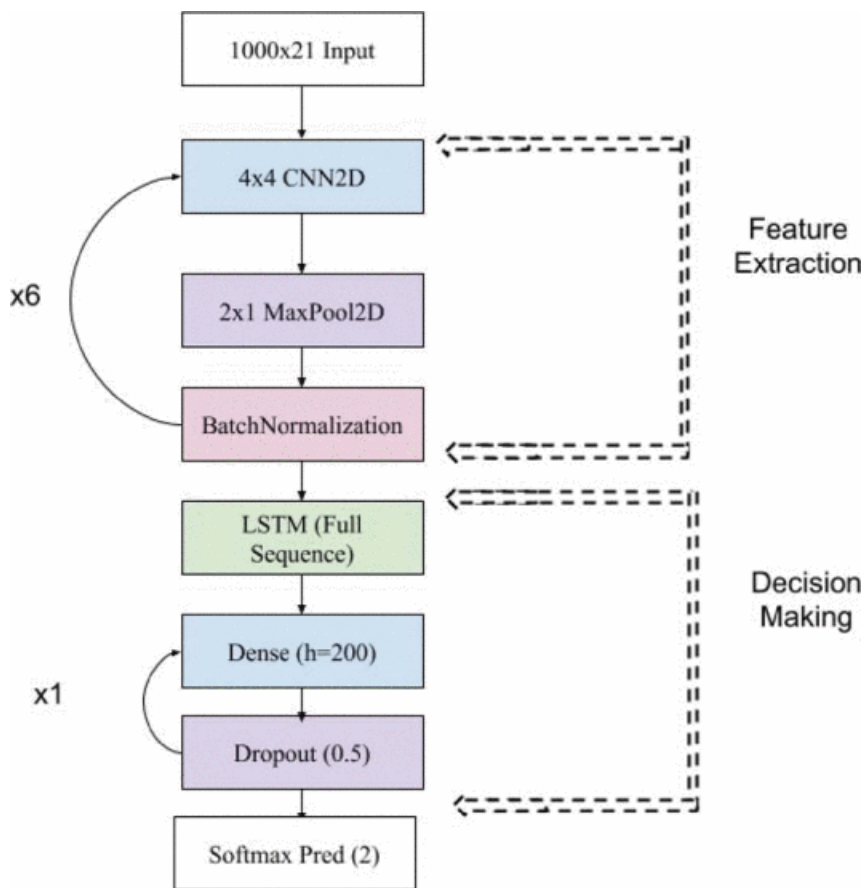
#### 4.4.1.2 Traditional Machine Learning Approach

In our traditional machine learning approach, we employed random forest, XGBoost, and logistic regression algorithms. We trained these models on hand-engineered features extracted from the entire 4-second segment. These features include frequency power derived from the Fast Fourier Transform (FFT), entropy, and coherence. Frequency power was computed by aggregating the total powers of frequency ranges corresponding to alpha, beta, theta, and delta waves. We calculated coherence across pairs of channels to determine the overall coherence sum. To optimize the model performance, we conducted a random search of the hyperparameter space, selecting the best hyperparameters based on validation set performance.

#### 4.4.1.3 Deep Learning Approach

For our deep learning approach, we focused solely on 4-second subsegments. We constructed our model using multiple layers of Convolutional Neural Networks (CNNs), followed by a Long Short-Term Memory (LSTM) layer, then completed by feedforward layers with dropout. We added Gaussian noise of 2  $\mu\text{V}$  as a standard data augmentation technique and applied re-referencing to each 4-second segment, which we found to be critical for deep learning, consistent with other previous approaches in literature [137]. We discarded outlier 4-second segments, defined by a standard deviation exceeding 100  $\mu\text{V}$ ,

due to their excessively noisy nature rendering them unsuitable for extracting discernable EEG patterns. The CNN-2D LSTM model is shown in Figure 14. As our data consisted of only 21 channels, we employed max pooling solely over time segments to avoid the loss of potentially useful channel information in upper layers of the CNN, meaning we refrained from reducing the channel dimension during feature extraction. To avoid overfitting, we



**Figure 14. Architecture of our CNN2D/LSTM. We ran multiple architectures but found that the following had consistently strong results. We found in practice that our networks were most likely to converge if we also applied a learning rate decay of 0.9 each epoch. We used an RELU activation function except for the final layer, where we used softmax. The network trains in parallel another session identification dense layer (not shown), which connects to the feature extraction layers of the network and which is prevented from updating weights to the network.**

implemented early stopping after 20 epochs and preserved the model yielding the best validation F1 score.

To balance the seizure classes, we implemented random under-sampling as part of the traditional machine learning algorithms pipeline. Seizure-positive 4-second segments constituted only 12% of the training set, leading us to maintain only 24% of the total data per epoch (12% seizure to 12% non-seizure). For the neural networks, we employed random resampling at each epoch to ensure class balance while providing the network with as many examples as possible.

#### 4.4.1.4 Adversarial Multi-Task Learning against Session Identification

Convolutional Neural Network (CNN) layers are recognized for their capacity to extract intricate features but remain susceptible to overfitting, much like other networks. We initially observed significant overfitting in our data and endeavored to confirm, as suggested by prior literature, if the CNN-extracted features were specifically tailored to a unique patient/session/system combination particular to a specific recording session [159]. To achieve this, we integrated a session identification layer into a parallel neural network that shared the same CNN feature extraction layers as the original network. This layer executed a 1 of 550 session identification task on all training set sessions. We opted for session identification over patient identification as a patient might undergo multiple recording sessions across various hospital wings and departments. Conversely, a session represents a unique blend of a specific patient, time, and hospital department, thus potentially capturing a wider range of biases.

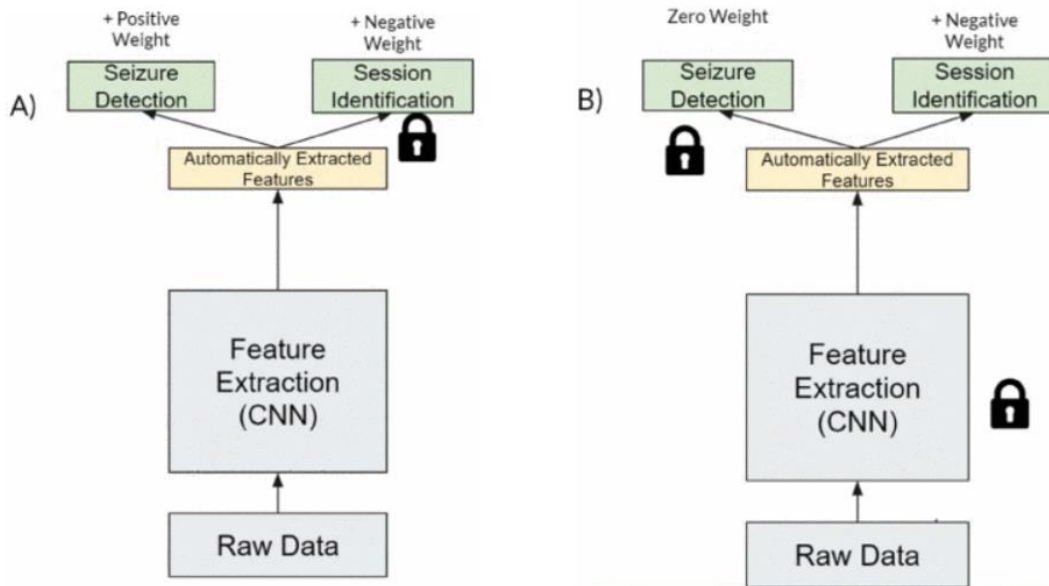
To avoid interference with the seizure detection task, all layers within the session identification network were "frozen," with the exception of the final identification layer. This design meant that feature extraction was primarily driven by the main network, while the secondary network served to measure incidental overfitting. We were able to directly evaluate whether the neural network-created feature set could predict exact training set sessions as the network underwent training. Exceptionally high session identification might indicate a failure in learning generalizable seizure detection features. This could suggest that we were extracting features more suited to a specific patient population rather than a general patient group outside of the training set. We have included the final train session identification accuracy for each model in our results.

We also devised an adversarial multitask learner by appending a session identification layer with a negative weight to the original seizure detection neural network, as illustrated in Figure 15. This layer's weights were updated using shared weights from the parallel, non-interfering session identification network for each minibatch. However, these were then backpropagated negatively through the network, thereby instigating adversarial training for each minibatch in relation to seizure detection and against session identification for the CNN feature extraction layers.

We tested varying seizure and negative session weightings to ascertain if this would decrease the overfitting observed in the network and to explore potential correlations between the generalizability of feature sets that could be used for session identification and those that could not. A higher seizure weight ratio should reduce the adversarial impact, whereas a lower ratio should promote feature extraction that cannot incidentally identify sessions. To our knowledge, this study is the first to attempt seizure detection in this

manner, countering patient-specific features to strive for generalizability across a broader patient population.

#### 4.4.1.5 Regularization Through Random Rescale and Random Rearrangement of Minibatch



**Figure 15. Adversarial multi-task learning approach for seizure detection. We used two neural networks with two task, weights that shared and updated between the networks, and “frozen” layers forced to remain the same after each minibatch update to create a multi-task adversarial network. A) The first network attempts to update the feature extraction for seizure detection and against session identification without updating the weights for the final session identification layer. B) The second network updates session identification weights by predicting positively for sessions, without changing the other weights.**

In order to enhance the learning capacity of our deep learning model and to prevent the deep neural network from overfitting on a specific order of EEG channels, we devised additional experiments based on our initial CNN/LSTM setup. Specifically, for each minibatch, we randomized the channel order. This reordering introduced a level of

variability which compelled the network to adapt to multiple spatial feature combinations as opposed to a fixed combination resulting from a single, consistent EEG channel order.

Another experiment involved randomly scaling the EEG channels within each minibatch, applying a consistent factor within a predetermined range. The scaling factor was selected from a defined interval,  $1/x$  to  $x$ , and applied uniformly to all data in the minibatch, thereby generating new, scaled examples.

These exploratory methods aimed to ascertain whether data augmentation could facilitate our network's generalizability to a more diverse, independent test set. We also sought to determine the impact of these methodologies on the performance of the network when applied to patients absent from the training set. If our augmentative measures led to an enhancement in performance, it would suggest that the regularization was effectively targeting patient- and session-specific aspects of EEG. Conversely, a decrease in performance could indicate that our modifications were inadvertently targeting features of the EEG data that were already generalizable across the population.

#### *4.4.2 Results*

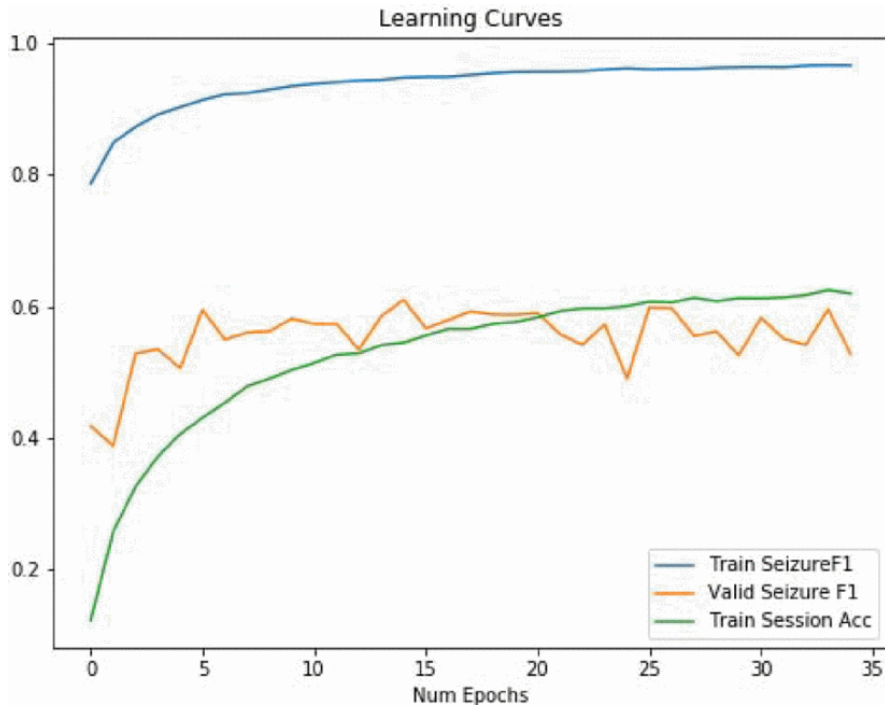
##### 4.4.2.1 Traditional Machine Learning Approach and Deep Learning Approach

As shown in Table 11, XGBoost achieves the best performance among all shallow learning models in AUROC, specificity and F1-score. However, it has much lower sensitivity than random forest and logistic regression models. The logistic regression model has very low specificity, indicating a high false positive rate. This is likely due to the fact

that logistic regression model fails to capture the complex non-linear relationship between hand-crafted features and seizure detection outcomes.

**Table 11. Results for 4 second segments using hand-crafted features and traditional machine learning models, as well as a CNN/LSTM model.**

	AUROC	Sensitivity	Specificity	F1 Score	Session Identification
Random Forest	0.653	0.715	0.653	0.559	N/A
Logistic Regression	0.401	0.743	0.127	0.384	N/A
XGBoost	0.752	0.502	0.906	0.580	N/A
CNN/LSTM	0.716	0.544	0.874	0.556	0.620



**Figure 16. Training/validation F1 score for a vanilla CNN/LSTM model. Session identification accuracy increases while valid seizure F1 fails to increase.**

The deep learning model achieves slightly worse performance than the XGBoost model. The session identification accuracy is 0.620, which indicates that our model learns well to differentiate between sessions. As shown in Figure 16, the deep learning model suffers from significant overfitting issues. Thus, we would like to tackle this issue with regularization and adversarial multitask learning.

#### 4.4.2.2 Adversarial Multitask Learning

In this experiment, we trialed a range of loss weight combinations for session identification and seizure detection, seeking to discern any significant discrepancies in seizure detection performance. The adversarial model demonstrated a capacity to leverage a higher relative adversarial weight, denoted by the negative session weight, which led to a decrease in session identification accuracy. As shown in Table 12, an increase in the relative adversarial weight is inversely proportional to the session identification accuracy.

**Table 12. Results for 4 second segments using a CNN/LSTM model with adversarial multitask learning.**

Seizure Weight	Session Weight	AUROC	Sensitivity	Specificity	F1 Score	Session Identification
50	-1	0.701	0.547	0.858	0.543	0.673
25	-1	0.687	0.440	0.884	0.483	0.251
10	-1	0.713	0.604	0.849	0.575	0.097
5	-1	0.717	0.578	0.864	0.571	0.059

However, the F1 score did not show substantial improvement with our approach compared to the strategy of randomly rearranging channels within each minibatch.

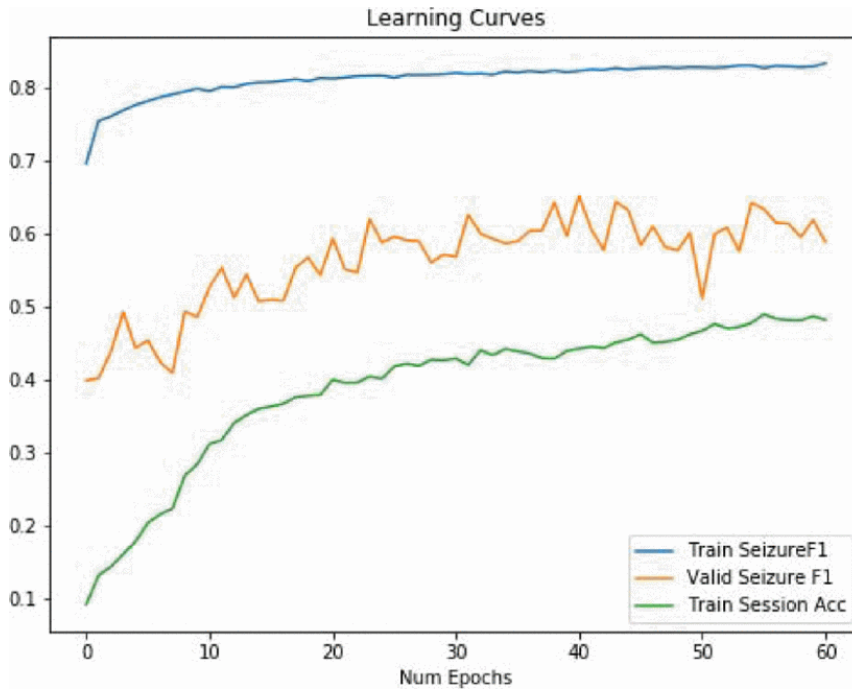
One notable outcome occurred when a seizure weight of 10 was paired with a session weight of -1, yielding a relatively high F1 score, which may represent an optimal equilibrium between these weights. This suggests that the network might be more adept at leveraging data augmentation achieved through random channel rearrangement, as opposed to using an adversarial multitask model. This area warrants further exploration to understand the precise dynamics between these two strategies.

#### 4.4.2.3 Adversarial Multitask Learning with Regularizations

Lastly, we implemented both random rescaling and random rearrangement of channels on each minibatch to modulate the F1 score performance, as delineated in Table 13. The incorporation of an adversarial multitask learning system did not bring about

**Table 13. Results for 4 second segments using a CNN/LSTM model with adversarial multitask learning, with randomly rearranged, randomly rescaled channels each minibatch.**

Seizure Weight	Session Weight	AUROC	Sensitivity	Specificity	F1 Score	Session Identification
50	-1	0.711	0.712	0.800	0.601	0.545
25	-1	0.746	0.719	0.844	0.644	0.482
10	-1	0.755	0.596	0.894	0.613	0.437
5	-1	0.718	0.662	0.832	0.598	0.358



**Figure 17. Training/validation F1 score for a CNN/LSTM model with randomly rearranged channels on each minibatch and random rescaling on each minibatch of 2 with a seizure weight of 25 and patient weight of -1.**

significant increases or decreases in performance. The overall results exhibited minor fluctuations depending on the weighting distribution between seizure detection and adversarial session identification. Notably, in some representative learning curves, there appears to be an interval during which session identification maintains relative suppression, as depicted in Figure 17. Nonetheless, the feature set, extracted by the CNN layers, gradually escalates the session identification score as the network converges. Remarkably, the network that achieved the highest F1 score, incorporating all three regularization strategies, exhibited a session identification accuracy of 0.482.

#### 4.4.3 Discussion

Our study is, to the best of our knowledge, pioneering in introducing several novel techniques aimed at enhancing seizure detection, including:

- Application of session identification as a measure of overfitting arising from biases specific to patients and sessions.
- Employing random rearrangement and random rescaling as data augmentation strategies.
- Utilizing adversarial multitask learning to target session identification.

In order to establish a baseline for seizure detection F1 scores and identify the most significant features, we implemented traditional machine learning models. These results corroborated existing literature, highlighting that coherence between pairs of channels may serve as an indicator of seizures, representing an aberrantly correlated state within certain brain regions. This concept emphasizes the relevance of spatial features reliant on channels, which in turn represent interactions among physically disparate brain areas. Furthermore, we observed that ensemble methods such as Random Forests (RF) and XGBoost outperformed linear approaches like Logistic Regression and our non-randomly rearranged Convolutional Neural Network (CNN). These ensemble learners achieved comparable high F1 scores to our other CNN models, likely due to their inherent ability to mitigate overfitting through training weak learners on different data subsets.

A particularly intriguing observation from our preliminary deep learning experiment was the ability of the CNN to extract features for a 1 out of 550 session identification tasks, despite being specifically trained for seizure detection. This indicated that the ability of CNN feature extraction layers to generalize to an external test set would

be a critical challenge. Our initial baseline network, for instance, was unable to generalize due to its extraction of session- and patient-specific features, as opposed to generalized seizure detection features. A high session identification accuracy of 0.602, coupled with an unimpressive performance for seizure detection with an F1 score of 0.56, substantiated our objective to reduce overfitting and informed our subsequent approaches.

Throughout our experiments, session identification training emerged as a crucial tool in minimizing the type of overfitting evident in the first neural network. Conventionally, overfitting in deep networks is detected when the validation curve plateaus; however, in our case, overfitting became apparent after a few batches due to our methodology. We speculate that our approach of reusing the same sessions and generating multiple 4-second segments from each might have triggered overfitting within a batch. In this scenario, the network is exposed to multiple examples from the same session, captured at different times but sharing similar identifying features. Despite these segments being non-overlapping and potentially representing varying aspects of EEG over time, they lack independence if derived from the same session.

Our most efficient models often demonstrated reduced session identification in exchange for enhanced generalization for seizure detection. The validation score can increase more with each additional epoch compared to the models without techniques targeting session identification. To achieve this, we employed strategies such as random rearrangement and random rescaling to minimize the session identification capability for the extracted feature set. We also discovered that adversarial multitask learning is a potent tool in directly affecting session identification. These strategies were deployed to target signal aspects that could uniquely correlate EEG segments to a specific session, and we

were particularly interested in assessing if altering an EEG with these methods would impact seizure detection performance.

The initial phase involved direct manipulation of the data to regularize the network. We began with an approach of random rescaling and random rearrangement of the channel input, intending to "disrupt" aspects of input data that led to overfitting, such as consistent channel order and maintained EEG magnitude. Although intuitively increasing the randomness in the signal should reduce both seizure detection and session identification performance, our experimental results revealed that the network was prompted to consider new examples outside the original dataset and removed identifying features potentially associated with a specific session. This resulted in improved seizure detection sensitivity and F1 score.

The notable reduction in session identification accuracy suggests that random rearrangement is a highly efficient regularization strategy, due to its presentation of spatial features that might not be inherent in the original data. A single patient may experience seizures localized as patterns in a limited number of channels, but random rearrangement facilitates replication of this pattern across various combinations, thereby enhancing generalization to other patients with seizures in different channels. This increases the sensitivity of our seizure detection classifiers in our test set, driving the increase in our F1 score.

Building upon this, we used randomly rescaling minibatches in combination with random rearrangement as an additional regularization strategy employing data augmentation. Our seizure detection F1 score increased from 0.63 to 0.65, implying a

synergistic enhancement in regularization. Furthermore, this approach reduced session identification to 0.08 with a rescale factor of 1.5. However, the performance metrics for session identification and seizure detection are highly sensitive to the rescale factor utilized. For instance, with a rescale factor of 2, the training session identification seemingly increased to 0.541 compared to lower values of rescale, concurrently reducing the seizure detection F1 score to 0.62. This implies that features for seizure detection might have been impacted more than those for session identification, thereby prompting the network to extract the latter.

We applied random rescale and random rearrangement as effective strategies to reduce overfitting, thereby creating spatially and session invariant features from the input data. On the other hand, our adversarial learning approach attempted to extract patient- and session-invariant features to bolster generalization. Rather than directly manipulating the input data, we influenced the training process of the neural network by penalizing increases in session identification. Initially, we hypothesized that these regularization methods could improve F1 seizure detection scores, but our findings revealed that F1 scores remained approximately 0.6 for most of our chosen hyperparameters. Nonetheless, adversarial multitask learning did enhance specificity compared to the use of random rescale and random rearrange, without a drastic reduction in F1 score.

The sensitivity and F1 scores of our regularized seizure detection models are competitive. We compare our work to Golmohammadi's work [158], as they pursued the same task using the same dataset. It has been noted that models trained on the TUSZ have lower performance compared to other datasets due to inherent issues in the data, thereby making a comparison to their results most appropriate. The researchers achieved a

sensitivity of 0.30 when optimizing for the highest specificity of 0.97 and achieved a sensitivity of 0.39 when accepting a lower specificity of 0.76. Our processing pipeline and application of regularization improved the sensitivity in intra-patient seizure detection to 0.86 compared to this baseline. However, our specificity was not as competitive, implying that our models might generate more false alarms than some of the approaches in the literature. This decrease in specificity might be attributable to the increased sensitivity for such a rare event or due to the use of shorter, 4-second segments instead of longer ones. We are confident that we can improve this in future experiments.

We anticipate that further exploration of other labels in the dataset, describing key aspects of the seizure events such as the seizure subtype and the text clinical notes, could offer additional approaches to counteract overfitting and address remaining challenges in the future. We must also consider the most effective strategy for enhancing specificity and, consequently, reducing false alarms for patients outside the training set, these discoveries and their applications in seizure detection will pave the way for future studies in EEG analysis, more specifically in seizure detection, especially with machine learning algorithms. We look forward to further enhancing the performance of seizure detection models, providing a better tool for clinicians, and enhancing the quality of life for individuals who experience seizures.

#### *4.4.4 Conclusion*

In our study, we underscored the significance of spatial dependencies as key features learned by traditional ML techniques during EEG seizure detection analysis. However, we observed that these features could potentially trigger overfitting in a deep

neural network by becoming aligned with specific patients. Our approach incorporated session identification during the training phase, offering an insightful perspective on overfitting based on session-specific features, even though it may not be the exclusive source of overfitting. To this end, we deployed regularization strategies while meticulously monitoring session identification.

Our unique approach resulted in the generation of spatially and session-invariant features. We specifically engineered our networks to rely less on precise combinations of channels and signal amplitudes, encouraging them to instead decipher novel dependencies instrumental for seizure detection. This study is the first of its kind to apply random rearrangement, random rescaling, and adversarial multitask learning as regularization techniques specifically for intra-patient seizure detection. Our efforts significantly increased the sensitivity to 0.86 from a previously established baseline study, although we noticed a marginal decrease in specificity. Further incorporation of random rescaling yielded a promising enhancement in performance. Nevertheless, the effectiveness of adversarial learning in combination with other regularization methods remains to be thoroughly investigated.

Our experimentation corroborates findings from extant literature indicating that augmented noise and high inter-session variability can induce a neural network to overfit on data rich in features. Specifically, the TUSZ continues to present itself as one of the more challenging datasets for learning due to its real-world sourced data. Despite these challenges, the prudent employment of regularization can alleviate numerous issues and bolster sensitivity for seizure detection, providing an exciting prospect for future research in this arena.

## 4.5 Case Study 2: Mitigating Patient-to-Patient Variation in EEG Seizure Detection using Meta Transfer Learning

One of the unresolved challenges in seizure detection is addressing the high patient-to-patient variation in EEG signals. Previous studies have proposed meta transfer learning for classification on cross-subject brain computer interfaces (BCIs). However, the implementation of this method in seizure detection is largely unexplored.

In this work, we introduce a novel pipeline for multi-class seizure detection, encompassing quality control, preprocessing, and meta transfer learning for feature extraction and classification. We provide an extensive analysis of different components in quality control preprocessing and the benefits and limitations of popular algorithms. Our findings demonstrate the effectiveness of meta transfer learning in minimizing inter-patient variations in EEG signals, leading to successful feature extraction and classification for multi-class seizure detection.

### 4.5.1 Method

#### 4.5.1.1 Meta Learning Basics

Meta learning aims to train a model across numerous learning tasks. Through optimization, this model can achieve superior performance on a variety of tasks from a given distribution, including unseen ones. Each task has its own dataset  $D = (x_i, y_i)$ , comprising both the feature vectors and the corresponding true labels. The parameter set of an optimal model is given by:

$$\hat{\theta} = \arg \min_{\theta} E_{D \sim p(D)} [\mathcal{L}_{\theta}(D)] \quad (11)$$

The probability for the classifier to classify a given data point  $x$  for a specific class  $y$  is,  $P_{\theta}(y|x)$ . Thus, the optimal parameter set shall maximize the probability of true labels:

$$\hat{\theta} = \arg \min_{\theta} E_{(x,y) \in D} [P_{\theta}(y|x)] \quad (12)$$

Meta-learning can also be interpreted as a two-step model update process:

- Training a classifier, referred to as the learner, on a specific task.
- Simultaneously, optimizing the learner's parameters based on its support set.

This update strategy is applicable to both the base learner and the meta learner, details of which will be further discussed in the following sections.

#### 4.5.1.2 Three Groups of Meta Learning Approaches

According to a presentation at the 2018 Conference on Neural Information Processing Systems (NIPS)<sup>12</sup>, meta-learning approaches can be divided into three primary categories: model-based, metric-based, and optimization-based.

---

<sup>12</sup> <http://metalearning-symposium.ml/files/vinyals.pdf>

Model-based meta-learning strategies do not make assumptions about the probability distribution  $P_\theta(y|x)$ . Instead, they rely on models specifically constructed for rapid learning. This swift parameter update can be facilitated through internal or external memory-associated architectures. For instance, the Memory-Augmented Neural Network (MANN) [160] incorporates an external memory storage buffer to quickly assimilate new information and adapt to new tasks with just a few example inputs.

Metric-based meta-learning strategies primarily focus on learning a robust kernel or distance metric. Notable methods within this approach include Matching Networks [161], Prototypical Networks [162], and Relation Networks [163].

Optimization-based meta-learning strategies aim to refine the optimization algorithm to enhance performance on new data or tasks after a few learning instances. One such robust optimization algorithm is the Model-Agnostic Meta-Learning (MAML) [164], which is compatible with models that use gradient descent for updates. Building upon this, Duan et al. proposed a Meta Update Strategy (MUPS) [165] that adapts MAML to optimize the model's performance on unseen subjects.

#### 4.5.1.3 Meta Update Strategy (MUPS)

MUPS comprises a base learner and a meta learner, each featuring a feature extractor and a one-layer classifier. The feature extractor can be any deep neural network capable of extracting features from raw EEG signals. In line with the original paper, and as implemented in this study, we use EEGNet [166] due to its compact design for rapid adaptation and its proven efficacy with EEG signals.

MUPS initiates the process by creating an ensemble of meta tasks  $\{T_1, T_2, \dots, T_M\}$  derived from the training set  $D$ , with each meta task  $T_i$  encompassing  $m$  data points.

The learning process involves cycles known as episodes, where each episode constitutes a meta update involving both the base learner and the meta learner. During an episode, a meta task  $T_i$  is sampled from the task ensemble  $\{T_1, T_2, \dots, T_M\}$ . The support set of this meta task is then divided into two subsets:  $p$  data samples to train the base learner and  $q$  data samples to train the meta learner.

Regarding optimization, the base learner and the meta learner employ distinct optimizers, each targeting the parameters of the feature extractor and the top layer classifier. The feature extractor parameters for both learners are initialized during the pretraining on EEGNet [166]. In subsequent training episodes, the parameter values for

Initialization through pretraining on EEGNet

```

while not done do
  Sample a meta task  $T_i$  from ensemble meta tasks
   $\{T_1, T_2, \dots, T_M\}$ 
  for all  $T_i, i \in 1, \dots, M$  do
    Split  $T_i$  into  $T_b$  and  $T_m$ 
    for base learner updates do
      Optimize base learner parameters using
      support set  $T_b$ 
    end
    Optimize meta learner parameters using sup-
    port set  $T_m$ 
  end
  Update base learner parameters
end

```

**Figure 18. Meta Transfer Learning for EEG Seizure Detection.**

both learners are derived from the meta learner of the preceding episode. The detailed algorithm of this work is shown in Figure 18.

## 4.5.2 *Experiments and Results*

### 4.5.2.1 Quality Control and Preprocessing

We adopt similar preprocessing pipelines as those in case study 1. Specifically, our and preprocessing stages involve several steps:

- Initially, we apply a third-order Butterworth bandpass filter to eliminate signal components below 1Hz or above 50Hz, to mitigate interference from line noise at 60Hz and low frequency muscle and heart noise at 1Hz.
- Subsequently, all signals are resampled to 250Hz in accordance with the Nyquist Theorem.
- Next, we generate 4-second, non-overlapping segments from the resampled signal recordings. Each segment comprises 1,000 data points. To increase computational efficiency for downstream tasks, we opt for non-overlapping segments.
- Segments that exceed the 500 $\mu$ V threshold are discarded due to excessive noise, considering the normal range of an EEG signal lies between 20 $\mu$ V and 100 $\mu$ V.
- We partition the segments into inter-patient training, validation, and testing folds, ensuring no patient overlap.

### 4.5.2.2 Baseline Models

To establish baseline results for seizure detection from EEGs, we conducted experiments on the defined folds. Our feature extraction pipeline involved entropy and variance measurements, as well as Fast Fourier Transform (FFT) features corresponding to alpha, beta, theta, and delta waves for each of the 22 montage channels. We selected the random forest as our baseline shallow learner.

In addition to the random forest, we also evaluated two deep learning models as baselines. The EEGNet model is a flexible neural network architecture designed specifically for Brain-Computer Interface (BCI) tasks [166]. MobileNet [167] is an efficient and effective CNN-based model utilized in various tasks, including feature extraction and classification on images and time-series signals.

#### 4.5.2.3 Meta Learning

The pretraining stage employs EEGNet for feature extraction, aiming to transfer the knowledge acquired from EEGNet as a warm start for meta-learning. During this

**Table 14. Results For Different Models. Macro-F1 Score and AUROC Score Are Evaluated On Four Classes On the Predefined Test Set.**

Model	F1	AUROC
Random Forest	0.256	0.709
EEGNet	0.175	0.502
MobileNet	0.176	0.496
Meta Learning	0.510	0.679

pretraining phase, we train the model for 200 epochs, employing a base learning rate of 0.0005 with a decay rate of 0.5, using Stochastic Gradient Descent (SGD) as the optimizer. For the meta-learning stage, we train the model for 250 epochs with a batch size of 64, employing Adam as the optimizer and a grid search to determine the optimal set of hyperparameters.

#### 4.5.2.4 Results

While our baseline shallow learner using a random forest approach yielded the highest Area Under the Curve (AUC) score, our meta-learning approach attained the highest macro-F1 score and the second-highest AUC score, as shown in Table 14. A high F1 score signals robust precision and recall performance across the non-seizure and three seizure types in classification tasks. Correspondingly, a high AUC score illustrates the competence of our meta-learning model in distinguishing among the four classes.

By taking into account both these metrics, it's clear that our meta-learning model surpasses all baseline models. As such, we convincingly demonstrate that meta transfer learning serves as a competitive and effective method for multi-class seizure detection in the feature extraction and classification of EEG signals.

#### 4.5.3 *Discussion and Conclusion*

In this study, we introduced a pipeline for multi-class EEG seizure detection, which encompasses preprocessing, automatic feature extraction, and classification. Importantly, we illustrated the potency of meta transfer learning in feature extraction and classification tasks for multi-class EEG seizure detection. The refined meta transfer learning algorithm

outperforms baseline shallow and deep learning models. Experimental results highlight that meta transfer learning offers a proficient means to mitigate patient-to-patient variation, leading to superior classification accuracy.

However, this work is not without limitations. A potential drawback stems from the utilization of EEGNet for pretraining and meta-updates. Primarily designed for brain-computer interfaces (BCI), EEGNet has traditionally been applied to various clinical tasks involving EEGs. While effective in BCI tasks, EEGNet's application to seizure detection tasks doesn't demonstrate equal competitiveness. Future research should consider implementing alternative deep neural network architectures to enhance feature extraction performance.

In addition to the Model-Agnostic Meta Learning (MAML) strategy, we intend to explore other meta-learning approaches. Given the success of these approaches in natural image classification, one viable option may involve transforming time-series EEG signals into spectrogram images before applying meta-learning algorithms. While this process may be time-intensive, we believe this presents a promising avenue to further enhance multiclass seizure detection results. We also aspire to explore alternative neural network architectures, comparing metrics both with and without a MAML strategy.

Another promising direction for future research lies in interpreting meta-learning results. We are particularly interested in identifying which segments of the EEG signals are most instrumental in distinguishing different seizure classes, including non-seizure types. If feasible, we plan to collaborate with clinical experts to validate these interpretation results in future studies.

#### **4.6 Case Study 3: Developing Novel CNN + Transformer Framework for Automated Seizure Detection on Multi-Channel EEGs**

The quest for automatic seizure detection has resulted in substantial advancements, with both traditional feature engineering and deep learning techniques coming into play. Yet, the inherent complexity of EEG signals has posed considerable challenges, often leading to sub-optimal performance in existing seizure detection methods. Patient-to-patient variations and the difficulty in extracting seizure-specific features have further resulted in significant overfitting problems.

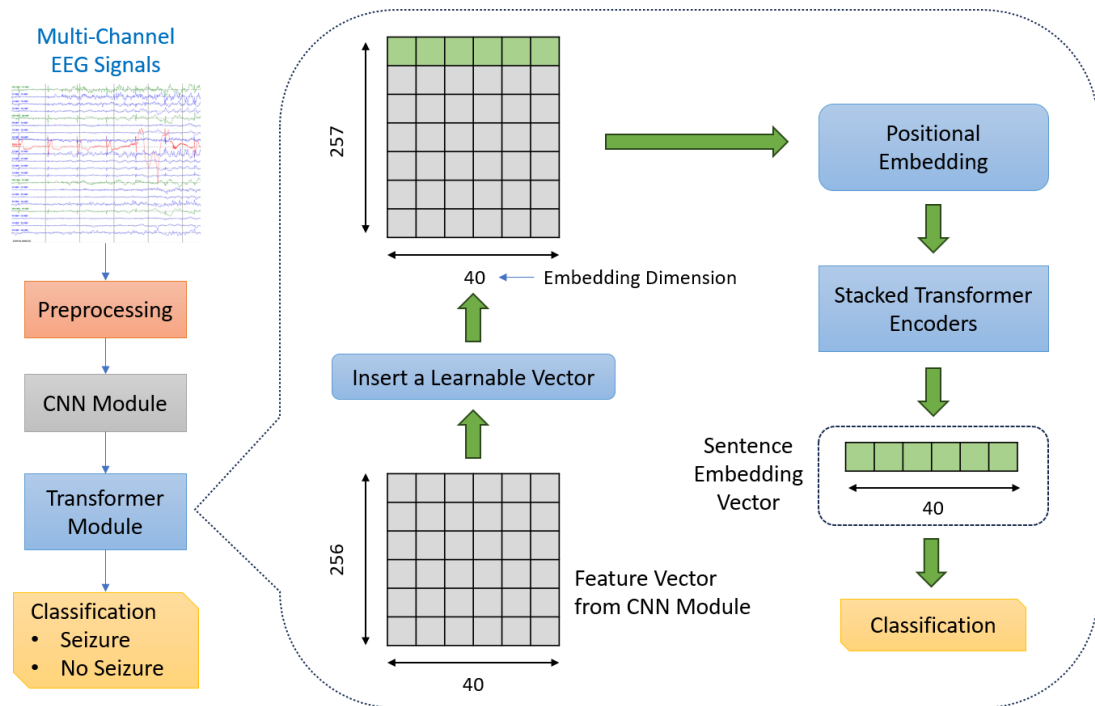
Given these challenges, there is an immediate need to introduce innovative AI methodologies that can facilitate more accurate, efficient, and dependable seizure detection. This case study introduces a groundbreaking deep learning framework for automatic seizure detection. The proposed model integrates a Convolutional Neural Network (CNN) module and a transformer encoder module. The CNN module adeptly extracts both temporal and spatial features from multi-channel EEG data, and the transformer encoder module further refines these by capturing long-term sequential information from the feature vectors. The efficacy of our approach is substantiated by an unweighted average F1 score of 0.731, an unweighted average precision of 0.724, and an unweighted average recall (sensitivity) of 0.744.

This case study has two primary contributions:

- We introduce a groundbreaking deep learning framework tailored for patient-specific seizure detection. The proposed model efficiently captures

spatial and temporal seizure-specific features from raw, multi-channel EEG signals while adeptly tracking long-term sequential information.

- We conduct a comprehensive series of experiments to evaluate our proposed model on the largest public EEG seizure detection dataset currently available. The results affirm that our model surpasses competitive state-of-the-art approaches in terms of performance.



**Figure 19. Overall flowchart diagram for our proposed approach for EEG seizure detection.**

#### 4.6.1 Methods

Our proposed approach, delineated in Figure 19, consists of a preprocessing module, a Convolutional Neural Network (CNN) module for feature extraction, and a Transformer module for classification.

#### 4.6.1.1 Data Preprocessing

We first extracted token-level EEG signals and time-stamped annotations from the TUSZ. To filter out extraneous noise and enhance signal clarity, we implemented a bandpass filter to maintain signals within the frequency range of 0.5 Hz to 100 Hz. To further minimize interference from heart rate and power line noise, notch filters were applied to discard signals at 1 Hz and 60 Hz, respectively.

Signals originally sampled at rates divergent from 250 Hz were resampled to this standard frequency. Furthermore, we classified any signal segment exceeding 500 microvolts as noise, excluding these segments from subsequent analysis.

For the binary classification task, seizure signals were segmented into four-second intervals with a 75% overlap to increase data variety, whereas non-seizure signals were divided into four-second segments without overlap.

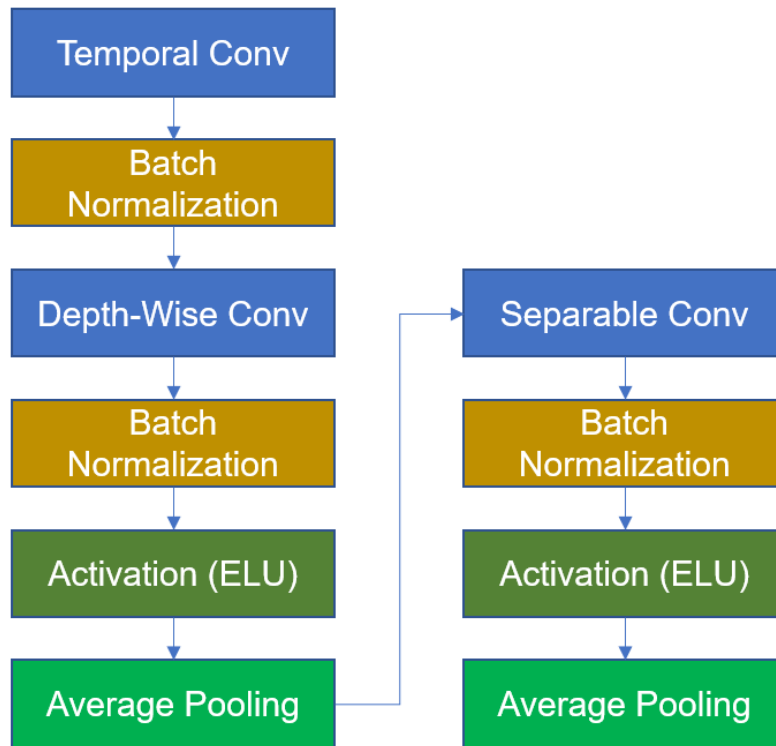
#### 4.6.1.2 CNN Module for Feature Extraction

As the first phase of the feature extraction process, we employed EEGNet, a compact CNN-based model optimized for capturing spatial and temporal features from multi-channel EEGs [166]. This CNN module primarily consists of three convolutional layers, along with several batch normalization, activation, and pooling layers (as shown in Figure 20).

The first convolutional layer extracts temporal features from the multi-channel EEGs by using  $F_1$  filters with a kernel size of  $(1, K_{C1})$ . The second convolutional layer harnesses  $F_1 \times D$  filters to draw out spatial information from the generated multi-channel

feature maps. The third convolutional layer, a separable one, learns a temporal summary for each feature map and merges these feature maps.

Complementing these, the batch normalization layers accelerate and stabilize the training process, the Exponential Linear Unit (ELU) activation layers introduce non-linearity, and the pooling layers serve to abstract temporal features.



**Figure 20. Using three convolutional layers, the CNN module effectively extracts both spatial and temporal features from the raw EEGs.**

#### 4.6.1.3 Transformer Model

Transformer models, with their multi-head self-attention modules, offer a powerful solution to the limitations that Recurrent Neural Network (RNN)-based models encounter when handling long sequences [33]. They enable accurate encoding and alignment of

words within a sequence, enhancing semantic accuracy and improving performance in downstream tasks.

The core component of the Transformer model is the self-attention module, which computes the self-attention score as a weighted sum of a Value matrix ( $V$ ), with coefficients derived from the dot product of Query ( $Q$ ) and Key ( $K$ ) matrices.

#### 4.6.1.4 Proposed Approach Summary

**Table 15. Model architecture and output shape.**

Module	Layer	# Filters	Kernel Size	Output Shape
Input EEG				(batch size, 22, 1000)
CNN	Temporal Conv2D	64	$(1, K_{C1})$	(batch size, 64, 22, 1000)
	Batch Norm			
	DepthWise Conv2D	256	$(C, 1)$	(batch size, 256, 1, 1000)
	Batch Norm			
	ELU			
	Average Pooling		$(1, 5)$	(batch size, 256, 1, 200)
	Separable Conv2D		$(1, 16)$	
	Batch Norm			
	ELU			
	Average Pooling		$(1, 5)$	(batch size, 256, 1, 40)
Transformer	Insert An Learnable Vector			(257, batch size, 40)
	Positional Encoding			
	Transformer Encoder Layers			
Classification	Linear Layer			(batch size, 2)

Inspired by [153], we treat each temporal feature map generated by the separable convolution layer as a word, with the length of the feature map representing the embedding size. We insert a one-dimensional, learnable vector at the beginning of the temporal feature map, similar to the sentence embedding learning approach of the BERT model [34]. After

the addition of positional embedding, the complete feature map is fed into the Transformer encoder for binary classification. The detailed architecture of the proposed model and the output data shape after each layer are presented in Table 15. Note that tensor permutation is not accounted for in this table.

## 4.6.2 *Experiments and Results*

### 4.6.2.1 Dataset Description

The Temple University Hospital EEG Corpus (TUH EEG), globally recognized as the largest public EEG recordings dataset [157], contains a specialized subset known as the TUH EEG Seizure Corpus (TUSZ). Specifically designed for seizure detection tasks, TUSZ is the largest publicly available dataset for this purpose.

The TUSZ's predefined training set comprises 1,185 EEG sessions from 592 patients, while the testing set includes 238 EEG sessions from 50 distinct patients. The EEG data was gathered using scalp electrodes arranged in the standard 10-20 system, providing 22 consistent channels across all EEGs within the dataset.

Given the variation in EEG sampling frequencies (250 Hz, 256 Hz, 400 Hz, or 512 Hz), a resampling to a uniform frequency was performed prior to utilization. The dataset is organized hierarchically: patients are allocated between a predefined training set and a predefined development/testing set. Each patient's data includes recording sessions that are further divided into token files. These token files correspond to time-stamped annotation files that specify the start and end times of each seizure event and its type.

One significant challenge TUSZ presents is the class imbalance. The dataset classifies each seizure event into one of eight distinct classes, with some classes being significantly underrepresented, leading to an uneven distribution across these categories.

We segregated the patients from the predefined training set into training and validation subsets at an 80% to 20% ratio, respectively. Critically, the predefined testing set comprised unique patients, unshared with the predefined training set, thus avoiding any data leakage at the patient level. This partitioning strategy enables the testing of our proposed methods on previously unseen patients, ensuring the model's generalizability.

#### 4.6.2.2 Experimentation and Results

We implemented our proposed deep learning framework using PyTorch and trained it on an Nvidia A100 80GB GPU. The training parameters incorporated a batch size of 1024, an initial learning rate of 0.0001, and the Adam optimizer. The loss function utilized was binary cross entropy loss. The framework was trained for 1000 epochs, with an early stopping mechanism triggered if no improvement in validation loss was observed over 50 consecutive epochs.

**Table 16. Classification report of the proposed approach.**

	<b>Precision</b>	<b>Recall</b>	<b>F1-Score</b>	<b>Support</b>
No Seizure	0.864	0.805	0.833	101,368
Seizure	0.584	0.683	0.630	40,650
Macro Avg	0.724	0.744	0.731	142,018
Weighted Avg	0.783	0.770	0.775	142,018

As Table 16 depicts, our proposed model achieved a macro-average F1 score of 0.731, a precision of 0.724, and recall (sensitivity) of 0.744. This demonstrates the efficacy of our proposed framework for seizure detection tasks, with the macro-average representing the unweighted average of the class-specific metrics.

#### 4.6.3 Discussion

To substantiate the effectiveness of our proposed methodology, we undertook replication of the deep learning pipelines featured in several pertinent EEG analysis studies. The selected studies, published between 2018 and 2022, were compared on equal footing: the same TUSZ dataset was employed, and a uniform training procedure was followed.

As illustrated in Table 17, the performance of our proposed approach surpasses that of other deep learning methodologies replicated from existing literature.

**Table 17. Performance comparison of our proposed approach against the replicated state-of-the-art deep learning approaches from the literature. All approaches are evaluated on the same TUSZ dataset for binary seizure detection task. MSA refers to multi-head self**

Approach/ Paper	Key Model Components	Macro F1	Macro Precision	Macro Recall
EEGNet, 2018	EEGNet	0.700	0.706	0.696
EEG-TCNet, 2020	EEGNet + TCN	0.689	0.695	0.738
ATCNet, 2022	EEGNet + MSA + TCN	0.707	0.706	0.707
Sun et al., 2022	Shallow CNN + Transformer Encoder	0.710	0.702	0.732
Ours	EEGNet + Transformer Encoder	0.731	0.724	0.744

One considerable challenge encountered when utilizing TUSZ is a significant class imbalance issue. Additionally, the dataset's low signal-to-noise ratio presents another substantial obstacle. Being derived from routine hospital recording sessions, the TUSZ

dataset experiences considerably less preprocessing than its counterparts, often resulting in significant noise in the recordings.

Further complications arise due to the variance in the environments where the EEGs are collected. These include numerous hospital departments, each potentially introducing additional sources of noise and inconsistency into the recordings. As identified by previous studies, these conditions often lead to diminished performance metrics when models are trained and evaluated using this dataset, compared to those trained with cleaner, more uniformly collected datasets [158].

One salient limitation of our study is the issue of overfitting, often instigated by patient-to-patient variability. Counteracting this overfitting can be achieved through adversarial learning against the patient subjects, a strategy that could potentially enhance seizure detection accuracy. As a future research direction, incorporating adversarial learning into the training process appears promising.

Another promising future direction lies in the integration of clinical notes as an additional data modality. From a clinical standpoint, focal non-specific seizures (FNSZ) and complex partial seizures (CPSZ) exhibit similar EEG characteristics, with the distinguishing factor being the patient's consciousness state during a seizure. As video recordings are not available due to privacy considerations, clinical notes may offer valuable insights regarding patients' states during the recording sessions. The multimodal integration of EEG features and word embeddings derived by Natural Language Processing (NLP) models has the potential to improve classification performance, particularly in multi-class classification scenarios.

#### 4.6.4 *Conclusion*

In this case study, we put forth a novel deep learning framework purposed for the automated detection of seizures. Our proposed system adeptly extracts spatial and temporal features specific to seizure occurrences from raw EEG signals, all while preserving the capability to capture long-term sequential dependencies.

Extensive experimentation has been carried out to assess the effectiveness of our proposed pipeline, harnessing a publicly available EEG seizure detection dataset. Our findings suggest that our model not only holds its ground against, but also potentially surpasses the performance of leading methodologies cited in current literature.

A notable advantage of our approach is the elimination of the need for laborious handcrafted feature engineering, while still maintaining the capacity to directly handle EEG data suffused with noise. This makes it a highly practical choice for real-world clinical applications. Moreover, our methodology has proven successful in detecting seizures in previously unseen patients, bypassing the need for patient-specific annotations.

With the potential to act as an effective clinical decision support system, our work paves the way for early and precise seizure detection, ultimately elevating the standard of patient care. We remain hopeful regarding the potential impact of our study on the broader realm of neurological disorder diagnosis and treatment and are excited for the positive changes it could bring.

## **CHAPTER 5. DISCUSSION AND CONCLUSION**

This dissertation has presented deep learning frameworks that enable precision medicine by adopting various NLP algorithms. The proposed deep learning frameworks have been applied and validated on multiple biomedical applications and data modalities, including electronic health records data for intelligent death reporting, free-text clinical notes and patient messages for disease identification, and EEGs for seizure detection.

In CHAPTER 2, I presented a novel deep learning framework with encoder-decoder models and optional preprocessing/ decoding modules to generate causal sequences of death. To the best of our knowledge, this is the first work of adopting such framework to promote intelligent and accurate death reporting. In addition, it was the first work to interpret the correlation between outcome sequence and the death-related conditions in the discharge records using attention visualization technique. The attention visualization confirmed the validity of the data driven approach in generating accurate causal sequences of death. Besides, this work includes the modified BLEU score, a common metric used in neural machine translation tasks, and three different task-specific accuracy metrics as evaluation metrics when reporting the experiment results. As such, people of technical background or clinical background can more accurately and comprehensively understand the accuracy and interpretability of the generated sequences. Lastly, I also delivered a proof of concept SMART-on-FHIR mobile application that enables data query, AI-powered sequence generation, human interaction and data upload.

In CHAPTER 3, I presented two cases studies for long COVID and COVID patient identification using unstructured free-text notes. Specifically, I designed and implemented

novel NLP frameworks with the backbone of pretrained BERT models (and its variants) for text classification. I showed that the pretrained BERT models are effective in identifying patients with long COVID on clinical notes. Among three BERT model variants, Bio\_ClinicalBERT achieves the highest recall score, due to the fact that it was pretrained on MIMIC III clinical notes so it has superior performance when handling clinical notes. Meanwhile, I demonstrated that distilBERT, a lightweighted BERT model variant, is the most desirable option in identifying patients with positive COVID symptoms given messages initiated by the patients. Comparing to other BERT models, distilBERT achieves similar performance when classifying patient messages, which are more similar to our daily languages, while is 40% more computational efficient. Thus, BERT models are effective in different classification tasks, a data-driven approach that provides the potential to overcome challenges in clinical practice and reduces workload of clinical team.

In CHAPTER 4, I explained the clinical challenges in manual seizure detection, introduced the technical task of automated seizure detection using EEGs, summarized & criticized the literatures, and presented three novel case studies. In the first case study, I designed a novel CNN + LSTM framework with adversarial learning and regularization techniques to mitigate the overfitting issues. In the second case study, I explore meta transfer learning with EEGNet model to expand the binary seizure detection task into a multi-class classification task. The goal is to mitigate patient-to-patient variation in EEG patterns while learning from limited seizure data for each individual patient. In the third case study, I designed and implemented a novel CNN + Transformer framework for more accurate seizure detection performance. To promote my novel contribution, I replicated several deep learning frameworks from most recent journal publications in the same field

as strong baselines, and demonstrated that my proposed framework outperformed their models. To summarize, I presented multiple novel NLP-based deep learning frameworks for automated seizure detection, along with different training strategies. As for future work, concept bottleneck models, outcome interpretation and transfer learning with private dataset can further improve the accuracy, interpretability and usability of deep learning-based automated EEG seizure detection frameworks, leading to their application into clinical practice.

To conclude, it is evident that NLP algorithms play an indispensable role within deep learning frameworks. They support the generation of causal sequences of death employing EHRs, enable the automatic identification of patients at risk from unstructured free-text clinical notes and promote seizure detection using time-series EEGs. These advanced tools are paving the way for precision medicine and refining the quality of patient care, underscoring their invaluable contribution to biomedical research and their transformational impact on clinical applications.

While our study has made significant strides, it is crucial to address several inherent limitations that, if tackled effectively, could potentially enhance the robustness of our performance. The first limitation originates from the complex nature of clinical notes. Standardization of clinical notes, even within a single healthcare facility, is often lacking, rendering it challenging to parse these notes into specific sections or modules through simple preprocessing steps. This necessitates meticulous manual inspections and carefully constructed steps to filter out noise and extraneous modules. However, some of these modules inevitably persist in downstream analysis, becoming a major concern when utilizing bag-of-word embeddings. Further, we grapple with the BERT model's constraint

of text inputs not exceeding 512 tokens, requiring us to truncate our clinical notes and possibly omitting pivotal information. Moreover, the interpretability of token-level and word-level attention scores remains far from optimal. It proves difficult to interpret classification results and convert them into practical clinical insights, given that top-ranking words often do not directly correspond to symptoms, thereby impeding clinicians' ability to accurately diagnose patients who may develop persistent symptoms.

The second limitation relates to the overfitting issues encountered in EEG seizure detection, partially attributable to the complexity of EEG data. The variability in EEGs from patient to patient often results in subpar seizure detection performance. Our experiments also highlighted significant class imbalance in the TUSZ dataset, compounded by challenges such as a low signal-to-noise ratio and signals collected from disparate hospital settings.

The third limitation arises from the interpretation of EEG seizure detection results. Identifying affected brain regions or corresponding EEG channels during a seizure is vital, yet poses a formidable task when utilizing deep neural networks. Additionally, it's important to uncover EEG biomarkers that neurologists can interpret in the context of automatic seizure detection, as this will foster their trust in the system.

Lastly, to summarize all my publications during my PhD, I decided to categorize them into four groups.

- Group 1: Electronical Health Records and NLP
  - Journal papers: [16], [17], [23] [19]
  - Conference papers: [18], [20], [22], [24]

- Group 2: EEG Seizure Detection
  - Conference papers: [25]–[27]
- Group 3: Medical Image Processing
  - Journal papers: [6], [7], [11], [12]
  - Conference papers: [8], [9], [13]–[15]
- Group 4: Other Health Informatics
  - Conference papers: [21], [56], [168]–[170]

In this dissertation, I only focused on NLP-related projects that address different challenges and clinical data modalities that promote precision medicine.

## REFERENCES

- [1] National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease, *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. Washington (DC): National Academies Press (US), 2011.
- [2] F. S. Collins and H. Varmus, “A new initiative on precision medicine,” *N. Engl. J. Med.*, vol. 372, no. 9, pp. 793–795, Feb. 2015.
- [3] F. S. Collins and V. A. McKusick, “Implications of the Human Genome Project for medical science,” *JAMA*, vol. 285, no. 5, pp. 540–544, Feb. 2001.
- [4] E. A. Ashley, “Towards precision medicine,” *Nat. Rev. Genet.*, vol. 17, no. 9, pp. 507–522, Aug. 2016.
- [5] M. R. Kosorok and E. B. Laber, “Precision Medicine,” *Annu Rev Stat Appl*, vol. 6, pp. 263–286, Mar. 2019.
- [6] W. Shi, L. Tong, Y. Zhu, and M. D. Wang, “COVID-19 Automatic Diagnosis With Radiographic Imaging: Explainable Attention Transfer Deep Neural Networks,” *IEEE J Biomed Health Inform*, vol. 25, no. 7, pp. 2376–2387, Jul. 2021.
- [7] F. Giuste *et al.*, “Explainable Artificial Intelligence Methods in Combating Pandemics: A Systematic Review,” *IEEE Rev. Biomed. Eng.*, vol. PP, Jun. 2022.
- [8] T. Naren, Y. Zhu, and M. D. Wang, “COVID-19 diagnosis using model agnostic meta-learning on limited chest X-ray images,” in *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, Gainesville Florida, 2021.
- [9] W. Shi, L. Tong, Y. Zhuang, Y. Zhu, and M. D. Wang, “EXAM,” in *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, Virtual Event USA, 2020.
- [10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.
- [11] A. Choudhary, L. Tong, Y. Zhu, and M. D. Wang, “Advancing medical imaging informatics by deep learning-based domain adaptation,” *Yearb. Med. Inform.*, vol. 29, no. 1, pp. 129–138, Aug. 2020.
- [12] F. O. Giuste *et al.*, “Explainable synthetic image generation to improve risk assessment of rare pediatric heart transplant rejection,” *J. Biomed. Inform.*, vol. 139, no. 104303, p. 104303, Mar. 2023.

- [13] Y. Zhu, L. Tong, S. R. Deshpande, and M. D. Wang, “Improved Prediction on Heart Transplant Rejection Using Convolutional Autoencoder and Multiple Instance Learning on Whole-Slide Imaging,” *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. 2019.
- [14] F. Giuste *et al.*, “Automated Classification of Acute Rejection from Endomyocardial Biopsies,” in *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, Virtual Event, USA, 2020, pp. 1–9.
- [15] A. Mirzazadeh *et al.*, “Improving heart transplant rejection classification training using progressive generative adversarial networks,” in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, Athens, Greece, 2021.
- [16] Y. Zhu, Y. Sha, H. Wu, M. Li, R. A. Hoffman, and M. D. Wang, “Proposing causal sequence of death by neural machine translation in public health informatics,” *IEEE J. Biomed. Health Inform.*, vol. 26, no. 4, pp. 1422–1431, Apr. 2022.
- [17] Y. Zhu, J. Venugopalan, Z. Zhang, N. K. Chanani, K. O. Maher, and M. D. Wang, “Domain adaptation using convolutional autoencoder and gradient boosting for adverse events prediction in the Intensive Care unit,” *Front. Artif. Intell.*, vol. 5, p. 640926, Apr. 2022.
- [18] Y. Zhu, H. Wu, and M. D. Wang, “Feature exploration and causal inference on mortality of epilepsy patients using insurance claims data,” *IEEE EMBS Int. Conf. Biomed. Health Inform.*, vol. 2019, May 2019.
- [19] F. O. Giuste *et al.*, “Early and fair COVID-19 outcome risk assessment using robust feature selection,” *Sci. Rep.*, vol. 13, no. 1, p. 18981, Nov. 2023.
- [20] J. Ben Tamo *et al.*, “Adolescent Idiopathic Scoliosis Patient Subphenotyping for Surgical Planning and Improved Patient Outcomes,” in *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, Houston, TX, USA, 2023, pp. 1–10.
- [21] A. Mahale *et al.*, “Automating Treatment Recommendations for Children with Cerebral Palsy Based on Multi-Modal Clinical Data,” in *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2022, pp. 1–4.
- [22] Y. Zhu *et al.*, “Using natural language processing on free-text clinical notes to identify patients with long-term COVID effects,” in *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, Northbrook, Illinois, 2022, pp. 1–9.

- [23] K. Mermin-Bunnell *et al.*, “Use of natural language processing of patient-initiated electronic health record messages to identify patients with COVID-19 infection,” *JAMA Netw. Open*, vol. 6, no. 7, p. e2322299, Jul. 2023.
- [24] W. Shi, Y. Zhuang, Y. Zhu, and H. Iwinski, “Retrieval-Augmented Large Language Models for Adolescent Idiopathic Scoliosis Patients in Shared Decision-Making,” *Proceedings of the 14th*, 2023.
- [25] M. Saqib, Y. Zhu, M. D. Wang, and B. Beaulieu-Jones, “Regularization of deep neural networks for EEG seizure detection to mitigate overfitting,” *Proc. COMPSAC*, vol. 2020, pp. 664–673, Jul. 2020.
- [26] Y. Zhu, M. Saqib, E. Ham, S. Belhareth, R. Hoffman, and M. D. Wang, “Mitigating patient-to-patient variation in EEG seizure detection using meta transfer learning,” in *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, Cincinnati, OH, 2020.
- [27] Y. Zhu and M. D. Wang, “Automated seizure detection using transformer models on multi-channel EEGs,” in *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, Pittsburgh, PA, USA, 2023.
- [28] J. Hirschberg and C. D. Manning, “Advances in natural language processing,” *Science*, vol. 349, no. 6245, pp. 261–266, Jul. 2015.
- [29] L. Medsker and L. C. Jain, *Recurrent Neural Networks: Design and Applications*. CRC Press, 1999.
- [30] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
- [31] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling,” *arXiv [cs.NE]*, 11-Dec-2014.
- [32] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv [cs.CL]*, 01-Sep-2014.
- [33] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *arXiv [cs.CL]*, 11-Oct-2018.
- [35] D. M. Korngiebel and S. D. Mooney, “Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery,” *NPJ Digit Med*, vol. 4, no. 1, p. 93, Jun. 2021.

- [36] R. Nabbout and M. Kuchenbuch, “Impact of predictive, preventive and precision medicine strategies in epilepsy,” *Nat. Rev. Neurol.*, vol. 16, no. 12, pp. 674–688, Dec. 2020.
- [37] H. A. Loomans-Kropp and A. Umar, “Cancer prevention and screening: the next step in the era of precision medicine,” *NPJ Precis. Oncol.*, vol. 3, no. 1, Jan. 2019.
- [38] N. Fujiwara, S. L. Friedman, N. Goossens, and Y. Hoshida, “Risk factors and prevention of hepatocellular carcinoma in the era of precision medicine,” *J. Hepatol.*, vol. 68, no. 3, pp. 526–549, Mar. 2018.
- [39] F. Xie, J. C. Chan, and R. C. Ma, “Precision medicine in diabetes prevention, classification and management,” *J. Diabetes Investig.*, vol. 9, no. 5, pp. 998–1015, Sep. 2018.
- [40] B. W. Stewart *et al.*, “Cancer prevention as part of precision medicine: ‘plenty to be done,’” *Carcinogenesis*, vol. 37, no. 1, pp. 2–9, Jan. 2016.
- [41] J. Xu, S. L. Murphy, K. D. Kochanek, B. Bastian, and E. Arias, “Deaths: Final data for 2016,” *Natl. Vital Stat. Rep.*, vol. 67, no. 5, pp. 1–76, Jul. 2018.
- [42] S. H. Woolf, D. A. Chapman, and J. H. Lee, “COVID-19 as the leading cause of death in the United States,” *JAMA*, vol. 325, no. 2, pp. 123–124, Jan. 2021.
- [43] H. K. Koh, A. C. Geller, and T. J. VanderWeele, “Deaths From COVID-19,” *JAMA: the journal of the American Medical Association*, vol. 325, no. 2. jamanetwork.com, pp. 133–134, 12-Jan-2021.
- [44] M. C. Castro, S. Gurzenda, C. M. Turra, S. Kim, T. Andrasfay, and N. Goldman, “Reduction in life expectancy in Brazil after COVID-19,” *Nat. Med.*, vol. 27, no. 9, pp. 1629–1635, Sep. 2021.
- [45] S. Elezkurtaj *et al.*, “Causes of death and comorbidities in hospitalized patients with COVID-19,” *Sci. Rep.*, vol. 11, no. 1, p. 4263, Feb. 2021.
- [46] P. Qin and L. Mehlum, “National observation of death by suicide in the first 3 months under COVID-19 pandemic,” *Acta Psychiatr. Scand.*, vol. 143, no. 1, pp. 92–93, Jan. 2021.
- [47] D. Contou, R. Cally, F. Sarfati, P. Desaint, M. Fraissé, and G. Plantefève, “Causes and timing of death in critically ill COVID-19 patients,” *Crit. Care*, vol. 25, no. 1, p. 79, Feb. 2021.
- [48] *International statistical classification of diseases and related health problems: Alphabetical index.* .
- [49] R. Khera, K. B. Dorsey, and H. M. Krumholz, “Transition to the ICD-10 in the United States: An emerging data chasm,” *JAMA*, vol. 320, no. 2, pp. 133–134, Jul. 2018.

- [50] A. M. Dai and Q. V. Le, “Semi-supervised sequence learning,” *arXiv [cs.LG]*, 04-Nov-2015.
- [51] T. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2015.
- [52] M.-T. Luong and C. D. Manning, “Achieving open vocabulary neural machine translation with hybrid word-character models,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, 2016.
- [53] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, “XLNet: Generalized Autoregressive Pretraining for Language Understanding,” *arXiv [cs.CL]*, 19-Jun-2019.
- [54] G. Lample and A. Conneau, “Cross-lingual Language Model Pretraining,” *arXiv [cs.CL]*, 22-Jan-2019.
- [55] D. C. Cowper, J. D. Kubal, C. Maynard, and D. M. Hynes, “A primer and comparative review of major US mortality databases,” *Ann. Epidemiol.*, vol. 12, no. 7, pp. 462–468, Oct. 2002.
- [56] W. Shi *et al.*, “A FHIR-compliant application for multi-site and multi-modality pediatric scoliosis patient rehabilitation,” in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Houston, TX, USA, 2021.
- [57] R. A. Hoffman, H. Wu, J. Venugopalan, P. Braun, and M. D. Wang, “Intelligent Mortality Reporting With FHIR,” *IEEE J Biomed Health Inform.*, vol. 22, no. 5, pp. 1583–1588, Sep. 2018.
- [58] H. Jiang, H. Wu, and M. D. Wang, “Causes of death in the United States, 1999 to 2014,” *IEEE EMBS Int. Conf. Biomed. Health Inform.*, vol. 2017, Feb. 2017.
- [59] H. Wu and M. D. Wang, “Infer cause of death for population health using convolutional neural network,” *ACM BCB*, vol. 2017, pp. 526–535, Aug. 2017.
- [60] R. A. Hoffman, J. Venugopalan, L. Qu, H. Wu, and M. D. Wang, “Improving validity of cause of death on death certificates,” *ACM BCB*, vol. 2018, pp. 178–183, Aug. 2018.
- [61] A. Blanco, A. Perez, A. Casillas, and D. Cobos, “Extracting cause of death from Verbal Autopsy with Deep Learning interpretable methods,” *IEEE J. Biomed. Health Inform.*, vol. 25, no. 4, pp. 1315–1325, Apr. 2021.
- [62] R. Yu, Y. Zheng, R. Zhang, Y. Jiang, and C. C. Y. Poon, “Using a multi-task recurrent neural network with attention mechanisms to predict hospital mortality of patients,” *IEEE J. Biomed. Health Inform.*, vol. 24, no. 2, pp. 486–492, Feb. 2020.

- [63] T. H. Lu, “Using ACME (Automatic Classification of Medical Entry) software to monitor and improve the quality of cause of death statistics,” *J. Epidemiol. Community Health*, vol. 57, no. 6, pp. 470–471, Jun. 2003.
- [64] K. Cho *et al.*, “Learning phrase representations using RNN encoder–decoder for statistical machine translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014.
- [65] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” *Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.
- [66] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: Encoder–decoder approaches,” in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, Doha, Qatar, 2014.
- [67] P. Koehn, “Pharaoh: A beam search decoder for phrase-based statistical machine translation models,” in *Machine Translation: From Real Users to Research*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 115–124.
- [68] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
- [69] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. Rush, “OpenNMT: Open-Source Toolkit for Neural Machine Translation,” in *Proceedings of ACL 2017, System Demonstrations*, Vancouver, Canada, 2017.
- [70] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, 2016.
- [71] H. Wu and M. D. Wang, “An Information Theoretic Learning for Causal Direction Identification,” in *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, Madrid, Spain, 2020.
- [72] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, “Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction,” *NPJ Digit Med*, vol. 4, no. 1, p. 86, May 2021.
- [73] S. Lopez-Leon *et al.*, “More than 50 long-term effects of COVID-19: a systematic review and meta-analysis,” *Sci. Rep.*, vol. 11, no. 1, p. 16144, Aug. 2021.
- [74] M. Nehme, O. Braillard, F. Chappuis, D. S. Courvoisier, I. Guessous, and CoviCare Study Team, “Prevalence of Symptoms More Than Seven Months After Diagnosis of Symptomatic COVID-19 in an Outpatient Setting,” *Ann. Intern. Med.*, vol. 174, no. 9, pp. 1252–1260, Sep. 2021.

- [75] L. Sigfrid *et al.*, “What is the recovery rate and risk of long-term consequences following a diagnosis of COVID-19? A harmonised, global longitudinal observational study protocol,” *BMJ Open*, vol. 11, no. 3, p. e043887, Mar. 2021.
- [76] R. C. Becker, “Anticipating the long-term cardiovascular effects of COVID-19,” *J. Thromb. Thrombolysis*, vol. 50, no. 3, pp. 512–524, Oct. 2020.
- [77] L. Huang *et al.*, “1-year outcomes in hospital survivors with COVID-19: a longitudinal cohort study,” *The Lancet*, vol. 398, no. 10302, pp. 747–758, 2021.
- [78] A. D. Proal and M. B. VanElzakker, “Long COVID or Post-acute Sequelae of COVID-19 (PASC): An Overview of Biological Factors That May Contribute to Persistent Symptoms,” *Frontiers in Microbiology*, vol. 12, 2021.
- [79] W. Shi, L. Tong, Y. Zhuang, and Y. Zhu, “EXAM: an explainable attention-based model for COVID-19 automatic diagnosis,” *Proceedings of the 11th*, 2020.
- [80] T. Naren, Y. Zhu, and M. D. Wang, “COVID-19 diagnosis using model agnostic meta-learning on limited chest X-ray images,” *Proceedings of the 12th ACM Conference*, 2021.
- [81] D. Jamiolkowski, B. Mühleisen, S. Müller, A. A. Navarini, A. Tzankov, and E. Roider, “SARS-CoV-2 PCR testing of skin for COVID-19 diagnostics: a case report,” *Lancet*, vol. 396, no. 10251, Elsevier BV, pp. 598–599, 29-Aug-2020.
- [82] P. Garg, U. Arora, A. Kumar, and N. Wig, “The" post-COVID" syndrome: How deep is the damage?,” *J. Med. Virol.*, 2020.
- [83] T. Greenhalgh, M. Knight, C. A’Court, M. Buxton, and L. Husain, “Management of post-acute covid-19 in primary care,” *BMJ*, vol. 370, p. m3026, Aug. 2020.
- [84] A. Sisó-Almirall *et al.*, “Long covid-19: Proposed primary care clinical guidelines for diagnosis and disease management,” *Int. J. Environ. Res. Public Health*, vol. 18, no. 8, p. 4350, Apr. 2021.
- [85] A. Sisó-Almirall *et al.*, “Prognostic factors in Spanish COVID-19 patients: A case series from Barcelona,” *PLoS One*, vol. 15, no. 8, p. e0237960, Aug. 2020.
- [86] M. Sivan and S. Taylor, “NICE guideline on long covid,” *BMJ*, vol. 371, p. m4938, Dec. 2020.
- [87] P. Venkatesan, “NICE guideline on long COVID,” *Lancet Respir. Med.*, vol. 9, no. 2, p. 129, Feb. 2021.
- [88] F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, and F. Rudzicz, “A survey of word embeddings for clinical text,” *J. Biomed. Inform.*, vol. 100S, no. 100057, p. 100057, Oct. 2019.

- [89] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv [cs.CL]*, 16-Jan-2013.
- [90] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Trans. Assoc. Comput. Linguist.*, vol. 5, pp. 135–146, Dec. 2017.
- [91] B. Athiwaratkun, A. Wilson, and A. Anandkumar, “Probabilistic FastText for Multi-Sense Word Embeddings,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, 2018.
- [92] J. Sarzynska-Wawer *et al.*, “Detecting formal thought disorder by deep contextualized word representations,” *Psychiatry Res.*, vol. 304, no. 114135, p. 114135, Oct. 2021.
- [93] J. Lee *et al.*, “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.
- [94] E. Alsentzer *et al.*, “Publicly Available Clinical BERT Embeddings,” *arXiv [cs.CL]*, 06-Apr-2019.
- [95] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, N. Rouf, and M. Mohi Ud Din, “Machine learning based approaches for detecting COVID-19 using clinical text data,” *Int. J. Inf. Technol.*, vol. 12, no. 3, pp. 731–739, Jun. 2020.
- [96] T. Wagner *et al.*, “Augmented curation of clinical notes from a massive EHR system reveals symptoms of impending COVID-19 diagnosis,” *Elife*, vol. 9, Jul. 2020.
- [97] J. Wang *et al.*, “COVID-19 SignSym: a fast adaptation of a general clinical NLP tool to identify and normalize COVID-19 signs and symptoms to OMOP common data model,” *J. Am. Med. Inform. Assoc.*, vol. 28, no. 6, pp. 1275–1283, Jun. 2021.
- [98] W. Ge, J.-W. Huh, Y. R. Park, J.-H. Lee, Y.-H. Kim, and A. Turchin, “An interpretable ICU mortality prediction model based on logistic regression and Recurrent Neural Networks with LSTM units,” *AMIA Annu. Symp. Proc.*, vol. 2018, pp. 460–469, Dec. 2018.
- [99] M. Jin *et al.*, “Improving hospital mortality prediction with medical named entities and multimodal learning,” *arXiv [cs.CL]*, 29-Nov-2018.
- [100] F. O. Giuste, L. L. He, M. Isgut, W. Shi, B. J. Anderson, and M. D. Wang, “Automated risk assessment of COVID-19 patients at diagnosis using electronic healthcare records,” in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, Athens, Greece, 2021.
- [101] A. V. Raveendran, R. Jayadevan, and S. Sashidharan, “Long COVID: An overview,” *Diabetes Metab. Syndr.*, vol. 15, no. 3, pp. 869–875, May 2021.

- [102] C. H. Sudre *et al.*, “Attributes and predictors of long COVID,” *Nat. Med.*, vol. 27, no. 4, pp. 626–631, Apr. 2021.
- [103] *Evidence of Persisting Autoreactivity in Post-Acute Sequelae of SARS-CoV-2 Infection.* .
- [104] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [105] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA, 2016.
- [106] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016.
- [107] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,” *PLoS One*, vol. 10, no. 7, p. e0130140, Jul. 2015.
- [108] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, 2017.
- [109] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks,” in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Tahoe, NV, 2018.
- [110] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” *arXiv [cs.LG]*, 03-Mar-2017.
- [111] Y. Gu *et al.*, “Domain-specific language model pretraining for biomedical natural language processing,” *ACM Trans. Comput. Healthc.*, vol. 3, no. 1, pp. 1–23, Jan. 2022.
- [112] H. J. Keselman and J. C. Rogan, “A comparison of the modified-Tukey and scheffe methods of multiple comparisons for pairwise contrasts,” *J. Am. Stat. Assoc.*, vol. 73, no. 361, p. 47, Mar. 1978.
- [113] H. Abdi and L. J. Williams, “Tukey’s honestly significant difference (HSD) test,” *Encyclopedia of research design*, vol. 3, no. 1, pp. 1–5, 2010.
- [114] B. Nath *et al.*, “Trends in electronic health record inbox messaging during the COVID-19 pandemic in an ambulatory practice network in New England,” *JAMA Netw. Open*, vol. 4, no. 10, p. e2131490, Oct. 2021.

- [115] A. Chavez, J. Bracamonte, M. Kresin, M. Yardley, and M. Grover, "High volume portal usage impacts practice resources," *J. Am. Board Fam. Med.*, vol. 33, no. 3, pp. 452–455, May 2020.
- [116] B. Rader *et al.*, "Use of at-home COVID-19 tests - United States, August 23, 2021-March 12, 2022," *MMWR Morb. Mortal. Wkly. Rep.*, vol. 71, no. 13, pp. 489–494, Apr. 2022.
- [117] M. Huang *et al.*, "Patient portal messaging for asynchronous virtual care during the COVID-19 pandemic: Retrospective analysis," *JMIR Hum. Factors*, vol. 9, no. 2, p. e35187, May 2022.
- [118] L. S. Rotenstein *et al.*, "Association between electronic health record time and quality of care metrics in primary care," *JAMA Netw. Open*, vol. 5, no. 10, p. e2237086, Oct. 2022.
- [119] P. Schwab *et al.*, "Real-time prediction of COVID-19 related mortality using electronic health records," *Nat. Commun.*, vol. 12, no. 1, p. 1058, Feb. 2021.
- [120] J. Hammond *et al.*, "Oral nirmatrelvir for high-risk, nonhospitalized adults with Covid-19," *N. Engl. J. Med.*, vol. 386, no. 15, pp. 1397–1408, Apr. 2022.
- [121] R. M. Cronin, D. Fabbri, J. C. Denny, S. T. Rosenbloom, and G. P. Jackson, "A comparison of rule-based and machine learning approaches for classifying patient portal messages," *Int. J. Med. Inform.*, vol. 105, pp. 110–120, Sep. 2017.
- [122] J. Chen *et al.*, "Detecting hypoglycemia incidents reported in patients' secure messages: Using cost-sensitive learning and oversampling to reduce data imbalance," *J. Med. Internet Res.*, vol. 21, no. 3, p. e11990, Mar. 2019.
- [123] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv [cs.CL]*, 02-Oct-2019.
- [124] B. Anderson, Z. Smith, S. Edupuganti, X. Yan, C. M. Masi, and H. M. Wu, "Effect of monoclonal antibody treatment on clinical outcomes in ambulatory patients with Coronavirus disease 2019," *Open Forum Infect. Dis.*, vol. 8, no. 7, p. ofab315, Jul. 2021.
- [125] A. Turchin and L. F. Florez Builes, "Using natural language processing to measure and improve quality of diabetes care: A systematic review," *J. Diabetes Sci. Technol.*, vol. 15, no. 3, pp. 553–560, May 2021.
- [126] C. Sinsky *et al.*, "Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties," *Ann. Intern. Med.*, vol. 165, no. 11, pp. 753–760, Dec. 2016.
- [127] D. Lazer *et al.*, "The COVID States Project #79: At-home COVID tests," 21-Jan-2022.

- [128] S. Noachtar and J. Rémi, “The role of EEG in epilepsy: a critical review,” *Epilepsy Behav.*, vol. 15, no. 1, pp. 22–33, May 2009.
- [129] R. S. Fisher *et al.*, “ILAE official report: a practical clinical definition of epilepsy,” *Epilepsia*, vol. 55, no. 4, pp. 475–482, Apr. 2014.
- [130] A. Van de Vel *et al.*, “Non-EEG seizure detection systems and potential SUDEP prevention: State of the art: Review and update,” *Seizure*, vol. 41, pp. 141–153, Oct. 2016.
- [131] A. Van de Vel *et al.*, “Non-EEG seizure-detection systems and potential SUDEP prevention: state of the art,” *Seizure*, vol. 22, no. 5, pp. 345–355, Jun. 2013.
- [132] V. Srinivasan, C. Eswaran, and N. Sriraam, “Approximate entropy-based epileptic EEG detection using artificial neural networks,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 11, no. 3, pp. 288–295, May 2007.
- [133] F. Mormann, K. Lehnertz, P. David, and C. E. Elger, “Mean phase coherence as a measure for phase synchronization and its application to the EEG of epilepsy patients,” *Physica D*, vol. 144, no. 3–4, pp. 358–369, Oct. 2000.
- [134] A. Shoeb, B. Bourgeois, S. Treves, S. C. Schachter, and J. Guttag, “Impact of patient-specificity on seizure onset detection performance,” *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 2007, pp. 4110–4114, 2007.
- [135] R. S. Fisher, H. E. Scharfman, and M. deCurtis, “How can we identify ictal and interictal abnormal activity?,” *Adv. Exp. Med. Biol.*, vol. 813, pp. 3–23, 2014.
- [136] E. Amorim *et al.*, “Performance of spectrogram-based seizure identification of adult EEGs by critical care nurses and neurophysiologists,” *J. Clin. Neurophysiol.*, vol. 34, no. 4, pp. 359–364, Jul. 2017.
- [137] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, “Deep learning-based electroencephalography analysis: a systematic review,” *J. Neural Eng.*, vol. 16, no. 5, p. 051001, Aug. 2019.
- [138] V. Shah, M. Golmohammadi, S. Ziyabari, E. Von Weltin, I. Obeid, and J. Picone, “Optimizing channel selection for seizure detection,” in *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, Philadelphia, PA, 2017.
- [139] F. Pauri, F. Pierelli, G. E. Chatrian, and W. W. Erdly, “Long-term EEG-video-audio monitoring: computer detection of focal EEG seizure patterns,” *Electroencephalogr. Clin. Neurophysiol.*, vol. 82, no. 1, pp. 1–9, Jan. 1992.
- [140] H. E. Ronner, S. C. Ponten, C. J. Stam, and B. M. J. Uitdehaag, “Inter-observer variability of the EEG diagnosis of seizures in comatose patients,” *Seizure*, vol. 18, no. 4, pp. 257–263, May 2009.

- [141] J. J. Halford *et al.*, “Inter-rater agreement on identification of electrographic seizures and periodic discharges in ICU EEG recordings,” *Clin. Neurophysiol.*, vol. 126, no. 9, pp. 1661–1669, Sep. 2015.
- [142] Y. Zhang, S. Yang, Y. Liu, Y. Zhang, B. Han, and F. Zhou, “Integration of 24 feature types to accurately detect and predict seizures using scalp EEG signals,” *Sensors (Basel)*, vol. 18, no. 5, p. 1372, Apr. 2018.
- [143] T. N. Alotaiby, S. A. Alshebeili, T. Alshawi, I. Ahmad, and F. E. Abd El-Samie, “EEG seizure detection and prediction algorithms: a survey,” *EURASIP J. Adv. Signal Process.*, vol. 2014, no. 1, Dec. 2014.
- [144] A. H. Ansari, P. J. Cherian, A. Caicedo, G. Naulaers, M. De Vos, and S. Van Huffel, “Neonatal seizure detection using deep convolutional neural networks,” *Int. J. Neural Syst.*, vol. 29, no. 4, p. 1850011, May 2019.
- [145] F. Achilles, F. Tombari, V. Belagiannis, A. M. Loesch, S. Noachtar, and N. Navab, “Convolutional neural networks for real-time epileptic seizure detection,” *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.*, vol. 6, no. 3, pp. 264–269, May 2018.
- [146] A. O’Shea, G. Lightbody, G. Boylan, and A. Temko, “Neonatal seizure detection using convolutional neural networks,” in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, Tokyo, 2017.
- [147] X. Zhang, L. Yao, M. Dong, Z. Liu, Y. Zhang, and Y. Li, “Adversarial representation learning for robust patient-independent epileptic seizure detection,” *IEEE J. Biomed. Health Inform.*, vol. 24, no. 10, pp. 2852–2859, Oct. 2020.
- [148] L. Vidyaratne, A. Glandon, M. Alam, and K. M. Iftekharuddin, “Deep recurrent neural network for seizure detection,” in *2016 International Joint Conference on Neural Networks (IJCNN)*, Vancouver, BC, Canada, 2016.
- [149] A. M. Abdelhameed, H. G. Daoud, and M. Bayoumi, “Deep convolutional bidirectional LSTM recurrent neural network for epileptic seizure detection,” in *2018 16th IEEE International New Circuits and Systems Conference (NEWCAS)*, Montreal, QC, 2018.
- [150] Y. Li *et al.*, “Automatic seizure detection using fully convolutional nested LSTM,” *Int. J. Neural Syst.*, vol. 30, no. 4, p. 2050019, Apr. 2020.
- [151] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, “Temporal convolutional networks: A unified approach to action segmentation,” *arXiv [cs.CV]*, 29-Aug-2016.
- [152] E. Aksan and O. Hilliges, “STCN: Stochastic Temporal Convolutional Networks,” *arXiv [cs.LG]*, 18-Feb-2019.
- [153] Y. Sun *et al.*, “Continuous seizure detection based on transformer and long-term iEEG,” *IEEE J. Biomed. Health Inform.*, vol. 26, no. 11, pp. 5418–5427, Nov. 2022.

- [154] J. Pedoeem, G. Bar Yosef, S. Abittan, and S. Keene, “TABS: Transformer based seizure detection,” in *Biomedical Sensing and Analysis*, Cham: Springer International Publishing, 2022, pp. 133–160.
- [155] H. Albaqami, G. M. Hassan, and A. Datta, “MP-SeizNet: A multi-path CNN Bi-LSTM Network for seizure-type classification using EEG,” *Biomed. Signal Process. Control*, vol. 84, no. 104780, p. 104780, Jul. 2023.
- [156] C. Li, X. Huang, R. Song, R. Qian, X. Liu, and X. Chen, “EEG-based seizure prediction via Transformer guided CNN,” *Measurement (Lond.)*, vol. 203, no. 111948, p. 111948, Nov. 2022.
- [157] I. Obeid and J. Picone, “The temple university hospital EEG data corpus,” *Front. Neurosci.*, vol. 10, p. 196, May 2016.
- [158] M. Golmohammadi, S. Ziyabari, V. Shah, I. Obeid, and J. Picone, “Deep architectures for spatio-temporal modeling: Automated seizure detection in scalp EEGs,” in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Orlando, FL, USA, 2018.
- [159] A. Melnik *et al.*, “Systems, subjects, sessions: To what extent do these factors influence EEG data?,” *Front. Hum. Neurosci.*, vol. 11, p. 150, Mar. 2017.
- [160] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, “Meta-Learning with Memory-Augmented Neural Networks,” in *Proceedings of The 33rd International Conference on Machine Learning*, 20--22 Jun 2016, vol. 48, pp. 1842–1850.
- [161] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, “Matching networks for one shot learning,” *arXiv [cs.LG]*, 13-Jun-2016.
- [162] J. Snell, K. Swersky, and R. Zemel, “Prototypical networks for few-shot learning,” *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [163] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, “Learning to compare: Relation network for few-shot learning,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 1199–1208.
- [164] C. Finn, P. Abbeel, and S. Levine, “Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks,” in *Proceedings of the 34th International Conference on Machine Learning*, 06--11 Aug 2017, vol. 70, pp. 1126–1135.
- [165] T. Duan, M. Chauhan, M. A. Shaikh, J. Chu, and S. Srihari, “Ultra efficient transfer learning with meta update for cross subject EEG classification,” *arXiv [cs.LG]*, 13-Mar-2020.

- [166] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, “EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces,” *J. Neural Eng.*, vol. 15, no. 5, p. 056013, Oct. 2018.
- [167] A. G. Howard *et al.*, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” *arXiv [cs.CV]*, 16-Apr-2017.
- [168] B. L. Marteau *et al.*, “Accelerating multi-site health informatics with streamlined data infrastructure using OMOP-on-FHIR,” *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 2022, pp. 4687–4690, Jul. 2022.
- [169] A. Hornback *et al.*, “Development of a generalizable multi-site and multi-modality clinical data cloud infrastructure for pediatric patient care,” in *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, Northbrook Illinois, 2022.
- [170] Y. Zhong *et al.*, “IDTVR: A novel cloud framework for an interactive digital twin in virtual reality,” in *2022 IEEE 2nd International Conference on Intelligent Reality (ICIR)*, Piscataway, NJ, USA, 2022.