Language Models for Temporal Decisions in Health Datasets

by

Ridam Pal
(PhD 19201)

Under the Supervision of Dr. Tavpritesh Sethi

Department of Computational Biology

Indraprastha Institute of Information Technology, Delhi

New Delhi - 110020

July, 2024

Language Models for Temporal Decisions in Health Datasets

by

Ridam Pal

A Thesis

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Computational Biology

Indraprastha Institute of Information Technology, Delhi

New Delhi - 110020

July, 2024

Certificate

This is to certify that the thesis titled "**Language Models for Temporal Decisions in Health Datasets**" being submitted by **Ridam Pal** to the **Indraprastha Institute of Information Technology Delhi**, for the award of the degree of **Doctor of Philosophy**, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

Dr. Tavpritesh Sethi
Associate Professor, Computational Biology
Founding Head, Center of Excellence in Healthcare
IIIT Delhi, 110020

Acknowledgements

Having completed my Ph.D. thesis, I reflect upon this significant milestone in my life with deep gratitude for the kindness and support bestowed upon me by numerous individuals. Each encounter I had along this journey provided valuable lessons and contributed to shaping the person I have become today. I am immensely thankful to all those who have played a role in my academic and personal development, for their guidance, encouragement, and belief in my abilities. The collective experiences and interactions have profoundly influenced my growth, and I am truly appreciative of the impact these kind individuals have had on my path.

I would like to commence by extending my heartfelt gratitude to my advisor, Dr. Tavpritesh Sethi, for his exceptional guidance, mentorship, and support throughout my academic journey. His expertise, invaluable insights, and unwavering dedication have been instrumental in shaping my research and professional growth. His constant encouragement, patience, and belief in my abilities have motivated me to push boundaries and strive for excellence. I am grateful for the countless hours he has devoted to guiding me, providing constructive feedback, and sharing their profound knowledge and expertise. I would like to express my heartfelt appreciation to the esteemed members of my PhD annual review and PhD comprehensive exam committees, Dr. Ganesh Bagler and Dr. Tarini Shankar Ghosh. Their insights, expertise, and guidance have significantly contributed to the successful progression of my doctoral journey. I am deeply grateful to my previous professor, Dr. Anjan Sarkar and Dr. Supratim Biswas for their invaluable contribution in shaping my early career and providing me with much-needed guidance. Their mentorship and encouragement played a pivotal role in helping me find my direction and develop a strong foundation in my field of study. Additionally, I would like to express my sincere gratitude to all my collaborators Dr. Piyush Mathur, Dr. Rakesh Lodha for their significant contributions to my study. Their commitment, knowledge, and assistance have been crucial in overcoming obstacles and attaining worthwhile results. They created a collaborative atmosphere that not only improved my research but also offered a space for mutual learning and development. Their suggestions, criticism, and collaborative attitude have greatly improved the caliber and significance of my work, and I am grateful for them. I would like to express my heartfelt gratitude to the management team at STL Digital for their exceptional support and understanding throughout my PhD journey. Their constant efforts to create a conducive work environment and reduce the burden on me have played a crucial role in allowing me to focus wholeheartedly on my research work. I extend my deepest appreciation to the company manager, Rupesh Prasad, for his remarkable leadership and guidance. His adept handling of various organizational responsibilities and continuous support ensured that my work obligations at the company were effectively managed, providing me with the necessary flexibility and freedom to pursue my PhD studies at IIITD. I am indebted to my professors, and the entire faculty at IIITD for their guidance, mentorship, and support throughout my doctoral journey. Their expertise, patience, and commitment to nurturing young researchers have been pivotal in shaping my

research work and academic development. I would like to convey my profound appreciation to IIIT-Delhi, which has served as my home for a number of years. The support offered by the Institute in terms of infrastructure, financial assistance, facilities, and helpful people continues to be among the finest I've encountered. Special thanks to Sudhanshu Tamta and Raju Biswas from the admin department for their support related to thesis submission.

I would like to express my heartfelt gratitude to my parents Prabir Kumar Pal, Jhuma Pal and my sister Rittika Pal for their constant support and encouragement throughout my academic journey. Their love, guidance, and sacrifices have been instrumental in shaping the person I am today. Their belief in my abilities and their constant motivation have been a constant source of strength for me. Their presence in my life has been invaluable, and I am deeply appreciative of the sacrifices they have made to ensure my success. Thank you for believing in me and for being my rock. This journey was never possible without their support and encouragement.

I would like to extend my heartfelt thanks to my dearest friends Onam Bhartia, Gaurang Mahawar, Soanak Deb, Hemang Sarkar, Soubhik Boral, Sargun Nagpal, Shaswat Patel, Harsh Bandhey, Saad Ahmed, Rohan Pandey, Akhil Jordia, Harleen Kaur, Gayatri Panda, Omkar Chandra, Sukriti Sacher, Vishakha Gautam, Sadiyah Afroz, Himanshi Agarwal, and Harsha for their indispensable companionship throughout this journey. Their presence has brought immense joy, laughter, and support during this complete journey. Their belief in me, encouragement, and uplifting words have been a source of motivation and inspiration. Whether it was celebrating accomplishments or providing a comforting shoulder during challenging times, their friendship has been a constant source of strength. I am truly grateful for their love, understanding, and the memories we have created together. Thank you, dear friends, for being there for me and for enriching my life in ways words cannot express. Special thanks to Gayatri, for being my constant pillar of support and being the best friend one could ask for on this remarkable journey of research. Your constant cheering and motivation have been invaluable, fueling my determination and inspiring me to overcome challenges.

I want to express my gratitude to my beloved seniors at IIIT-D, Dr. Aditya Nagori, Dr. Raghav Awasthi, Dr. Kaushik Biswas, Dr. Shreya Mishra, and Dr. Hridoy Shankar Dutta ,who have served as a beacon of light for me during my research career. Their guidance and readiness to divulge their own difficulties and experiences have been priceless. Their suggestions and insights were quite helpful in helping me develop my research strategy and get over challenges. I appreciate their advice and the expertise they have shared with me because it has helped me develop as a researcher. I extend my sincere thanks to Aditya and Raghav for their invaluable role as mentors and friends, whose presence has made a profound impact on my growth and success. Their presence has been akin to that of older brothers, always there to lend a helping hand and provide insightful answers to my queries.

I want to sincerely thank my junior PhD colleagues Pradeep Singh, Akshaya Devagada, Jasmine Kaur, Alok, Ayushi, Manas, Aakansha, Varsha and Pallawi who have contributed greatly to my research endeavors. Their energy, fresh-perspectives, and commitment to their work have been genuinely motivating. Special thanks to Pradeep Singh and Akshaya for keeping a constant check on me throughout my PhD journey. I am grateful for the support both of you have provided me, sharing insights, discussing ideas, and offering encouragement along the way.

I would like to express my gratitude to Hardik Garg, Vihaan Mishra and Sanjana Srinidhi and Deepak Mahato, my co-authors from undergraduate batches who have played a significant role in my research journey. Their willingness to listen, understand, and accurately implement my research guidance has been invaluable. Their dedication, attention to detail, and commitment to excellence have greatly contributed to the successful execution of our collaborative work. Additionally, their energy and involvement have made my research trip delightful, encouraging an amiable and exciting atmosphere. Beyond their academic contributions, I am also thankful for the friendships that have developed during our collaboration. I would also like to express my sincere gratitude to all other team members Akshala Bhatnagar, Prakriti Garg, Kriti Aagarwal, Gopal Mengi, Harshita Chopra, and many more who have been a part of my research projects. Their involvement and contributions have been integral to the successful outcomes we have achieved together. The collective effort, collaboration, and diverse perspectives brought forth by each collaborator have enriched the research process and expanded the scope of our work. I am thankful for their active participation and the collaborative environment we fostered, which made our research endeavors more impactful and fulfilling.

I would like to extend my thanks to my colleagues from the department, Dr. Neteesh Pandey, Indra Prakash Jha, Abhishek Halder, Madhu Sharma, Subhadeep Duari, and Swarnava Samanta, who have been a source of encouragement and support during challenging times. Their kind words, gestures, and positive energy have lifted my spirits and provided much-needed motivation. I am grateful for their empathy and understanding, which have helped me navigate through difficult periods.

Finally, I would like to express my heartfelt gratitude to all the colleagues and collaborators whose names may have been missed out. Each and every person I have interacted with throughout my research journey has contributed in their own unique way. Whether it was through engaging discussions, insightful feedback, or simply lending a listening ear, their contributions have been invaluable. I am deeply appreciative of the contributions made by all my colleagues and collaborators, and I am honored to have been a part of such an incredible community.

ABSTRACT

Healthcare has been undergoing a data-driven transformation, further accelerated by the COVID-19 pandemic. A significant amount of healthcare data is unstructured and underutilized. The success of Large Language Models (LLMs) in achieving human-like conversations has unlocked their potential in healthcare. For example, language models can help improve patient outcomes through temporal decision support, early warning systems, and clinical risk assessment. Through our work, we have explained how language models can assist in pandemic preparedness and support decision-making processes in critical care. Integrated frameworks incorporating machine learning, deep learning, and language models have been developed to effectively track and analyze temporal changes in unstructured healthcare data, to make informed decisions, and to enhance patient outcomes in a dynamic healthcare landscape.

In this thesis, my first contribution was a deep learning based language model for modeling the spike region of COVID-19 genome sequences. This led to novel knowledge discovery and real-world implementation for predicting pandemic progression, *StrainFlow,* which successfully captured COVID-19 caseloads two months ahead of their occurrence. The integrative framework for language models, statistical features and machine learning to capture the temporal changes in the semantics of the genomic sequence was deployed as a publicly available web-application.

In my second contribution, I constructed language models on COVID-19 scientific literature to track and predict emerging scientific evidence. The findings of this contribution illustrated that temporal changes in unsupervised word embeddings of scientific literature effectively captured and tracked new knowledge. Additionally, my work leveraged machine learning techniques and predicted emerging themes based on evolving word associations. This was also implemented as an openly available web application called *EvidenceFlow*.

In my third contribution, I developed language models on unstructured clinical notes data from intensive care units (ICU) for prognosticating critical outcomes. Shock Index (SI) is a commonly employed prognostic indicator used in intensive care units (ICU) and emergency settings to assess patient outcomes. We developed a comprehensive multimodal early warning system (EWS) utilizing an integrated framework combining machine learning, deep learning, and language models. The framework leverages routinely available vital signs and clinical notes data to detect abnormal shock index and provide timely alerts for potential deteriorations in patient health. This model is planned to be evaluated prospectively for real-world clinical decision making, which is outside the scope of my thesis.

In our final contribution, I contributed to the development and deployment of an end-to-end language model pipeline and android application, *WashKaro*, for raising WASH awareness during the COVID-19 pandemic. This was one of the first AI-based information dissemination applications built during COVID-19, which provided both Hindi and English bite sized text and

audio based upon text summarization, word embedding similarities and text-to-speech technologies using advanced NLP methods. The application and research publication also demonstrated the user-feedback based improvement of our AI model, providing pointers for designing public health intervention systems for pandemic preparedness.

Overall, my thesis contributed to the development, evaluation, and deployment of language model based technologies in ICU and pandemic preparedness settings, specifically in the setting of future predictions and early warning systems using temporal data. The findings contribute to advancing knowledge and methodologies while assisting medical practitioners and policymakers in effectively responding to disease outbreaks and formulating data and AI-augmented policy for healthcare settings.

List of Figures

Sample Entropy dimension 32 and monthly new cases for USA, indicating that the entropy in dimension 32 has a leading relationship with the cases. (C) Detrended Cross-correlation coefficient values for different lags between Entropy dimension 27 and new cases for India. (D) Line plot for Sample Entropy dimension 27 and monthly new cases for India, indicating that the entropy in dimension 27 has a leading relationship with the cases. (E) Feature importance scores from the Boruta algorithm for predicting cases in the month following the next month.

3.6 Prediction of new COVID-19 cases with Sample Entropy values of the latent dimensions. (A) Line plot showing the Entropy values of the selected features and new COVID-19 cases for the USA. (B) Actual and predicted cases based on the entropy values of selected features for the USA. The model predicts a rise in cases for July and August 2021. (C) Entropy of selected features and new cases for Japan. (D) Actual and predicted cases for Japan. A spike in cases is predicted for July and August 2021. (E) Entropy of selected features and new cases for India. (F) Actual and predicted cases for India.

3.7. Potential association of codons observed in SARS-CoV-2 Delta variant (lineage B.1.617.2) with their corresponding entropy features, and the trend of caseloads with the entropy features. (A) Absolute latent space weights of the codons associated with the entropy features linked to the Delta variant. Line plots showing the entropy features and cases in countries, (B) England, (C) India, and (D) USA. The entropies show an increasing trend in the months April - June 2021 for India and USA, indicating a possible surge in the delta variant in these countries. (E) Predicted and Actual cases for India. The region shaded in grey represents the months for which the case prediction model was prospectively validated. (F) Entropy and Caseloads for India. Explore-exploit patterns in the genomic feature-space can be observed.

4.1 Graphical representation of proposed framework explaining the complete workflow. The pipeline takes abstracts as input from which entities are extracted using Named Entity Recognition. Embeddings are generated, which are used as features for longitudinal networks. These networks are used for visualizing the trends using alluvial diagrams, link prediction, and predicting top-k influential modules for theme prediction.

4.2 (A) Showing the number of articles occurring each month. The curve depicts that there has been a rampant increase in the number of articles across each month since February 2020. (B) Latent space of word embeddings of diseases visualized around the keyword 'post-covid syndrome', displaying 100 isolated points nearest to it. (C) Bar plot (left) showing the frequency of top diseases in the corpus of abstracts extracted using NER. (D) Bar plot (right) showing the frequency of top chemicals in the corpus of abstracts extracted using NER.

4.3 (A) Alluvial diagram for tracking the trends in 2020, from the networks of March, August and December. (B) Alluvial diagram for monitoring the trends in 2021, from the

networks of January, March and June. The alluvial diagram eases tracing the temporal dynamics of literature across different time intervals.

4.4 (A) Evaluation of Mean Squared Error (MSE) between original and predicted proximity scores for the network of April 2021, May 2021, June 2021. (B) Confusion Matrix with normalized values of results from AdaBoost classifier across the months of April 2021l, May 2021, and June 2021. AdaBoost has been the best-performing model across all three months. (C) Results of link prediction between disease entities from March 2021 to June 2021, with a margin of error for 95% Confidence Intervals. The mean value of metrics has been recorded by testing the models on a resampled test set.

4.5 Alluvial diagram for tracking the trends of chemical entities from the networks of February 2020 to November 2021. Threshold used for assigning links between nodes was set at 70th percentile of cosine similarity between pairs of top-100 entities in respective months.

4.6 Distribution of errors in the prediction of proximity scores between node pairs (used as features in model training) for the month of June 2021. The white marker depicts the median and the black marker depicts the interquartile range. The mean of errors were found to be close to zero.

4.7 Temporal trends of PageRank centrality of (A) ‘statins’, (B) ‘glucocorticoids’, (C) ‘depressive’, (D) ‘thromboembolic’. The annotations denote the module/cluster index (lesser meaning more central).

4.8 Temporal evolution of the context of the term “vaccine” across alternate months. Top-10 most similar words based on cosine similarity using monthly Word2Vec embeddings are plotted. Origin and evolution of drug repurposing in early months, hesitancy and vaccine candidates in later months are highlighted.

5.1 Overview of ShockModes pipeline. ShockModes takes physician notes and vitals data as input, clinical entities (Therapeutics & History of Patient Illness) are extracted using MED7 Named Entity Extraction(NER) whereas Time-Series (TS) features are extracted using tsfresh python module. Unsupervised Embeddings fused with TS features serve as input features for the machine learning models for predicting the onset of abnormal SI with 24-hours lead time. SHAP analysis reveals interpretable features for global and patient-specific notes.

5.2 a) Plots illustrating the frequency of different categories of treatment notes. b) Frequency of Physician note types. c) Word Cloud of Therapeutics present in the corpus. d) Top-20 most common therapeutics encountered during periods with abnormal SI.

5.3 Leverage gained with addition of History of Present illness and length of context. Efficacy of various machine learning models a) exclusive and b) inclusive of HOPI as features. Therapeutic embeddings were generated from BioBERT while HOPI embeddings were generated from DocBERT. c) Comparison of different contextual window lengths (512, 256, 128) for the embeddings generated from BioBERT. These embeddings were compared against different metrics for the best pipeline exclusively of textual data(Gradient Boosting+DocBERT embedding+BioBERT contextual embedding). The plot depicts that sequence length of 512 has achieved the best AUC-ROC demonstrating that with an increase in the sequence length, there has been an increment in the AUC-ROC.

5.4 Interpretability of model prediction using SHAP analysis. Sections such as chief complaints, past medical history, and treatment plan along with specific therapeutics such as anticoagulants and antibiotics contain important information for predicting abnormal SI in global feature importance analysis (a) as well as understanding the influence each feature has on the outcome for the textual pipeline. (c) Vitals features were predominant in determining the abnormal SI for the multimodal pipeline as well as certain therapeutics such as Heparin Sodium Prophylaxis turned out to be significant in global feature importance analysis. (b,d) An example of an individual level feature importance for a record is shown as a waterfall plot for textual and multimodal pipeline respectively.

5.5 Evaluation of models based on clinical segregation. a) Efficacy of ShockModes on top diagnosis. b) Efficacy of only text-based pipelines on top drugs.

6.1 Proposed workflow of the App based upon Identify, Simplify, Amplify and Quantify framework as specified WHO's EPI-WIN strategy.

6.2 NLP Pipeline. The pipeline takes in news articles and the World Health Organization (WHO) reports and constructs two-level sentence similarity between titles and the full-text to build a similarity score. Finally, the relevant texts are subject to translation and text to speech conversion for local language consumption (Hindi).

6.3 Self Assessment Tool Flowchart. Based on the World Health Organization(WHO) Interim Guidance, a questionnaire and flowchart were developed to classify the responders as `Suspects' or `Non-suspects'. Here SOB refers to Shortness Of Breath, and ARI refers to Acute Respiratory Infection.

6.4 Request-Response cycle in the chatbot. This is a schematic diagram depicting how the answer is displayed whenever a query is asked to the chatbot by a user.

6.5 Analysis of Natural Language Processing(NLP) pipeline. The graph shows relevance as a

function of user feedback functionality in the app. Relevance is seen to increase with cumulative feedback over time. From day 0 onwards, the Relevant Count's angular coefficient is 1.39 (±0.488), the angular coefficient of Irrelevant Count is -0.99 (±0.602), with an average slope difference of about 2.29.

6.6 Analysis of Public Health Survey. Distribution graphs showing the distribution of gender among Hindi and English Users. It clearly shows skewness in gender for English users whereas in the case of Hindi users it shows an approximate normalization among the genders.

6.7 Analysis of Self Assessment. A simple user-level self-assessment has been deployed to enable the general population to perform self-assessment and identify the population at risk, which can be used as an effective screening. A higher trend for a positive COVID-19 report in people who reported cough was observed. The symptoms of the disease have been known to change with strains, hence this approach of crowdsourcing information provides an agile approach to screen patients with specific symptoms.

List of Tables

5.4 Results of SI (abnormality) prediction combining textual and vitals data on a 24-hour cohort of MIMIC III dataset with a margin error of one standard-deviation from the mean. Random Sampling(N=100) of the test dataset, with Bootstrap iterations of 100 has been recorded for the mean value.

List of Publications

Publications and Preprints

1. Sargun Nagpal[+], **Ridam Pal**[+], Ashima, Ananya Tyagi, Sadhana Tripathi, Aditya Nagori, Saad Ahmad, Hara Prasad Mishra, Rintu Kutum, Tavpritesh Sethi. *Genomic Surveillance of COVID-19 Variants with Language Models and Machine Learning.* **Frontiers in Genetics**. https://doi.org/10.1101/2021.05.25.445601
2. **Ridam Pal**, Harshita Chopra, Raghav Awasthi, Harsh Bandhey, Aditya Nagori, Tavpritesh Sethi. *Predicting Emerging Themes in Rapidly Expanding COVID-19 Literature with Dynamic Word Embedding Networks and Machine Learning*. **J Med Internet Res (JMIR).** https://doi.org/10.1101/2021.01.14.21249855
3. **Ridam Pal**, Shaswat Patel, Akshala Bhatnagar, Hardik Garg, Pradeep Singh, Ritesh Singh Soun, Aditya Agarwal, Aditya Nagori, Ashish Khanna, Rakesh Lodha, Piyush Mathur, Tavpritesh Sethi. *ShockModes: A Multimodal Model for Prognosticating Intensive Care Outcomes from Physician Notes and Vitals.* **(In Review: Computer Methods and Programs in Biomedicine).**
4. Rohan Pandey, Vaibhav Gautam[+], **Ridam Pal**[+], Harsh Bandhey, Lovedeep Singh Dhingra, Tavpritesh Sethi. *A Machine Learning Application for Raising WASH Awareness in the Times of Covid-19 Pandemic*. **Nature Scientific Reports.** https://doi.org/10.48550/arXiv.2003.07074

Other Publications and Preprints

1. Harshita Chopra[*], Aniket Vashishtha[*], **Ridam Pal**, Ashima Ashima, Ananya Tyagi, Tavpritesh Sethi. *Mining Trends of COVID-19 Vaccine Beliefs on Twitter With Lexical Embeddings: Longitudinal Observational Study*. **JMIR Infodemiology.** https://doi.org/10.2196/34315
2. Raghav Awasthi[*], **Ridam Pal**[*], Pradeep Singh, Aditya Nagori, Suryatej Reddy, Amogh Gulati, Ponnurangam Kumaraguru, Tavpritesh Sethi. *CovidNLP: A Web Application for Distilling Systemic Implications of COVID-19 Pandemic with Natural Language Processing.* **medRxiv**. https://doi.org/10.1101/2020.04.25.20079129
3. **Ridam Pal**, Hardik Garg, Shaswat Patel, Tavpritesh Sethi. *Bias Amplification in Intersectional Subpopulations for Clinical Phenotyping by Large Language Models.* **medRxiv (***In Review: AMIA Symposium***).** https://doi.org/10.1101/2023.03.22.23287585
4. **Ridam Pal**, Sanjana Srinidhi, Deepak Mahto, Kriti Agrawal, Gopal Mengi, Sargun Nagpal, Akshaya Devadiga, Tavpritesh Sethi. *Characterizing the Emotion Carriers of COVID-19 Misinformation and Their Impact on Vaccination Outcomes in India and the United States.* **arXiv(In Review)**. https://doi.org/10.48550/arXiv.2306.13954

Contents

Chapter-1

Introduction

1.1 AI in Healthcare

Recent advances in artificial intelligence and machine learning (AI/ML) have garnered significant attention due to their potential to transform healthcare delivery. Areas such as medical imaging analysis, clinical decision support, early disease diagnosis, and unstructured data analysis have been at the forefront of such transformation (Rajpurkar et al. 2017; Adams et al. 2022; Henry et al. 2022; Nagori et al. 2019). These technologies have demonstrated potential in simplifying workflows, and helping medical practitioners to analyze large volumes of complex healthcare data from electronic health records (Adams et al. 2022; Henry et al. 2022) that include medical images (Nagori et al. 2019; Rajpurkar et al. 2017), vitals data (Nagori et al. 2021), treatment charts, genomic sequences (Hie et al. 2021), etc. AI/ML models can learn from vast amounts of data, enabling them to recognize subtle patterns and associations that may not be apparent to human observers. Explainable models also have the potential to provide more in-depth insight into a patient's health reports and improve the accuracy of healthcare processes, leading to better patient outcomes, reduced healthcare costs, and enhanced overall healthcare quality (Davenport and Kalakota 2019; Bajwa et al. 2021).

A major social challenge in achieving optimum development of AI/ML in healthcare is the lack of collaboration between healthcare professionals and data scientists. Once bridged, this collaboration can be transformative for constructing holistic models that combine the vast amount of unstructured and structured data from electronic health records (EHR) ("Natural Language Processing in Healthcare Medical Records," 2023.). However, it is very time-consuming for a physician to aggregate these modalities of data and draw meaningful conclusions. Structured data on its own contains incomplete information about a patient leading to inaccurate or inefficient treatment. Extracting information from unstructured data to complement the structured data is a rigorous manual process leading to the physician's burnout which puts a greater burden on the already resource-starved healthcare industry. Also, Big Data Analytics demonstrated that a significant amount of healthcare data is stored in an unstructured format (ex. like medical reports and clinical notes), which often remains unutilized by health systems as the extraction of relevant information from such data sources is resource-intensive ("Natural Language Processing in Healthcare Medical Records," 2023.). With the advent of Natural Language Processing (NLP), AI/ML algorithms are designed to process these unstructured data into a computer-readable format for extracting valuable information in a computationally efficient manner. This can assist physicians in reducing the burden and drawing better conclusions based on detailed insights generated from unstructured data. Although NLP

has significantly contributed to other domains, the use cases in real-world healthcare applications are minimal. Some popular use cases for NLP techniques in healthcare include clinical documentation, computer-assisted coding, clinical decision support, and clinical trial matching (Ariwala 2019; Garvin 2019). Clinical Documentation can be done using voice recognition tools which allow medical practitioners to use transcriptions for documenting patient records(“Natural Language Processing 101: A Guide to NLP in Clinical Documentation” 2022). It sets free clinicians from the laborious task of complying data with EHR systems, thereby allowing them to give more time to patient care. Another implementation of NLP in healthcare is computer-assisted coding, which allows synthesizing of relevant information from long treatment charts. Primarily this process of distilling information took hours of tedious effort, which was subsequently reduced with the usage of NLP techniques. Although this is a popular use case of NLP, its adaptation rate in healthcare has been low (30%) due to its poor accuracy (Garvin 2019). Clinical decision support is another potential area where efforts have been made to employ NLP techniques for the early detection and diagnosis of diseases. These implementations of NLP technologies facilitate the exploration of unstructured datasets from a healthcare perspective.

1.2 Types of Data in Healthcare

Health datasets contain information related to an individual's or group of individual health and prior medical history. The data can be collected, stored, analyzed and modeled for studying the impact of a particular disease (“Health Data,” 2023.). It encompasses a wide range of data such as Electronic Health Records (EHRs), clinical trial data, public health data, scientific data, genetic data, etc. Public health data refers to information collated by public health agencies, such as disease surveillance data, vaccination rates, and health behavior surveys (van Panhuis et al. 2014). This data provides insights into individuals' responses to a specific disease (pandemic), including their adherence to public health guidelines and treatment policies. Clinical data are gathered from EHRs or EMRs, typically in the form of vital signs, medical notes, nursing notes, and discharge summaries (Pendergrass and Crawford 2019). This data can be utilized to improve patient outcomes in critical care by developing models based on data-driven methods. Analysis of critical care data can assist medical practitioners to gain a more comprehensive understanding of patient conditions, which can inform treatment decisions. Genetic data is a sort of health data that offers insights into an individual's propensity for particular diseases and potential response to treatments (Sariyar, Suhr, and Schlünder 2017). With the development of genetic testing, medical professionals may now gather genetic variant information, suggesting details about their susceptibility to particular diseases and the likelihood of responding to various treatments. Medical professionals can personalize treatment recommendations based on a patient's unique genetic makeup by analyzing an individual's genetic data, perhaps resulting in better treatment outcomes. Scientific data in healthcare is information collected from sources such as clinical trials, patient records, and biomedical research studies (White 2020). It is crucial for advancing medical knowledge, developing new treatments, and improving patient outcomes. This data can

identify patterns and trends in patient health, track the efficacy of treatments, and develop predictive models for disease diagnosis and prognosis. Healthcare providers can use this data to make informed decisions about patient care, identify effective treatments, adjust treatment plans, and monitor patient progress. Sharing and analyzing scientific data across different research groups can accelerate scientific discoveries and facilitate collaboration between researchers.

Healthcare data exists in the format of both structured and unstructured. An example of a system that contains both structured and unstructured data in healthcare is Electronic Health Record (EHR) (Pendergrass and Crawford 2019). EHRs contain structured fields and data elements that capture and organize patient health information in a standardized format. These fields contain information related to patient demographics (e.g., name, age, gender), medical history, diagnoses, medications, laboratory results, and vital signs. While unstructured data in EHR refers mostly to textual data which consists of clinical notes, nursing notes, medication charts, patient genome sequences, etc. These datasets originate from diverse sources such as hospitals, clinics, insurance providers, government agencies, and research institutions. Hence, healthcare data contain a wealth of knowledge on various aspects of health, including disease prevalence, treatment efficacy, and health outcomes, which are crucial for advancing medical research. Researchers, medical practitioners, and policymakers can identify patterns and trends that can intervene in medical decision-making and improve patient care by analyzing vast amounts of health data through proper tools and techniques.

The growth of health datasets across various institutions highlights the potential of data in the healthcare domain. However, the healthcare industry faces challenges in fully harnessing this valuable resource. Previous studies have indicated that a significant portion (80%) of healthcare data exists in unstructured formats, posing difficulties in effectively utilizing it (Kong 2019). Unstructured data originates from sources such as electronic health records (EHR), patient data, diagnostic summaries, progress records, medical imaging, and prescriptions, which are not readily machine-readable. Consequently, processing and organizing this data into structured, machine-readable formats pose significant time complexity. My dissertation focuses on developing pipelines to analyze and derive insights from unstructured healthcare data, specifically textual data.

1.3 Textual (Unstructured) Data in Healthcare

In healthcare, unstructured data refers to information that does not have a predefined format or organization. It includes data that is not easily searchable or categorized, such as free-text clinical notes, narrative reports, scientific literature, images, audio recordings, genome sequences, and social media posts. Here are some examples of textual (unstructured) data for the healthcare domain that I have used in my studies:

- *Clinical data*: These are written or dictated narratives created by medical practitioners during patient encounters (Pendergrass and Crawford 2019). The data are generated in the form of physician notes, nursing notes, discharge summaries, and lab reports, among others. Clinical notes and discharge summaries contain detailed information about symptoms, medical history, physical examination findings, treatment plans, and other relevant clinical observations. Lab reports generate unstructured reports based on the examination of tissue samples, such as biopsies or surgical specimens. These reports describe the findings, including the presence of abnormalities or diseases.

- *Social Media Data*: Patient-generated content on social media platforms, such as Facebook, Twitter, or online forums, can provide insights into patient experiences, health beliefs, and opinions about treatments. Analyzing these unstructured text data can help identify trends and understand patient perspectives.

- *Scientific Medical Literature*: Scientific articles, research papers, and publications in medical journals often contain unstructured data in the form of text, tables, and figures (White 2020). Extracting relevant information from these sources can contribute to evidence-based decision-making and research.

- *Genomic data*: Genomic data does not conform to the structure of natural language text. It comprises information extracted from an individual's DNA in the form of A, T, G, and C; which encompasses genome sequence, genetic makeup, genetic variations, and gene expression patterns ("Genomic Data," 2023.). These forms of unstructured data enable clinicians and researchers to better understand the genetic basis of diseases, identify potential therapeutic targets, and develop tailored treatment approaches.

1.4 Predictive Modeling in Healthcare

Predictive modeling can be leveraged across various domains within healthcare settings for improving patient care, enhancing operational efficiency, and optimizing resource allocation. Here are some key areas where predictive modeling have be applied in the healthcare setting:

- Early Disease Prediction and Patient Monitoring: Predictive models have been developed to identify individuals at risk of developing specific diseases or health conditions. When integrated with remote monitoring systems, these models serve as an Early Warning System, continuously analyzing real-time patient data to detect temporal changes indicating clinical deterioration (Adams et al. 2022; Henry et al. 2022; McGaughey et al. 2021; Smith et al. 2014). Early disease prediction models utilize a broad spectrum of data such as demographic characteristics, medical history, physiological data, vital signals, early symptoms and laboratory test results, for predicting the patient outcome in advance

(Alanazi 2022). These models have the capability to alert medical practitioners about potential health threats, enabling them to deliver timely interventions to prevent adverse outcomes, consequently reducing the occurrence and severity of diseases.

- Resource Allocation and Capacity Planning: Predictive models can help healthcare organizations optimize resource allocation and capacity planning by forecasting patient demand, bed occupancy, and staffing needs (Tello et al. 2022). By analyzing historical data and trend analysis, predictive models support proactive decision-making and resource allocation to ensure efficient use of resources and timely access to care for patients (Barros, Weber, and Reveco 2021). One such utility of predictive models for resource planning has been used to identify patients at high risk of hospital readmission. These models can be utilized to analyze the patient data which include clinical variables, discharge summary, patient history, socio-demographic factors, and healthcare utilization patterns. It can help healthcare providers in prioritizing resources to prevent unnecessary readmissions and improve critical care coordination.

- Population Health Management: Population health management is a critical component of public health policy, aiming to improve the health outcomes of entire populations. Predictive models play an important role in this by providing insights into population health trends and predicting disease outbreaks (Jonkmans, D'Acremont, and Flahault 2021; Yadav and Akhter 2021), aggregating and analyzing patient data from various sources, such as electronic health records (EHRs), physician notes, and medical imaging, to identify patterns and make predictions. This information helps in understanding the prevalence of specific diseases, identifying risk factors, and recognizing health disparities within the population. It empowers policymakers to introduce preventive measures, such as vaccination campaigns or public health interventions, to mitigate the impact of disease outbreaks. Furthermore, predictive modeling can be used to allocate resources to high-risk populations, prioritize health education and prevention programs, and distribute funding for public health initiatives.

1.5 Leveraging Unstructured Datasets for Predictive Modeling

Numerous researchers have developed predictive models by harnessing structured data, employing various techniques across healthcare settings. Yet, the utilization of unstructured datasets has emerged as a pivotal strategy for advancing predictive modeling techniques in this domain. Leveraging such datasets presents a multiple challenge due to their inherent complexity and heterogeneity. However, innovative approaches, such as natural language processing (NLP) (H. Wu et al. 2022; B. Zhou et al. 2024), computer vision (Esteva et al. 2021; A. Sharma et al. 2023), audio processing (Rana et al. 2023), and language models (Kai He et al. 2023; Clusmann et al. 2023) have enabled researchers to extract valuable insights from unstructured healthcare

data sources. Previously, researchers have worked to build models utilizing images and video collated from the radiological departments, Intensive Care Units, etc (Nagori et al. 2019; Vats et al. 2022; Z. Zhang and Sejdić 2019; Varoquaux and Cheplygina 2022; Ullah et al. 2023; Gourdeau et al. 2022) for predicting the onset of disease. By harnessing the latent information embedded within unstructured datasets, predictive modeling algorithms can enhance their predictive accuracy, uncover hidden patterns, and facilitate more nuanced decision-making across various domains. Despite the growing interest in utilizing textual datasets, only a limited number of studies have explored their potential for enhancing healthcare decision-making. In this thesis, I have demonstrated the effectiveness of textual data in the healthcare domain to construct predictive models employing language models.

However, the dynamic nature of healthcare data necessitates predictive modeling techniques that account for temporal changes. Temporal predictive modeling helps in understanding how health-related events evolve over time, thus enabling us to forecast disease trajectories and patient outcomes based on the dynamic changes occurring in the data.

1.6 Temporal predictive modeling in Healthcare

Early detection and prompt treatment decisions play a crucial role in clinical management (Adini et al. 2019). Human capabilities have limitations which may result in the oversight of subtle patterns. Hence, the use of decision support systems allows to capture these intricate patterns and aid in making informed clinical decisions. Clinician decision support refers to the use of technology, specifically AI and computational tools, to provide healthcare professionals with valuable insights at the point of care (Sanchez-Martinez et al. 2021). It aims to assist clinicians in making well-informed and evidence-based decisions regarding patient diagnosis, treatment, and management. Temporal modeling in the clinical decision support system holds potential, as it involves treatment decision-making over time and continuously reassessing ongoing care to enhance clinical outcomes for individual patients (Bennett et al. 2014; Xie et al. 2022; Ofstad et al. 2014). Health datasets often contain a large amount of textual patient data, including medical history, lab results, and treatment plans. By analyzing this data over time, decision support systems can identify and track evolutionary trends in a patient's health, predict potential health risks, and suggest appropriate interventions (Cosgriff and Celi 2020). This can improve the accuracy of diagnoses and treatment plans, reduce the risk of adverse events, and ultimately improve patient outcomes. Additionally, temporal decision support can aid health management organizations by identifying changes in disease incidence and treatment outcomes over time, which can inform public health policies. Overall, incorporating temporal decision support into health datasets can lead to better health outcomes for individual patients and populations alike.

The implementation of language models for temporal decision support has gained momentum due to their potential to analyze and mine information from various unstructured data sources. It

can be used to develop predictive models that can identify patterns and predict critical outcomes such as pandemic progression and abnormal shock index based on temporal changes. These advancements have significant implications for improving the efficiency, accuracy, and accessibility of healthcare services. Significant progress has been achieved in the temporal modeling of healthcare data, but there are practical hurdles in translating these advancements from controlled laboratory or theoretical settings to real-world applications involving actual patients. Scarcity of research involving the implementation of language models for temporal decision-making is also one of the limiting factors. Overcoming these challenges is crucial to harness the full benefits of temporal modeling in improving patient care. Therefore, the study presented in this dissertation is important as it provides insights into the various ways in which language models can be utilized to address critical challenges in the healthcare domain. It also highlights the need for continued research to optimize the performance of language models in health datasets and ensure their safe and ethical deployment in real-world settings.

1.7 Language models

Historically, the process of incorporating salient linguistic information into natural language processing (NLP) systems involved a manual feature engineering approach. This method required substantial expertise and effort to determine appropriate numerical functions that would accurately reflect the textual data. In contrast, with the advent of language models, the concept of word embeddings emerged as a streamlined solution for learning text corpora without the need for manual labeling, feature extraction, or feature engineering. A real-valued vector representing a single word based on the context in which it appears is described as word embedding (Almeida and Xexéo 2019). Language models facilitate word vector representation which enables to map words in vocabulary to a point in a vector space. According to the "distributional theory," words that appear in related or comparable settings have similar meanings (Sahlgren 2008). As a result, we anticipate that in vector space, the embeddings for words that are semantically or syntactically connected will be closer to one another than for words that are unrelated. This relatedness is solely dependent on the corpus or set of text data used for creating embeddings. The unsupervised nature of word embeddings enables their application to any text data, which includes news articles, clinical text from electronic health records (EHRs), social media posts, genomic sequences, public healthcare data in the form of research articles, blogs, surveys, and so on. The unsupervised embedding synthesizes insightful information from textual data which can be utilized for decision-making and prediction tasks, such as text categorization, sentiment analysis, and entity recognition. This enhanced the efficacy of numerous NLP applications, including chatbots, recommendation engines, fraud detection tools, etc.

Language models started to emerge back in the early 1950s when researchers made the first attempts to utilize statistical techniques in the field of NLP (Sahlgren 2008). The Markov model was one of the simplest generative models developed in the early stages of this discipline, which

could generate text based on statistical patterns in a training corpus (Rabiner 1989). In the upcoming decades, researchers continued to improve and build upon these early models, creating complex architectures such as hidden Markov models, recurrent neural networks, convolutional neural networks, etc. In the late 2000s, deep learning methods such as the long short-term memory (LSTM) architecture (Hochreiter and Schmidhuber 1997) and the transformer architecture paved the way for significant breakthroughs in the field of NLP (Vaswani et al. 2017). These techniques allowed the development of potent language models that could produce text with a high level of fluency and coherence. Over the last decade, there has been a resurgence of interest in language models. This led to the development of large-scale pre-trained models such as BERT (Devlin et al. 2018) and GPT-3 (Brown et al. 2020), which achieved state-of-the-art results across numerous NLP tasks. These models have been trained on a massive corpus of textual data which allowed them to learn a broad spectrum of linguistic information from the corpus. Advanced computational power and easy accessibility to large textual corpus have expedited research around language models. It is conceivable that we will see much more sophisticated language models in the near future, which will enable even more potent and versatile NLP applications as academics continue to investigate new methodologies and architectures. Some of the applications of language models have been discussed in the next section.

1.8 Applications of language modeling approaches

Language models (LMs) have been shown to improve the field of natural language processing due to their ability to perform complex tasks such as machine translation (L. Wang et al. 2023), speech recognition (K. Hu et al. 2023), and question-answering (Y. Tan et al. 2023). These tasks involve understanding and processing human language, which is often complex and ambiguous for a machine to interpret. Language models exploit deep learning algorithms to learn semantic relationships between words and sentences; hence extracting information from the linguistics. It has demonstrated its utility in a wide range of applications including language generation tasks such as summarization (Kryściński et al. 2018), dialogue generation (Y. Cao et al. 2020), question answering (Zaib et al. 2022), and story generation (M. Chen and Gimpel 2021). In addition, it has been used for sentiment analysis (Kumawat et al. 2021), topic modeling (Eklund, Forsman, and Drewes 2022), and named entity recognition (R. Sharma, Morwal, and Agarwal 2022). Moreover, language models have been implemented for various other applications such as fraud detection (Dong 2014), recommendation systems (J. Lin et al. 2023), and search engines (Ziems et al. 2023). For instance, the power of language models can be harnessed to understand multilingual language, facilitating language barrier reduction, enhancing information accessibility, and promoting effective communication among individuals from different linguistic backgrounds (Doddapaneni et al. 2021). Furthermore, language models could aid in the development of automated systems that can track and monitor social issues, such as political unrest during election campaigns (Y. Hu et al. 2022), hate speech (Tekiroğlu et al. 2022),

misinformation (Taboubi, Brahem, and Haddad 2022), etc. As we continue to refine language model's architecture, their long-range applications will become increasingly important for addressing some of the most pressing challenges faced by the research community.

Language models have seen use in healthcare for tasks like analyzing medical records (J. Huang, Xu, and Vydiswaran 2016) and research articles (Akkasi and Moens 2021), generating medical codes (K. Patel et al. 2017; K. Lee et al. 2017) and clinical abbreviations (Yue Liu et al. 2015). Ongoing research focuses on refining models tailored for healthcare contexts (Yang, Chen, PourNejatian, Shin, Smith, Parisien, Compas, Martin, Costa, et al. 2022), enhancing generalizability, advancing predictive modeling (Lau et al. 2021), and addressing privacy concerns (Abdalla et al. 2020). These continuous efforts aim to enhance the effectiveness and applicability of language models in the field of healthcare.

1.8 Expansion of Language Models and their application to Healthcare

Recent advancement in language models has demonstrated their potential application in the healthcare domain. Availability of a large corpus of unstructured datasets in the form of clinical notes, research articles, social media posts, and patient records, has facilitated the development of language models trained on these diverse and heterogeneous data sources. Numerous studies have shown its application in various tasks such as clinical prediction (Steinberg et al. 2021), drug discovery (Z. Liu et al. 2021), medical question answering (Yang, Chen, PourNejatian, Shin, Smith, Parisien, Compas, Martin, Flores, et al. 2022), and named entity recognition (NER) (Yang, Chen, PourNejatian, Shin, Smith, Parisien, Compas, Martin, Costa, et al. 2022). Additionally, researchers have depicted that these models can be valuable for building predictive models (Steinberg et al. 2021; Amrollahi et al. 2020), which can identify patients at risk for certain conditions or forecast pandemic/disease outbreaks. It can also aid in the development of new therapies, and diagnostic tools for interventions, ultimately leading to better health outcomes for individuals and populations. It has also been extended to the field of bioinformatics for genomic analysis (H.-L. Li, Pang, and Liu 2021), protein sequence analysis (Z. Lin et al. 2022; Madani et al. 2020), gene expression analysis (Avsec et al. 2021), and targeted drug design (Uludoğan et al. 2022).

The release of Large Language Models (LLMs) such as BERT (Devlin et al. 2018), and GPT3 (Brown et al. 2020) has opened some interesting directions for exploration. While BERT architecture holds advantages for advanced fine-tuning focussed on specific domains (Y. Zhou and Srikumar 2022), GPT-3 models can be trained on very few instances of data related to a specific domain. Incorporating these LLMs into the healthcare sector has the potential to enhance performance in complex tasks like clinical document summarization, medical question answering (Shao et al. 2023), and extraction of medical entities (Tarcar et al. 2019). As research in this area continues to evolve, language models are expected to play an important role in

transforming healthcare delivery and advancing medical knowledge. However, these applications are limited by challenges such as managing the unstructured data, like clinical notes and handwritten documents, due to their complexity and variability (Polnaszek et al. 2016). Moreover, the real-world application of processing and modeling such data presents considerable challenges. Issues such as healthcare data quality, stringent privacy regulations, and the requirement for specialized domain knowledge hinder their widespread adoption.

1.9 Limitations and Challenges of language models in Healthcare

Despite the capabilities of language models, they also have inherent limitations that need to be carefully considered. This section highlights a few such limitations of language models in the context of healthcare applications, shedding light on the challenges they pose for patient care and clinical decision-making.

Temporal decision-making based on Contextual Understanding: Language models excel in generating coherent text but struggle with contextual understanding. More often the models were rarely used to extract relevant features based on the semantics of textual data. In healthcare, where precise interpretation is crucial, language models fail to capture subtle clinical nuances, leading to inaccurate or misleading predictions. I have tackled these challenges in my dissertation by developing multiple robust pipelines trained on diverse datasets, thereby facilitating temporal decision-making.

Lack of Explainability: Language models, particularly advanced deep learning models, can be perceived as black boxes due to their complex framework and intricate computations. Understanding the inner workings and reasoning behind its predictions can often be challenging. This lack of transparency can hinder trust and acceptance among healthcare professionals, who require explainable and interpretable models to understand the reasoning behind the generated outputs. In this dissertation, I have addressed this shortcoming and demonstrated techniques for the explainability of the models and pipelines.

Bias and Generalization: Language models heavily rely on training data, which may contain inherent biases present in the data sources. These biases can lead to biased language generation, perpetuating healthcare disparities and inequalities. Additionally, language models may struggle with generalization across diverse patient populations. It predominantly learns from the available data, which may not adequately represent all demographic groups.

Ethical Concerns: The deployment of language models in healthcare raises ethical concerns related to patient privacy, consent, and data protection. Language models require access to vast amounts of patient data, which must be handled with utmost care to ensure confidentiality and

compliance with data privacy regulations. The potential misuse or unintended disclosure of sensitive health information poses significant ethical challenges.

Hallucinations: Large Language models (LLMs) may exhibit hallucinations, generating outputs that deviate from accurate and contextually relevant information. This tendency for hallucination poses a critical limitation in healthcare applications, where precision and reliability are required. The generation of erroneous or fictional content can lead to severe consequences, impacting clinical decision-making and patient outcomes.

While language models have tremendous potential in healthcare, it is crucial to acknowledge the limitations and address the challenges for responsible deployment in clinical practice. By recognizing and addressing these limitations, we can leverage the power of language models while upholding patient safety, privacy, and equitable healthcare delivery.

1.10 Objective of the thesis

We propose the implementation of language models for temporal Decisions in Health Datasets with the following objectives:-

To implement language models,

1. On COVID-19 genome sequences for predicting strain emergence.
2. On COVID-19 scientific literature to capture and predict emerging scientific evidence.
3. On unstructured clinical notes data from intensive care unit (ICU) for prognosticating critical outcomes
4. And deploy end-to-end language model pipeline and application for raising WASH awareness in COVID-19 pandemic.

1.11 Outline and Structure of the thesis

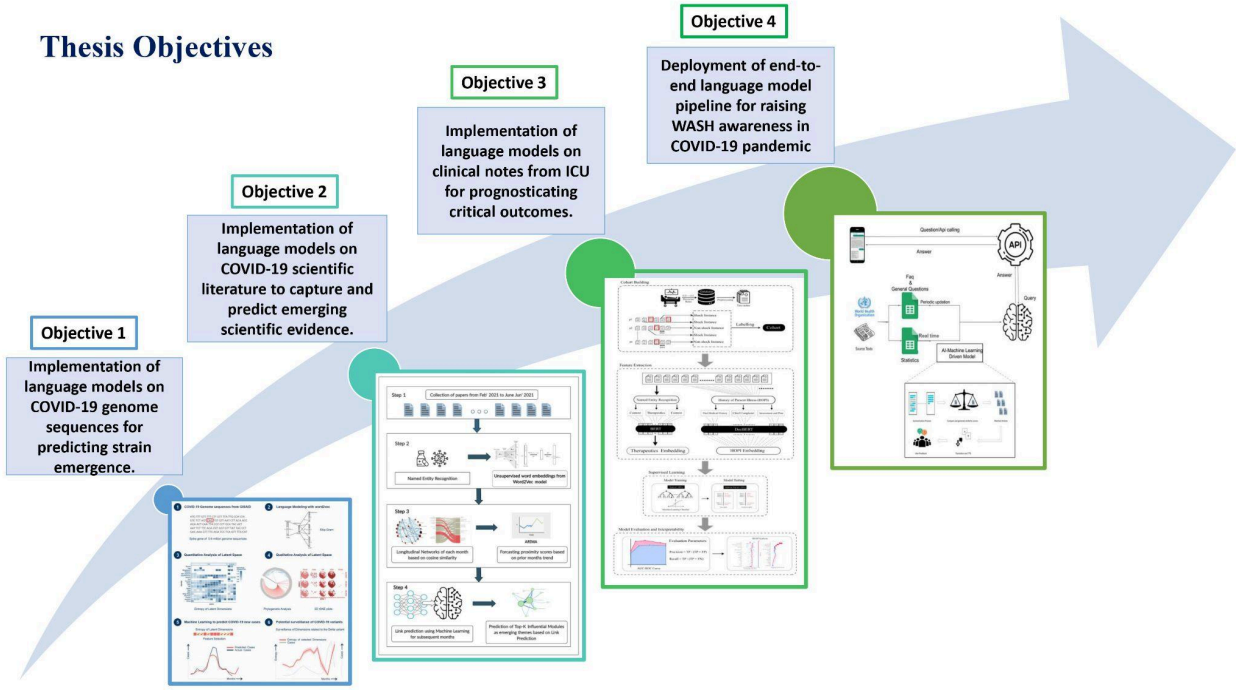


Figure 1.1: ***Objective of the thesis****. The thesis comprises seven distinct chapters, including the introduction chapter, a chapter on the background and related literature, and four chapters on each objective presenting a case study published or under review in peer-reviewed journals employing natural language processing, machine learning, and artificial intelligence to tackle issues related to public health. The common thread in the thesis among chapters is the effective utilization of language models for temporal decision support.*

My dissertation elucidates the scope and applicability of language models in the healthcare domain. This dissertation is a comprehensive study on the implementation of state-of-the-art language models for temporal decision support using healthcare datasets including clinical, public health, and genomic datasets. Specifically, the research enfolded ideas on the development of natural language processing techniques for distilling valuable information from unstructured data sources, such as electronic medical records (EMRs), clinical notes, research articles, genomic sequences, social media data, etc. This thesis involves the development of multiple frameworks leveraging language models integrated with machine learning and deep learning

algorithms. Unique demonstrations of techniques have been presented for pre-processing data to extract meaningful information from textual jargon. The dissertation also covered techniques for fine-tuning language models (LM) on domain-specific tasks. A comprehensive evaluation of the proposed pipelines has been presented, showcasing the challenges and limitations of applying language models in healthcare domains. The primary objective of this dissertation is to establish methodologies for building frameworks that efficiently leverage language models to provide temporal decision support by harnessing information stored in health datasets. Also, the explainability of the proposed frameworks has been discussed, elucidating the process through which the pipelines arrive at decision-making. The dissertation aims to assist healthcare professionals and researchers in making informed decisions and providing practical solutions to challenges faced in clinical and public health settings. Following a comprehensive background on the problem statement and thesis objectives, the paragraph below provides an overview of the remaining thesis's organizational structure and the noteworthy contributions for attaining the stated objectives.

Chapter 2 of this dissertation presents an extensive literature review focusing on language models and their application in health datasets. The first section explains the architecture of language models utilized in my dissertation. Subsequently, the following section highlights the practical implementation of language models across genomic, critical care, and scientific literature data. The literature review aims to establish a comprehensive understanding of the language models in healthcare research, emphasizing their potential impact and significance in diverse healthcare datasets.

Chapter 3 unravels the idea of understanding and deciphering the pandemic progression using language models. The objective of this study was to learn the semantics and grammar of genomic sequences through language models while predicting COVID-19 caseloads 2 months ahead of time. The chapter talks about the implementation of language models for analyzing the emerging SARS-Cov2 spike strains based on the latent space of spike protein-coding nucleotide sequences. In this work, a Word2Vec model was trained on the genomic sequences considering the virus DNA sequences as documents and codons as words. We investigated the role of Sample Entropy in representing mutation sets in viral genomic sequences. Our analysis showed that changes in the word patterns (codons) across the document (sequences), result in changes in the entropy of the underlying dimensional space. Utilizing this insight, we developed a Machine Learning-based approach to extract special dimensions, referred to as Dimension of Concerns (DoCs), which were able to predict the spread of viral strains. The use of DoCs provides a practical approach for country-specific prediction of the emergence and spread of viral strains, making it a valuable tool for genomic surveillance efforts. Our findings demonstrate the potential for using Sample Entropy and DoCs as a means of early detection and monitoring of viral outbreaks.

Chapter 4 explains the utility and implementation of language models in literature mining for temporal decision-making. As the world continued to grapple with the COVID-19 pandemic, it was essential to have access to accurate and updated information from peer-reviewed literature thereby designing effective treatment and responses. A comprehensive computational approach, *EvidenceFlow* was designed to compute multiple aspects such as network topological features, community dynamics, and temporal evolution from literature which aided in the effective navigation of large corpus. To assimilate reliable information from literature, our study was hypothesized which involved tracking the temporal change in unsupervised word embeddings to predict emerging themes relevant to COVID-19 literature, which was one of the first efforts apprehending semantic knowledge. Through this approach, the process of reviewing and synthesizing the literature was made more efficient, thereby providing valuable insights for policymakers and clinical practitioners to design effective treatment strategies.

In chapter 3 and chapter 4, I will discuss the implementation and utilization of language models in the context of public healthcare with COVID-19 pandemic as a use case. Different data-driven approaches were undermined using language models to utilize public datasets such as scientific literature, and genomic datasets for tackling pandemic progression. The following chapter (Chapter-5) of this dissertation discusses the implementation of language models in Intensive Care Units (ICUs), with a focus on predicting critical outcomes. The main hypothesis of the study was to demonstrate the effective utilization of textual data using language models in critical care settings. The study was designed to create an explainable multi-modal early warning system (EWS) for 24-hour ahead prediction of abnormal shock index (SI) in critically ill patients. The models were built by fusing routinely available vital signs and clinical notes collated from Electronic Medical Records (EMR). A total of 3117 features from vitals time-series combined with BERT-based features extracted from clinical notes, were used to train a series of machine learning models. The best multimodal pipeline (*ShockModes*) was then assessed for clinical interpretability using SHAPLEY analysis. The development of explainable multi-modal Early Warning Systems, such as *ShockModes*, holds promise in providing medical practitioners with a reliable and interpretable tool for predicting and monitoring critical care disorders.

Lastly, chapter 6 focuses on real-time deployment of AI-suite for effective dissemination of accurate information to combat infodemic. The COVID-19 pandemic has demonstrated the influence of misinformation on the internet and its impact on global health. To address this challenge *WashKaro*, an AI-powered mobile application was designed which provides accurate COVID-19 information aligned with WHO recommendations. This multifaceted mobile application used an effective combination of advanced technology such as conversational AI, machine translation, and NLP. The primary objective of this study was to assess the utilization of text summarization and machine learning techniques in providing accurate COVID-19 information from the World Health Organization (WHO), with an aim to mitigate the spread of

misinformation. The findings of the study demonstrated the potential of a mHealth platform by highlighting the importance of providing accurate information in a user-friendly and accessible format (local languages), which can serve as a template for other applications in the future.

Chapter-2

Background & Related Work

Language models are one of the most significant advancements in Natural Language Processing (NLP). Their ability to effectively model human language is exemplified by the success of GPT-4 (B. Peng et al. 2023) and ChatGPT (Yiheng Liu et al. 2023), albeit their capability to achieve human-level comprehension is still questionable. A majority of language models learn a compressed representation that maps billions of tokens (e.g. letters, words, sentences) to lower-dimensional vector representations referred to as embeddings. This chapter presents an overview of the most prominent language models used in my dissertation. Furthermore, it provides a comprehensive review of the existing literature on the application of language models in healthcare.

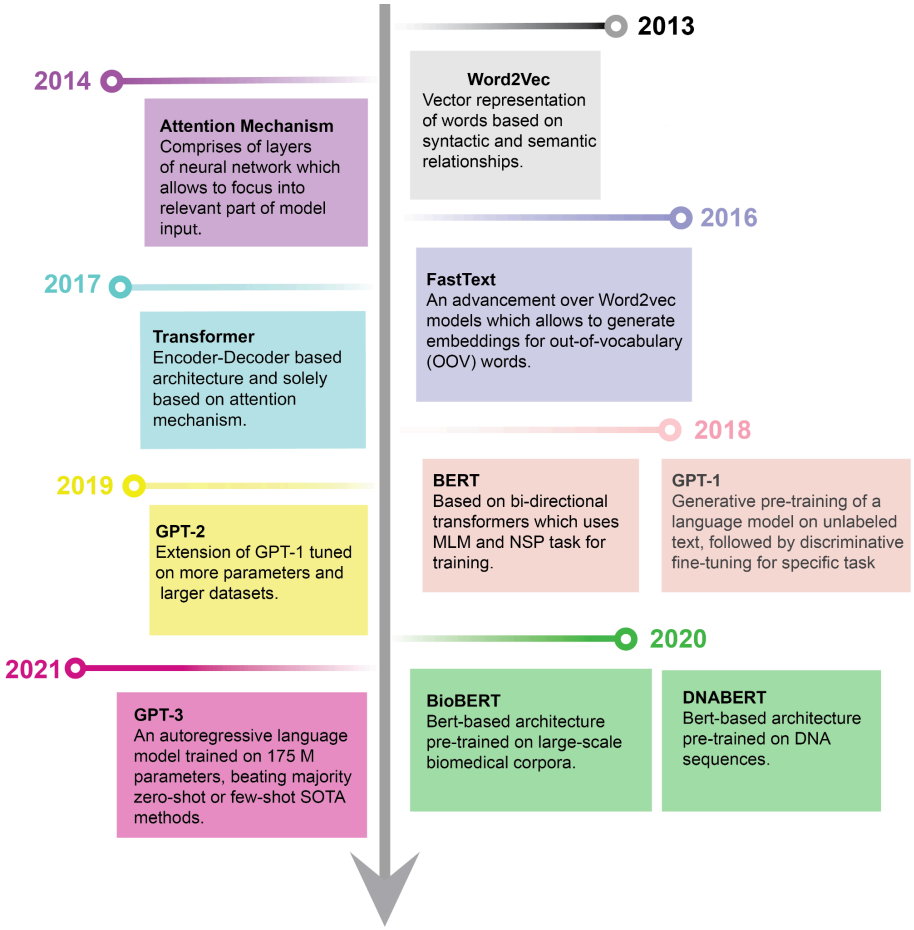


Figure 2.1. ***An overview of important works related to Language Models for Biomedical applications****. The depicted architectures are arranged in ascending order based on their timeline, highlighting their progressive development. The figure provides a*

succinct description of each architecture, showcasing their relevance and contributions to the biomedical domain.

2.1 Word2vec & Doc2vec

The concept of unsupervised learning of word representations was initially proposed in 2013 by Mikolov et al (Mikolov, Sutskever, et al. 2013). It demonstrated the efficacy of distributed word representations by contextualizing the semantics from literature. Word2vec was trained on a large corpus using two algorithms Continuous Bag-of-Words (CBOW) and Skip-gram respectively. The algorithms use two-layered neural networks to train the embedding model, the key distinction in the algorithms being different loss functions. Whereas CBOW algorithm uses the surrounding context to predict the focal word, the skip-gram predicts the surrounding word based on the focus word.

CBOW algorithm utilizes log-linear architecture to learn word vectors by calculating the probability of focus words based on the surrounding context. It maximizes the probability of the target word by minimizing the softmax loss function through gradient descent optimization, thereby estimating the learning parameters. The following equation derives the probability of target words.

$$P(w_f | w_c) = \frac{\left(exp(w_f^T w_c)\right)}{\sum_{i=1}^{V}\left(exp(w_i^T w_c)\right)}, \qquad (2.1)$$

where w_f is the focal word, w_c is the context (one or more words), and V is the size of the vocabulary.

In the simplest scenario, the hidden layers correspond to vector representation of the context word, when the context has only one word. The dot product between vectors of context and target word is calculated for measuring a score. The objective of the model is to maximize the score for focus words which maximizes the likelihood chance of occurrence in the given context. Alternatively,the objective function of the skip-gram model maximizes the probability of the observed context words given the focal word.

$$P(w_c | w_f) = \sum_{c=1}^{C} \frac{\left(exp(w_c^T w_f)\right)}{\sum_{i=1}^{V}\left(exp(w_c^T w_i)\right)}, \qquad (2.2)$$

where w_f is the focal word, w_c is the context (one or more words), and V is the size of the vocabulary.

In case of single word context, the algorithm exhibits similar behavior like CBOW architecture, the hidden layer representing the vector of the focal word. For multi-word context, the model predicts C context words minimizing the model loss or objective function, where model loss is

equivalent to addition of respective context-specific loss functions. There will be C inner-product scores where each score corresponds to individual focal-word context pairs, assigning maximum value to those pairs which have a likable chance of appearance.

Doc2vec and paragraph2vec is an extension of Word2vec to learn vectors for a document or paragraph. It predicts words within the document by considering the entire document. There are two types of Doc2vec: Paragraph Vector-Distributed Memory (PV-DM) and Paragraph Vector-Distributed Bag of Words (PV-DBOW), prior architecture being similar to the skip-gram model and later to the CBOW model respectively. These models are designed to learn high-quality document or paragraph embeddings, which can be useful for a broad spectrum of downstream tasks, such as document classification, text summarization, etc.

2.2 FastText

FastText (Bojanowski et al. 2017) addresses a limitation of popular vector embedding models, including Word2vec and GloVe, by enabling the handling of out-of-vocabulary (OOV) terms. This is achieved through the incorporation of internal sub-word information into the word representation process; the vector for a given word is derived from its morphological components. Specifically, a word is viewed as a combination of character-level n-grams, which enables the sharing of morphological information across words and the construction of vector representations for rare or unseen words in the training corpus. If character-level n-grams cannot form a given word, the model assigns a 0-vector to that word. The learning objective of the model predicts context words given a focal word, similar to the Word2vec Skip-gram (SG) model. However, unlike the SG model, both the vectors for focal words and the context words are represented as the composition of their character-level n-grams in FastText. The objective function is:

$$\sum_{t=1}^{T}\left[\sum_{c\in C_t} l\left(s\left(w_t, w_c\right)\right) + \sum_{n\in N_{t,c}} l\left(-s\left(w_t, n\right)\right)\right], \qquad (2.3)$$

Where w_t is a word in the text and w_c is a context word, $N_{t,c}$ is the set of negative examples sampled from the vocabulary, C_t is the set of context words for words at position, *t* is the position in the text.

And *l* is the logistic loss function defined as:

$$l(x) = log(1 + e^{-x}), \qquad (2.4)$$

And *s* is the scoring function that computes the similarity between a word *w* and a context *c*:

$$s(w, c) = \sum_{g\in G_w} z_g^T v^c, \qquad (2.5)$$

where G_w is the set of character n-grams in the word w and z_g is the vector representation of n-gram g, and v^c is the context vector.

The model employs a logistic loss function to compute similarity between words and context. The objective function is composed of two terms for accounting the context words as both positive and negative examples respectively. Additionally, the FastText model has a hyper-parameter character n-gram size which needs to be mentioned during training. Empirically, researchers found that an n-gram size of 3-6 extracts sufficient information about the sub-word, and produces better results during evaluation.

2.3 Transformers

Significant breakthrough happened in the direction of language models with the publication of the paper titled "Attention is all you need" (Vaswani et al. 2017). In this paper, authors proposed Transformers, a transduction model built upon simple neural network architecture for sequence-to-sequence learning, consisting solely of self-attention layers and removing RNNs. The Transformer model was trained on WMT 2014 English-German dataset with 4.5 million sentence pairs and WMT 2014 English-to-French dataset consisting of 36 million sentences. The autoregressive architecture consisted of an encoder and a decoder, both composed of a stack of six identical layers, each layer containing two sub-layers: a multi-head self-attention mechanism, and a position-wise fully connected feed-forward network. The multi-head self-attention mechanism learned the importance of different words in the input sequence based on their relevance to the current output word. Both encoder and decoder had a fully connected feed-forward network with two consecutive linear transformations in which after the first transformation the activation function used was ReLU. This provided non-linear transformations to the output of the self-attention mechanism. Residual connections were generously utilized at each sub-layer to avoid the vanishing gradient problem, analogous to residual connections in ResNet architecture (Kaiming He et al. 2015). Motivated from Information Retrieval (IR), attention can be mathematically represented as a triple of Query, Key and Value, i.e.,

$$Attention\ (Q,\ K,\ V)\ =\ softmax\ (\frac{QK^T}{\sqrt{d^k}})V\ , \qquad (2.6)$$

$$FFN\ (x)\ =\ max\ (0,\ xW_1,\ b_1)W_2\ +\ b_2\ , \qquad (2.7)$$

Where K, Q, V represents Key, Query and Value matrix respectively in equation (2.6); W_1 & W_2 represents the weight matrix with b_1 and b_2 as bias in equation (2.7).

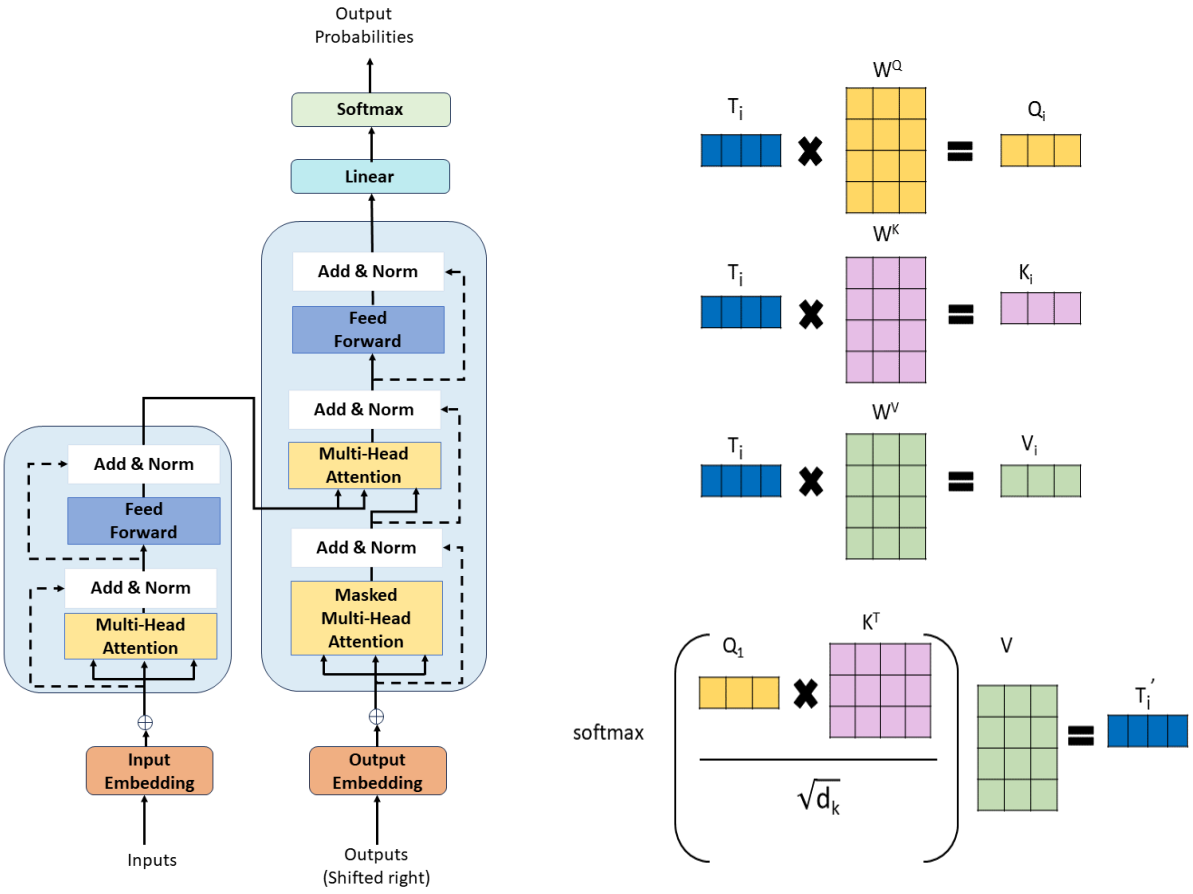


Figure 2.2. ***Illustration of Transformer Architecture and Self-Attention Mechanism.*** *(a) The Transformer model architecture, showcasing its components and flow of information through the encoder-decoder structure. (b) The Key-Query-Value paradigm, representing the self-attention mechanism within Transformers. This paradigm allows the model to effectively process and establish relationships among various elements in the input sequence, empowering it to make contextually informed predictions and generate meaningful information.*

The input to the model is first transformed into an embedding, which is then processed by the encoder layers. The output sequence is produced by decoder in a sequential manner, with each token being generated based on the encoder outputs and the prior output tokens. To add information about the position of the word in the sequence a sinusoidal positional encoding was employed with the advantage of dealing with sequences of variable length. During training, the model is optimized using a modified form of the cross-entropy loss function. The key highlights of this structure are parallelization, lesser training time and improved performance. Hence, the Transformer architecture introduced a novel way of modeling the sequence of tokens by relying entirely on self-attention mechanisms.

2.4 BERT & DocBERT

BERT, or Bidirectional Encoder Representations from Transformers (Devlin et al. 2018) was built on top of the Transformer architecture to solve the limitations of the traditional language models that process text input only in a left-to-right or right-to-left direction. BERT marked a significant departure from traditional word embedding models, as it was the first to use a Transformer-based architecture for learning word representations. BERT is a language model that comprises a multi-layer bidirectional transformer encoder. It stands out from other models due to its bidirectional pre-training, which involves training on the unsupervised tasks of masked language modeling and next sentence prediction. To make BERT suitable for various tasks, a task-specific output layer can be added to it for fine-tuning. BERT uses the *WordPiece* tokenizer, which employs an iterative approach to update the vocabulary with the most frequent combinations of existing words. The pre-training process involves masked language modeling and binarized next sentence prediction tasks, and the model is trained bidirectionally to avoid the problem of each word "seeing itself." The masked language modeling task involves randomly masking 15% of the tokens with the "[MASK]" token, and the model is then trained to predict these masked tokens. BERT has two pre-trained model sizes: BERT-Base, which comprises 12 transformer blocks, 12 attention heads, and 110 million parameters, and BERT-Large, which has 24 transformer blocks, 16 attention heads, and 340 million parameters. This approach allowed BERT to capture bidirectional relationships between words, leading to representations that were more context-aware. BERT has since been shown to perform well on a wide range of NLP tasks, becoming the state-of-the-art in many cases.

2.5 Language Models for Genomics: Representing DNA Sequences

One of the fundamental techniques in computational methods has been to analyze biological sequences. A biological sequence is a single, continuous molecule of nucleotides or amino-acids. Genomic sequence is a type of biological sequence which consists of a long sequence of nucleotides consisting of four characters; A (adenine), T (thymine), C (cytosine), and G (guanine). The grammar of a genome sequence refers to the arrangement of nucleotides and the patterns of their interactions, such as the formation of secondary structures, the presence of repeats, and the occurrence of mutations. Understanding the grammar of genomic sequences can be crucial across various scenarios, such as identifying genetic variations, predicting gene expression, predicting structural and functional properties of a gene and studying evolutionary relationships between species. Biological sequence alignment provides necessary information related to common evolutionary descent or common structural function. It helps in identification of similar sequences, generation of phylogenetic trees, and construction of homology models of protein. Several NLP tools and techniques have been developed to understand and analyze large chunks of genomic sequences and language models are one among those tools. It aids in deciphering the biological grammar of genomic sequences by identifying the patterns and relationships among the aligned sequences. Overall, the ability of language models to

comprehend and extract meaningful information from genomic sequences makes them a valuable tool in understanding the genetic basis of diseases and developing personalized treatments.

Analyzing biological sequences involves implementation of computational, mathematical and statistical algorithms to biological data for extracting meaningful information about their structure, function and evolutionary history. These analyses provide insights into gene expression, protein structure and function, genetic variation, and evolutionary relationships among species. Traditionally, mathematical formulae for consensus sequence, weight matrices, and rank alignment were employed to interpret the role of nucleotide sequences. The GC content, bendability, flexibility, and free energy are some of the meaningful descriptors used by several groups to describe the nucleotide sequence (Łozinski et al. 1989). Previously, the computational methods subjected for biological sequence analysis relied heavily on the frequency of k-mers (short subsequences of nucleic-acids), which is limited in capturing the semantic relationships (J.-H. Choi and Cho 2002; Bussi, Kapon, and Reich 2021). Existing Deep learning architecture such as convolutional neural networks (CNNs), have limitations to capture long-range semantic dependency (Tang et al. 2018). These models can contextualize local features based on the filter size. To capture long-range relationships in sequences, recurrent neural network (RNN)-based models, including long short-term memory (LSTM) and gated recurrent units (GRU), were developed (Butte 2001).
However, due to their low level of parallelization, these models could not execute large-scale learning. Given the abundance of unlabeled genomic sequences, Language models were trained in an unsupervised learning paradigm to find latent patterns that could be fine tuned to specific outcomes. Pre-trained transformer-based language models were developed and have emerged as a promising approach for DNA sequence analysis (Ji et al. 2021). These models have shown significant improvement over traditional and deep learning methods and have garnered increasing attention in the field of bioinformatics. The use of self-attention mechanism enables global and transferrable representations of genomic DNA sequences by capturing important dependencies and relationships across long genomic sequences in both directions. This allows robust transfer of semantic knowledge gained from one genomic sequence to another, reducing the amount of data needed to train new models while accelerating the pace of genomic research. In the subsequent paragraphs, I explain studies where pre-trained language models have been implemented for sequence analysis.

Gene2Vec (Du et al. 2019) explored the concept of gene embedding, which is a distributed representation of genes similar to word embedding. The authors trained a 200-dimensional vector representation of all human genes utilizing gene co-expression patterns in 984 data sets from the GEO databases, in a data-driven approach. t-SNE was performed to produce a gene co-expression map that showed local concentrations of tissue-specific genes. Also, the embedded gene vectors were found to be useful in predicting gene-gene interactions based solely on gene names. On similar lines, Patrick Ng created Dna2vec by training DNA sequences (Ng 2017). He

divided DNA sequences into variable length k-mers and trained the Word2Vec model for distributed representations of variable-length k-mers. The experiments conducted in this paper show that the summing of Dna2vec vectors is similar to nucleotides concatenation. Additionally, there is a correlation between cosine similarity of Dna2vec vectors and Needleman-Wunsch similarity measure. Protein sequence analysis is another subset of sequence analysis, which allows establishing relationships between protein sequences and their spatial structure. Also, it can serve as a theoretical foundation for future research on protein structure and function. A language model called ProtVec was trained to represent and mine features from protein sequences (amino-acid sequences) (Asgari and Mofrad 2015). The authors of this study demonstrated the effectiveness of protein vector representation in a variety of bioinformatics applications which includes assigning family, visualization, predicting structure, and interaction among themselves. Various studies have built hybrid architecture fusing language models and deep learning architecture. DeepARC (L. Cao et al. 2022), a deep learning-based architecture, fused convolutional neural network (CNN) and recurrent neural network (RNN) to effectively predict transcription factor binding sites (TFBS) in DNA sequences. It works by feeding the positional embedding obtained from Dna2Vec into a CNN-BiLSTM-Attention-based framework to predict the TFBS. Incorporating an attention mechanism in the framework enables the architecture to selectively attend to crucial segments of the sequence, facilitating enhanced access to valuable information about the motif. The positional embedding method used in DeepARC outperforms the traditional OneHot encoding and Dna2Vec methods, demonstrating the effectiveness of the approach. Another framework named Motif oriented DNA (MoDNA) (An et al. 2022), to learn semantic level genome representations from vast amounts of unlabeled genomic data, was pre-trained on human genome data using self-supervised learning. It was further fine-tuned for different downstream tasks, such as promoter prediction and transcription factor binding site prediction. The self-learning approach allowed the model to learn useful features and patterns from the data without being limited by the availability or quality of labels. It was found to be more computationally efficient than previous methods, making it more scalable and accessible.

In a recent study, a neural language model (Hie et al. 2021, 2020) was trained to acquire knowledge of both semantics and grammar of viral protein sequences, with an aim to predict mutations that may lead to viral escape mutations that may lead to viral escape. Independent unsupervised language models were trained for three viruses which includes influenza A hemagglutinin, HIV-1 envelope glycoprotein, and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) spike glycoprotein. The model trained on semantic landscapes for each virus which were used to predict viral escape mutations. These mutations result in sequences which conserves grammatical and syntactical correctness, while exhibiting semantic dissimilarity compared to the original virus sequence, allowing them to evade the immune system. Another study (Bepler and Berger 2019) has been conducted to infer protein's structural properties from its amino acid sequence based on sequence embeddings. The study introduced a framework to

assign any protein sequence to sequence of vectors, where each vector corresponds to an amino acid position. The approach outperformed other classical sequence-based methods when predicting structural similarity and improved the performance of transmembrane domain prediction.

Owing to the popularity of transformers, language models integrated with these architectures have shown greater promise in learning the grammar of sequence data. They have achieved enhanced performance across various tasks related to sequence analysis. Prior studies have documented numerous instances of employing transformer-based models in sequence analysis, with a subset of such studies being highlighted in this paragraph. Among them, DNABERT, trained on DNA sequences using the BERT architecture, achieved exceptional performance on diverse downstream tasks, establishing new benchmarks (Ji et al. 2021). By training on large genomic datasets, DNABERT can capture the grammar of DNA sequences and identify patterns that are important for understanding biological processes. The key innovation in DNA-BERT is the design of a new tokenization scheme that transforms DNA sequence into k-mers representation which can be processed by BERT. DNABERT tokenized sequences into kmers, which were fed as an input to the BERT architecture. It accomplished advanced results in promoter prediction and identification of transcription factor binding sites.

An alternative approach entails leveraging a multi-language BERT model to transform DNA sequences into a fixed-size numerical matrix, enabling the prediction of enhancers (N. Q. K. Le et al. 2021). This methodology exhibited superior performance compared to other advanced bioinformatics tools, with improvements observed in evaluation metrics by 5-10%. A Nucleotide Transformer model was developed using a comprehensive dataset of 3,202 diverse human genomes and 850 genomes from various species, to accurately predict molecular phenotype in low-data settings (Dalla-Torre et al. 2023). BERT-m7G (L. Zhang et al. 2021), a transformer model based on BERT, was designed for determining RNA N7-methylguanosine (m7G) sites (RNA post-transcriptional modification) using RNA sequence information. It works by taking RNA sequences as input and passing them through the pre-trained transformer models. The output of the models is then fed into a stacking ensemble layer that combines the predictions of the base models to produce a final prediction for each input sequence. The BERT-m7G model performed better than existing prediction methods for identifying m7G sites in RNA sequences. The accuracy of the model was increased by 3-20.7%, and the MCC (Matthews correlation coefficient) was improved by 0.06-0.415. This demonstrates the potential of deep learning models like BERT-m7G to automate and improve the accuracy of RNA sequence analysis.

Transformer-models have also been extended to protein sequences and proteomics. BERT4Bitter, another deep learning based model, was designed for prediction of bitter peptides from their original amino acid sequences without requiring any structural information of protein (Charoenkwan et al. 2021). It employs feature encoding to extract meaningful features from the

amino acid sequence, including dipeptide composition and physicochemical properties, for predicting bitter peptides and thereby accelerating the discovery of new bitter compounds in the food and pharmaceutical industries. Elnaggar et al. (Elnaggar et al. 2022) developed a transfer learning technique that employed transformer-based language models like Transformer-XL, XLNet, BERT, ALBERT, ELECTRA, and T5 to capture constraints relevant to protein structure and function. They found that ProtT5 refined on UniRef50 without MSA outperformed ESM-1b. Researchers developed ProteinBERT (Brandes et al. 2022) to capture both local and global representations of proteins, by blending language modeling techniques with gene ontology annotation prediction. Another pre-trained BERT model called EpiBERTope (Park et al. 2022) was developed to predict both linear and structural epitopes. It builds global dependencies for each amino acid in protein sequences via a multi-headed attention technique. These models have produced encouraging outcomes when used to predict protein shapes, among other tasks linked to proteins. The use of Transformer-based pre-trained models has also been extended to genome analysis for predicting interactions between regulatory elements, as seen in GeneBERT (Mo et al. 2021). Traditional methods typically fail to consider the interconnection among multiple regulatory elements in the regulatory genome. To address this issue, GeneBERT was pre-trained in a self-supervised and multi-modal manner using large-scale genomic data. The pre-training involved three tasks: sequence pre-training, region pre-training, and sequence-region matching, aiming to advance the model's generalization ability and robustness.

2.6 Language Models for Evidence Mining: Representing Scientific Literature

Scientific literature such as research papers, blogs, articles play a critical role in natural language processing (NLP). It provides a vast amount of text data that can be used for training and evaluating NLP models, which can then be applied to various scientific domains. Moreover, scientific literature contains specialized language and technical terminology that presents unique challenges for NLP models, and thus, it serves as a valuable testbed for advancing the state-of-the-art in NLP. Additionally, scientific literature offers opportunities to develop NLP applications for scientific research, such as text mining, information extraction, and automatic summarization, etc. By leveraging the wealth of information contained in scientific literature, NLP models can assist researchers in discovering new knowledge, identifying trends, and making discoveries in various fields, including biomedicine, chemistry, and engineering, and more. Therefore, the ability to process scientific literature effectively is essential for advancing scientific research and accelerating the pace of scientific discovery. Briefly, I have discussed some studies where Transformer-based language models have been trained or fine-tuned in the scientific literature. One such development has been SciBERT (Beltagy, Lo, and Cohan 2019), a transformer-based language model that is pre-trained on a large corpus of scientific text. The architecture was trained from scratch on 1.14 M papers from semantic scholars out of which 18 percent research articles from the computer science domain and 82 percent from the biomedical literature. This led to improvement in downstream tasks such as sequence tagging, sentence

classification and dependency parsing with respect to the scientific domain. In a similar context, PubMedBERT (Gu et al. 2020) was trained from scratch on a large corpus of biomedical literature. In this work, authors have highlighted the advantage of training models from scratch in contrast to continuous pretraining. Differences between mixed-domain pre-training and domain specific pre-training has been demonstrated in reference to downstream tasks such as question answering, document classification, sentence similarity and others. BioBERT (J. Lee et al. 2020) was designed on similar lines, where the architecture was trained on large-scale biomedical corpora for domain specific language models. It outperformed previous state-of-the-art models and BERT in various biomedical text mining tasks such as biomedical named entity recognition, biomedical relation extraction and biomedical question answering. Analysis shows that pre-training BERT on biomedical corpora enhances its ability to understand complex biomedical texts. The Biomedical Language Understanding Evaluation (BLUE) benchmark was introduced to facilitate advanced research on the biomedical domain (Y. Peng, Yan, and Lu 2019). This benchmark consisted of ten datasets covering both biomedical and clinical texts with varying levels of difficulty across five tasks. The efficacy of pre-trained language representations was examined by evaluating previous baseline models based on BERT and ELMo on this dataset. The evaluation found that the BERT model pre-trained on PubMed abstracts and MIMIC-III clinical notes achieved superior results compared to the baseline models. Hence, this benchmark serves as a valuable dataset for developing and evaluating pre-trained language representations in the biomedical domain.

2.7 Language Models for Clinical Text: Representing Clinical Notes

In recent years, the application of Natural Language Processing (NLP) has become increasingly important in the field of critical care. It distills relevant information from medical records quickly and efficiently, thereby assisting resource-starved health sectors. In healthcare, large amounts of information are generated every day in the form of electronic health records (EHRs), medical notes, discharge summaries, and research articles, among others. The information in these textual sources are valuable for decision-making, research, and improved patient care. For example, NLP-based algorithms have been developed to extract patient data from electronic health records (EHRs) and to identify patients at risk for sepsis, shock, and many other life-threatening complications that can occur in critically ill patients.

However, the implementation of NLP algorithms in clinical settings is not without its challenges. Building NLP driven models that can accurately process clinical text requires a close collaboration between clinicians, who have expertise in medical jargon, and data scientists, who have expertise in NLP techniques. It requires a human-in-loop intervention for better processing of data into relevant information. One of the main challenges for implementing NLP in critical care requires for experts collaboration in creating gold standard training corpora, which can be both expensive and time-consuming. These gold standard corpora can be treated as benchmark

datasets for modeling predictive outcomes. Additionally, current tools lack provisions for domain experts to inspect the outcomes algorithms and incorporate corrections which would enhance these results. This lack of user-centered development has been identified as one of the barriers to NLP adoption in the clinical domain. While there are significant challenges in building NLP models that can accurately process clinical text, the existing literature has shown the potential implementation of NLP in critical care, particularly in the areas of patient monitoring, diagnosis, and treatment planning. Future investigations need to be conducted to tackle the obstacles associated with the implementation of natural language processing (NLP) in the clinical realm, as well as to devise NLP tools that prioritize the preferences of end-users, facilitating seamless integration into clinical practice. Notably, the widespread integration of electronic health records (EHRs) in healthcare has led to the vast accumulation of patient data, including free-text clinical notes. These contain valuable demographic and clinical information such as medical history, treatment, and progress. Processing such vast resources of data helps healthcare organizations to improve patient outcomes through advanced data science and analytics. Healthcare organizations are increasingly adopting Machine Learning and Artificial Intelligence (AI/ML) for task automation, predictive modeling, and knowledge discovery, gleaning relevant information from textual data remains to be a riddle. The real challenge lies in representing the unstructured and domain-specific language of free-text clinical notes into a format suitable for ML modeling. Language models aid in converting clinical text to vector representation which can be exploited as training features for machine learning and deep learning architecture.

The procedure for training one's own word embedding model requires learning to a specific pre-processed corpus, and tuning the model hyper-parameters based on empirical observations. Experiments have been conducted to investigate the impact of dataset size and domain on the performance of models across a range of tasks. For example, Zhu et al. (Y. Zhu, Yan, and Wang 2017) conducted experiments using medical data derived from PubMed abstracts, specifically focusing on tasks related to semantic relatedness and similarity. Their findings demonstrated a significant improvement in performance until the vocabulary size reached 4 million distinct words. However, other researchers sought to investigate the significance of domain-specific datasets. Zhao et al. (Zhao, Masino, and Yang 2018) showed that architecture trained on a smaller domain-specific clinical dataset outperformed a larger generic domain dataset in tasks such as relatedness assessment, neighborhood coherence, outlier detection, and drug name recognition. Similarly, Wang et al. (Y. Wang et al. 2018) examined the impact of embeddings learnt on biomedical corpora (including EMR data from the Mayo Clinic and MedLit articles from PubMed Central) compared to generic corpora (such as Google News, Wikipedia) for semantic relatedness. Their study revealed that embeddings derived from clinical notes demonstrated a higher degree of agreement with human assessments of word similarity and showed greater efficacy in predicting fractures when compared to embeddings trained on other datasets. However, the results were less robust for other downstream tasks specific to the

biomedical domain, such as predicting smoking patients, extracting drug interactions and information retrieval.

Pre-trained word embeddings might be useful in some cases, yet training on domain-specific text can improve performance in many tasks. Examples of such datasets across healthcare domains include MIMIC-III, i2b2, etc. MIMIC-III contains approximately 40,000 de-identified patient data, consisting of clinical texts, nursing notes, lab reports and ICD-9 codes;while the i2 center's data was in the form of clinical notes and discharge summaries. Existence of these freely available large clinical and healthcare datasets have often assisted researchers in training their own embedding attributing to specific tasks. Notably, pre-trained clinical embeddings have been generated by training various publicly-available resources such as PubMed abstracts, PMC full-text documents, Wikipedia, and a combination of these sources. For instance, Glove architecture was trained on 27 million anonymized notes collected from 1.2 million patients during 49 million visits (Dubois et al. 2017). Other research groups have trained Word2vec on MedHelp online forums, PubMed text, Wikipedia or on concatenated datasets of i2b2, MedTrack, and CLEF's 2013 to create domain-specific embeddings for healthcare (J. Huang, Xu, and Vydiswaran 2016).

Often there is lack of clinical knowledge in word embeddings extracted from clinical data, which is essential for medical applications. Domain specific word embeddings can be incorporated with clinical knowledge from ULMS database, ICD-9, ICD-10 codes, and MeSH thesaurus during the time of training. It refines word embeddings to improve their representation of clinical knowledge. Boag et al. added clinical knowledge from UMLS features into the word embeddings trained on MIMIC dataset (Boag and Kané 2017). Patel et al. (K. Patel et al. 2017) added medical information from medical claim datasets and ICD-9 code to the pre-trained word embeddings for automated medical coding review. Yu et al. (Yu et al. 2016) used the clinical concepts of MeSH taxonomy for fitting clinical knowledge into the pre-trained embeddings using Faruqui et al. 's (Faruqui et al. 2014) approach. Zhao et al. (Zhao, Masino, and Yang 2018) extracted data from PubMed and DrugBank to generate word embeddings for drug entities, while tuning the hyperparameters of Word2vec model based on their correlation with word frequency. Some researchers represented medical concepts from UMLS and ICD-9 codes in low dimensional vectors, using medical claim datasets (Y. Choi, Chiu, and Sontag 2016). Cui2vec model was developed by training GloVe and Word2Vec embeddings on the co-occurrence matrix of UMLS identifiers. Benchmark techniques were proposed to assess the embedding similarity for relationships related to comorbidity, drug conditions, etc. These techniques depicted how clinical knowledge was represented in word embeddings. Although in these studies, the context behind these concepts were neglected, instead it only focused on relationships between clinical concepts. The study performed by Mencia and group (Mencía, de Melo, and Nam 2016) involved the implementation of an all-in-text method to learn embeddings of labels and documents jointly, utilizing the manual labels for the document. Zhu et al. (H. Zhu, Paschalidis,

and Tahmasebi 2018) presented an approach to extract clinical concepts using contextual word embeddings. The downstream performance of embeddings trained on medical text was better than that of embeddings trained on bigger, more general corpora, such Google News. Incorporating clinical knowledge into word embeddings improved their representation of clinical knowledge, which is essential for medical applications.

Prior studies in literature presented extrinsic evaluation of clinical word embeddings based on specific tasks or applications. The studies illustrated the effectiveness of word embeddings for automating tasks related to predictive outcomes. Textual data in clinical and healthcare space can be effectively represented as matrices which act as an input to the machine learning classifier across different tasks. Some common tasks where such embeddings can be used for extrinsic evaluation include identifying clinical concepts from text, predicting unplanned readmission after discharge, ICD code prediction, phenotyping, and various other predictive tasks. Usage of publicly available datasets allowed for reproducible and effective comparisons between different systems. Also, researchers could use the same dataset and compare the performance of different machine learning models that used word embeddings as a feature. The study conducted by Wang et al. (Y. Wang et al. 2018) investigated the impact of diverse corpora on the tasks of fracture detection and smoking status prediction. The i2b2's 2006 dataset was utilized as the benchmark dataset for evaluating the performance of the models. Patient phenotypes have been predicted from the "2008 Obesity Challenge" dataset (Dubois et al. 2017); some other groups made efforts to extract clinical information from the “2010 i2b2” dataset (Kholghi et al. 2016; De Vine et al. 2015). Word embeddings have been implied to retrieve information from clinical text in the form of medical entity extraction, adverse drug reaction and so on. An ELMO model (H. Zhu, Paschalidis, and Tahmasebi 2018) combined with RNN architecture and conditional random field (CRF) was trained to perform named entity recognition (NER), achieving an improvement of 3.4% in the score on the i2b2 dataset compared to other state-of-the-architecture (SOTA) architectures. Several studies explored the efficacy of different types of recurrent neural networks (RNNs) for drug name classification. RNN architectures built to classify drug names considering word embedding as learning parameters, performed better than the existing models (Zhao, Masino, and Yang 2018). Researchers depicted that ClinicalBERT performed better than prior models, for predicting hospital readmission within 30 days from both discharge summaries and early clinical notes (K. Huang, Altosaar, and Ranganath 2019). Other groups (Alsentzer, Murphy, Boag, Weng, Jindi, et al. 2019) evaluated performance of ClinicalBERT on three common natural language processing (NLP) tasks (i) named entity recognition (NER), (ii) de-identification, and (iii) natural language inference (NLI). The results depicted that ClinicalBERT performed better than BERT-base and BioBERT on NER and NLI tasks. However, there was a drop in the performance for de-identification tasks due to the de-identified nature of the MIMIC dataset. Finally, Beltagy et al. (Beltagy, Lo, and Cohan 2019) assessed SciBERT on a range of NLP tasks and found that it often outperformed BERT-Base without requiring fine-tuning or task-specific architectures. Si et al. (Si et al. 2019) conducted a study to compare

traditional word embeddings against contextualized embeddings for identifying clinical concepts from text. The inference suggested that BERT-based architecture, learnt on MIMIC-III data, further fine-tuned by appending an extra Bi-LSTM layer to base architecture. After a specific number of iterations, the performance of the model for extracting clinical concepts decreased due to overfitting of the model. Additionally, the quality of low dimensional word representation generated from EMRs can also be assessed by predicting unplanned readmission following patient discharge. This can be achieved using convolutional neural networks (Craig, Arias, and Gillman 2017; Nguyen et al. 2017) or dynamic memory models (Pham et al. 2016). ICD code prediction is another common task where implementation of word embeddings have been studied. Studies showed that adapting word embeddings (trained on generic corpus) to the clinical space enhances performance for ICD prediction (K. Patel et al. 2017). Additionally, patient phenotyping can be accomplished to predict various conditions such as advanced cancer, depression using word embeddings (Gehrmann et al. 2018). The results for the prediction were promising as it achieved a F1-scores ranging in the mid-80s. These approaches can smoothen the workflow of medical practitioners by recommending proper ICD codes during decision making. Notably, these studies demonstrated the effectiveness of word embeddings for predictive tasks by extracting relevant semantic information from clinical text with lower computational expense.

In conclusion, the development of word embedding models has been a significant milestone in the field of NLP. Word2vec, FastText, GloVe, and BERT are some of the most prominent architectures which shaped the direction of NLP research, and made important contributions to the field. In this chapter, a comprehensive literature survey was conducted to understand the implementation of language models in health datasets, and examine the effectiveness of such implementation across wide range downstream tasks. It has revolutionized the field of natural language processing, enabling researchers and healthcare practitioners to extract meaningful insights from vast amounts of clinical text, genomic data, and scientific literature. The implementation of language models has significantly improved the accuracy and efficiency of various health-related downstream tasks, such as predicting disease outcomes, patient risk stratification, clinical trial matching and adverse event detection. Despite advancements in implementation language models for healthcare datasets, there are still gaps in understanding the optimal use of these models for temporal decision support. Additionally, their prospect has not been properly explored in various tasks which include identifying emerging scientific themes, predicting critical outcomes (such as abnormal shock index), predicting pandemic progression, etc. The subsequent chapters of this dissertation aim to investigate the potential of language models in these areas to advance medical decision-making by tracking, analyzing and monitoring patient outcomes.

Chapter-3

Language models for understanding and predicting pandemic progression: Genomic Surveillance of COVID-19 Variants with Language Models and Machine Learning

3.1 Introduction

The COVID-19 pandemic has brought unprecedented challenges to the scientific community. Consortia across the globe were formed for the advancement of research related to COVID-19. Eventually, this led to the launch of numerous open-source datasets, including genomic sequences, scientific literature, and social media data. The abundance of open-source data sources have provided researchers with valuable resources to advance our understanding about the pandemic and develop effective strategies to mitigate its impact. In this chapter of my dissertation, I have explored how genomic sequences can be utilized by language models to make temporal decisions on pandemic progression. Previously, genomic datasets have been extensively analyzed using machine learning models to understand the virus's transmission and mutation patterns. In this chapter, we have presented the use of language models to investigate the latent space of the spike gene region of SARS-CoV-2, enabling early prediction of caseloads. Our approach accurately anticipated the increase in cases caused by emerging variants.

New variants of SARS-CoV-2 continue to rage across the globe causing devastating waves of the pandemic. Such waves may continue to occur and many lives can be saved through early preparedness. COVID-19 is reported to have claimed 5.45 million lives as of Jan 10, 2022 ("WHO Coronavirus (COVID-19) Dashboard," 2023.). A large number of these deaths are attributed to unexpected surges in infections caused by new strains with higher pathogenicity such as the Delta variant of SARS-CoV-2, prompting international health organizations such as the CDC and WHO to declare these as variants of concern (CDC 2023). The most recent surge of Omicron across the globe with its potential for escaping immunity has seriously undermined the efficacy of global vaccination programs. Most studies around the globe have focussed on forecasting case time series using traditionally reported administrative data. Standard epidemiological approaches such as compartmental and agent-based modeling have been used extensively for forecasting COVID-19 caseloads (P. Arora, Kumar, and Panigrahi 2020). Additionally, numerous studies have used time series analysis, social media mining and multimodal approaches have been utilized for case predictions (Ayan et al. 2021; Kapoor et al. 2020; Melin et al. 2020; Qin et al. 2020; "Modeling COVID-19 Scenarios for the United States" 2020; Rodríguez et al. 2021). Earlier, initiatives such as *Nextstrain (Hadfield et al. 2018)* have focused on providing high-quality tracking information for the strains and lineages as these emerge without forecasting or predictions. Hence early prediction of caseloads and emerging variants through genomic signals remains an open challenge for COVID-19.

Unsupervised embeddings have been shown to capture highly nonlinear and contextual relationships (Mikolov, Sutskever, et al. 2013). Biological sequences contain a plethora of information that can be exploited for genomic surveillance. However, there is a paucity of studies that explore the use of unsupervised embeddings for machine learning based prediction of surges in infections. In these models, codons (tri-nucleotides, 3-mers) translations represent a natural basis for word representations and have been utilized in the past for learning embedding models for modeling various outcomes such as mutation susceptibility and gene sequence correlations (Yilmaz 2020; F. Wu et al. 2021). Recently, Hie et. al used machine learning along with word embedding techniques to model the semantics and grammar of amino acids corresponding to antigenic change to predict the mutations which might lead to viral escape (Hie et al. 2021). Similarly, Maher et. al predicted emerging mutations of SARS-CoV-2 variants and evaluated biological and neural network based predictors of emerging mutations (Maher et al. 2022). Here, we propose *Strainflow (*Figure 3.1*)*, a prospectively validated pipeline with prediction and prospective validation of caseload two months ahead of time. Our empirical experiments demonstrate interpretable features based on Entropy of the latent space of SARS-CoV-2 spike region, thus aiding an early warning system for emergence of new variants of concern and caseload.

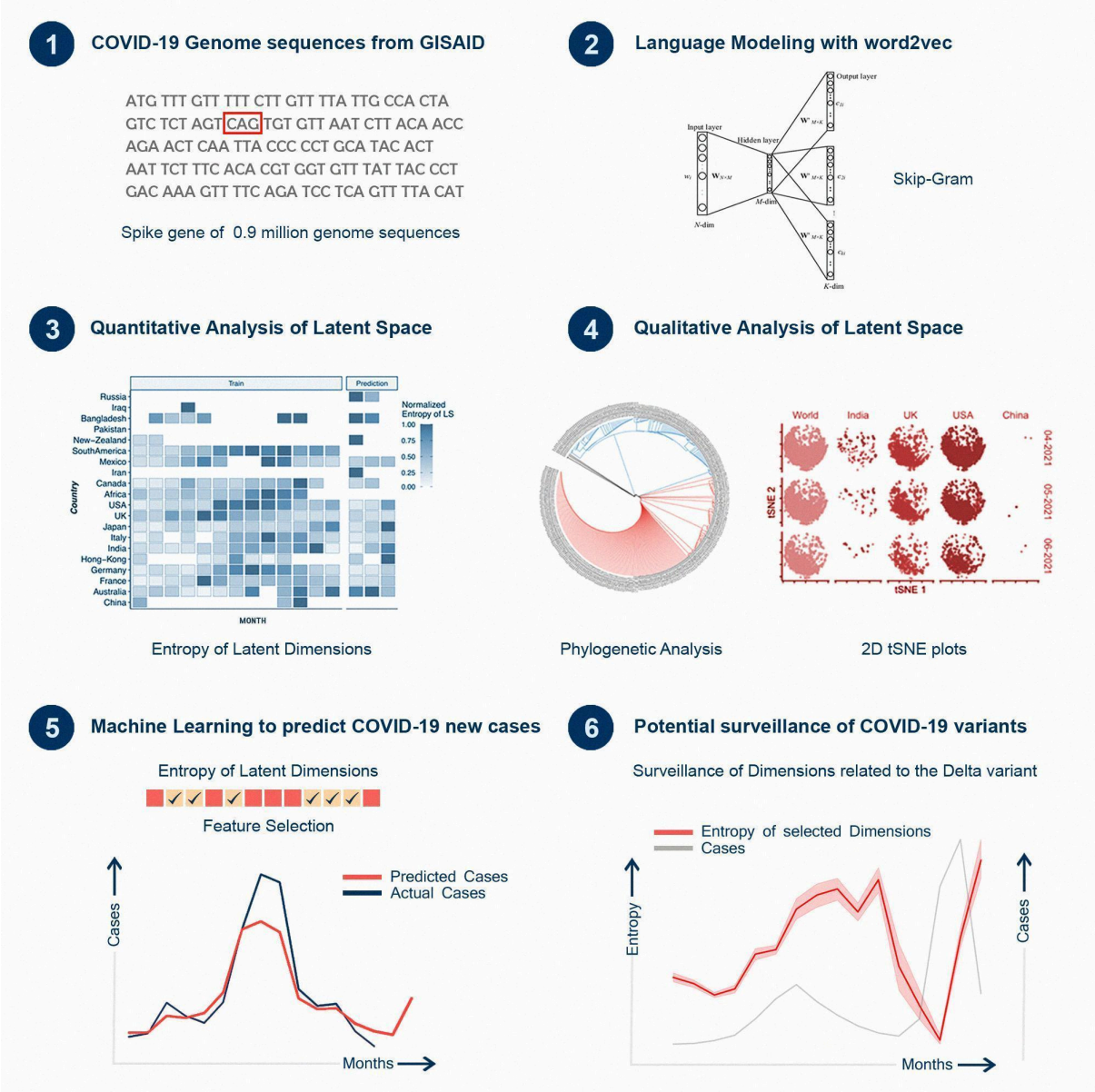


Figure 3.1. *Architecture of the Strainflow pipeline*.

3.2 Methods

3.2.1 Datasets

Training dataset: The dataset was downloaded from GISAID EpiCoV (April 8, 2021 release) (Shu and McCauley 2017). 0·36 million genome sequences (December, 2019 - June, 2021) with high nucleotide completeness, coverage, complete temporal information, and presence of less than 5% non-identified nucleotide bases (N) were downloaded. The sequences included 63 countries, including India, United Kingdom, USA, Australia, New Zealand, Germany, Russia, Italy, France, Mexico, Canada, China, Japan, Pakistan, Bangladesh, Iran, Iraq, the continent of South America, and Africa. Duplicate samples were removed, and whole genome sequences were parsed using CoV-Seq to extract nucleotide sequences corresponding to each of the 12

Coding DNA Sequences (B. Liu et al. 2020). Accession IDs that did not cover 12 coding regions were discarded, yielding 0·31 million high-quality SARS-CoV-2 genome sequences for language modeling. The spike gene region of each sequence was filtered and used for all subsequent analysis. We downloaded country-wise COVID-19 data for new cases from a publicly available repository maintained by Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE).

Evaluation dataset: We downloaded around 0·6 million genome sequences submitted to GISAID from April 2021 to June 2021. We used our trained model to predict the latent representations for these sequences.

3.2.2 Word Embeddings in Strainflow pipeline

In our Strainflow pipeline, we have adopted a word2vec model (Mikolov, Sutskever, et al. 2013). Low dimensional representations for the genome sequences were learned using the word2vec model. Non-overlapping sequences of 3-mers (codons) were considered as words for training the model, which was implemented in Gensim (Řehůřek and Sojka 2010). The skip-gram algorithm was used, with a fixed window size of twenty and vector size of thirty-six. For generating a consensus embedding for a particular strain, genomic sequences were represented by taking the mean of each codon occurring in the sequence dimension-wise. The mean was calculated by summing across all the k-mers over each dimension and then dividing it by the total number of codons present in the sequence. For selecting the dimension size for our word embeddings, we calculated the PIP (Pairwise Inner Product) loss (Yin and Shen 2018). PIP loss is a metric used for calculating the dissimilarity between two word embedding matrices. For the embedding matrix of strains (E), the PIP matrix is defined as the dot product of the embedding matrix with its transpose ($E.E^T$). The PIP loss between two embedding matrices is defined as the norm of the difference between their PIP matrices.

$$||PIP(E1) - PIP(E2)|| = \left|\left|E_1E_1^T - E_2E_2^T\right|\right| = \sqrt{\sum_{i,j}\left((v_i^{(1)}, v_j^{(1)}) - (v_i^{(2)}, v_j^{(2)})\right)^2}, \quad (3.1)$$

Where, $v_i^{(1)}, v_j^{(1)}$ are elements at indices i and j of the first vector, $v_i^{(2)}, v_j^{(2)}$ are elements at the same indices in the second vector.

Various word2vec models were trained on the dataset with different vector sizes varying in multiples of three. Empirically based on the PIP loss calculations, we found out that word embeddings with 36 dimensions deviated from its linear form in the curve by a small change (change in straight line), due to which we selected this to be the dimension of the word embeddings.

3.2.3 Phylogenetic analysis using the latent dimensions of the spike genes

To evaluate the phylogenetic properties based on the latent dimensions of the spike gene, we computed the cosine distances among spike genes of SARS-CoV-2 with the 36 latent dimensions. The pairwise distance was further used for hierarchical clustering using the 'hclust' function in R statistical programming language. This analysis was performed using 400 random sequences of spike genes from 16 countries. The visualization of the phylogenetic tree derived based on the latent dimensions was done using 'iTOL' software (Figure 3.2) (Letunic and Bork 2021) .

3.2.4 Entropy of the latent dimensions

To quantify the properties of latent dimensions, we have used a well-known information theory based algorithm suitable for time series datasets, called 'Fast Sample Entropy' (Pan et al. 2011). To compute Fast Sample Entropy, we have used the 'FastSampEn' function in the 'TSEntropies' package in R (Tomcala 2018). Fast Sample Entropy can be computed as follows.

$$FastSampEn(x, m, r) = log\left(\frac{\sum_{i=1}^{N_m} |s_{i,m}|}{\sum_{i=1}^{N_{m+1}} |s_{i,m+1}|}\right), \quad (3.2)$$

where,

$$s_{i,m} = \{\xi | (\|y_i - y_\xi\| \leq r, \xi \neq i) \wedge (\xi \notin s_{j,m}, j < i)\} \ , \quad (3.3)$$

$$y_i = [x_i, x_{i+1}, \ldots, x_{i+m-1}], \quad (3.4)$$

$s_{i,m}$ is a set of sub-sequences of length m belonging to the i-th neighborhood, and N_m is the number of these neighborhoods.

In our case, 'x' is the latent dimension of the spike genes of the SARS-CoV-2 strains per month for a given country with default values of 'm' and 'r'. Entropy was computed for each latent dimension on a monthly basis for each country. The default value for m and r in the equation were 2 and 0.15 * sd(TS) respectively, where m refers to the dimension of a given time series and r refers to the radius of searched areas. A minimum sample size of 30 is necessary for the accurate computation of entropy across each country. To compare geographies across months, we used average entropy derived from 36 latent dimensions, followed by normalization using all the monthly entropies for a given country (Figure 3.2). To compare the entropy of the latent dimensions among countries, we used the total entropy of the country for each dimension and visualized it with line graphs (Figure 3.3).

3.2.5 Detrended Cross Correlations Analysis (DCCA)

To investigate the entropy of the latent dimensions with the new cases observed for COVID-19, we used the Detrended Cross Correlation Analysis (Prass and Pumi 2019). Here, DCCA captures the long-range cross correlation between time series (entropy of the months and caseloads for a

given country). We tested both time series for stationarity using Augmented Dickey-Fuller (ADF) test (Mushtaq 2011). The ADF test was implemented using the function 'adf.test' available in the 'tseries' package in R ("Time Series Analysis and Computational Finance [R Package Tseries Version 0.10-54]" 2023). Due to the non-stationary distribution of the estimated entropies and the caseloads for a given country, we used the 'DCCA' R package ("Detrended Fluctuation and Detrended Cross-Correlation Analysis [R Package DCCA Version 0.1.1]" 2020). Cross-correlation was calculated between the entropy dimensions at time t+h and new cases at time t, where h = 0, ±1, ±2, ±3 … ±10.

3.2.6 Machine Learning based identification of significant predictive features

Country-wise monthly total new cases data was taken at the end of each month. Total new cases data for each month was merged to monthly entropy dimensions data from March, 2020 to June, 2021. We used a regression based machine learning approach called 'Boruta', a wrapper algorithm around a random forest algorithm to select the most relevant entropy dimensions for the prediction of subsequent two months' new cases ("Wrapper Algorithm for All Relevant Feature Selection [R Package Boruta Version 8.0.0]" 2022; Kursa, Jankowski, and Rudnicki 2010). We used the default parameters with the modification of the maximum runs as 1000. We selected the confirmed entropy dimensions as the most relevant predictive features for the prediction of new-cases.

3.2.7 Model development and evaluation for prediction of new cases in subsequent months

To predict the new cases in the next to next months, we used a regression based random forest model using the most relevant predictive features using the 'Boruta' R package (Kursa, Jankowski, and Rudnicki 2010). The model training was performed using entropy data from March, 2020 to February, 2021; and the fitted model was validated on entropy data from March, 2021 to April, 2021. The regression modeling was performed using 1000 decision trees using the 'randomForest' package in R (Liaw and Wiener 2002).

3.2.8 Strainflow algorithm

The algorithm for the Strainflow pipeline has been described below:

1. We have collected the SARS-CoV-2 sequences from the GISAID EpiCoV database. High quality sequences with complete temporal information were filtered.
2. We extracted the spike gene region of these sequences from FASTA files using the CoV-Seq tool. A CSV containing these sequences and other metadata such as country names and dates was created.
3. The sequences were splitted into chunks of three characters (codons). A splitted sequence represents a document with three-letter words.

4. We trained a word2Vec model on the spike gene sequences for learning 36-dimensional word embeddings. The average of all word embeddings in a given sequence was treated as the embedding of the sequence.
5. We calculated the sample entropies of each dimension of our embeddings for each month and country.
6. New COVID-19 cases for each country in each month were calculated using data from the JHU CSSE repository.
7. A feature selection algorithm (Boruta) was used for selecting the entropy dimensions predictive of caseloads two months in advance.

Random Forest regression algorithm was used for predicting new cases two months ahead of time. The inputs to the model are the country names and important features extracted from the Boruta algorithm. The predictor variable is the caseload two months ahead of time for each country.

3.2.9 Strainflow Dashboard

Implementation: The strainflow dashboard web application is primarily built using *ReactJS* and other accompanying libraries for UI needs and *GraphJS* for graphical needs. Python libraries such as numpy, pandas, matplotlib and seaborn were used to pre-process and infer the dataset. The Random Forest regression model was implemented using the R library *randomForest*. The web application is available for use on http://strainflow.tavlab.iiitd.edu.in and works on all modern browsers.

Functionalities: The application has three tabs: Cases Plots, Entropy Plots, and Paper. The Cases Plot tab exhibits two graphs; one compares the actual number of cases with our predicted cases, with a two-month lead time, while the second shows entropy against the caseload for a given country. The Entropy Plots tab displays the sum of sample entropy across all the latent dimensions for each pair of countries. The toggler present above the graph can be used to change countries to compare their entropies. Lastly, the "Paper" tab presents a graphical abstraction of our paper.

Discussion: COVID-19 had a devastating impact on our health systems, thus with caseload predictions made two months in advance, we provide a data-driven handle on epidemiological surveillance to warn about potential upcoming cases, so that people can be prepared in advance and appropriate preemptive steps can be taken by policymakers to prevent the spread of COVID-19.

3.2.10 Rationale for the study

Data Quality, Temporal Information and Spike-gene extraction: The study prioritizes the collection of high-quality SARS-CoV-2 sequences with complete temporal information from the GISAID EpiCoV database. Ensuring data integrity and temporal context is crucial for accurate

analyses. The utilization of the CoV-Seq tool to extract the spike gene region emphasizes the importance of focusing on specific genomic regions relevant to the virus's virulent behavior and evolution.

Word Embeddings and Representation: Word2vec language model was trained on the spike region of the COVID-19 sequences to capture the changes (mutations) occurring within these sequences. It encoded the sequence into vector representations by capturing the grammar of the underlying genomic space. The selection of the word2Vec model over more advanced transformer-based models is justified in light of specific study requirements, emphasizing computational efficiency and interpretability. Large language models such as BERT and GPT require extensive datasets for training due to their data-intensive nature and significantly have higher time complexity compared to training the simpler neural network model such as Word2vec. Another crucial aspect of our methodology involves the dynamic updating of word2Vec models in response to the increasing genomic sequence data. In the initial stages of pipeline processing, the genomic sequences data were relatively limited, progressively expanding over subsequent months. This necessitated the iterative updating of word2Vec models over distinct time frames. Hence, employing Bert-based models for this purpose would pose computational challenges, considering their resource-intensive nature and prolonged training periods.

Entropy Analysis: We calculated sample entropies for each dimension of the embeddings across months and countries to capture the variability occurring in the genomic sequence due to mutations. We used Fast Sample Entropy because the computation algorithm calculates sample entropy in a faster way. This step contributes to understanding the changing grammar of the virus sequence over time and across geographical locations.

Integration with Epidemiological Data: Incorporating new COVID-19 cases data from the JHU CSSE repository establishes a connection between genomic features and real-world epidemiological outcomes, allowing for a holistic analysis.

Predictive Feature Selection and Modeling: Boruta feature selection algorithm was used for identifying entropy dimensions predictive of caseloads two months in advance. It captured the non-linear relationship between entropy dimensions (input feature variable) and case surges (predictor). Shapley analysis allowed us to identify dimensions with potential early warning indicators for forthcoming outbreaks. Subsequently, the important dimensions extracted from the feature selection process were employed into the RandomForest model for predicting caseloads two months in advance. RandomForest was chosen for predictive modeling due to the non-linearity of the data. The selection of RandomForest aligns with the feature selection method, as Boruta also utilizes RandomForest as a wrapper class in its process.

Statistical Analysis: Statistical analyses involved employing both Pearson (parametric) and Spearman (nonparametric) correlation tests to thoroughly assess the associations between

predicted cases and original cases. The Pearson correlation coefficient was used to quantify the linear relationship between two variables and can be seen as a scaled version of the regression coefficient. Therefore, we used this to understand the linear trend between our predictions and actual caseloads. On the other hand, the Spearman correlation coefficient calculates correlation on ranks and captures if a variable monotonically increases/decreases with respect to another variable, and can be used even when the relationship is non-linear.

3.3 Results

3.3.1 Genomic sequence-based language modeling captures emerging diversity in the SARS-CoV-2 spike gene.

Our results validate the idea that a complex combination of codon weights may confer evolutionary advantage to the variant. The combinations of weights were learned using state-of-the-art unsupervised embeddings for capturing the latent space of spike DNA sequences of SARS-CoV-2. The framework of *Strainflow* is depicted in the figure below (Figure 3.2A*)*. The global tSNE plot represents dynamic emerging patterns derived from latent space representations of spike genes of SARS-CoV-2 (Figure 3.2B) from September, 2020 to March, 2021, along with specific geographic locations (country-level) such as India, UK, USA, and China.

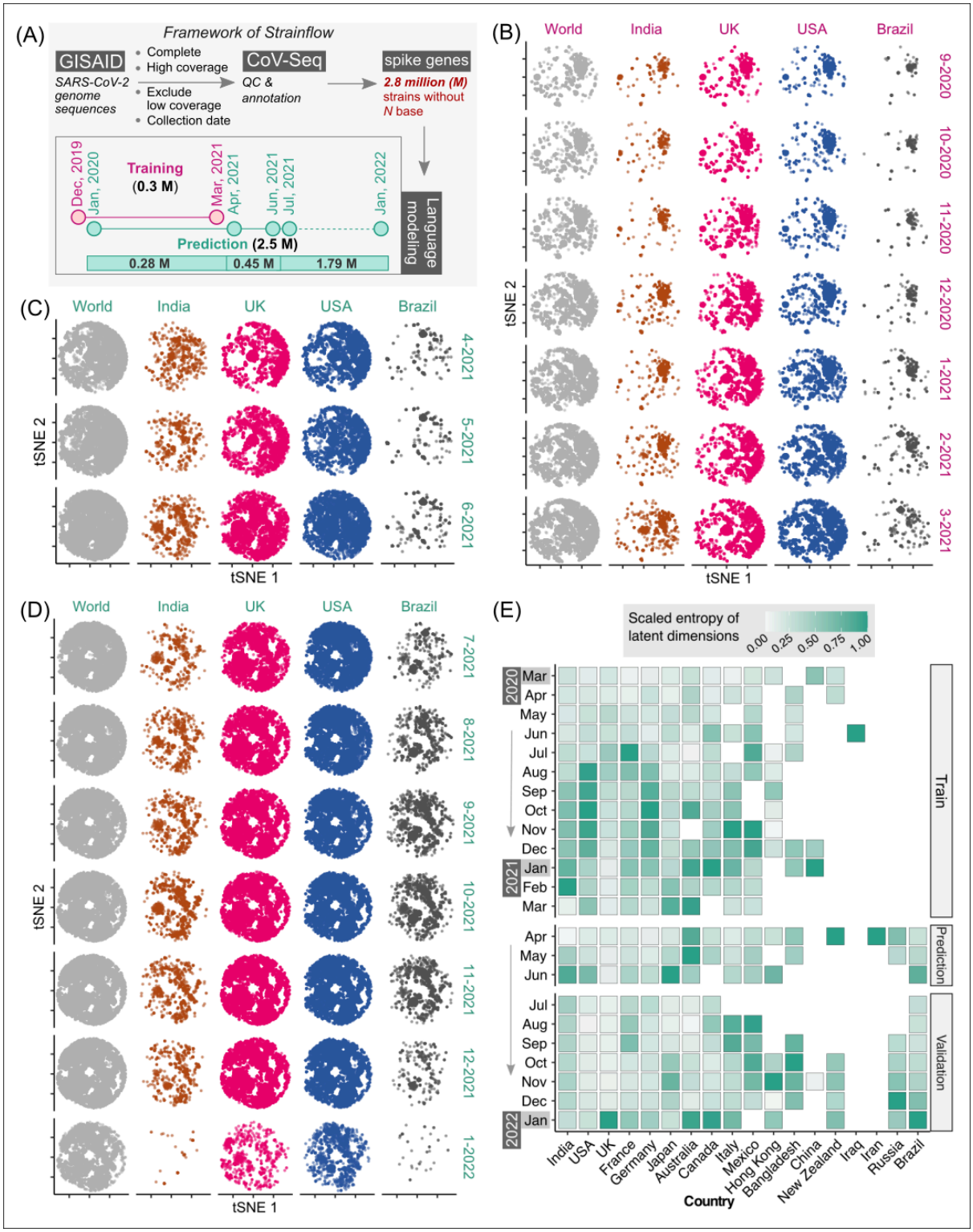


Figure 3.2. *Latent space of spike genes derived using Strainflow preserves spatiotemporal information of SARS-CoV-2 spread. (A) The implementation framework of Strainflow (details described in the method section) (B) tSNE plot showing distinct spatio-temporal relationship based on the latent space learned from the spike gene of 0·308 million SARS-CoV-2 genomes*

collected till 31 March 2021 (world), India, UK, USA, and Brazil. (C) Embeddings estimated or predicted from the Strainflow model for 0·45 million SARS-CoV-2 spike genes from the month of April, 2021 to June, 2021. (D) Embeddings estimated or predicted from the Strainflow model for 1.79 million SARS-CoV-2 spike genes from the month of July, 2021 to January, 2022. (E) Heatmap showing the scaled entropy for 18 countries from March, 2020 to January, 2022 (showing data for a. training: March, 2020 to March, 2021, b. prediction: April, 2021 to June, 2021, and c. validation: July, 2021 to January, 2022). The entropies for each country were scaled to the same range to visualize the temporal trends within the country.

To investigate the change in entropy in the latent space of the spike gene learned by our *Strainflow* pipeline, we performed qualitative and quantitative analysis on 2.7 million SARS-CoV-2 spike genes collected from December, 2019 to January, 2022. Qualitative analysis was performed by performing dimensionality reduction with a fast tSNE method called Flt-SNE (Linderman et al. 2019). We compared the 2D t-SNE plot of the world with four countries (India, United Kingdom, United States, Brazil) from September, 2020 to January, 2022, which clearly highlights the dynamic changes in the spike genes across countries in different months (Figures 3.2 B,C,D). Additionally, quantitative analysis of the latent space was performed by calculating the fast sample entropy of each latent dimension. To compare the monthly entropy of the latent dimensions of different geographical regions, the mean entropy was calculated and normalized across the months for each country. We observed the highest entropy for India, United Kingdom, United States and Brazil in the months of February-2021, January-2022, August-2020, and January-2022 respectively. Interestingly, we observed high entropy for 4 months from August, 2020 to November, 2020 in the United States (Figure 3.2E). This highlights that the spike protein latent space representation learned by Strainflow could be used as a proxy to capture the spatiotemporal entropy or diversity in the emerging SARS-CoV-2 strains across different countries.

3.3.2 Preservation of spatiotemporal information of SARS-CoV-2 spread depicted with phylogenetic analysis.

Sequence-level embeddings were obtained from the codon embeddings and investigated for the presence of genomically meaningful characteristics. The phylogenetic tree derived from the embeddings for the United Kingdom (Figure 3.3A) shows two clear temporally split clusters for 2020 and 2021 sequences, which may be indicative of different strains in these time periods. The temporality of the collected sequences was found to be preserved in the two clusters, although the model was trained only on genome sequences.

The phylogenetic tree with globally collected sequences (Figure 3.3B) demonstrates that geospatial information is also preserved in the sequence embeddings. The dendrogram constructed using cosine distance between embeddings revealed clear clusters of geospatially close regions. Embeddings from geographically close locations were clustered together (Figure

3.3), and countries closer geographically had similar embedding patterns (Figure 3.4). This highlights that our de-novo embeddings captured these similarities without the need for standard alignment methods or expert knowledge of lineages. Clusters for China (purple), Australia (green), and England (magenta) are highlighted in Figure 3.3B. Strains from Italy, France, Brazil, Japan, Canada, USA, Scotland, and India were found to be dispersed with other countries. Overall, *Strainflow* captures the temporal emergence of strains and geographic information in a country-specific manner.

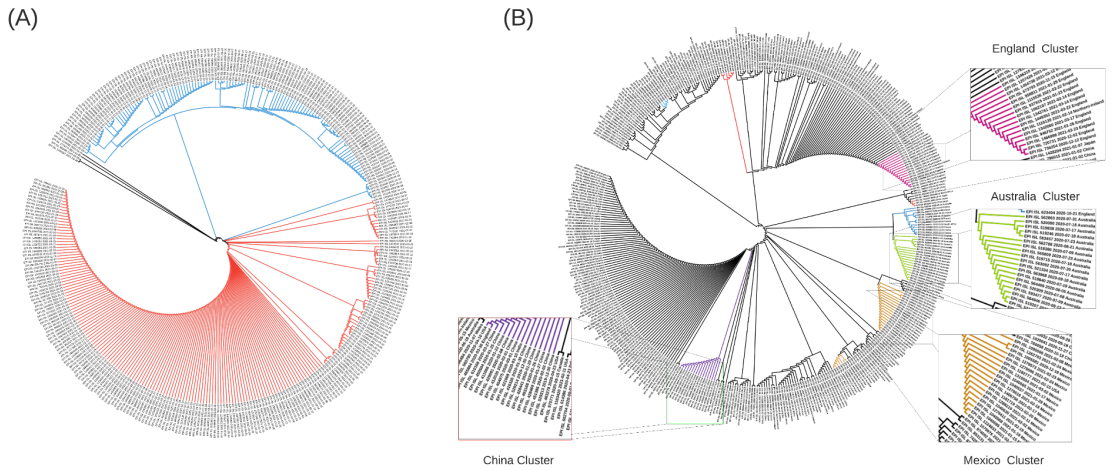


Figure 3.3. *Phylogenetic trees constructed using cosine similarities between 400 randomly sampled sequence embeddings. (A) Dendrogram for strains from the U.K.: Cluster 1 (blue) contains strains from the period Oct 2020 - Dec 2020, while Cluster 2 (orange) contains strains collected between Jan 2021 - Mar 2021. (B) Dendrogram for 16 countries across the globe: Chinese, Australian and England strains form tight clusters (marked in purple, green, and magenta), while strains from Italy, France, Brazil, Japan, Canada, USA, Scotland, and India are dispersed with other countries.*

3.3.3 Entropy in the latent space dimensions captures variability in the spike gene.

Entropy of a latent dimension has biological significance as it intuitively captures the variation in codon level changes during a certain time window. Each latent dimension encodes a combination of codon weights and increase in entropy represents frequent changes to these weights. Temporal changes in entropy are therefore expected to uncover the explore-exploit cycles of SARS-CoV-2 spike gene changes, hence biologically indicative of future trends. To compare different geographical regions, the sum of sample entropy was computed for each latent dimension across all the months. This revealed that certain geospatial regions such as France and Germany (Figure 3.4A) and USA and Canada (Figure 3.4B) have similar total entropies across the latent dimensions, indicating that strains in these regions have been accumulating similar genomic changes.

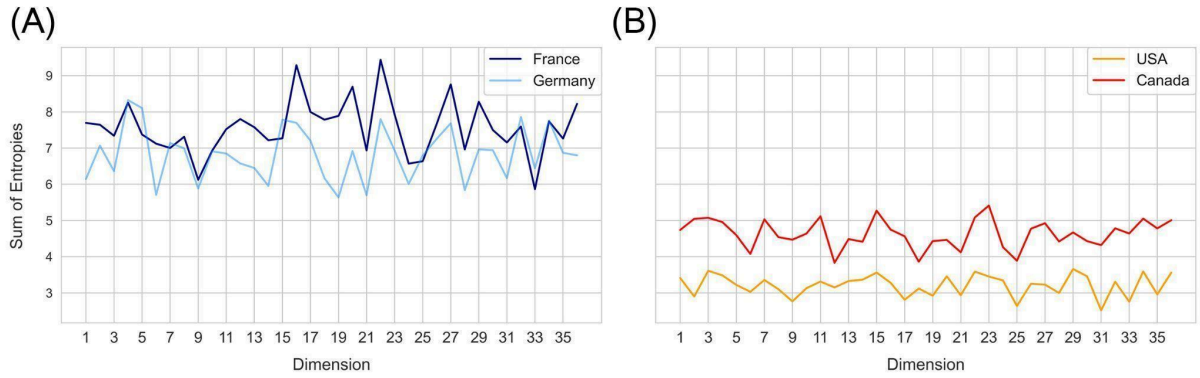


Figure 3.4. *Sum of sample entropy for each latent dimension for different countries. Country pairs (A) - France and Germany, (B) USA and Canada show a similar distribution of total sample entropy across dimensions, while each pair differs from the other.*

3.3.4 Entropy dimensions are predictive of new COVID-19 caseloads.

We then attempted to decipher the relationship between monthly sample entropy and monthly new COVID-19 cases in different countries. Detrended cross-correlation coefficient was calculated at different lag values, which revealed that entropy dimensions have a leading relationship with new cases (Figure 3.5A, 3.5B). This suggests that the genome sequence data in a given month can be used to predict new cases in subsequent months. A lead period of two months was chosen and Boruta algorithm was employed to assign feature importance scores to different dimensions, which revealed that dimension 32 is the most significant predictor of new cases **(**Figure 3.5C**)**. Significant dimensions from Boruta analysis were used for further modeling. Random forest based regression modeling on the predictive features achieved a total R-squared of 73% on the validation set. The predicted cases were found to be highly correlated with the actual cases (Table 3.1), which suggests that our model can indicate the directional change of cases for different countries. Further, the predicted relative change in cases between successive months was found to be correlated to the actual relative changes (Table 3.1) for delta variant, which suggests that our model can also indicate the magnitude of change that we expect to observe in the cases.

Country	Pearson Correlation	p value	Spearman's Correlation	p value
Cases prior to Delta variant				
USA	0·97	$8{\cdot}41 \times 10^{-9}$	0·94	0·00
India	0·91	$6{\cdot}13 \times 10^{-6}$	0·97	0·00
Germany	0·91	$6{\cdot}78 \times 10^{-6}$	0·87	7.57×10^{-6}
France	0·86	$7{\cdot}35 \times 10^{-5}$	0·97	0·00
England	0·82	$2{\cdot}89 \times 10^{-4}$	0·66	$1{\cdot}22 \times 10^{-2}$
Japan	0·71	$4{\cdot}38 \times 10^{-3}$	0·63	$1{\cdot}92 \times 10^{-2}$
Brazil	0·48	$8{\cdot}61 \times 10^{-2}$	0·45	$1{\cdot}12 \times 10^{-1}$
Cases at the onset of Delta Variant				

USA	0.93	2.81×10^{-6}	0.94	0.00
England	0.82	5.60×10^{-4}	0.71	8.14×10^{-3}
France	0.8	9.36×10^{-4}	0.55	5.25×10^{-2}
Germany	0.76	2.47×10^{-3}	0.77	2.92×10^{-3}
India	0.71	7.05×10^{-3}	0.59	3.60×10^{-2}
Japan	0.68	1.08×10^{-2}	0.57	4.73×10^{-2}
Brazil	0.67	1.22×10^{-2}	0.78	2.62×10^{-3}

Table 3.1: *Pearson and Spearman's correlation coefficients between predicted and actual cases in different countries pre- and post-onset of Delta variant.*

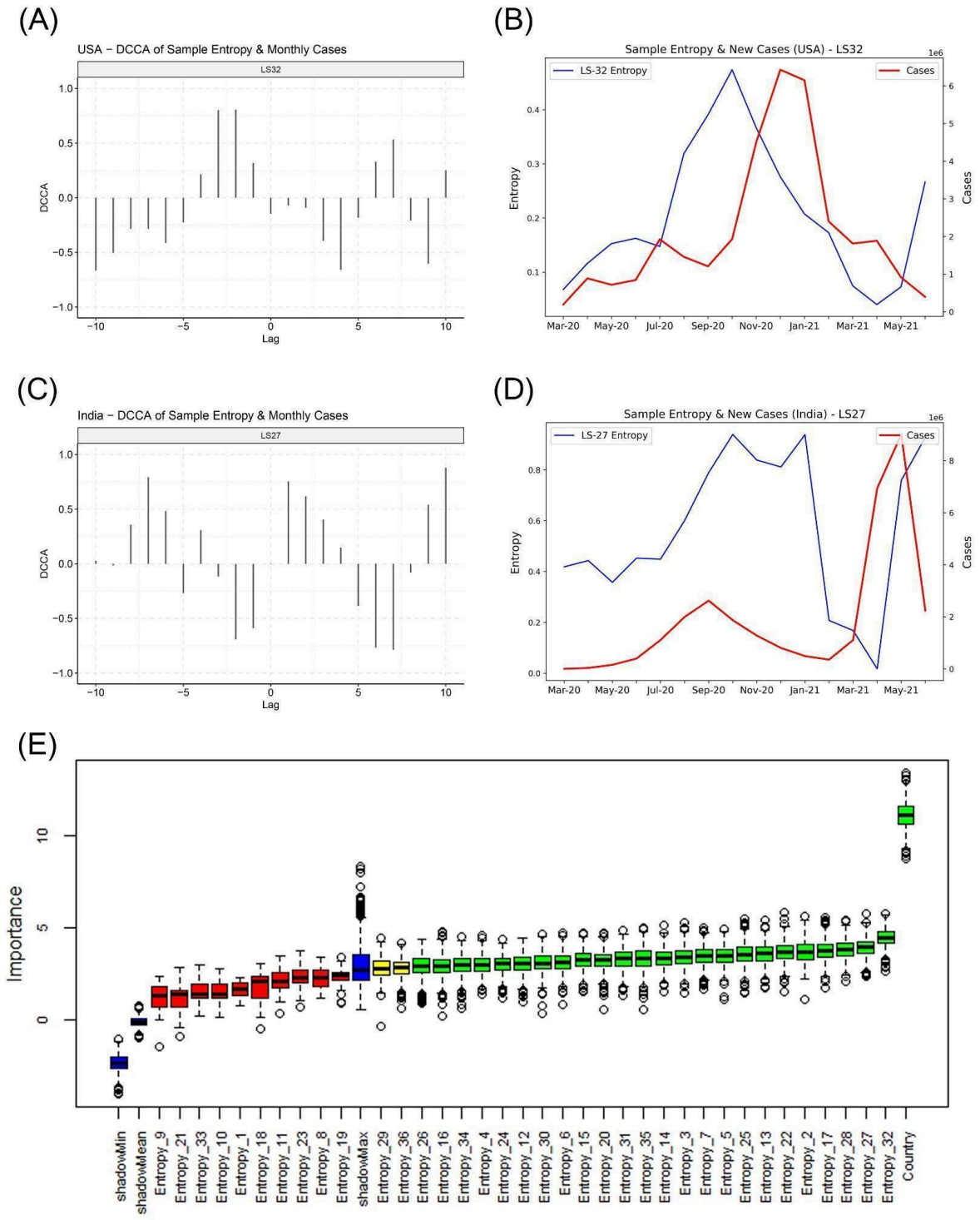


Figure 3.5. *Relationship of the entropy of latent space dimensions with COVID-19 caseloads. (A) Detrended Cross-correlation coefficient values for different lags between Entropy dimension 32 and new cases for USA. High values are observed for a lead of 1 and 2 months. (B) Line plot for Sample Entropy dimension 32 and monthly new cases for USA, indicating that the entropy in*

dimension 32 has a leading relationship with the cases. (C) Detrended Cross-correlation coefficient values for different lags between Entropy dimension 27 and new cases for India. (D) Line plot for Sample Entropy dimension 27 and monthly new cases for India, indicating that the entropy in dimension 27 has a leading relationship with the cases. (E) Feature importance scores from the Boruta algorithm for predicting cases in the month following the next month.

Our model can be therefore used to predict the COVID-19 caseloads in several countries. Both USA (Figure 3.6A) and Japan (Figure 3.6C) show an increase in the sample entropy across the time period April - June 2021, concurrent with the respective spreads in these countries. Our model predicts new caseloads with a two-month lead time, which strongly predicts a spike in new cases both in USA (Figure 3.6B) and Japan (Figure 3.6D) in the months of July and August, 2021. For India our model predicted a decline in the number of cases for the month of July and August, 2021 (Figure 3.6E, 3.6F). Therefore our model may be used as an epidemiological early warning system to predict new caseloads.

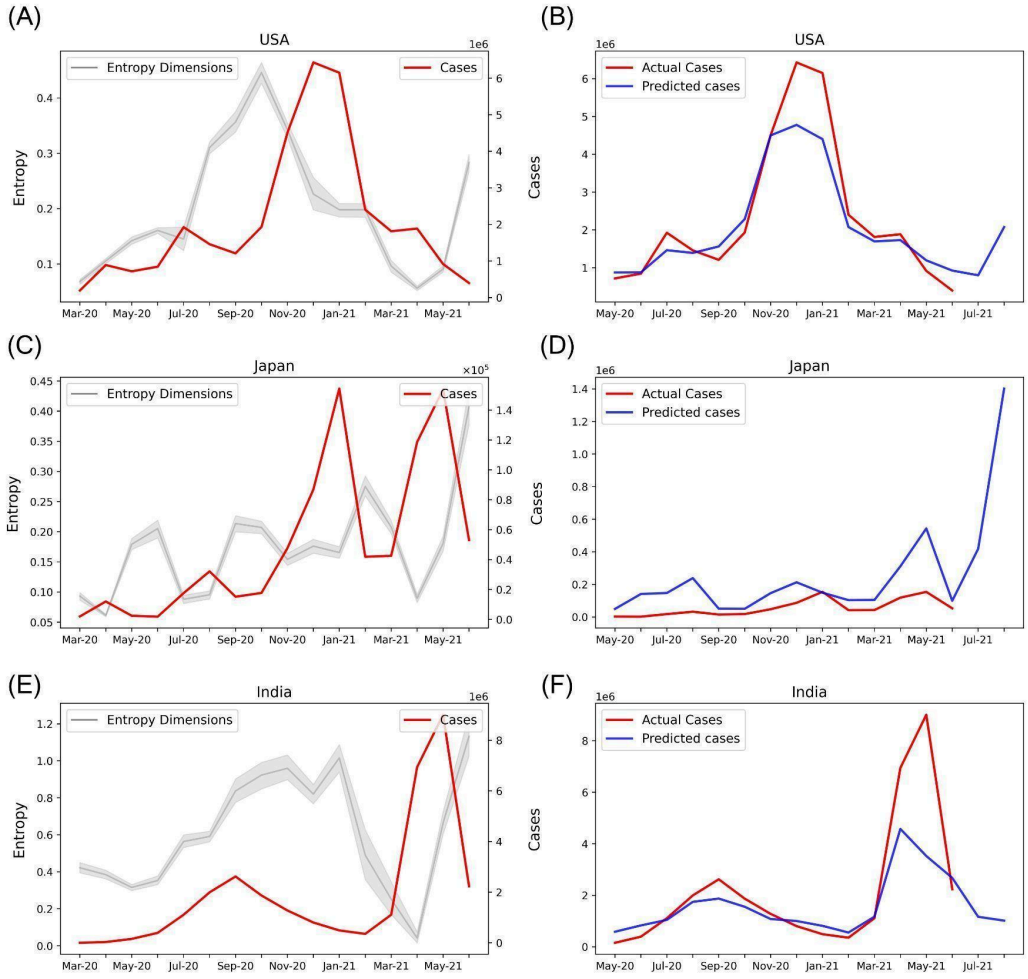


Figure 3.6. *Prediction of new COVID-19 cases with Sample Entropy values of the latent dimensions. (A) Line plot showing the Entropy values of the selected features and new COVID-19 cases for the USA. (B) Actual and predicted cases based on the entropy values of selected features for the USA. The model predicts a rise in cases for July and August 2021. (C) Entropy of selected features and new cases for Japan. (D) Actual and predicted cases for Japan. A spike in cases is predicted for July and August 2021. (E) Entropy of selected features and new cases for India. (F) Actual and predicted cases for India.*

3.3.5 Prospective validation of the model in the Delta and Omicron surges revealed interpretable predictive features.

For investigating the potential of our predictive features to track the spread of SARS-CoV-2, we used the codon level information of the SARS-CoV-2 delta variant for the spike gene and extracted the weights of these codons specific to each feature. We selected Dimensions 3, 4, 12, 13, 15, 16, 25, 28, 30, 32 with high absolute weights for the codons related to the delta variants (Figure 3.7A). The entropy of these features was contrasted with the caseloads in England (Figure 3.7B), India (Figure 3.7C), and USA (Figure 3.7D). Overall, the temporal tracking of these features may be used as a surrogate to track the spread of various SAR-CoV-2 variants.

Our case prediction model was frozen in June, 2021 and prospectively predicted the caseloads from July, 2021 to December, 2021. Our model predicted the case upsurge in India due to the Omicron variant in November, and December, 2021 (Figure 3.7E) two months ahead of time. Although the model fails to predict the exact values of cases, it is useful as a trend indicator. Further, we observe explore-exploit cycles in the entropy-space of India prior to the case peak due to the Delta variant in May, 2021 (Figure 3.7F). A similar exploration phase can be observed for the months from September - November, 2021, which may be indicative of an upcoming case peak driven by the Omicron variant.

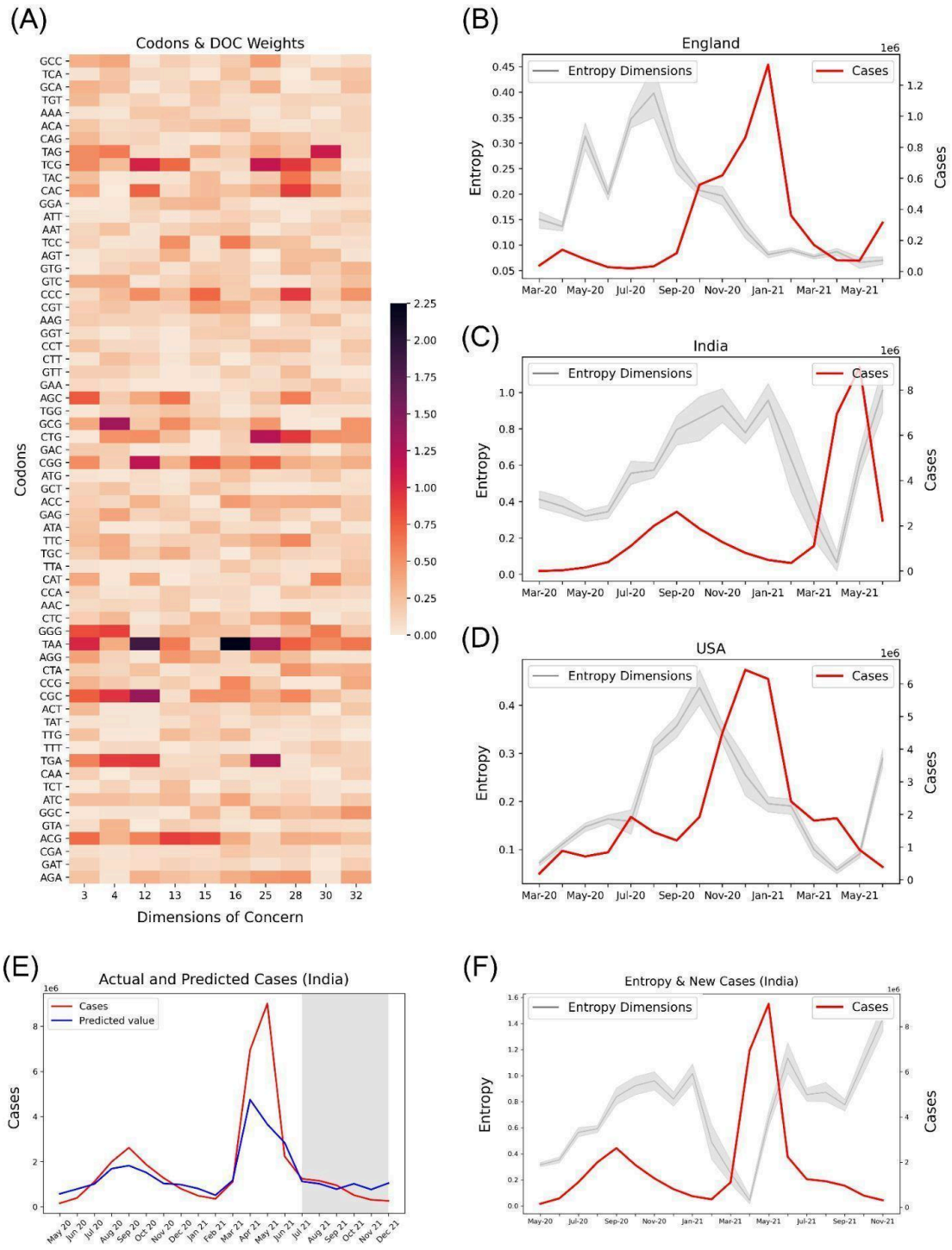


Figure 3.7. *Potential association of codons observed in SARS-CoV-2 Delta variant (lineage B.1.617.2) with their corresponding entropy features, and the trend of caseloads with the entropy features. (A) Absolute latent space weights of the codons associated with the entropy features*

linked to the Delta variant. Line plots showing the entropy features and cases in countries, (B) England, (C) India, and (D) USA. The entropies show an increasing trend in the months April - June 2021 for India and USA, indicating a possible surge in the delta variant in these countries. (E) Predicted and Actual cases for India. The region shaded in grey represents the months for which the case prediction model was prospectively validated. (F) Entropy and Caseloads for India. Explore-exploit patterns in the genomic feature-space can be observed.

3.4 Discussion

We have implemented an approach for analyzing the emerging strains based on the latent space of spike protein coding nucleotide sequences. We chose the nucleotide sequences instead of proteins in order to capture and track the variations that may not have immediate functional consequences. Our approach has two main underlying tenets: (i) long-range interactions are known to modulate the functional interaction between receptor binding domain and ACE2 receptors, hence may be captured in the NLP models that capture 3-mer changes and context, and (ii) latent dimensions may be differentially correlated with indicators of spread, thus providing a data-driven handle for tracking and predicting variants of concern and variants of interest (Mugnai et al. 2020). The pipeline takes advantage of temporal changes in the semantics of mutating sequences. Preservation of phylogenetic structure based upon the similarity matrix obtained using the embeddings validated that the latent dimensions capture spatio-temporal information. Analyzing the dynamic patterns and underlying correlations in the 30,000 base pair long sequence of SARS-CoV-2 is important to highlight the mechanistic understanding of mutations (Shishir, Naser, and Faruque 2021). SARS-CoV-2 seems to show a particularly high frequency of recombinations arising due to the absence of a proof-reading mechanism and sequence diversity, which calls for urgency in studying its transmission pattern (Rouchka, Chariker, and Chung 2020; Mandal, Roychowdhury, and Bhattacharya 2021). Therefore predicting mutations in the spike protein, which binds to ACE2 receptors can help us estimate the spread of disease and the efficacy of therapeutic treatments and vaccines (Q. Li et al. 2020).

While most research studies have attempted to predict the exact number of cases and have failed, our work is focussed on early prediction of trends from a non-obvious source of data. Unlike obvious data sources, the inter-relationships between codons in genome sequences are complex and less likely to be influenced or biased. Furthermore, sequencing data are made routinely available via various national and global consortia for genomic surveillance of SARS-CoV2. Our study also highlights the potential for triangulating insights from completely unrelated datasets, an approach that is expected to eliminate systematic biases in reporting by independent organizations. Further studies may triangulate insights from disparate, heterogeneous datasets such as mobility, genome surveillance, testing and case predictions to partially solve the problem of biases in individual datasets.

Entropy is a measure of the disorder of a system. We hypothesized that mutations increase the chaotic dynamics in the latent space of spike genes. To calculate entropy, we used the accelerated

versions of the Approximate Entropy and Sample Entropy algorithms, called Fast Approximate Entropy and Fast Sample Entropy (Tomcala 2018). Both algorithms aim to quantify how often different patterns of data are found in a time series. Fast Approximate Entropy, however, is a biased statistic and depends on the length of the series. Since we could have different counts of genome sequences collected each month, we preferred Sample Entropy, which is independent of the length of the series. Entropy values were calculated for each latent dimension in each month. Thereafter, Detrended Cross-Correlation Analysis (DCCA) was performed between the entropy dimensions and the new cases (Prass and Pumi 2019). DCCA is a modification of the standard cross-correlation analysis for finding relationships between non-stationary time series. High cross-correlation for different lead periods revealed that the entropy values in a given month could be used to predict the new cases in different countries in subsequent months. Different countries had different lead times at which the highest cross-correlation was observed between the entropy dimensions and the cases, ranging from 1-6 months. Overall, a lead time of two months was chosen to model the new cases. An empirical analysis was also done with daily values of entropy and new cases. Entropy was calculated in rolling windows, and cross-correlation analysis was performed between entropy and new cases at different lead periods. Although the cross-correlation values were found to be significant, the values were low and ranged between -0·1 to 0·1. Therefore, we decided to use the monthly entropy values for the modeling exercise.

To predict new COVID-19 cases, a random forest regression model was trained on the monthly entropy data. With sample entropy, we achieved an R-squared value of 73% on the validation set, while with approximate entropy, the value was only 10%. Therefore the model trained on sample entropy was selected. The predictions from the model were found to be highly correlated with the actual cases, indicating that our model can be used for preemptive warning signals for the rise in cases in different countries. Further, the actual and the predicted difference in the number of cases in consecutive months was found to be correlated, which suggests that the relative change in the cases in consecutive months predicted by our model is linked to the relative change in the number of cases. Overall, we recommend that our model be used to predict dangerous trends and not the actual number of cases. Further, the mapping from latent dimensions to Variants of Concern (VOCs) and Variants of Interest (VOIs) may help us track the country-specific spread of different variants.

The COVID-19 pandemic has been a dynamically evolving scenario, with new strains emerging and vaccines being developed. With the SARS-CoV-2 genome constantly mutating, we anticipate an underlying change in the grammar of the sequence, underpinning the need to update our language model every few months. Further, the regression model for caseloads needs to be periodically retrained too. An empirical analysis led us to discover that the Random forest model used for prospective validation from November, 2021 onwards performed better in terms of predicting the number of cases than the model used for prospective validation from July 2021. However, both models indicate similar trends in cases for most countries.

Further, models trained on genomic sequences can be used for predicting infection severity based on Co-associations between the SNPs of Co-morbid Diseases and COVID-19 (R. Y. Wang et al. 2020). The machine learning models can also be trained on genomic sequences for COVID-19 classification (Arslan 2021). Although the variance explained by our model is low, however, we were able to compute the variability associated with spike protein mutations. So our method showed a potential way to estimate the new cases variability associated with spike protein mutations. Our methods can be incorporated with the epidemic projections model to better predict the epidemic trajectories. The latent dimensions may further be employed to predict the clinical consequences of emerging strains. The currently available vaccines are intended for early SARS-CoV-2 strains, but with new emerging variants, immune responses triggered by these vaccines may be weaker and short-lived. As seen in the devastating second wave of the pandemic in India, newer SARS-CoV-2 variants have acquired an increased pathogenic potential resulting in rapid clinical progression and overwhelmed health systems. Mitigating such events in the future will require stronger surveillance systems in place. Our study offers a promising solution in this direction and lays the foundation for proactive genomic surveillance of COVID-19.

Our study has the following limitations. Our approach of codon embeddings does not indicate the position where the codon change may have happened in the spike gene. This is because low-dimensional embeddings do not preserve the positional encoding of words. However, we are investigating advanced approaches such as complex-valued word embeddings with positional encodings and transformer models such as BERT to overcome our current limitations (J. Lee et al. 2020; R. Y. Wang et al. 2020; Wolf et al. 2019). The latter are considered expensive and data-hungry models and it will remain to be evaluated if the gain of positional information may be countered by the loss of prediction accuracy for forecasting new cases in the future. However, we believe that the availability of sequences for a wide variety of viral pathogens presents an exciting opportunity to train data-hungry models that may be able to transfer insights across pathogens and yet remain interpretable. Further, our Strainflow model is trained only on the spike gene of the viral genome, which does not represent the complete variation spectrum of the virus. To mitigate this shortcoming, we will develop a genome-level *Strainflow* pipeline for SARS-CoV-2. Furthermore, the present study does not consider the interaction between the spike gene and other genes in the SARS-CoV-2 genome. We have not considered the interaction between the ACE2 receptor sequence for the human and the spike gene sequences due to the unavailability of such large-scale paired data. However, we believe this is a strength of our study as we were able to extract relevant features as well as make valid predictions using the spike region of the SARS-CoV-2 gene alone.
Finally, a relatively small number of samples were used to construct the supervised predictive model for case prediction. As more data becomes available in subsequent months, we can produce more confident case predictions. An empirical validation depicted that we require a

minimum of 100 samples per month for calculating the sample entropy. This also underscores the need for a more reliable and agile approach to deposit country-level datasets on repositories such as GISAID. We make an appeal to the countries to facilitate the sharing of such data in order to be prepared for any future waves of the current pandemic and for preventing the new emergence of strains. We believe our study is an instance of the new paradigm of pathogen surveillance using a novel language modeling approach that is potentially scalable to infectious disease surveillance and antimicrobial resistance.

In conclusion, *Strainflow* provides a promising approach for predicting the rise in cases based on the mutation landscape of SARS-CoV-2 genomic sequences. By using unsupervised latent space features of genomic sequences, Strainflow accurately predicted the rise in cases of Delta and Omicron surges for most of the countries including India with a lead time of two months. The entropy analysis of *Strainflow* unsupervised embeddings provides interpretability to the deep learning-based language model and the codon-level analysis ensures the biological validity of the unsupervised features. Owing to the next objective, the extensive rise in COVID-19 scientific literature has created an opportunity for researchers to leverage language models to make informed decisions and policies. While this chapter demonstrated the application of language models in predicting the progression of COVID-19 based on genomic datasets, the subsequent chapter will explore the use of language models in mining scientific literature to identify key themes and topics related to COVID-19.

Chapter-4

Language models for mining COVID-19 themes in scientific literature and social media: Predicting Emerging Themes in Rapidly Expanding COVID-19 Literature with Unsupervised Word Embeddings and Machine Learning

4.1 Introduction

COVID-19 pandemic has proved to be an enigma with its diverse clinical presentation, controversial evidence for treatment, fast-tracked vaccine development, and unclear systemic implications (Singh et al. 2020; Aouissi, Ababsa, and Gaagai 2021). In accordance with the previous chapter, the application of a language model for predicting the progression of COVID-19 pandemic was demonstrated based on a genomic dataset. However, the determination of the key themes and topics related to the pandemic, such as the development of vaccines, the effectiveness of different public health measures, and the impact of the pandemic on mental health is also crucial. Notably, language models can also be utilized to analyze text data, specifically from scientific literature and social media articles related to COVID-19 which could aid in developing public health interventions and understanding the social and psychological impacts of the pandemic. The literature around COVID-19 is growing exponentially, with more than 150 thousand COVID-19 articles vetted by the World Health Organization ("Global Research on Coronavirus Disease (COVID-19)," 2023.). Understanding evolving themes in a context such as in COVID-19 is essential as knowledge synthesis from peer-reviewed literature becomes increasingly difficult for researchers, clinicians, and policymakers alike. Methods such as topic modeling and sentiment analysis have been previously carried out comparing pre-print with peer-reviewed literature only over a short period. Ebadi et al (Ebadi et al. 2021) studied the temporal patterns of sentiments and similarity between publications from different sources over time, using document embeddings. High-level research topics like oncology, personal protective equipment (PPE), analytics, rehabilitation-panic, high-risk groups and genomics were uncovered using structural topic modeling. Although such analyses reflect an abstract overview of the broad areas of research, they lack to capture the evolving context between distinct domain-specific entities. The objective of our study is to analyze and track word-level semantic similarity among biomedical entities to uncover emerging themes.

Abstracts of articles hold a substantial amount of information about the literature. Named entities within the abstracts play a crucial role in deducing valuable information from large amounts of text and influencing literature trends (Cho and Lee 2019). Models pre-trained on biomedical, scientific, and clinical benchmark datasets have been used to extract various clinical entities such as diseases, symptoms, chemicals, and adverse drug reactions from the continuous text. The

relative context of these entities changes over time, leading to a shift in similarity with other words (Kutuzov et al. 2018). Unsupervised word embeddings have previously been used to capture complex science concepts using the semantic relationship signified by cosine similarity (Tshitoyan et al. 2019).

Predicting links between “medical terms” is of high significance to understand the underlying themes within the literature and the phenomenon. Link Prediction is the task of predicting the existence of links between two nodes in a complex network based on a set of topological features. The problem of link prediction in real-world temporal networks has been explored a lot in recent years (Bu et al. 2019) primarily in online social media networks where nodes are represented by users and edges by the relationship between them. Supervised learning methods based on topological proximity measures have been vastly used to capture the shifting of links across time within networks (Özcan and Öğüdücü 2017; Güneş, Gündüz-Öğüdücü, and Çataltepe 2016). Our study builds upon prior research by pioneering the application of diachronic word embeddings and link prediction within dynamic networks of biomedical entities (Tshitoyan et al. 2019; Bu et al. 2019), utilizing machine learning to predict emerging themes in the rapidly expanding COVID-19 literature. While prior studies (Ebadi et al. 2021; Cho and Lee 2019) offer abstract overviews or lack contextual evolution among domain-specific entities, this work tracks word-level semantic similarity among biomedical entities, providing a deeper understanding of the pandemic's progression within scientific literature. The methodologies in our work demonstrated a significant advancement in the analysis and understanding of the dynamic landscape of COVID-19 research. This work also studies the evolution of literature based on changing cosine similarity between extracted entities in weighted temporal networks and predicts future emerging trends using link prediction. While Jake et. al (Lever and Altman 2021) utilized machine learning to organize the extensive SARS-CoV-2 literature, categorizing it by topics and article types, aiding in content analysis. Conversely, EvidenceFlow forecasts emerging themes within the evolving COVID-19 literature using diachronic word embeddings and link prediction, providing proactive insights into the literature's future trends and evolution. Another study (Ioannidis et al. 2022) evaluated the impact of COVID-19 publications on scientific citations in 2020-2021, revealing that 20% of citations were for COVID-19 papers. It highlights the shift in citation patterns across disciplines, affecting researchers' citation profiles, with COVID-19 publications dominating citations. In contrast, EvidenceFlow employs diachronic word embeddings and link prediction to forecast emerging themes in the COVID-19 literature, offering proactive insights into future trends and its evolving nature, which extends beyond mere citation analysis.

We propose for the first time the use of diachronic word embeddings, link prediction in dynamic networks of entities, and machine learning to predict emerging themes literature and make these publicly available as a web application. We have primarily focused on the fast emerging COVID-19 literature to train and validate our architecture for this study. We forecasted semantic and topological proximity features of named entity pairs generated from their temporal trends in

prior months. Further, we have used these forecasted features to predict links between clinical entities extracted from textual data over the forecasted time interval using machine learning algorithms. Furthermore, these links were used to create a network weighted by forecasted cosine similarity for detecting communities of entities that tend to reflect on themes of the articles published in that month. To assess the efficacy of our predictive modeling, we validated the proximity features of entity pairs forecasted from ARIMA using Mean Squared Error. We have also evaluated the machine learning algorithm's performance for predicting the links over a time span of three months.

The schematic representation of workflow has been demonstrated (Figure 4.1). The interactive analysis and results of emerging themes is available publicly at our web application called EvidenceFlow. The details about its working can also be found in the methods section. This study proposes a framework for capturing and tracking imminent themes formed by medical entities in the temporal space based on networks constructed using word embeddings trained upon the evolving COVID-19 literature. Both Strainflow and EvidenceFlow provide insights into the evolving nature of the COVID-19 pandemic and its impact on public health. While Strainflow focused on predicting and validating COVID-19 cases based on new variants, EvidenceFlow aims to capture and track the evolving themes in COVID-19 research. Together, these pipelines can provide a more comprehensive understanding of the pandemic progression and aid in the development of effective strategies to mitigate its impact.

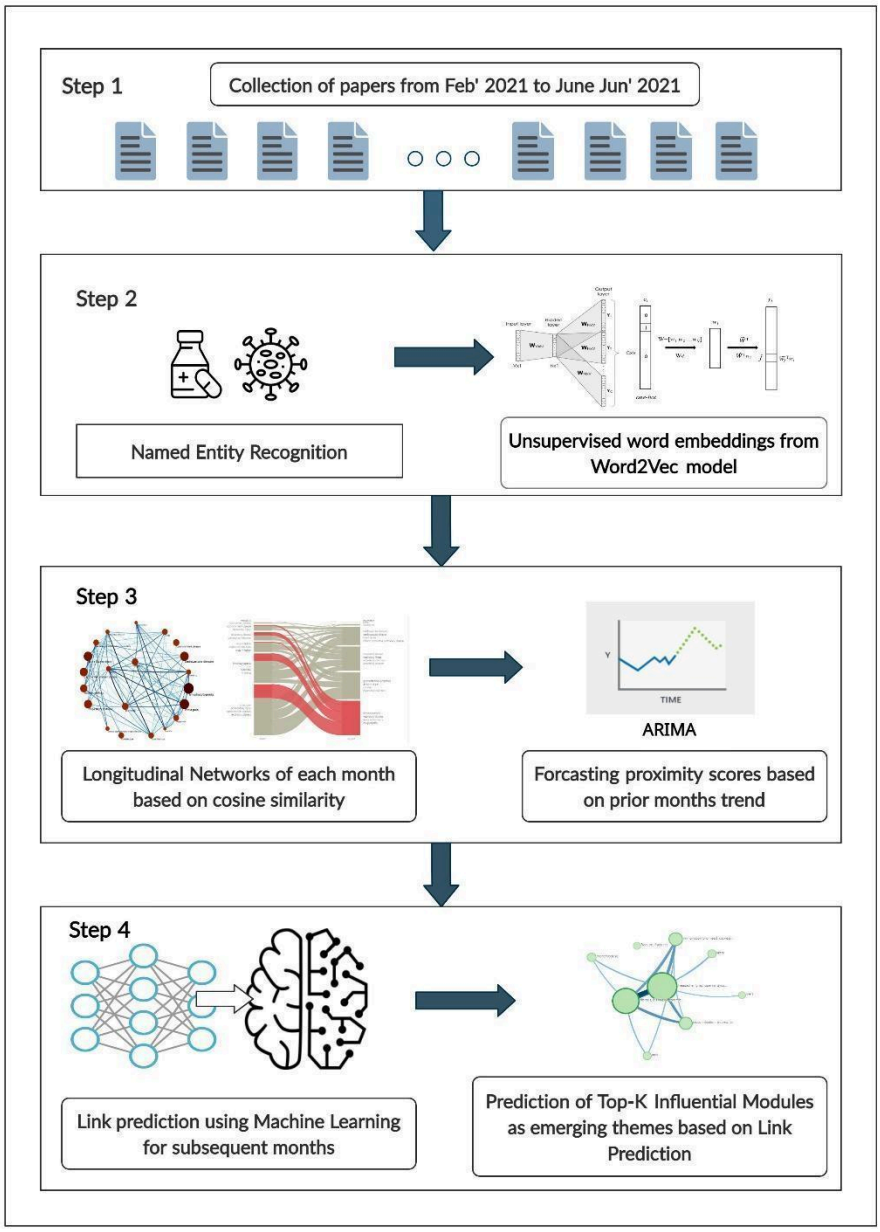


Figure 4.1: *Graphical representation of proposed framework explaining the complete workflow. The pipeline takes abstracts as input from which entities are extracted using Named Entity Recognition. Embeddings are generated, which are used as features for longitudinal networks. These networks are used for visualizing the trends using alluvial diagrams, link prediction, and predicting top-k influential modules for theme prediction.*

4.2 Methods

4.2.1 Dataset and Text Pre-processing

The dataset was created from abstracts of ~150,000 COVID-19 articles published in the publicly available *WHO Database ("Global Research on Coronavirus Disease (COVID-19)," 2023.)* from February 2020 to June 2021 (Figure 4.2A). For every research article, the database contains the corresponding title, authors, source of publication, journal, database, language, type of publication, entry date, country and full text URL. We queried the database on all Full Text articles in the English language, keeping the rest of the fields unfiltered. Formatting of text and removing white spaces, punctuation, digits, and stop words were carried out on lowercase converted text using the NLTK package (Loper, 2023.).

4.2.2 Named Entity Recognition

Named Entity Recognition (NER) was used to extract two types of entities: diseases and chemicals, from the original abstracts of vetted research articles using a model pre-trained on BC5CDR corpus by SciSpacy, an open-source project for Biomedical Natural Language Processing (Neumann et al. 2019). The model identifies the entities with an F1 score of 84.49% ("Scispacy," 2023.). The words extracted under the category of diseases also contained symptoms, adverse effects, conditions, disorders, and syndromes. All of these are collectively referred to as diseases in the other sections. Entities were further used to create networks to study the trends through alluvial diagrams and predict links between nodes across past and upcoming months.

4.2.3 Unsupervised Word Embeddings

In the previous chapter, embeddings of genomic sequences were learnt using the Word2vec model. We considered words as codons and sequences as paragraphs for training the model. In this chapter, word embeddings were trained upon the abstracts obtained from the WHO Database, updated with new publications and pre-prints as these become available every month. A low-dimensional representation (d=100) for the words present in the corpus of abstracts was learned using the Word2Vec model with the skip-gram algorithm and fixed window size of five, implemented in Gensim (Ma and Zhang 2015; Mikolov, Sutskever, et al. 2013; "Gensim," 2023.). Cosine distance between the word vectors of the extracted entities was calculated to analyze the dis(similarity) between entity pairs. Visualization of the word vectors was carried out using TensorFlow Embedding Projector (Smilkov et al. 2016) to allow interactive exploration of relationships between diseases and chemicals. To create each month's network of entities, separate Word2Vec models were trained to capture shifts in word similarities in the literature published over time.

4.2.4 Longitudinal Entity Networks and Communities

High cosine similarity represents strong relationships between words. We used diachronic word embeddings to capture the evolving contextual similarities between various diseases and studied its evolution over time. Weighted networks were constructed using similarity between word vectors of extracted entities as edge weights. From each month's corpus of abstracts,

top-N(=100) most frequently occurring diseases were extracted, and the pairs having greater than the 90th percentile of cosine similarity based on the corresponding month's word embeddings were used to create a union set of entities across months, preserved as nodes in the temporal networks. Therefore, every month's network had a fixed set of nodes with varying links, labeled as 0 or 1 based on the threshold of cosine similarity, and varying weights, calculated based on the evolving semantic closeness. The mentioned threshold has been chosen empirically based on experimentation; a high threshold has been selected to depict contextual similarity between two words present in the same latent space. For training and evaluation, the fixed set of entity pairs was created from the diseases identified in the abstracts of the papers published from February 2020 to February 2021, using the mentioned procedure. For the subsequent months, the word embedding models were trained on the respective corpora of abstracts, and the links between the fixed set of node pairs were assigned if they appeared in the vocabulary and weighted by the cosine similarity between their word vectors. Community detection was performed over the monthly networks using the Infomap algorithm (Bohlin et al. 2014). Semantic change in the word embeddings led to the formation of communities, which shifted as emerging themes over months. The importance of each node (entities) was tracked using an alluvial visualization based on PageRank values which changed across different months (Rosvall and Bergstrom 2010). Detailed steps with parameters have been explained below.

We considered the top N(100) entities from the abstracts of papers published each month based on the frequency. Corresponding networks were constructed using the following algorithm:

1. For every month τ in T months,
1. Top N entities extracted from the corpus of abstracts using NER.
2. All possible pairs of entities ($^{N}C_2$ combinations) are created to represent node pairs in a network. The weight of the edges is valued by the Cosine Similarity between the entities generated from the Word2Vec model trained on the corpus of month τ.
3. Edges weighing more than the M^{th} percentile of the weight are preserved as links.
2. A union of node pairs ($\Delta = \cup(u,v)$) from all the months were taken to have a common set of nodes in each. Every month τ was depicted by a network G_{τ}.

4.2.5 Time Series Forecasting of Proximity Scores

In order to predict the existence of links between nodes in the networks of subsequent months, we computed five neighborhood proximity scores for the network of each month. Jaccard Similarity, Common Neighbors, Preferential Attachment (Barabasi and Albert 1999), and Adamic Adar Similarity (Adamic and Adar 2003) were used as topology-based features, and Cosine Similarity between the entities represented by the nodes was used as a semantic feature. These proximity scores based upon network topology were calculated using the NetworkX package (Hagberg, Swart, and S Chult 2008). The range of Adamic Adar Similarity, Common Neighbors, and Preferential Attachment values lies between $(0.00 - \infty)$, while that of Jaccard

Similarity and Cosine Similarity lies in the range of (0.00 − 1.00). To scale the values, we normalized the former three scores in each network to bring them in the range of (0.00 − 1.00). Every proximity score was modeled as a time series for each node pair, and the value was predicted for the subsequent month using the Auto-Regressive Integrated Moving Average or the ARIMA model (G. P. Zhang 2003). Stationarity of the time series was assessed using Augmented Dickey-Fuller test. First-order autoregressive model (p=1, d=0, q=0) was used for stationary series and non-stationary time series were passed through the random walk order of the model (p=0, d=1, q=0). For validating, proximity scores for the network at timestamp $\tau+1$ were predicted based on their respective past values in the networks till timestamp τ. The model's performance was assessed by comparing the predictions with the original proximity scores in the $\tau+1$ time using the Mean Squared Error. MSE is one of the robust indicators to measure the closeness of forecast outputs to actual values in the time-series setting. To assess its sensitivity to outliers, we analyzed the distribution of errors (Figure 6). It was seen that the median of errors was close to zero with minimal influence from outliers. Steps followed for conducting this experiment have been explained below.

The time series forecasting was performed using the ARIMA model from the statsmodels package of Python. Each pairwise association in the network can vary temporally. The time series for each association (proximity score) may have different parameters. We tested each time series for stationarity using the Augmented Dickey-Fuller test, considering stationarity for $P<.05$. An ARIMA approach was then taken. If the series were stationary, a first-order autoregressive model (AR) model (p,d,q=1,0,0) was fit, else a random walk order (p,d,q=0,1,0) was used.

4.2.6 Link Prediction between Entities

The proximity scores predicted using the ARIMA model were further used to identify the occurrence of a link between entities in network $G_{\tau+1}$ based on the proximity scores and links in all previous networks [G_1, G_2, G_3, …, G_τ], using supervised machine learning. We experimented with the proposed link prediction approach using Logistic Regression (LaValley 2008), Random Forests (Breiman 2001), SVM (Hearst et al. 1998), AdaBoost (Freund, Schapire AT, and Labs 1999), XGBoost ("XGBoost Documentation — Xgboost 1.7.6 Documentation," 2023.). For training the models, four proximity scores, Jaccard Coefficient, Preferential Attachment, Adamic Adar Index, and Common Neighbors, were used as features of node pairs at each timestamp till τ. For validation, the forecasted proximity scores of the network at timestamp $\tau+1$ were used to predict links between nodes. Due to the high imbalance between the labels, Area Under Receiver Operator Characteristic Curve (AUROC) score was evaluated to select the optimal threshold for binary classification. While training, validating and testing the model, we did not use Cosine Similarity as a feature as it was the identifier variable for the link. Validation of the model was performed on the predicted proximity scores of April 2021 to June 2021. For Logistic Regression, evaluation of the key assumptions was done using Variance Inflation Factor (VIF) for measuring the degree of multicollinearity, Cook's Distance for detecting the presence of

strongly influential outliers, and the scatter plot of log-odds for checking the linearity of independent variables. These tests were not satisfied for the data of most months, hence Logistic Regression was not our preferred model and we did not consider it further in results. Welch's t-test was performed for comparing performance of the machine learning models, followed by Bonferroni correction (Bitnun 2015). The full details of the algorithm and features have been demonstrated below.

The algorithm used to predict links in the network at timestamp τ+1 has been demonstrated below.

1. For each node pair $(u,v) \in \Delta$ in G_t, $t \in \{1, 2,\ldots,\tau)$, five proximity scores were calculated based on the topological features of the graph and semantic similarity between entities.
    - Cosine Similarity, from the Word2vec model trained on the corpus of month τ.
    - Jaccard Coefficient (JC)
    - Number of Common Neighbors (CN)
    - Preferential Attachment (PA)
    - Adamic-Adar Index (AA)
    - Since the range of CN, PA and AA lies between $0.00 - \infty$, we normalized the respective scores in the range of $0.00 - 1.00$ in each network G_t.
2. For each proximity score, a ($\Delta\tau$) matrix was created where Δ represents the number of node pairs, and τ represents the number of months taken in the training set. This matrix stores the value of the proximity score for node pair $(u,v) \in \Delta$ at timestamp t.
3. For each node pair, the value of proximity score was forecasted at timestamp τ+1 using the AR159IMA model (p=1, d=0, q=0) if the series was stationary, else random walk order was used (p=0, d=1, q=0). The Mean Squared Error was calculated for the predicted proximity score for April 2021, May 2021, and June 2021.
4. Training of the classification model was done using four topological proximity scores as features (excluding cosine similarity as it is an identifier variable) from the networks $G_1, G_2,\ldots, G_\tau$.
5. Testing set features represent the four predicted proximity scores $G_{\tau+1}$ for all the node pairs $(u,v) \in \Delta$.
6. Due to a high imbalance between positive and negative labels, the Receiver Operator Curve (ROC) was used to obtain an optimal threshold for the binary classification, using Youden's J Statistic in the following formula:
    - $J = TPR - FPR$

where TPR = True Positive Rate and FPR=False Positive Rate.
J represents an array of differences between TPR and FPR of different points on the ROC curve. The index of the maximum value of J, argmax(J), was used as a criterion for selecting the cut-off that may represent the optimum threshold.

7. The optimum threshold was used to binarize the predicted probabilities into 0 and 1.

8. Links predicted from the model were verified across the ground truth links of $G_{\tau+1}$. The average performance metrics of the model were obtained by resampling the test set repeatedly for 100 times and testing 1000 samples in each iteration. The margin of error for 95% confidence intervals was also calculated.

4.2.7 Community Detection on Predicted Networks

The links between node pairs predicted by the best performing model were used to create networks weighted by cosine similarity scores predicted by the ARIMA model. Infomap algorithm was applied on the predicted and original test network to cluster the nodes into ten modules. The modules were compared using Intersection Over Union (IOU) using the following formula:

$$IOU = \frac{|A \cap B|}{|A \cup B|}, \qquad (4.1)$$

where A represents a set of nodes in the predicted i^{th} module, $i \in \{1, 2, \ldots, 10\}$, and B represents a set of nodes in the original j^{th} module, $j \in \{1, 2, \ldots, 10\}$.

4.2.8 EvidenceFlow

EvidenceFlow is an open-source interactive web application built upon COVID-19 specific literature vetted by the WHO, for tracking literature trends using alluvial diagrams, projection of influential entities, and network analysis across different months. The dashboard assists the user to understand the current and upcoming trends in the literature. The *StrainFlow* dashboard in the previous chapter helped in forecasting pandemic surges while tracking the mean entropy over the monthly timeline.The functionality of each tab has been mentioned below.

Alluvial Diagram: Select two or more months from the Add Month section, one by one. Once the months are selected, you can drag and change the order of these months in the Node/Month list that appears below. Click on the Create Diagram button to visualize the alluvial diagram. The alluvial diagram helps in tracking the trends in the literature between the selected months. It eases tracing the temporal dynamics of literature across different time intervals. In the Module Explorer panel on the right, there are multiple features that can help in better visualization, e.g., by painting all nodes of a selected module.

Multi-level Network: Select a specific month from the "Select Month" button. This illustrates the communities in the networks that are formed across entities extracted from the literature. More information about this network is given in the paper.

Source-level Network: Select a specific month from the "Select Month" button. This illustrates the source networks that are formed across diseases entities extracted from the literature. The link between two diseases suggests an association between two entities based on the cosine similarity.

Embedding Projector: Once we click on the "Embedding Projector" tab, it illustrates the latent space of the low dimensional word embeddings trained on the literature of COVID-19. The search option allows the user to query the nearest entity present across it. Isolate point allows the user to isolate N nearest points present around it.

Emerging Trends: This tab demonstrates the forecasted trends for the upcoming months based on the Link Prediction of entities.

Overview: This demonstrates the architecture of our current study. We have also attached the link to the paper in that tab.

Extra Features: This tab cumulates two features, "Text Summarisation" and "Word Algebra". Text summarisation allows text summary of keywords from the abstract of literature. It highlights important points related to searched keywords from the extensive corpus of abstracts. The "Word Algebra" facilitates the linking of dimensional space based on vector algebra. It instantiates an intuition related to the vector space of the corpus for which the language model has been trained.

4.3 Results

46885 distinct diseases and 53375 unique chemicals were identified and top entities are shown in Figure 4.2 (C, D). Anxiety, depression and hypertension were found to be present in top-20 most discussed medical conditions in the research articles. Oxygen and Hydroxychloroquine were followed by nucleic acid and Angiotensin, a peptide hormone that causes vasoconstriction, in the most discussed chemicals. The latent space of word embeddings around the keyword 'post-covid syndrome' visualized using t-SNE plot (Figure 4.2 B) depicted 'chronic fatigue', 'debilitating', 'neurodegenerative disorders' and 'vascular complications' among the closest medical entities in terms of cosine distance.

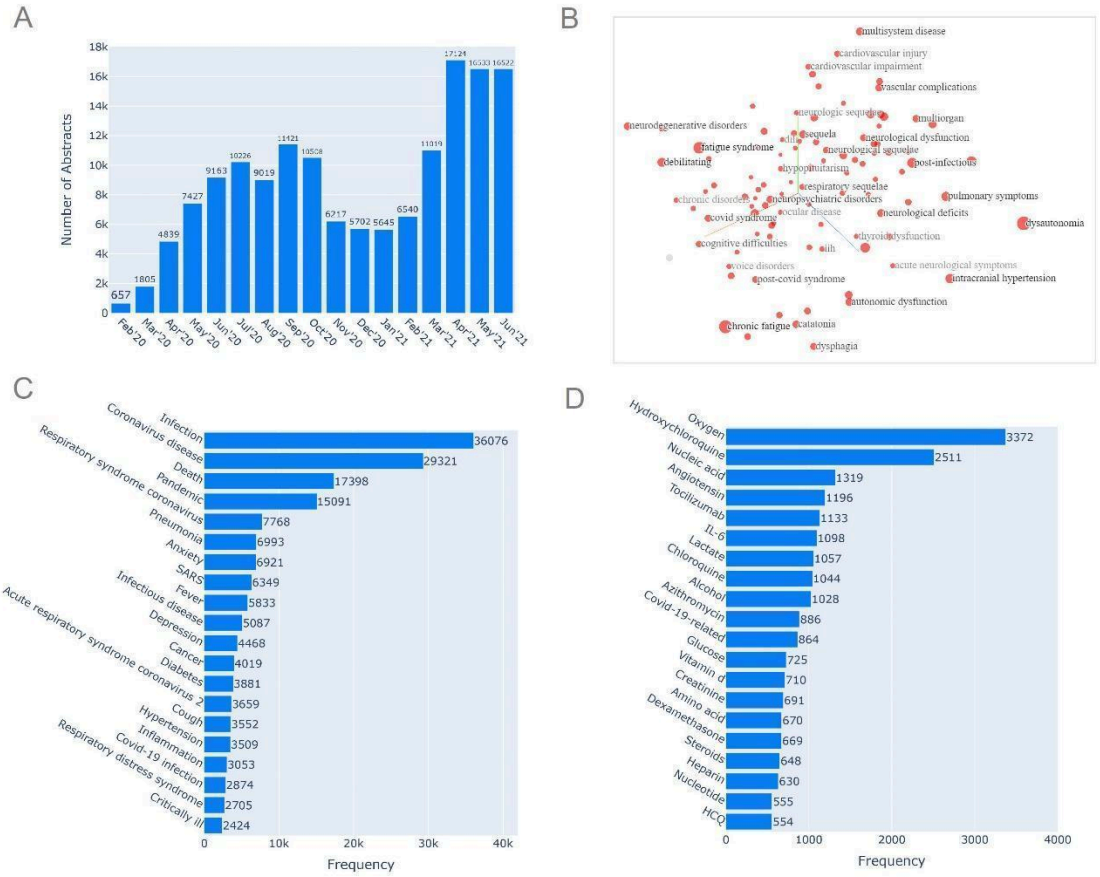


Figure 4.2: *(A) Showing the number of articles occurring each month. The curve depicts that there has been a rampant increase in the number of articles across each month since February 2020. (B) Latent space of word embeddings of diseases visualized around the keyword 'post-covid syndrome', displaying 100 isolated points nearest to it. (C) Bar plot (left) showing the frequency of top diseases in the corpus of abstracts extracted using NER. (D) Bar plot (right) showing the frequency of top chemicals in the corpus of abstracts extracted using NER.*

We conducted detailed inference of the alluvial diagram across different months to graphically explore the temporal trends in the literature based on dynamic and homogeneous networks of prevalent medical entities and their associated cosine similarities. Figure 4.3A represents the flow of themes found in the literature published in 2020. For March 2020, the dominant themes noted were chest pain, acute kidney injury, and lymphocytopenia. While there were lesser traces of thromboembolic complications in literature of early months, it emerged as the most significant theme in August 2020 (Figure 4.3A). Myocardial injury and cardiovascular diseases surfaced as a crucial cluster of entities in December 2020. Mental health factors such as depression, loneliness, anxiety and burnout gained significance in the literature of the last quarter of 2020. Figure 4.3B represents the flow of themes found in the literature published in 2021. While thromboembolism, hypoxemia and myocardial infarctions remained major concerns till January

2021, a significant transition towards Long COVID symptoms was found to be a major theme in March 2021. In June 2021, central modules including post effects and neurological complications, stroke, headache and anosmia were found to gain importance, along with newer themes around immunocompromised and chronic diseases. Cross infection related entities gained focus due to the second wave of COVID-19 cases in multiple countries around the world. The importance of mental health effects transitioned from lesser importance in the first quarter to more emerging and prominent links in the second quarter as highlighted in the alluvial diagram.

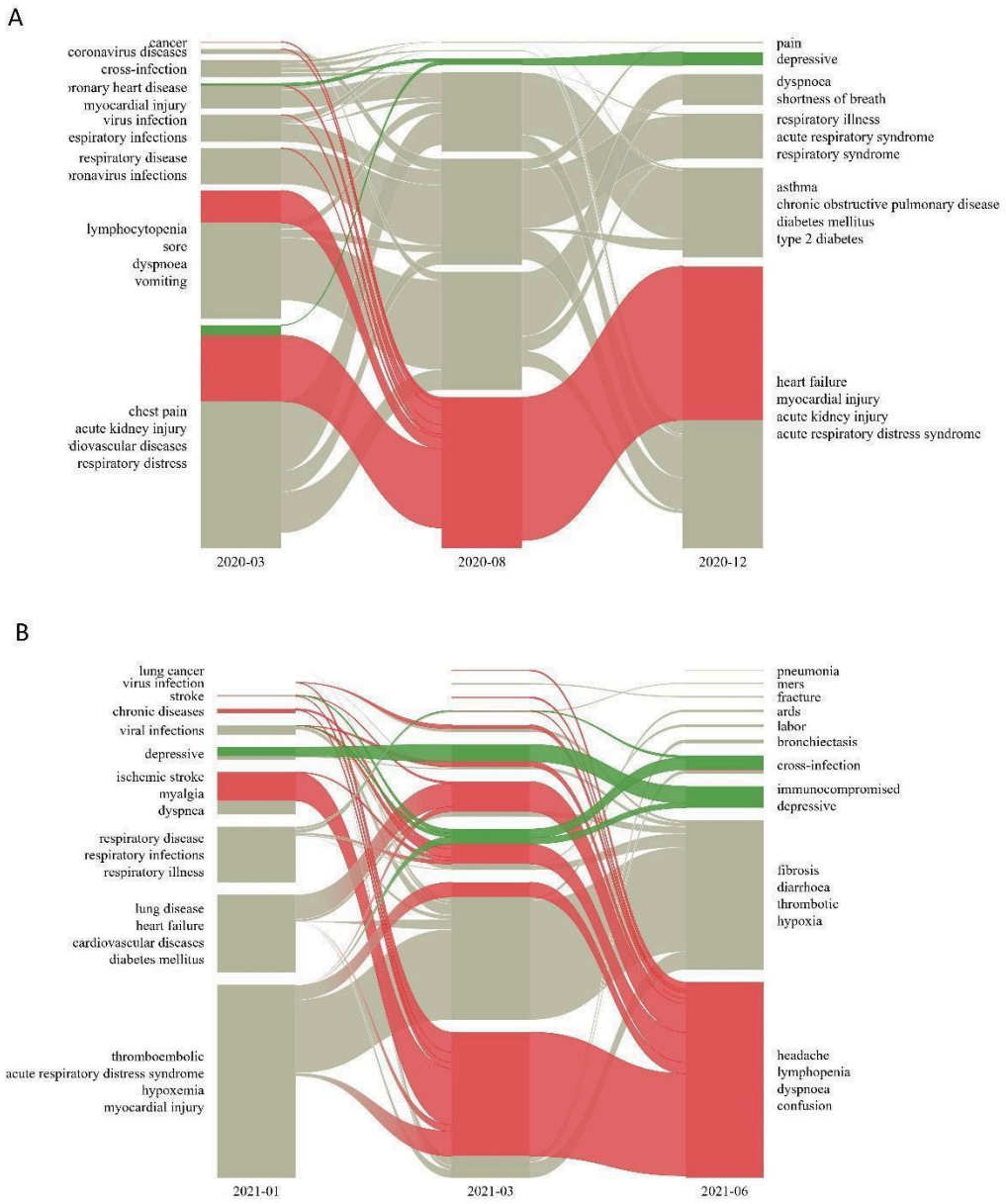


Figure 4.3: *(A) Alluvial diagram for tracking the trends in 2020, from the networks of March, August and December. (B) Alluvial diagram for monitoring the trends in 2021, from the networks*

of January, March and June. The alluvial diagram eases tracing the temporal dynamics of literature across different time intervals.

We further advanced the analysis of trends to predicting links between entity pairs for the upcoming months. Our proposed framework for Temporal Link Prediction effectively forecasted five proximity scores, including semantic and topological measures, between node pairs by modeling its time series using the ARIMA model. Mean Squared Error in the predictions of each proximity score for April 2021, May 2021, and June 2021 was shown in Figure 4.4A. The associations between diseases for the successive month were predicted as links using supervised learning based on dynamic networks belonging to the previous months. Our results show that among the four classifiers (Table 4.1), the AdaBoost model with 50 estimators and learning rate of 0.1 classified links with a mean AUROC score of 0.871 (all P<.001, statistically significant at a Bonferroni-corrected significance level of .0167) in the test data of June 2021 (Figure 4B, 4C). The predicted links weighted by forecasted cosine similarity showed a high intersection with the original modules, hence validating the proposed architecture (Table 4.4). Table 4.2 shows the clusters detected in the original network vs the predicted network. The ARIMA model was used for forecasting proximity scores for subsequent months based on the trends in node pair proximity measures retrieved from the prior months (Feb 2020 - June 2020). Our findings suggest that themes of predisposing conditions and risk factors, studies on cross-infection and neuro-psychiatric manifestation will assume a higher centrality in the upcoming quarter of 2021 (Table 4.3).

Model	**April 2021**		**May 2021**		**June 2021**	
	AUC ROC	**Acc.**	**AUC ROC**	**Acc.**	**AUC ROC**	**Acc.**
Random Forest	0.58 ± 0.0014	0.77 ± 0.0008	0.74 ± 0.0012	0.70 ± 0.0009	0.80 ± 0.0013	0.74 ± 0.0008
Support Vector Machine	0.51 ± 0.0017	0.75 ± 0.0009	0.79 ± 0.0014	0.79 ± 0.0009	0.85 ± 0.0012	0.86 ± 0.0007
ADABoost	0.52 ± 0.0016	0.74 ± 0.001	0.81 ± 0.0011	0.77 ± 0.0009	0.87 ± 0.0009	0.83 ± 0.0008
XGBoost	0.58 ± 0.0015	0.65 ± 0.0009	0.79 ± 0.0012	0.75 ± 0.0008	0.84 ± 0.0011	0.83 ± 0.0008

Table 4.1. *Results of temporal link prediction between entities for the month of April 2021, May*

2021, and June 2021, with a margin of error for 95% Confidence Intervals. The mean value of metrics has been recorded by testing the models on a resampled test set.

Module ID	Top nodes in Predicted network	Module ID	Top nodes in Original network
1	headache, lymphopenia, dyspnea, confusion, encephalitis, nausea	1	vomiting, nausea, headache, diarrhea, dyspnea, lymphopenia
2	fibrosis, coagulopathy, thrombotic, hypoxia, inflammation, delirium	2	Fibrosis, myocarditis, coagulopathy, hypoxemia, thromboembolic, shock, sepsis
3	comorbidity, asthma, COPD, hypertension, dementia, diabetes	3	Confusion, immunocompromised, traumatic, panic, scaly, cross-infection, cancer
4	traumatic, anxiety, depression, loneliness, burnout, insomnia	4	Comorbid, asthma, COPD, diabetes, hypertension, obesity
5	immunocompromised, chronic diseases like tuberculosis	5	Anxiety, traumatic, depressive, depression, burnout, panic, insomnia, anger

Table 4.2. *Community detection results from the predicted and actual network for June 2021, a subset of nodes from different modules have been shown for both predicted and actual networks.*

Module ID	Theme	Subset of nodes from different modules
1	Adverse events (predisposing conditions and risk factors)	Myocarditis, coagulopathy, thromboembolic, hypoxemia, fibrosis, respiratory distress, immunocompromised
2	Symptoms (complications and symptoms of diseases)	Lymphopenia, dyspnoea, vomiting, diarrhea, dyspnea, nausea, headache, myalgia, anosmia

3	Respiratory illness	Respiratory infections, respiratory illness, respiratory infection, respiratory disease, coronavirus infection, mers
4	Neuro-psychiatric manifestation, Cross-infection	Confusion, psychiatric, cross-infection, trauma, pain, panic, labor, cross-infection, contagion
5	Psychological conditions	Traumatic, depression, depressive, anxiety, burnout, insomnia, psychological distress

Table 4.3. *Results of community detection from the predicted subsequent network based on training data till June 2021. A subset of nodes was mentioned, which broadly signifies a theme for the given module.*

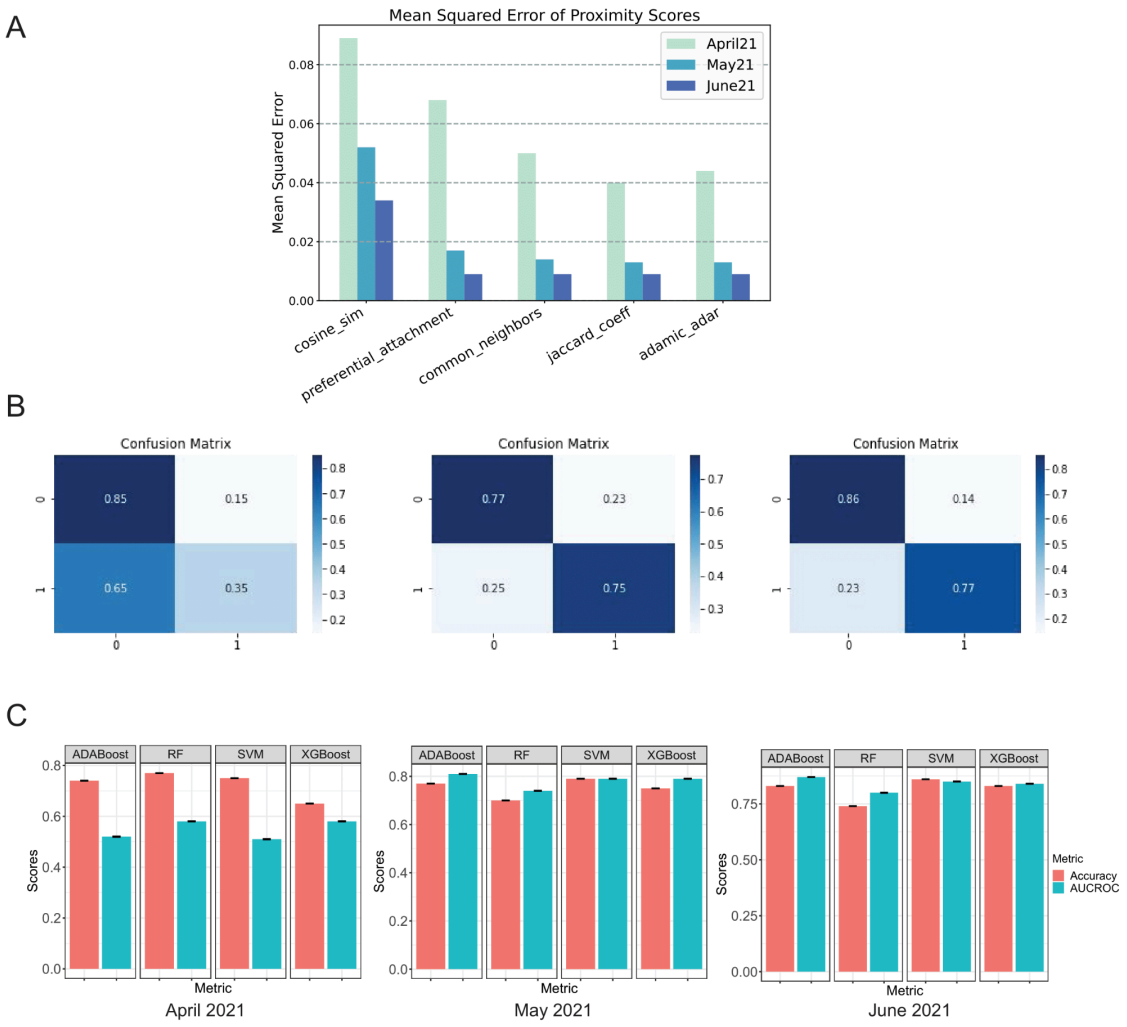


Figure 4.4: *(A) Evaluation of Mean Squared Error (MSE) between original and predicted proximity scores for the network of April 2021, May 2021, June 2021. (B) Confusion Matrix with normalized values of results from AdaBoost classifier across the months of April 2021l, May 2021, and June 2021. AdaBoost has been the best-performing model across all three months. (C) Results of link prediction between disease entities from March 2021 to June 2021, with a margin of error for 95% Confidence Intervals. The mean value of metrics has been recorded by testing the models on a resampled test set.*

The intersection of nodes between predicted and original modules was analyzed to prospectively validate the effectiveness of the proposed prediction framework. Table 4.4 depicts the top nodes in the different modules along with their respective IOU scores for January and June 2021. The collection of intersecting nodes have been interpreted to represent broad themes. Organ damage like acute kidney injury and pulmonary embolism associated with COVID-19 was the most central theme in literature from January 2021, followed by cardiovascular diseases, respiratory infections and psychological effects. Interestingly, major themes in June 2021 shifted towards

conditions related to Long COVID and neurological symptoms. Headache, encephalitis and confusion were predicted to be the central nodes, and showed a high IOU score when compared with the original network. A subset of nodes belonging to different modules from both predicted and true networks have been demonstrated in Table 4.2.

Module ID	**January 2021**		**June 2021**	
	Top Nodes	***IOU***	***Top Nodes***	***IOU***
1	acute kidney injury, ARDS, coagulopathy, myocardial injury, pulmonary embolism	0.45	headache, lymphopenia, dyspnea, confusion, encephalitis, nausea	0.71
2	cardiovascular disease, diabetes mellitus, COPD, hypertension	0.66	fibrosis, coagulopathy, thrombotic, hypoxia, inflammation, delirium	0.70
3	respiratory infection, MERS, respiratory diseases	0.55	comorbidity, asthma, COPD, hypertension, dementia, diabetes	0.64
4	depression, insomnia, anxiety, loneliness	0.71	traumatic, anxiety, depression, loneliness, burnout, insomnia	0.81
5	myalgia, lymphopenia, headache, anosmia, dyspnoea	0.43	immunocompromised, chronic diseases like tuberculosis	0.33

Table 4.4: *Clusters or Modules of diseases from the predicted network of January 2021 and June 2021. The given Intersection over Union (IOU) was computed between clusters of predicted and original networks of the respective months. A subset of top intersecting nodes in each cluster is mentioned, which collectively signify themes.*

Analysis of networks constructed upon chemical entities revealed the evolution of various drugs studied in the COVID-19 literature. During February 2020, the major module contained entities such as paracetamol, tofacitinib, thalidomide, vitamins, zinc and other linked chemicals. Another relevant module included central entities such as doxycycline, ruxolitinib, heparin and ivermectin, which were discussed in the scientific research on treatment and prevention of COVID-19. In contrast, our recently updated models showed the emergence of evidence for various immunosuppressive drugs such as Tacrolimus and anti-inflammatory drugs such as Glucocorticoids and Colchicine during November 2021 (Figure 4.5). These relatively less important entities in earlier months started to become more prominent as the literature expanded. Evidence around ‘statins’ also gained centrality over recent months. Our findings show that the proposed framework captures the dynamic changes in the importance of entities based on their evolving relationship with neighboring entities.

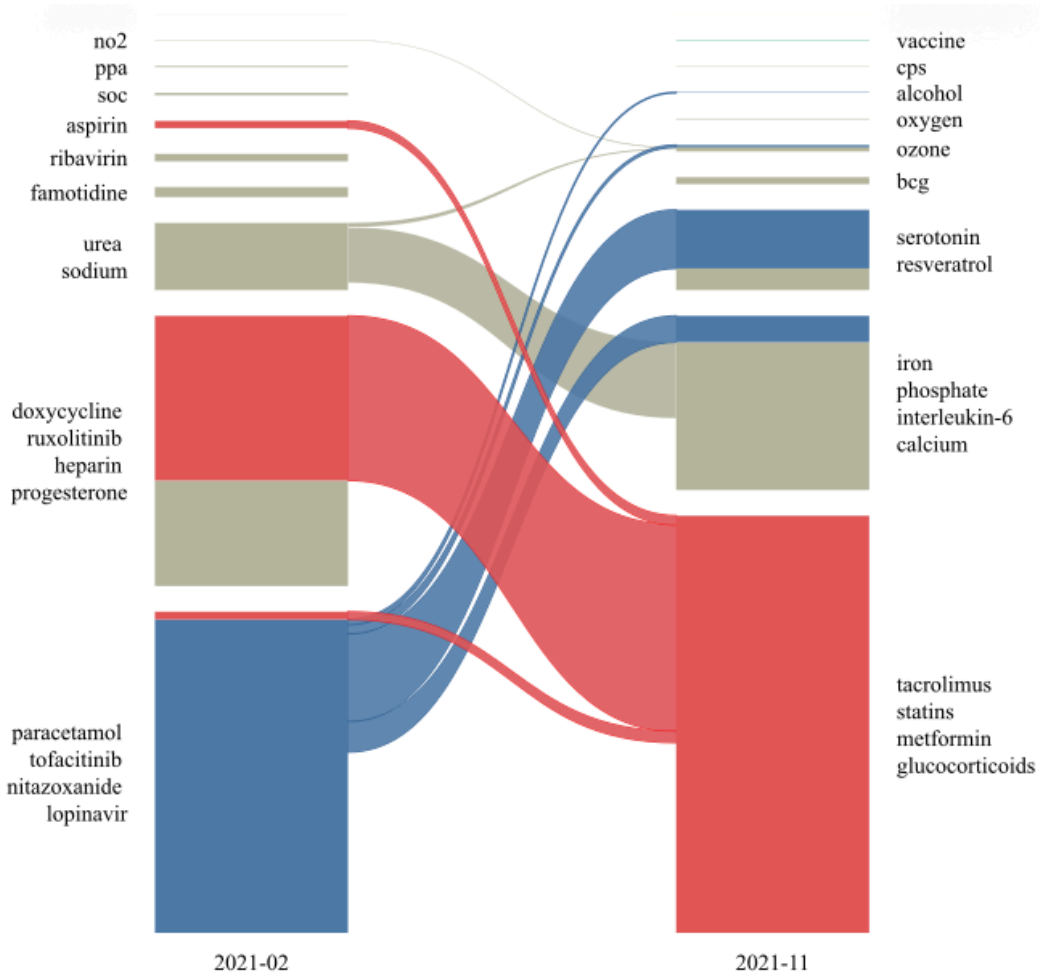


Figure 4.5. *Alluvial diagram for tracking the trends of chemical entities from the networks of February 2020 to November 2021. Threshold used for assigning links between nodes was set at 70th percentile of cosine similarity between pairs of top-100 entities in respective months.*

4.4 Discussion

4.4.1 Principal Findings

In this study, we demonstrate a computational approach, EvidenceFlow, in which a user will interact with the rapidly expanding COVID-19 literature to derive and predict emerging themes. The proposed framework tracks patterns of changing semantic and topological proximity between entity pairs across months. Further, it predicts links and network communities that may emerge in the future months. Hence users can follow the papers that contribute to emerging communities of themes, e.g. literature around thromboembolic complications which was captured as early as August 2020 and mental health factors during the end of 2020. Interacting with the clusters on the interactive interface of the EvidenceFlow model revealed that symptoms of Long-COVID, such as fatigue, headache, myalgia, cough, and anosmia, were found to be

forming a central cluster during March 2021. This early signal for accumulating evidence was later validated in large prospective and retrospective cohorts of COVID-19 patients (Taquet et al. 2021; Lopez-Leon et al. 2021; Blomberg et al. 2021). Another way in which users can interact with EvidenceFlow is to gain an understanding of the evolution of themes that goes beyond the current approaches such as topic modeling and sentiment tracking (Ebadi et al. 2021). An example is the early finding of imminent themes around neurological complications, such as confusion, psychiatric illness, stroke, and mental health factors such as anxiety, depression, PTSD, burnout and insomnia, in June 2021. Our violin plot analysis (Figure 4.6) showed that despite the mean error being centered on zero, there were some outlier node pairs whose predicted associations deviated from the ground truth. Future scope of this work will involve an analysis of such associations and insights gained by an interactive analysis of such pairs on the *EvidenceFlow* application.

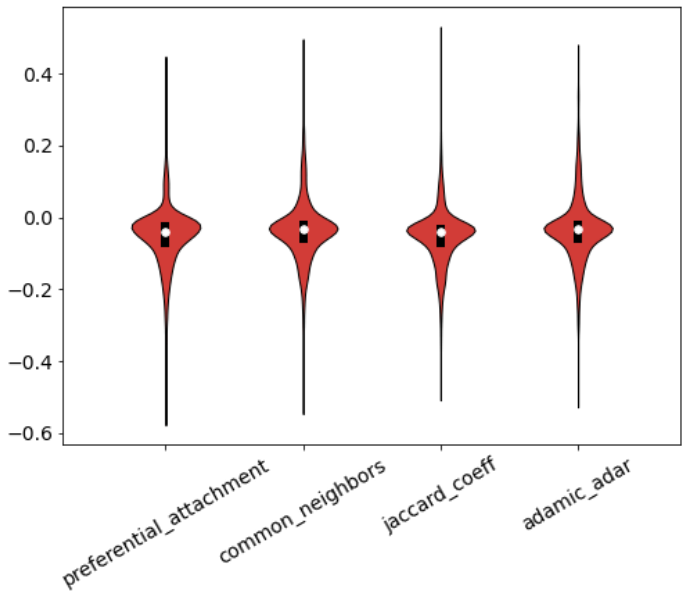


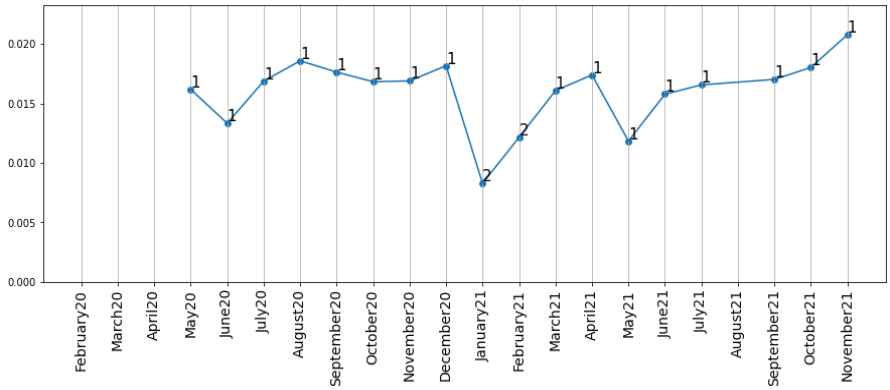
Figure 4.6. *Distribution of errors in the prediction of proximity scores between node pairs (used as features in model training) for the month of June 2021. The white marker depicts the median and the black marker depicts the interquartile range. The mean of errors were found to be close to zero.*

Prediction of the themes represented by rising centrality of entities can assist in formation of promising research hypotheses. The dynamics of literature reveal the emergence of central themes as a combination of pre-existing themes in recent times (Tshitoyan et al. 2019). For example, the alluvial diagram (Figure 4.3A) demonstrated how entities from multiple modules in March 2020 merged into a major cluster of thromboembolic complications. Similarly, the flow of importance of psychological disorders over the months indicate their contemporary relevance in the COVID-19 literature and their links with other entities in the cluster. Our framework can

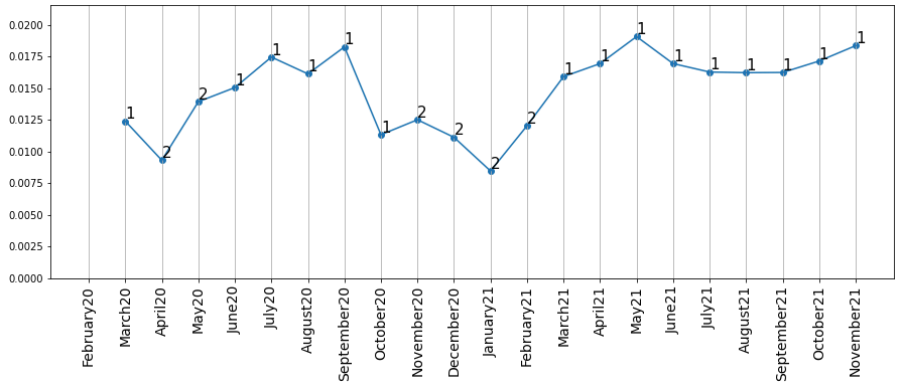
potentially help researchers in monitoring the existing themes and directing their studies based on the trends and predictions.

We conducted an analysis on the trends of PageRank centrality of selected chemical and disease entities. Statins, a class of lipid-lowering medications, were found to be gaining centrality in late 2021 as compared to earlier values (Figure 4.7A). Numerous studies discussed statins for having anti-inflammatory and immunomodulatory effects that may reduce the severity of COVID-19 (Daniels et al. 2021; Peymani et al. 2021). Glucocorticoids, a class of steroid hormones that reduce inflammation and suppress the immune system, also emerged as an entity with rising entity Figure 4.7B). Depression and other mental health disorders started becoming a prominent topic of research during the middle of 2020 and gained higher importance in subsequent months (Figure 4.7C). COVID-19 has also been largely discussed in the context of a thromboembolism and our model captured its emerging evidence as a theme till late 2020. However, the trends showed that its centrality in the literature relatively decreased in 2021 (Figure 7D). Discovering such trends from a large corpus is indeed possible using manual curation and analysis by experts. However, our EvidenceFlow pipeline provides an efficient lens to discover, track and predict emerging trends. This framework will enable faster synthesis of evidence, which then can be validated by experts.

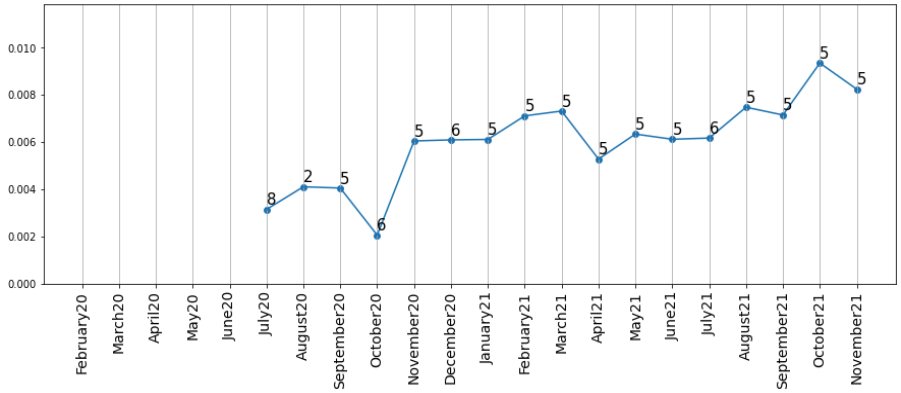
A



B



C



D

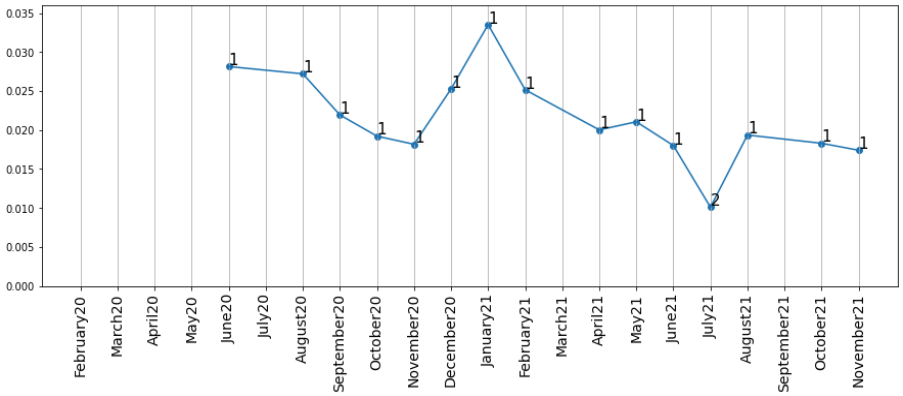


Figure 4.7. *Temporal trends of PageRank centrality of (A) 'statins', (B) 'glucocorticoids', (C) 'depressive', (D) 'thromboembolic'. The annotations denote the module/cluster index (lesser meaning more central).*

To explore the potential of unsupervised word embeddings and changing cosine similarity among words, we analyzed the trends of terms having maximum similarity with selected keywords. For example, we analyzed the temporal shift in the context of "vaccine" over the months by finding

the top-10 terms most similar to *vaccine* in the latent space of word embeddings trained on the abstracts from each month (Figure 4.8). From February to August 2020, the research on COVID-19 vaccines was underway and the studies revolved around "therapeutics", "prophylactics", "drug-repurposing" and associations with MMR (measles-mumps-rubella) vaccine and BCG (Bacillus Calmette–Guérin) vaccine. As the clinical trials of certain vaccine candidates became prominent after August 2020, a theme of vaccine *hesitancy* emerged in October 2020 and gained higher similarity in subsequent months. Additionally, as the literature evolved in 2021, a wide range of COVID-19 vaccines such as BNT162b1, Pfizer-BioNTech, AstraZeneca, ChAdOx1, mRNA-1273 or Moderna were found to be majorly discussed in the context of research on vaccines. Terms such as *immunogenicity* and *efficacy* further suggested high association with vaccine trials and rollouts. Recently updated models showed the emergence of 'booster' doses from August 2021 onwards. Such retrospective evaluation of development of evidence from literature over time can assist the research community in deriving detailed insights leveraging the applications of word embeddings.

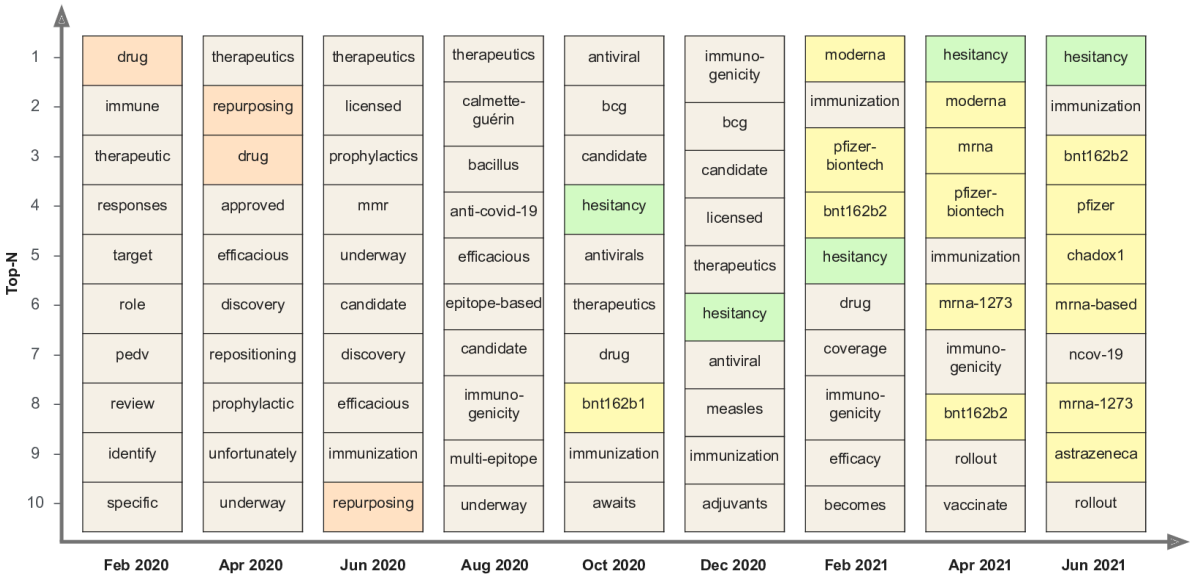


Figure 4.8: *Temporal evolution of the context of the term "vaccine" across alternate months. Top-10 most similar words based on cosine similarity using monthly Word2Vec embeddings are plotted. Origin and evolution of drug repurposing in early months, hesitancy and vaccine candidates in later months are highlighted.*

4.4.2 Limitations

Our study has some limitations. Firstly, although the WHO database is built using a detailed search strategy for COVID-19 literature, it does not explicitly report the exact purpose or

accuracy of the search and decision process. The documentation mentions screening done by expert reviewers and an attempt to remove duplicates, but further details are lacking. For example, the process doesn't clarify if redundancy across various publishers was taken care of. Further, the frequent use of 'OR' combination of keywords may have led to inclusion of less relevant articles, while other forms of literature, such as patent applications, which can add value to the study were not included in this database. Nonetheless, we chose the WHO COVID-19 database as it provides a large collection of articles that are updated regularly from searches of multiple bibliographic databases ("Global Research on Coronavirus Disease (COVID-19)," 2023.). This, combined with curated expert-referred scientific articles which wouldn't be readily accessible on a custom search, was useful for building the EvidenceFlow pipeline. Future work with this framework will include potential extension to databases curated both through generic queries and expert vetting, thus facilitating targeted evidence synthesis from a variety of databases.

Further, we are currently using abstracts of research articles to extract named entities and may be missing on the details contained in the full-text of the article while training word embeddings. Therefore, future work may build upon the framework to include the full text of articles and full text, wherever available. The NER model used in our study has been reported to achieve an F1 score of 84.49% on a benchmark dataset (Cho and Lee 2019). Despite the limitations of F1 score such as equal weightage given to precision and recall (Hand and Christen 2018; Powers 2020), F1 remains one of the most widely reported performance indicators. We chose this metric in the absence of other metrics reported for this NER model. For forecasting, we utilized a relatively basic model such as an AR approach as our goal was to capture robust patterns, however, further research is possible for the use of more complex time-series approaches with higher order difference and lags. Moreover, as the number of timestamps and data points increase, advanced architectures such as RNN and LSTM (Sherstinsky 2020; Greff et al. 2017) can be used for handling complex trends in the time series efficiently. Further experiments with larger networks can reveal themes that were not found with top-100 entities. Importantly, our model is supporting early detection of emerging trends, but it cannot capture themes on which no evidence has been accumulating.

4.4.3 Conclusion

The global attention has led to widespread increase in the scientific literature to study and prevent the disease from spreading, hence understanding the disease from multiple perspectives. We introduced a framework built upon COVID-19 specific literature vetted by the WHO and deployed as a dashboard called EvidenceFlow. The dashboard allows the user to unravel the literature with an interactive map of embeddings based on the visualization provided by Tensorboard. It aims to track literature trends using alluvial diagrams, multi-level community detection, and projection of influential entities through network analysis across different months. This study presented how machine learning-based prediction of emerging links can contribute

towards analyzing research by capturing themes represented by groups of medical entities, based on patterns of semantic relationships over time.

Hence, in this chapter another application of the language model has been demonstrated which involved mining literature related to COVID-19 disease. It highlighted how language models can be implemented to decipher the emerging research themes from literature based on the temporal semantic drift. It also helped in identifying trends and patterns in literature while guiding future research directions. This chapter addressed how language models can act as a potent tool to significantly advance our understanding of a particular disease and guiding decisions about public health policies.

EvidenceFlow, the proposed framework can be extended to several other diseases such as Tuberculosis, Sepsis where rapid progression is made in scientific literature at regular intervals. This would effectively allow us to understand the emerging trends and themes related to the disease based on the contextual changes hidden in the vast amount of available unstructured data. Expanding the scope of my research, I explored another important application of language models in the healthcare domain. In recent years, there has been a surge in the application of language models for clinical prediction in clinical settings. One particular use case is in the analysis of textual data such as physician notes and nursing notes in the early detection of critical outcomes such as Shock and Sepsis. These conditions are time-sensitive and require prompt intervention in order to prevent complications and improve patient outcomes. Language models are used to extract meaningful features from clinical notes which are given as input to the machine learning models to predict such critical outcomes. In the subsequent chapter, I will delve into the implementation of these language models in clinical prediction for critical care settings. Various techniques and approaches used in the development of these models as well as the challenges that come with implementing them in real-world settings have also been discussed.

Abbreviations

AUROC: Area Under Receiver Operator Curve

IOU: Intersection Over Union

ARIMA: Autoregressive Integrated Moving Average

NER: Named Entity Recognition

MSE: Mean Squared Error

Chapter-5

Language models for clinical prediction in ICU: A Multimodal Model for Prognosticating Intensive Care Outcomes from Physician Notes and Vitals

5.1. Introduction

Language models are showing promise as general purpose learners for predictive modeling and information retrieval from clinical text (L. Zhou, Suominen, and Gedeon 2019; Luo et al. 2022). In this chapter, we extend the scope of our research to explore the use of language models in predicting critical outcomes in critical care settings. In this chapter, I will discuss one such implementation of language models that have been used to predict critical outcomes and its impact on the medical community. Data and predictive models are being used increasingly to augment decision making in Intensive Care Units (ICUs). Language models have been extensively employed in the critical care domain to facilitate clinical predictions, such as inpatient mortality, 30-day readmission, prolonged hospital stay, ICU transfer, and critical care outcomes (Steinberg et al. 2021). In addition to these predictions, language models have shown promising results in predicting critical outcomes such as Shock, Sepsis, and Acute Kidney Injury (Sun et al. 2019; Vats et al. 2022; Wardi et al. 2021). In this chapter, I focus on a specific use case of predicting Shock Index abnormality in critical care settings. Shock index (SI) which is defined as the heart rate (HR) divided by systolic blood pressure (SBP) serves as a clinical outcome predicting tool for bedside assessment, risk stratification and prognostication (Koch et al. 2019; Cheng et al. 2020). The use of shock index has been examined in predicting outcomes in emergency and low resource settings for various conditions including sepsis, hemodynamic shock, acute myocardial infarctions, stroke, advanced cancer, pulmonary embolism, pneumonia and trauma (Cheng et al. 2020; Maheshwari et al. 2020; B. Huang et al. 2014). These outcomes in their early stages are more likely to respond to therapy, whereas once irreversible end-organ damage sets in, chances of mortality are high (Shoemaker 1996). Thus, prediction and early identification of deterioration in shock index can support timely life-saving interventions (Bonanno 2011; Seymour et al. 2016).

Early warning signals to predict physiological decompensation in ICU patients can trigger early interventions, improve outcomes and potentially save patient lives in the ICUs (Adams et al. 2022; Ehwerhemuepha et al. 2021; Kennedy and Rudd 2022; Henry et al. 2022). The widespread adoption of electronic medical records (EMRs) has accelerated the development of early warning systems by making large volumes of data available, potentially on a real-time basis. Accessibility to EMRs enables clinicians to make use of numerous modalities of information such as patient history, vitals, laboratory investigations, treatment response and other diagnostic modalities to evaluate patient status in the ICU. This assists in designing risk scores such as the Sequential Organ Failure Assessment (SOFA) score for diagnosing sepsis-related organ failure. The scoring

criteria often include heterogeneous and non-simultaneous measurement such as vitals, clinical evaluation and laboratory investigations, typically not available simultaneously in real-time. Moreover, most of these scoring criteria require subjective clinical assessment and have changing definitions which make these difficult to be consistently implemented. While sepsis remains one of the most commonly modeled outcomes in the ICUs the nebulous definition of sepsis and the related scoring criteria limits the utility of models that have been developed using a combination of structured, unstructured and vitals data (Hammoud et al. 2020; Amrollahi et al. 2020; Goh et al. 2021). Most of these studies concluded that there is an additional value of including unstructured clinical data to improve the prediction of models (Amrollahi et al. 2020; Goh et al. 2021).

Data-driven scoring approaches in the ICUs, on the other hand, allow continuous scoring using real-time vitals that can be combined with daily lab investigations and clinical notes. The utility of Shock Index extends beyond the clinical evaluation of Shock, although that is the most frequently modeled use case. Previously, thermal imaging (Nagori et al. 2019) and time series data from the vitals (Nagori et al. 2021) has been used to predict abnormal shock index across age, ICU specialties and geographies. Similarly, a multivariate combination of vitals was used to create machine learning models for shock prediction (Desautels et al. 2016; Gultepe et al. 2014; Potes et al. 2017). However, none of these studies has addressed Shock Index using a multi-modal approach. The current study addresses this gap and combines structured and unstructured data present in EMRs for prognosticating deterioration in the Shock Index. We hypothesized that (i) physician notes encapsulate the sum total of underlying diseases (e.g., diabetes), medications, and other indicators of infection, e.g. fever, use of antibiotics, etc. These information can be mined into features using representation learning with the help of language models. (ii) The strong association of vital signals, such as heart rate and BP with Shock index makes this an important feature for abnormal shock prediction.

The objectives of this study were (i) Model Shock Index as an objective indicator of outcomes, (ii) Assessment of improvement, if any, in model performance and clinical interpretability of the multi-modal models as compared to vitals or physician notes alone, (iii) Efficacy of model performance for diagnosing diseases across clinically segregated cohorts. To this end, we created *ShockModes*, an end-to-end multi-modal pipeline by fusing vitals signals with medications and contextual information such as history of present illness (HOPI), for 24-hour ahead prediction of abnormal shock index (Figure 5.1).

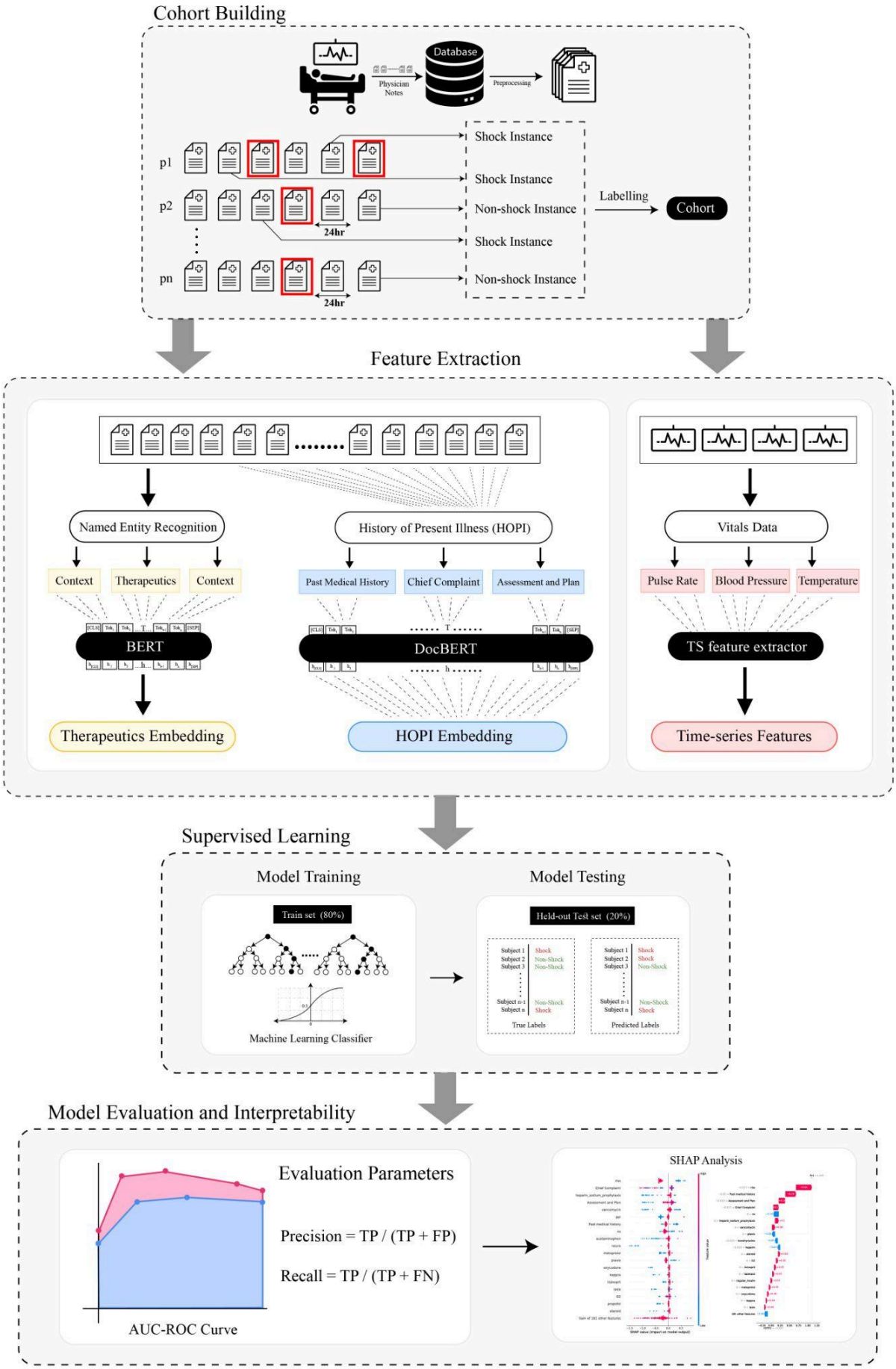

Cohort Building
Database
Physician Notes
Preprocessing
p1
p2
pn
24hr
24hr
Shock Instance
Shock Instance
Non-shock Instance
Shock Instance
Non-shock Instance
Labelling
Cohort
Feature Extraction
Named Entity Recognition
Context
Therapeutics
Context
BERT
Therapeutics Embedding
History of Present Illness (HOPI)
Past Medical History
Chief Complaint
Assessment and Plan
DocBERT
HOPI Embedding
Vitals Data
Pulse Rate
Blood Pressure
Temperature
TS feature extractor
Time-series Features
Supervised Learning
Model Training
Model Testing
Train set (80%)
Machine Learning Classifier
Held-out Test set (20%)
True Labels
Predicted Labels
Model Evaluation and Interpretability
Evaluation Parameters
Precision = TP / (TP + FP)
Recall = TP / (TP + FN)
AUC-ROC Curve
SHAP Analysis

Figure 5.1: Overview of ShockModes pipeline. ShockModes takes physician notes and vitals data as input, clinical entities (Therapeutics & History of Patient Illness) are extracted using MED7 Named Entity Extraction(NER) whereas Time-Series (TS) features are extracted using tsfresh python module. Unsupervised Embeddings fused with TS features serve as input features for the machine learning models for predicting the onset of abnormal SI with 24-hours lead time. SHAP analysis reveals interpretable features for global and patient-specific notes.

5.2. Methods

5.2.1 Dataset

Our study has been conducted using the publicly available Medical Information Mart for Intensive Care (MIMIC) III v.1.4 database (Johnson et al. 2016). The MIMIC-III database provided critical care data for over 40,000 patients admitted to intensive care units at the Beth Israel Deaconess Medical Center (BIDMC). The Noteevents data contained textual information of physician notes corresponding to different categories. We curated this clinical database to create 24-hour patient encounter cohorts of vitals and text data from physician notes. We have also extracted the patients' heart rate (HR), systolic blood pressure (SBP), respiratory rate (RR), and oxygen saturation (SpO_2) for every 24 hours recorded in the critical care setting.

5.2.2 Cohort Construction

Shock index was defined as the heart rate (HR) divided by systolic blood pressure (SBP), and has been used for assessment of hemodynamic instability in suspected shock patients. Continuous vitals data corresponding to each patient was summarized at 1-minute resolutions and utilized to calculate the SI. Epochs of 30-minute length were generated. The onset time of abnormal shock index was defined as the starting time of an epoch in which the median shock index value was greater than 0.7. The corresponding abnormal shock index labels were calculated using a cut-off value of 0.7 for the shock index ($SI \geq 0.7$ corresponds to abnormal shock index, $SI<0.7$ corresponds to normal shock index and NA corresponds to missing label) (B. Huang et al. 2014; Seymour et al. 2016). Days for which the final label was missing were discarded. A new instance of abnormal SI was defined as a >30 minutes length episode of abnormal SI preceded by at least 24 hours of normal SI. Only new instances of abnormal SI were included in the cohort as ongoing abnormal SI may be a trivial predictor for future abnormal SI. For text, we have considered only the top seven frequencies of physician note types. The seven notes from the physician category included notes from physician residents, intensivists, the physician attending, and other ICU notes which encompassed the majority of the clinical data that could be extracted from the physician notes. Physician notes were mapped using date, subject ID and ICU stay ID to their next-day labels generated using continuous vitals data. Our cohort exclusively for textual data comprised 355 instances of normal SI patients and 87 abnormal SI patients. Our vitals data and multi-modal cohort comprised 337 normal instances and 84 abnormal SI patients. Table 5.1 summarizes the patient characteristics for the extracted

multimodal cohort.

5.2.3 Preprocessing

Text formatting was initially done to extract therapeutics from physician notes using Named Entity Recognition(NER) (Ramesh et al. 2021). The encrypted information ([**word**]) along with single-letter words (example F, M etc), present in the MIMIC data were removed from the text. Text was converted to lowercase followed by standard pre-processing such as removal of whitespaces, stopwords, punctuations, digits, and words with less than or equal to two letters. The misspelled words were rectified using Levenshtein distance and the resulting vocabulary was also vetted by a clinician. While training the language model, the numeric values were considered for minimizing the loss of contextual information present in the form of dosage, frequency, and other critical pieces of information.

5.2.4 Feature Extraction

Lately, I have used the SciSpacy model in the previous chapter to extract entities (diseases and chemicals) from scientific literature. In this chapter, I was required to extract therapeutics from clinical notes. Named Entity Recognition was performed using MED7 (Kormilitzin et al. 2021) to extract therapeutics from physician notes. Therapeutics consisting of more than one word were concatenated with an underscore character.(e.g. metoprolol tartrate => metoprolol_tartrate).

5.2.5 Word2Vec and Transformer based Therapeutics embeddings.

In previous chapters, I have elucidated on techniques to extract embeddings from entities and codons from genomic sequences and scientific literature. In this chapter, I have depicted the process for extracting embeddings of therapeutics and history of present illness (HOPI). A low-dimensional representation (100 dimensions) for the therapeutics present in the corpus of the physician notes was learned using the Word2Vec model with skip-gram algorithm, one-hot encoding, and a fixed window size of five using the Gensim library (Mikolov, Chen, et al. 2013). For Transformer-based models, WordPiece tokenization (Song et al. 2020) was carried out to avoid out-of-vocabulary issues. BioBERT (J. Lee et al. 2020) and BioClinical BERT (Alsentzer, Murphy, Boag, Weng, Jin, et al. 2019) were then tuned using a sliding window of 512 tokens/subwords with the therapeutic entity in the middle of the window. The resulting embeddings for BioClinical BERT and BioBERT were 768 and 1024 dimensional, respectively. Since each word may have multiple embedding vectors based on their context within the document, for each document a mean embedding vector for each therapeutic was generated.

5.2.6 History of Present Illness (HOPI) Embeddings.

To incorporate clinical context we considered i) Chief complaint, ii) Assessment and plan, and iii) Patient history, together known as History of Present illness(HOPI). The chief complaint section constituted the present issues of the patient with suspicion of shock, time of onset, duration, and the progression of the clinic sign & symptoms in detail. It also includes previous

vitals and initial management done(if available) at the time of evaluation during admission. Assessment & plan comprises the provisional management scheduled based on the initial clinical evaluation. Past medical history along with the HOPI contributes to the initial management by identifying risk factors, etiological causes and presenting stages of shock. Each component contains sequence tokens which can potentially exceed the input sequence constraint of 512 tokens hence models that are capable of addressing this challenge were explored (Beltagy, Peters, and Cohan 2020; Mulyar et al. 2019; Pappagari et al. 2019). Low-dimensional representation of HOPI components was learnt using Doc2vec (Q. Le and Mikolov 22--24 Jun 2014) with distributed memory algorithms.

For the Transformer based approach, a DocBERT (Pappagari et al. 2019) was utilized for extracting low-level representation for each of the above-mentioned components. The entire textual input was split into multiple chunks and provided to the model for feature extraction. All the [CLS] tokens from each chunk were aggregated to obtain a vector representing the complete textual input.

A hybrid architecture that combined embeddings generated from medications and HOPI was developed in which the embeddings were fused together to serve as inputs for classification models.

5.2.7 Vital Features

A collection of 3117 time series features (TS features) was extracted using the "tsfresh" Python module, which includes linear and nonlinear physiological characteristics such as autocorrelation, sample entropy, linear trends, statistics, etc (Christ et al. 2018).

5.2.8 Model Development and Validation

The cohort was randomly sampled into a stratified training test split in the ratio of 80:20. All the patient ids with multiple instances were present either in the training or testing dataset. Each drug in the corpus along with HOPI has been considered as a feature, where the mean of these embeddings has been treated as the feature value for the classifier. Words directly indicative of shock were masked. Vitals features combined with mean embedding of therapeutics and HOPI have been considered as input features for *ShockModes*. We had an extensive set of features for the classifier, which would have led to the overfitting of the models and increased the time complexity. In order to reduce features, feature selection was performed using Extra Tree classifiers (Geurts, Ernst, and Wehenkel 2006) present in the scikit-learn library (Pedregosa et al. 2012). Since the dataset was highly imbalanced, we have performed oversampling using SMOTE (Chawla et al. 2002) present in the scikit-learn library. Machine learning models including Logistic Regression, Random Forest, GradientBoost, AdaBoost, and XgBoost (LaValley 2008; Breiman 2001; Friedman 2002; T. Chen and Guestrin 2016; Freund and Schapire 1997) were trained on the vital time series features combined with learned embeddings of therapeutics and HOPI for predicting the onset of the abnormal SI. While implementing the algorithms in the training dataset, we have initialized most of the parameters with the default value, and

hyper-tuned some features based on our task. The algorithms were tuned to have higher sensitivity to be used as a screening approach. Model performance indicators were evaluated using bootstrap sampling for 100 iterations to record the mean value. AUC-ROC score and F1-scores were used to select the best pipeline from each category.

5.2.9 SHAP Analysis for Interpretability.

Shapley Additive Explanations (SHAP) utilize Shapley values to calculate feature importance while explaining the model's output and providing an interpretability interface (Lundberg and Lee 2017).[40]

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z, \qquad (5.1)$$

Where g is the classifier, z' is the simplified features of therapeutics and HOPI, M is the maximum number of simplified features of therapeutics and HOPI, and $\phi_j \in R$ is the Shapley values associated with feature j. The training dataset has acted as prior knowledge from which prior probability is calculated, the model's output probability can then be explained with respect to the prior probability. Specific instances of patients were explained using waterfall plots for depicting the most important features which lead to the prediction of abnormal/normal SI. Global feature importance for the entire dataset has been visualized using the bar plots and beeswarm plots. The Shapley values were calculated for therapeutics and HOPI and provide a window into the inner workings of the classifiers used for predictions.

5.2.10 Implementation of Pipeline.

We have cited the software and algorithms that have been used for developing the novel architecture. Bi-directional transformer models (BioBERT, BioClinical BERT) have been implemented using the Hugging Face library (Wolf et al. 2020). The models were trained using Tesla P100-SXM2 with 512GB RAM, and 16GB Nvidia GPU.

5.3 Results

5.3.1 Exploratory Data Analysis.

The multimodal cohort was constructed from 17,294 ICU-stays in MIMIC-III data comprising 334 normal and 87 abnormal instances of SI. The median of heart rate and respiratory rate for abnormal SI was recorded as 72/min and 18.75/min respectively, greater than the values recorded for normal SI. Other characteristics of the cohort were reported in Table 5.1.

The textual modality was constructed from 2,083,180 clinical notes in MIMIC-III, the categories with the top seven frequencies in the physician notes (141,624) were taken to understand the context of the notes (Figure 5.2). Among the Physician notes, physician resident progress note (62,682) was the most common category followed by intensivist notes. A WordCloud and frequency plots of most(N=20) common therapeutics shows antibiotics, heparin, lasix/furosemide among the most frequently occurring in the data.

Variables	Median (IQR)		Statistical Test
	Normal Shock Index (n=334)	Abnormal Shock Index (n=87)	p-value
Age	69.87 (21.97)	68.99 (18.67)	0.943 (w)*
Length of stay(LOS)	5.16 (11.26)	4.40 (8.77)	0.183 (w)*
Heart Rate, per min	67.31 (10.00)	72 (10.25)	5.82×10^{-07} (w)*
Pulse Rate, per min	67 (11.00)	71 (10.25)	4.41×10^{-06} (w)*
Respiratory rate, per min	18 (4.94)	18.75 (4.00)	0.119 (w)*
Oxygen Saturation	97 (3.00)	97 (3.63)	0.899 (w)*
Gender (Female %)	38.28%	38.1%	0.999 (c)*

*Table 5.1. Multimodal Cohort characteristics. Values are median (IQR) unless indicated, *signifies Wilcoxon (W) Ranksum test (non-parametric) or Student's t-test (t) (parametric) which were used after testing for the normality assumption. c - Chi-squared test of proportions.*

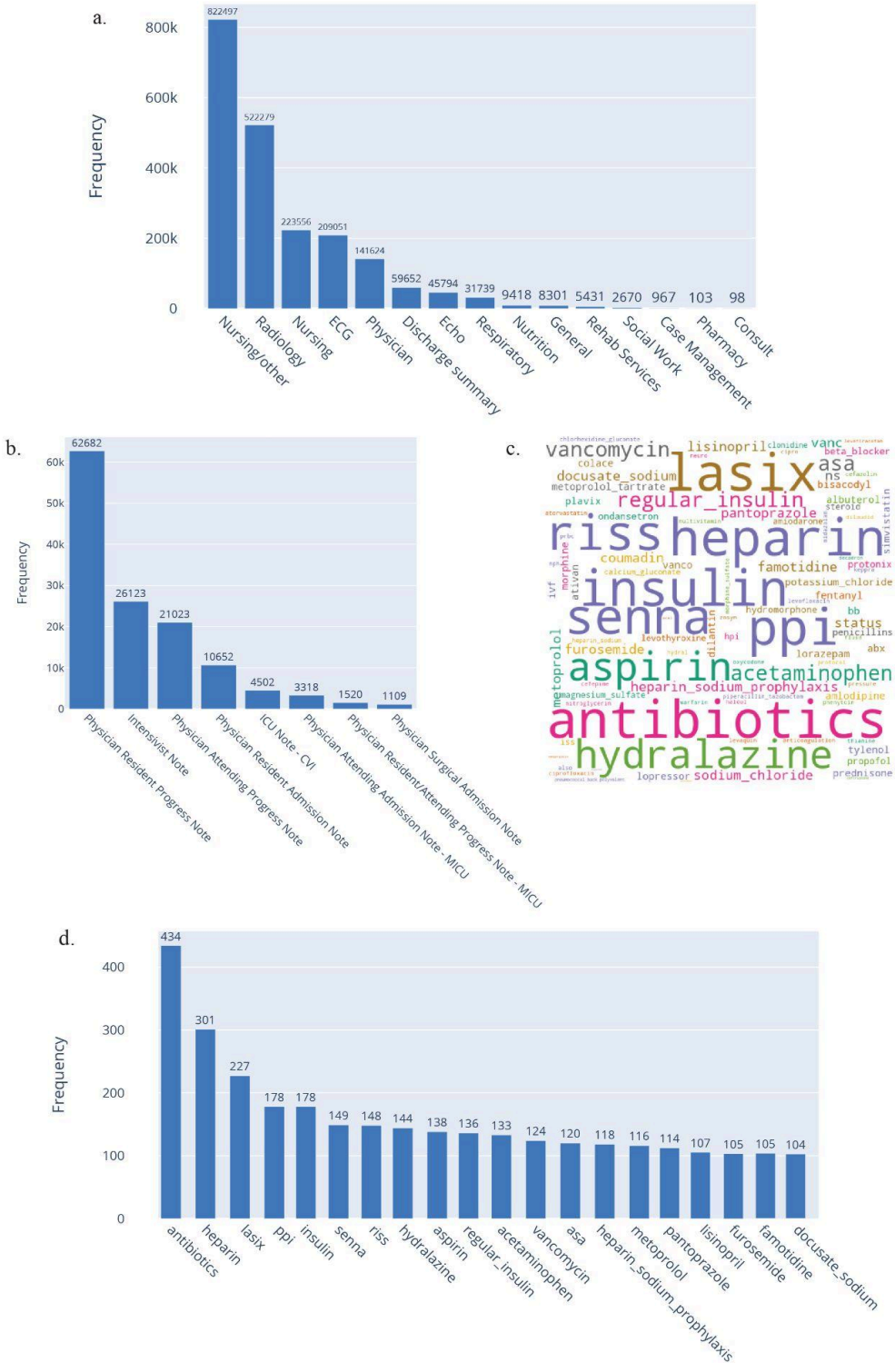


Figure 5.2: a) Plots illustrating the frequency of different categories of treatment notes. b) Frequency of Physician note types. c) Word Cloud of Therapeutics *present in the corpus. d)*

Top-20 most common therapeutics *encountered during periods with abnormal SI.*

5.3.2 Embedding features and model performance

The model performance indicators for the variety of language models developed indicate that there was no single combination that outperformed all the others on all metrics. The transformer based BioBERT and DocBERT embeddings combined with Gradient Boosted model (learning rate of 0.1, n_estimators of 100, maximum depth of 3 and minimum samples split of 2) achieved the best performance based on combined AUC-ROC and F1-score of 0.68 and 0.76 respectively. Taking sensitivity as a metric for screening, Adaptive Boosting trained on Word2Vec and Doc2Vec achieved the highest sensitivity of 81% with F1-score of 74%. While taking specificity as screening metric, Logistic Regression trained on Word2Vec and Doc2Vec achieved the highest specificity (0.84). The high specificity of the models is important to prevent alarm fatigue in the ICU settings, hence emphasized. This depicts that simpler embeddings capture the context within the history of present illness and medications and performed comparably to transformer models. We believe that this may be because of the preliminary step of NER which simplifies the complex task instead of working with the raw text. Model parameters and model performance indicators for all combinations are described in Table 5.2.

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC	Specificity	Weighted Specificity
Word2Vec + Doc2Vec							
Logistic regression (w/o multicollinearity)	0.56 ± 0.0103	**0.84 ± 0.0083**	0.56 ± 0.0103	0.60 ± 0.0104	**0.75 ± 0.0144**	0.48 ± 0.0117	**0.84 ± 0.0185**
Random Forest	0.80 ± 0.0073	0.67 ± 0.0116	0.80 ± 0.0073	0.73 ± 0.0098	0.60 ± 0.0158	0.98 ± 0.0030	0.16 ± 0.0315
Gradient boosting	0.79 ± 0.0077	0.74 ± 0.0111	0.79 ± 0.0077	**0.75 ± 0.0092**	0.63 ± 0.0149	0.94 ± 0.0052	0.26 ± 0.0276
Adaptive Boosting	**0.80 ± 0.0070**	0.73 ± 0.0145	**0.81 ± 0.0070**	0.74 ± 0.0095	0.65 ± 0.0160	**0.98 ± 0.0031**	0.2 ± 0.0304
Extreme Gradient Boosting	0.78 ± 0.0081	0.73 ± 0.0115	0.78 ± 0.0081	0.75 ± 0.0097	0.56 ± 0.0166	0.93 ± 0.0058	0.26 ± 0.0287
BioClinical BERT + DocBERT							
Logistic regression (w/o multicollinearity)	0.62 ± 0.0099	**0.77 ± 0.0100**	0.62 ± 0.0099	0.66 ± 0.0091	0.67 ± 0.0125	0.62 ± 0.0109	**0.64 ± 0.019**

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC	Specificity	Weighted Specificity
Random Forest	**0.80 ± 0.0072**	0.72 ± 0.0137	**0.80 ± 0.0072**	0.74 ± 0.0099	0.60 ± 0.0143	**0.97 ± 0.0029**	0.2 ± 0.0316
Gradient boosting	0.71 ± 0.0084	0.70 ± 0.0108	0.71 ± 0.0084	0.70 ± 0.0091	0.62 ± 0.0140	0.83 ± 0.0079	0.28 ± 0.026
Adaptive Boosting	0.79 ± 0.0073	0.74 ± 0.0108	0.79 ± 0.0073	0.75 ± 0.0094	**0.65 ± 0.0157**	0.94 ± 0.0044	0.28 ± 0.0291
Extreme Gradient Boosting	0.79 ± 0.0073	0.74 ± 0.0101	0.79 ± 0.0073	**0.76 ± 0.0090**	0.64 ± 0.0143	0.93 ± 0.0050	0.29 ± 0.0281
BioBERT + DocBERT							
Logistic regression (w/o multicollinearity)	0.54 ± 0.0108	**0.78 ± 0.0107**	0.54 ± 0.0108	0.58 ± 0.0101	0.62 ± 0.0157	0.49 ± 0.0118	**0.70 ± 0.02**
Random Forest	**0.81 ± 0.0072**	0.67 ± 0.0116	**0.81 ± 0.0072**	0.73 ± 0.0099	0.60 ± 0.0147	**0.99 ± 0.0021**	0.18 ± 0.007
Gradient boosting	0.79 ± 0.0079	0.75 ± 0.0111	0.79 ± 0.0079	0.76 ± 0.0096	**0.68 ± 0.0139**	0.93 ± 0.0053	0.31 ± 0.0172
Adaptive Boosting	0.78 ± 0.0080	0.73 ± 0.0112	0.78 ± 0.0080	0.74 ± 0.0096	0.65 ± 0.0151	0.93 ± 0.0054	0.31 ± 0.0134
Extreme Gradient Boosting	0.79 ± 0.0083	0.76 ± 0.0107	0.79 ± 0.0083	**0.77 ± 0.0096**	0.61 ± 0.0157	0.92 ± 0.0059	0.33 ± 0.0277

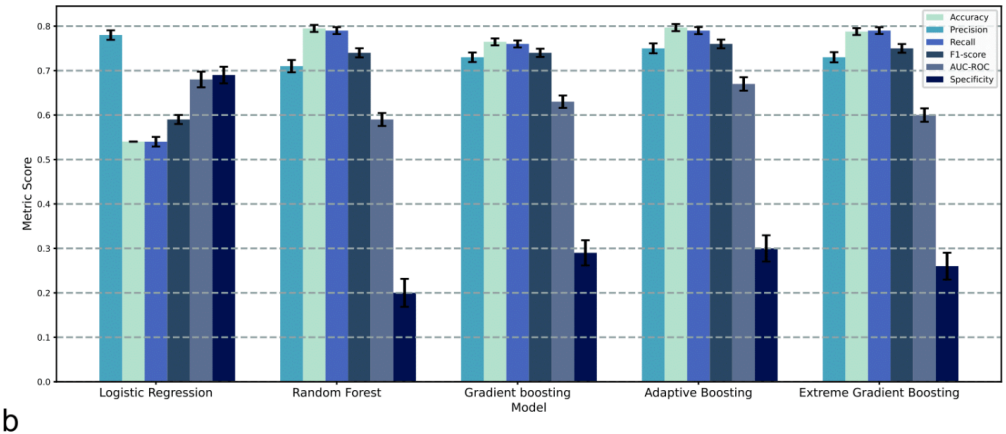
Table 5.2. Using the clinical notes features alone, results of SI (abnormality) prediction on a 24-hour cohort of MIMIC III dataset with a margin error of one standard-deviation from the mean. Random Sampling(N=100) of the test dataset, with Bootstrap iterations of 100 has been recorded for the mean value. Only Therapeutics and HOPI embeddings were considered as input features for the Machine learning models. We considered the Gradient Boosting model with BioBERT+DocBERT embedding as the best model since it recorded a high F1-score and AUC-ROC score.

5.3.3 Leverage gained from History of Present illness and length of context.

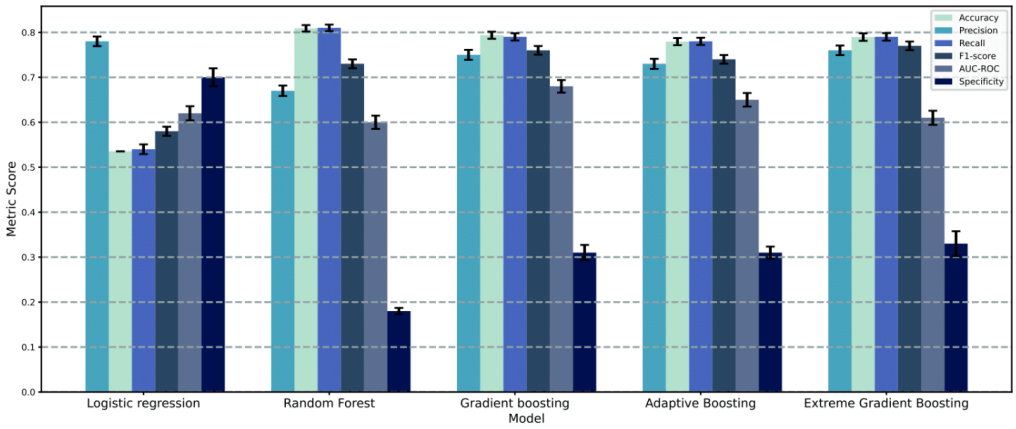
Since HOPI is expected to capture the patient's trajectory, an ablation experiment to understand the importance of the HOPI in the Physician notes was carried out. Comparison among the models (with/without HOPI as feature) is shown in Figure 5.3a, 5.3b. We observe a significant increase (p-values ≤ 0.05) in AUC-ROC score with the incorporation of HOPI embeddings for three models. Therefore, we find that the context encapsulated in HOPI is important for

performance of models predicting the outcomes in patients admitted to the ICU and must be included. We have also empirically vetted the fallout in performance of 768 sized embeddings for HOPI against its mean embeddings. For Transformer based models, we hypothesized that the length of context (window length) would be positively correlated with model performance indicators (Figure 5.3c).

a



b



c

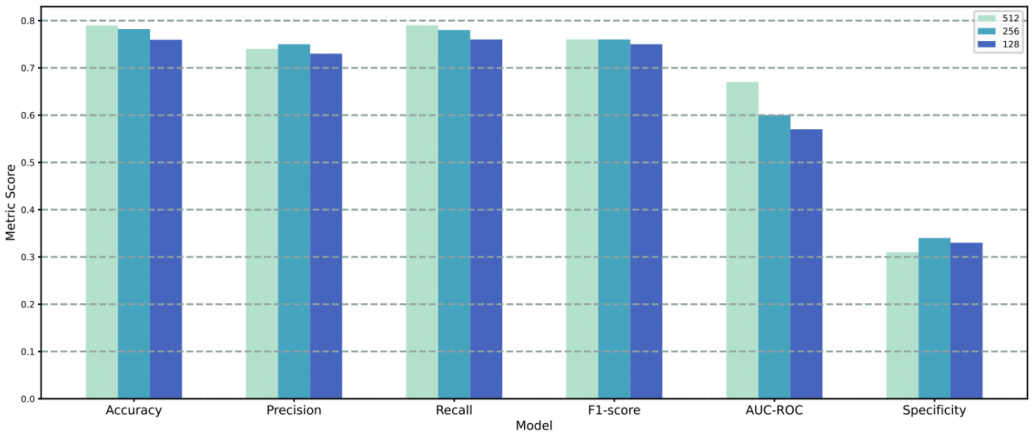


Figure 5.3. Leverage gained with addition of History of Present illness and length of context. Efficacy of various machine learning models a) exclusive and b) inclusive of HOPI as features. Therapeutic embeddings were generated from BioBERT while HOPI embeddings were generated from DocBERT. c) Comparison of different contextual window lengths (512, 256, 128) for the embeddings generated from BioBERT. These embeddings were compared against different metrics for the best pipeline exclusively of textual data(Gradient Boosting+DocBERT embedding+BioBERT contextual embedding). The plot depicts that sequence length of 512 has achieved the best AUC-ROC demonstrating that with an increase in the sequence length, there has been an increment in the AUC-ROC.

5.3.4 Analysis using only Vitals Data

We trained a battery of machine learning models exclusively on top 200 TS features processed from vitals data to verify its performance over clinical notes (Table 5.3). Considering both AUC-ROC and F1-score as screening parameters we observed that vitals data alone outperforms textual embeddings data by 6% in terms of AUC-ROC. Adaptive Boosting was the best model for vitals data, as compared to Gradient Boosting model having BioBERT+DocBERT embeddings as features. Notably, Random Forest achieved the highest sensitivity of 0.81, while Adaptive Boosting demonstrated the highest specificity of 0.45.

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC	Specificity	Weighted Specificity
Logistic Regression (w/o multicollinearity)	0.60 ± 0.0093	0.67 ± 0.0110	0.60 ± 0.0093	0.63 ± 0.0092	0.42 ± 0.0128	0.69 ± 0.0100	0.27 ± 0.0146
Random Forest	**0.81 ± 0.0076**	**0.80 ± 0.0087**	**0.81 ± 0.0076**	**0.80 ± 0.0080**	0.73 ± .0122	**0.92 ± 0.0066**	0.44 ± 0.028
Gradient boosting	0.72 ± 0.0078	0.73 ± 0.0088	0.72 ± 0.0078	0.72 ± 0.0078	0.59 ± 0.0132	0.82 ± 0.0079	0.36 ± 0.021
Adaptive Boosting	0.81 ± 0.0077	0.79 ± 0.0090	0.81 ± 0.0077	0.79 ± 0.0083	**0.74 ± 0.0122**	0.91 ± 0.0064	**0.45 ± 0.025**
Extreme Gradient Boosting	0.76 ± 0.0074	0.74 ± 0.0092	0.76 ± 0.0074	0.75 ± 0.0088	0.60 ± 0.0148	0.88 ± 0.0067	0.33 ± 0.023

Table 5.3. Results of SI (abnormality) prediction exclusively using vitals data on a 24-hour cohort of MIMIC III dataset with a margin error of one standard-deviation from the mean. Only TS-features extracted from vitals data were considered as input features for the Machine learning models. Random Sampling(N=100) of the test dataset, with Bootstrap iterations of 100

has been recorded for the mean value.

5.3.5 Multimodal Analysis inclusive of Structured (Vitals) and Unstructured (Textual) Data

Embeddings from therapeutics and HOPI were combined with TS features from vitals data to effectively understand the significance of multimodal features over individuals. We observe that Random Forest trained on Word2Vec and Doc2Vec embeddings inclusive of vitals data achieved an AUC-ROC and F1-score of 0.76 and 0.81 respectively outperforming only textual and only vitals based pipelines (Table 5.4). Taking sensitivity as screening parameter, Random Forest trained on BioBERT and DocBERT achieved the highest sensitivity of 0.83. In general, all the best performing models have a high sensitivity, thus indicating suitability for sensitive detection, yet limiting alarm fatigue in the early warning settings. This indicates models can leverage further information when both structured and unstructured data are fused together for shock prediction.

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC	Specificity	Weighted Specificity
Word2Vec + Doc2Vec							
Logistic Regression (w/o multicollinearity)	0.60 ± 0.0093	0.67 ± 0.0110	0.60 ± 0.0093	0.63 ± 0.0092	0.41 ± 0.0127	0.69 ± 0.0100	0.27 ± 0.0214
Random Forest	0.83 ± 0.0068	0.81 ± 0.0092	0.83 ± 0.0068	0.81 ± 0.0081	**0.76 ± 0.0116**	**0.94 ± 0.0050**	0.42 ± 0.0201
Gradient boosting	0.74 ± 0.0072	0.74 ± 0.0085	0.74 ± 0.0072	0.74 ± 0.0074	0.60 ± 0.0144	0.84 ± 0.0071	0.37 ± 0.0211
Adaptive Boosting	0.82 ± 0.0074	**0.81 ± 0.0085**	0.82 ± 0.0074	0.81 ± 0.0079	0.71 ± 0.0143	0.92 ± 0.0061	**0.48 ± 0.0229**
Extreme Gradient Boosting	0.76 ± 0.0078	0.75 ± 0.0090	0.76 ± 0.0078	0.75 ± 0.0081	0.56 ± 0.0151	0.87 ± 0.0075	0.36 ± 0.0227
BioClinical BERT + DocBERT							
Logistic Regression (w/o multicollinearity)	0.60 ± 0.0093	0.67 ± 0.0110	0.60 ± 0.0093	0.63 ± 0.0092	0.41 ± 0.0127	0.69 ± 0.0100	0.27 ± 0.0214
Random Forest	0.79 ± 0.0073	0.76 ± 0.0091	0.79 ± 0.0073	0.77 ± 0.0081	0.74 ± 0.0119	0.92 ± 0.0065	0.34 ± 0.0239

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC	Specificity	Weighted Specificity
Gradient boosting	0.75 ± 0.0079	0.74 ± 0.0094	0.75 ± 0.0079	0.74 ± 0.0084	0.66 ± 0.0122	0.85 ± 0.0071	0.38 ± 0.0238
Adaptive Boosting	0.79 ± 0.0074	0.77 ± 0.0086	0.79 ± 0.0074	0.78 ± 0.0079	0.74 ± 0.0128	0.90 ± 0.0066	0.41 ± 0.0228
Extreme Gradient Boosting	0.76 ± 0.0076	0.75 ± 0.0087	0.76 ± 0.0076	0.75 ± 0.0078	0.63 ± 0.0136	0.87 ± 0.0073	0.37 ± 0.0218
BioBERT + DocBERT							
Logistic Regression (w/o multicollinearity)	0.60 ± 0.0093	0.67 ± 0.0110	0.60 ± 0.0093	0.63 ± 0.0092	0.42 ± 0.0128	0.69 ± 0.0100	0.27 ± 0.0214
Random Forest	**0.83 ± 0.0065**	0.81 ± 0.0088	**0.83 ± 0.0065**	**0.81 ± 0.0077**	0.72 ± 0.0147	**0.94 ± 0.0050**	0.42 ± 0.0236
Gradient boosting	0.74 ± 0.0077	0.73 ± 0.0095	0.74 ± 0.0077	0.73 ± 0.0082	0.57 ± 0.0128	0.85 ± 0.0072	0.33 ± 0.023
Adaptive Boosting	0.78 ± 0.0077	0.76 ± 0.0089	0.78 ± 0.0077	0.77 ± 0.0082	0.71 ± 0.0118	0.90 ± 0.0068	0.37 ± 0.0231
Extreme Gradient Boosting	0.75 ± 0.0073	0.72 ± 0.0090	0.75 ± 0.0073	0.73 ± 0.0079	0.64 ± 0.0110	0.88 ± 0.0068	0.29 ± 0.0227

Table 5.4. Results of SI (abnormality) prediction combining textual and vitals data on a 24-hour cohort of MIMIC III dataset with a margin error of one standard-deviation from the mean. Random Sampling(N=100) of the test dataset, with Bootstrap iterations of 100 has been recorded for the mean value.

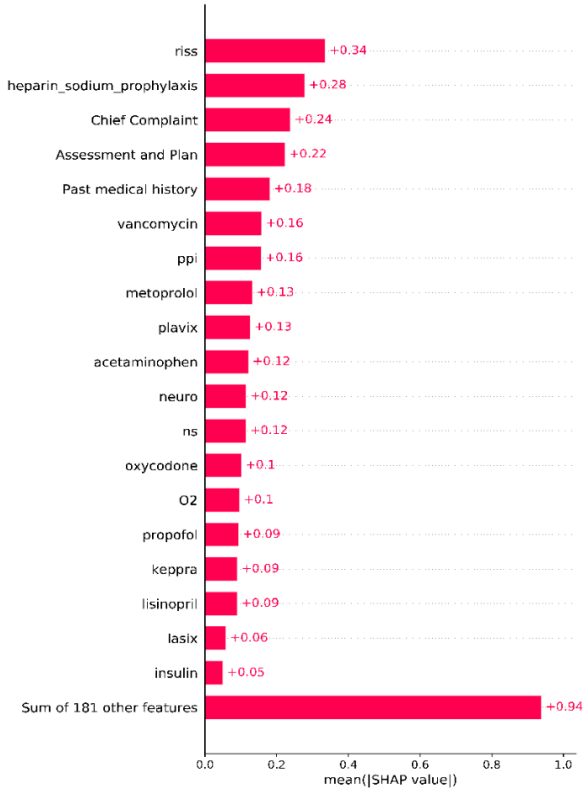
5.3.6 Interpretability of model predictions using Shapley plots.

Figure 5.4 shows Shapley plots that have been used to analyze the explainability of the models. For interpretability analysis, Gradient Boosting model trained on BioBERT and DocBERT embeddings (textual pipeline) and Random Forest model trained on clinical notes and vital features (multimodal pipeline) were chosen as these performed the best in terms of F1 score and AUC-ROC.

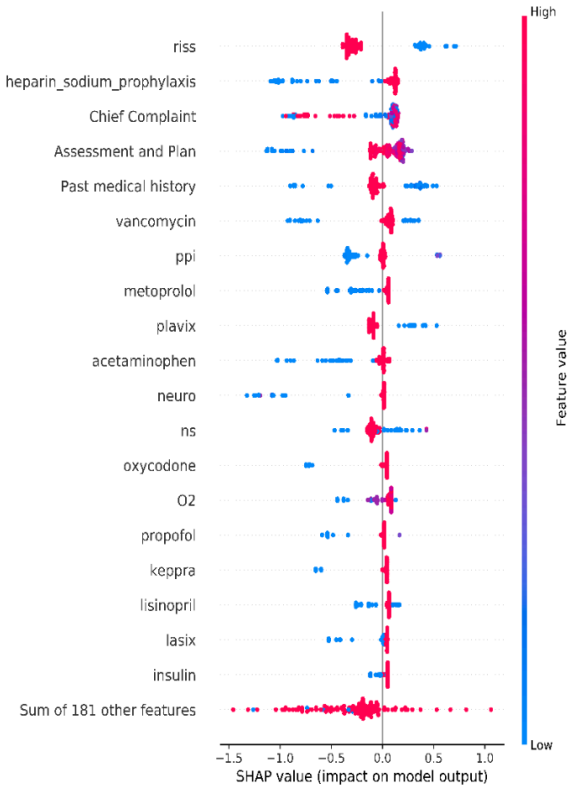
Higher absolute Shapley values are indicative of higher importance. The global feature importance is an aggregate indicator of overall importance of the word across all clinical notes. Figure 5.4a, 5.4c depict the global feature importance in the test dataset for textual and

multimodal pipelines respectively. RISS (Regular Insulin sliding scale), Heparin sodium prophylaxis and Chief complaints sections were seen as the top ranking global features for predicting worsening of shock index in the textual pipeline. For the multimodal pipeline, SHAP analysis illustrated that most of the top predictors were Fourier-features of heart rate, which is a primary indicator of worsening SI. However, Heparin Sodium Prophylaxis still remained a significant feature for this pipeline, thus leveraging the information contained in clinical notes to improve model predictions when combined with vitals data. Figure 5.4b, 5.4d illustrates the impact of these features on the model's output. It can be observed from the graph that the presence of RISS in a patient's treatment chart increases the likelihood of downstream development of abnormal SI while the absence of such increases the likelihood of normal SI.

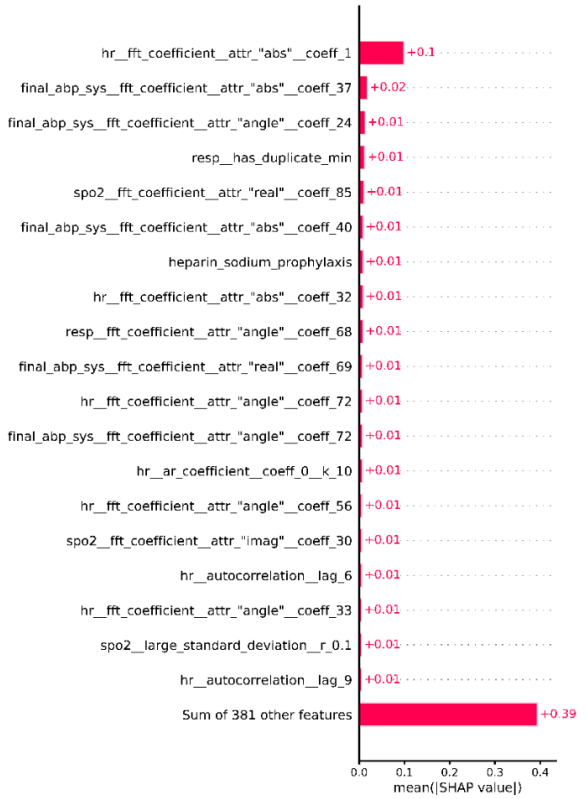
a.


riss +0.34
heparin_sodium_prophylaxis +0.28
Chief Complaint +0.24
Assessment and Plan +0.22
Past medical history +0.18
vancomycin +0.16
ppi +0.16
metoprolol +0.13
plavix +0.13
acetaminophen +0.12
neuro +0.12
ns +0.12
oxycodone +0.1
O2 +0.1
propofol +0.09
keppra +0.09
lisinopril +0.09
lasix +0.06
insulin +0.05
Sum of 181 other features +0.94
0.0 0.2 0.4 0.6 0.8 1.0
mean(|SHAP value|)

b.


riss
heparin_sodium_prophylaxis
Chief Complaint
Assessment and Plan
Past medical history
vancomycin
ppi
metoprolol
plavix
acetaminophen
neuro
ns
oxycodone
O2
propofol
keppra
lisinopril
lasix
insulin
Sum of 181 other features
−1.5 −1.0 −0.5 0.0 0.5 1.0
SHAP value (impact on model output)
High
Low
Feature value

c.


hr__fft_coefficient__attr_"abs"__coeff_1 +0.1
final_abp_sys__fft_coefficient__attr_"abs"__coeff_37 +0.02
final_abp_sys__fft_coefficient__attr_"angle"__coeff_24 +0.01
resp__has_duplicate_min +0.01
spo2__fft_coefficient__attr_"real"__coeff_85 +0.01
final_abp_sys__fft_coefficient__attr_"abs"__coeff_40 +0.01
heparin_sodium_prophylaxis +0.01
hr__fft_coefficient__attr_"abs"__coeff_32 +0.01
resp__fft_coefficient__attr_"angle"__coeff_68 +0.01
final_abp_sys__fft_coefficient__attr_"real"__coeff_69 +0.01
hr__fft_coefficient__attr_"angle"__coeff_72 +0.01
final_abp_sys__fft_coefficient__attr_"angle"__coeff_72 +0.01
hr__ar_coefficient__coeff_0__k_10 +0.01
hr__fft_coefficient__attr_"angle"__coeff_56 +0.01
spo2__fft_coefficient__attr_"imag"__coeff_30 +0.01
hr__autocorrelation__lag_6 +0.01
hr__fft_coefficient__attr_"angle"__coeff_33 +0.01
spo2__large_standard_deviation__r_0.1 +0.01
hr__autocorrelation__lag_9 +0.01
Sum of 381 other features +0.39
0.0 0.1 0.2 0.3 0.4
mean(|SHAP value|)

d.

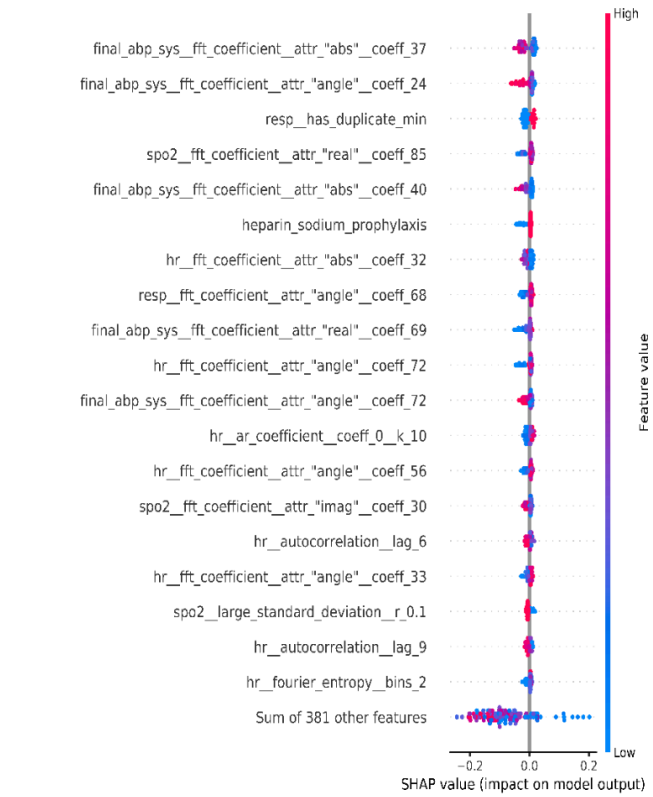

final_abp_sys__fft_coefficient__attr_"abs"__coeff_37
final_abp_sys__fft_coefficient__attr_"angle"__coeff_24
resp__has_duplicate_min
spo2__fft_coefficient__attr_"real"__coeff_85
final_abp_sys__fft_coefficient__attr_"abs"__coeff_40
heparin_sodium_prophylaxis
hr__fft_coefficient__attr_"abs"__coeff_32
resp__fft_coefficient__attr_"angle"__coeff_68
final_abp_sys__fft_coefficient__attr_"real"__coeff_69
hr__fft_coefficient__attr_"angle"__coeff_72
final_abp_sys__fft_coefficient__attr_"angle"__coeff_72
hr__ar_coefficient__coeff_0__k_10
hr__fft_coefficient__attr_"angle"__coeff_56
spo2__fft_coefficient__attr_"imag"__coeff_30
hr__autocorrelation__lag_6
hr__fft_coefficient__attr_"angle"__coeff_33
spo2__large_standard_deviation__r_0.1
hr__autocorrelation__lag_9
hr__fourier_entropy__bins_2
Sum of 381 other features
−0.2 0.0 0.2
SHAP value (impact on model output)
High
Low
Feature value

Figure 5.4. Interpretability of model prediction using SHAP analysis. Sections such as chief complaints, past medical history, and treatment plan along with specific therapeutics *such as anticoagulants and antibiotics contain important information for predicting abnormal SI in global feature importance analysis (a) as well as understanding the influence each feature has on the outcome for the textual pipeline. (c) Vitals features were predominant in determining the abnormal SI for the multimodal pipeline as well as certain therapeutics such as Heparin Sodium Prophylaxis turned out to be significant in global feature importance analysis. (b,d) An example of an individual level feature importance for a record is shown as a waterfall plot for textual and multimodal pipeline respectively.*

5.3.7 Model performance across clinically segregated categories

We have analyzed the predictive power of *ShockModes* (multimodal pipeline) by clinically segregating the cohort based on clinical categories (Figure 5.5). Notably, the best pipeline for *ShockModes* (Random Forest+Word2Vec+Doc2Vec+vitals time-series) was screened on the basis of F1-score and AUC-ROC, whose efficacy was studied across a broad range of disease-related outcomes. *ShockModes* achieved the highest sensitivity (93%) and F1-score (91%) across acute respiratory failure, followed by Chronic kidney disease (92% and 90%, respectively). It achieved better specificity for instances with acute respiratory failure and kidney-related terms (chronic kidney disease), while comparing against (p-value ≤ 0.05) terms suggestive of congestive heart failure. This demonstrates the broad applicability of *ShockModes* across critical care illnesses. We have also analyzed the efficacy of text-based models by segregating the instances of patients based on drugs present in physician notes. Lasix achieved the best F1-score of 0.84 with a sensitivity of 0.89 respectively.

Figure 5.5: Evaluation of models based on clinical segregation. a) Efficacy of ShockModes on top diagnosis. b) Efficacy of only text-based pipelines on top drugs.

5.4 Discussion

In this study, we developed and evaluated a novel multimodal pipeline, *ShockModes,* for prognosticating 24-hours before the onset of abnormal shock-index. *ShockModes* utilizes retrained BERT based representations in combination with time series features from vitals to create a machine learning model. The multimodal approach surpassed individual modalities by nearly 2% in AUC-ROC while adding features from physician notes such as history of present illness and medication in a manner similar to clinical reasoning. Hence, *ShockModes* is an advance over standard approaches for early detection of Shock Index deterioration, and can enable timely interventions.

Early detection of hemodynamic instability is life-saving as it allows medical practitioners to take immediate action for restoring proper blood circulation and oxygenation to vital organs (Rahman et al. 2021). In a previous study where 36 variables (uni-modal), including vitals and laboratory investigations, were used to predict the onset of abnormal shock index 12-hours ahead in pediatric patients, the AUC and sensitivity reported were 0.70 and 0.50 respectively (Potes et al. 2017). On the contrary, our model surpassed these values with an AUC and sensitivity of 0.76 and 0.81 respectively, while predicting the outcomes with 24-hours lead time. The incorporation of textual data from clinicians notes extracts information related to therapeutics, HOPI, potential risk factors and complications, which aids in decision-making and management of patient care, thereby improving the prediction of abnormal shock index. Despite the difference in study data, this comparison is valuable as it assesses the performance of *Shockmodes* against unimodal approaches for predicting the hemodynamic stability. With paucity of available benchmarks, our study provides baseline models for predicting onset of abnormal SI on a 24-hours cohort.

In Early Warning Systems, higher sensitivity of prediction is an important component for designing screening systems of ICUs to prevent alarm fatigue due to false negatives. This can substantially help in reducing the overhead charges for medical organizations. Drawing conclusions based on sensitivity values can optimize medical resources as clinicians would otherwise need to perform additional follow-up examination and treatment. Further, a high sensitivity in a clinical setting is a strong indicator of the robustness of the system as it correctly identifies abnormal SI, whereas a high specificity assists in detecting patients with normal SI. Our notes-based pipeline achieved high specificity values (above 80%) indicating the probability of false positive cases to be minimal. It also demonstrated that the simpler models (Logistic Regression+Word2Vec+Doc2Vec) had higher specificity and can be used to screen patients with normal SI, thereby reducing the chances of false positive cases. This leads to conservation of crucial resources which can be deployed more sensibly in treatment of patients, thereby reducing mortality. The sensitivity also touched 80% in the best case for all three versions of the pipeline. This indicates that a vast proportion of shock patients, who require immediate medical attention, were correctly classified by the model. Hence, this supplements and accelerates the decision making for clinicians, enabling them to make timely interventions and overcoming alert fatigue. Furthermore, *ShockModes* also achieved high sensitivity values across clinically segregated cohorts, demonstrating the robustness and clinical adaptability of our pipeline. Interpretability of models in clinical settings is very important as it delivers a clear understanding on how the decisions have been made based on input features. Shapely determined Fourier-features of heart rate and heparin sodium prophylaxis as clinically meaningful features of ShockModes for predicting abnormal SI; elucidating the significance of heart rate on our development cohort.

There are few limitations to our current pipeline. Firstly, our pipeline has been exclusively trained on the MIMIC-III dataset which is constrained to adult ICU of specific regions. However, the resulting pipelines can be fine-tuned to other datasets arising from critical care settings for

wider adaptability and robustness. Secondly, for interpretability purposes, embeddings for each drug were supplied to the machine learning model which led to sparsity in the matrix. Although this poses the danger of overfitting, a cross-validation approach would partially mitigate it. In the future, we propose to create an interpretable approach using the dense embeddings themselves. Thirdly, we have not integrated demographic data from EHR with vitals and text data. Effectively, we propose to combine such datasets, along with the text and vitals based features for improving the predictions further. Lately, a major concern has been the failure of AI pipelines at bedside. We essentially plan to deploy our pipeline in a real-world setting for retrospective validation. Also, to demonstrate the impact of our architecture on long-term patient outcomes, we would like to extend our study in cohorts based on different time stamps (48 hours, 72 hours) and using additional outcome data such as mortality and length of ICU stay.

In conclusion, clinician notes are the closest representation of a clinician's thought process for decision-making. Hence, they contain rich contextual information which has been underutilized for predictive settings. Also vitals data contains a large chunk of information, a subset of which is disjoint to the contextual information contained in clinical notes. This leads to better prediction when both the modalities of data (physician notes and vitals data) are utilized in critical care settings. The technical novelty of our pipeline and interpretable predictions can be leveraged to yield predictions for other critical outcomes through a transfer learning mechanism. Finally, we believe that work enables decision making in a variety of settings including resource constrained ICUs because of broader availability of clinician notes and vitals data.

Over the last three chapters, I have introduced some key applications of language models on healthcare settings. In this dissertation, COVID-19 and Abnormal Shock Index have been considered as used-cases for demonstrating the applications of language models in healthcare. Some key takeaways from the previous chapters were (i) Semantics captured from the genomic sequences of COVID-19 allowed language models to predict case loads two months ahead of time. (ii) Knowledge distilled from COVID-19 research helped in undermining the trend of current research scenarios. It also helped to decipher the emerging themes for future research directions. (iii) Unstructured data in the form of clinical notes can be independently used as features for predictive modeling of critical outcomes such as Abnormal Shock Index. Language models helped in extracting such features from clinical text and infusing those features into machine learning models. (iv) Overall, Language models aided in making temporal decision support by learning the semantics of the dataset. Furthermore in the next chapter, I have demonstrated the potential of language models to enhance support in mobile applications. Utilizing natural language understanding techniques allows algorithms to understand user input and provide accurate and relevant responses in real-time. This has numerous applications, including customer service chatbots, language translation, and voice assistants. As these technologies continue to evolve, we can expect to see language models play an increasingly important role in enhancing user experience and improving access to information. In the next

chapter, I will explain more about our mobile application launched during COVID-19 to raise WASH awareness and curb the spread of infodemic. The chapter deciphers how language models have been effectively utilized to provide adequate support to users based on defined objectives.

Chapter-6

An end-to-end LM pipeline and application for raising WASH awareness in COVID-19: A Machine Learning Application for Raising WASH Awareness in the Times of COVID-19 Pandemic

6.1 Introduction

In previous chapters, few of the major aspects where language models can intervene in temporal decision making have been highlighted. Retrospective evidence on such intervention has been reported based on the analysis conducted through such studies. The chapters depicted the feature extraction technique from a diverse dataset which included genomics, scientific literature and clinical notes using language models. These features were further processed using machine learning algorithms to build pipelines that helped in providing extensive decision support during the pandemic progression (COVID-19) and critical care setting (ICU). In the previous chapters, studies related to the COVID-19 pandemic were accompanied by interactive dashboards that enabled researchers and policymakers to efficiently analyze and visualize large amounts of data collected from publicly available databases such as GISAID and PubMed. The dashboards facilitated the identification of complex patterns and relationships in the data, leading to new discoveries and insights. Additionally, they allowed for real-time monitoring of pandemic progression, including the prediction of caseloads and the tracking of evolving themes and emerging topics in scientific literature. Overall, the dashboards served as a powerful tool for data exploration and analysis in the context of the COVID-19 pandemic. However, during the onset of the pandemic a mobile health platform was required to directly interact with users and raise awareness among the general public related COVID-19. As this was the first pandemic that hit on a global scale, knowledge pertaining to the disease was much limited. Hence, dissemination of right information at the right time was the urge of the hour. Effectively, a mobile health application can bring solutions in multiple regards. (i) A mobile-health application can provide a platform for direct communication with the general public and dissemination of accurate information to combat the COVID-19 infodemic. (ii) A mobile health app can reach a wider audience than a dashboard, as it can be easily downloaded and used by anyone with a mobile device. This can be particularly important for reaching underserved populations, who may not have access to other sources of health information or support. (iii) It can provide more targeted interventions and reminders than a dashboard by sending notifications to users to remind them to wash their hands or wear a mask, etc.

Healthcare misinformation is a growing menace in digital societies (Ioannidis et al. 2017; Tasnim, Hossain, and Mazumder 2020). This is clearly highlighted by the COVID-19 pandemic that has affected over 3.8 million people worldwide, causing a widespread loss in all aspects of

daily life (Rezaei 2021). Digital consumption has increased manifolds, creating both an opportunity and a danger in terms of information dissemination. Infodemic has been defined as an overabundance of information, some accurate and some not, making it hard for people to find trustworthy sources and reliable guidance when they need it (Rezaei 2021). The spread of the COVID-19 infodemic was faster than the pandemic itself and posed a threat to public health (Kouzy et al. 2020; Pian, Chi, and Ma 2021; M. P. Patel et al. 2020; Cinelli et al. 2020). Further, mitigation of misinformation is also vital for raising correct awareness for the primary prevention of most communicable and non-communicable diseases. Mobile health (mHealth), coupled with verified health information, can serve as an information dispensing tool to tackle the spread of misinformation. Clear and effective communication of preventive measures and updated information is essential. To achieve this goal, designing a trustworthy app that helps navigate the information deluge can be crucial. Therefore, recognizing the potential of mHealth platforms, we developed WashKaro, a multi-pronged AI approach for Infodemic Management. WashKaro was driven by the imminent need to raise Water, Sanitation, and Hygiene (WASH) awareness and combines English (WASH) with vernacular (Karo, meaning "Do" in Hindi) for mitigating the spread of COVID-19. OnAir is a feature on the WashKaro app which combines Natural Language Processing (NLP) to match news articles with WHO guidelines. Conversational AI (Satya, meaning "Truth" in Hindi) reaches out to the community as audio-visual content in local languages. To keep the information relevant, WashKaro provides daily news matched with WHO guidelines (World Health Organization 2020), WHO directive-based Symptom Self-Assessment tool, and human-vetted information delivering these in Hindi, the most widely understood local language across India. Since India is one of the largest and fastest-growing markets for digital consumers, with 560 million Internet subscribers in 2018 (Kaka 2019), and about 60% using mHealth technologies, this offered a unique opportunity to test WashKaro. The study is based on the WHO's Information Network for Epidemics (EPI-WIN) ("EPI-WIN Updates," 2023.) strategy, covering four strategic areas of work to tackle the infodemic, as shown in Fig. 6.1.

Prevention of disease using interventions of Artificial Intelligence, Machine learning, and NLP has been a significant breakthrough in the era of Covid-19. There has been a rampant spread of misinformation related to oxygen, availability of oxygen beds, vaccines, drugs, and many other things. For authentication and validation of such information web search engines have been created (Shams et al. 2021), various IoT and AI-based tools have been created to raise awareness for concerns related to handwashing, social hygiene, maintaining social distancing, and wearing masks (Samyoun et al. 2021; Kodali et al. 2020).

The methodology used for this study is centered around the WHO EPI-WIN strategy, which covers four strategic areas of work to respond to the infodemic. The first area of coverage is focused on identifying the problem at hand, given the current evidence and information to promote and form public policies strategically. We have identified the context-specific

community problems and the potential of mHealth, AI, and NLP in order to acknowledge possible solutions. This is followed by simplifying the enormous amount of information currently available across multiple sources to disseminate accurate information in a simplified manner. In order to achieve this objective, multilingual support is provided in the form of audio visual-based content. The spread of misinformation is tackled by providing information such as Mythbusters and government updates along with meta-information in the form of geographic coordinates of essential facilities. We also offer periodic hand washing reminders. In case of such an unforeseen event, it is vital to amplify the intervention by means of establishing two-way communication with the intended audience to tailor the advice and messages. This has been catered to by engaging users in active feedback based involvement, participating in enhancing the AI proposed model along with any generic feedback in an audio format. A public health survey and symptom self-assessment are crucial components in amplifying our study. In order to devise constantly evolving strategies, it is essential to validate the methodology and quantify the infodemic. WashKaro application statistics, demographic analysis of public health surveys, health analysis of at-risk population using symptom self-assessment, and user agreement on AI-based intervention is critical to quantify and evaluate.

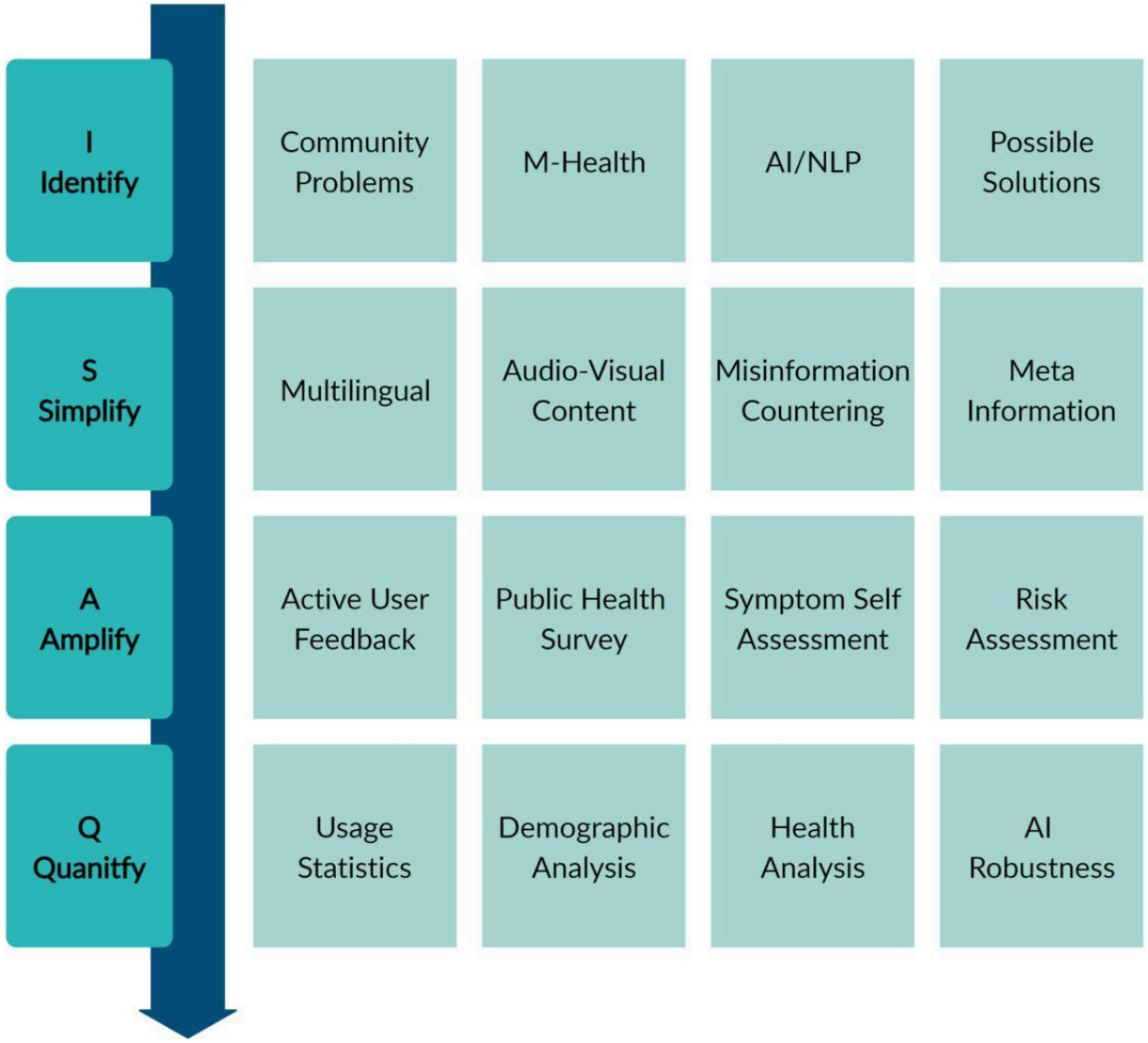


Figure 6.1: *Proposed workflow of the App based upon Identify, Simplify, Amplify and Quantify framework as specified WHO's EPI-WIN strategy.*

6.2 Methods

WashKaro was developed as a holistic mHealth solution that could serve as a one-stop AI-powered infodemic management suite during the current COVID 19 pandemic. The underlying strategy utilized was Identify-Simplify-Amplify-Quantify, as deployed by the Information Network for Epidemics (EPI-WIN) established by the WHO (Rezaei 2021). The main idea of our application was to provide unsolicited information as bite-sized text and audio in Hindi and English. The mobile application was made available to the general public through Google Play Store and was downloaded by more than 5000 users. We did not select a cohort, and all the responses received from the general public were analyzed to gather real-world evidence

about the effectiveness of our machine learning-based messaging intervention. The choice of the Android platform reflects the predominant usage of the platform among smartphone users in India.

The machine learning algorithms helped in filtering correct information from the deluge of news, which were then vetted by medical experts. Gathering raw data from credible sources such as the WHO and consumer-centric daily news articles, we used NLP approaches and Machine Learning (ML) to identify authentic and pertinent information. The information thus extracted was simplified and presented as audio-visual content in Hindi (the most widely understood local language across India), English, and various other vernaculars. By garnering feedback on the relevance of the WHO information provided along with the news pieces, the advice to the individual was tailored according to their personal needs, thus amplifying the reach of appropriate messages. We also offered a WHO directive-based Symptom Self-Assessment tool and numerous categories of human-vetted information in the form of Infographics, MythBusters, geographic information, etc. Forming real-time, on-the-ground, multidisciplinary research partnerships is essential to mitigate the infodemic. Therefore, our entire methodology and infodemic suite is open-source (https://github.com/tavlab-iiitd/WashKaro/tree/master/washkaro-textmatching) and available for the whole of the scientific community to build upon.

6.2.1 Datasets for the Application

The news articles dataset was scrapped from vernacular Hindi News Portals (DainikJagran) online publication. Articles presented to the annotators were in English which were translated from Hindi using GoogleTranslateAPI. WHO guidelines related to COVID-19 were scrapped manually from the WHO website. WHO COVID-19 guidelines and NewsArticles were collected starting from February 2020 to May2020. These news articles were filtered to remove all the non-COVID articles. The distillation was done using the following keywords ‘COVID’, ‘COVID-19’, ‘Corona’, ‘Coronavirus’, ‘Lockdown’, and ‘Pandemic’. The application also has a feature named “More Info” which provides information and guidelines related to COVID-19 for raising awareness among common people. The feature consisted of Infographics based on WHO recommendations, mythbusters and government advisories credibly sourced from the official government website of the Ministry of Health and Family Welfare.

6.2.2 NLP in healthcare

In the current situation, timely delivery of tenable content to the masses is exceptionally crucial to counter the spread of misinformation. WashKaro targeted this requirement using Natural Language Processing techniques to dispense information sourced through highly trusted WHO outlets such as EPI-WIN, which may not reach the appropriate audience or be too complicated for them (Zarocostas 2020).

The NLP pipeline (Fig. 6.2) involved two datasets: the WHO guidelines and the news articles. Multiple pairs of WHO guidelines and News Articles are generated as an input for the Machine Learning System, extractive ML summarization techniques were used to abbreviate the text. Articles were refreshed on a daily basis (using an automated web scraper) and from the Indian vernacular news source: Dainik Jagran and the WHO website. This data was collated in a csv file which has been used for the modeling task. The previous chapters unravel various perspectives of word embeddings in context to diverse dataset including genomics, scientific literature and clinical notes. These word embeddings have been effectively used for supplementing decision making. Both neural network and transformer-based architectures were explored in language models for generating word embeddings. Pre-Trained Word2Vec Embedding (Mikolov, Sutskever, et al. 2013) was used to generate embedding vectors for each word in the two documents. These word vectors were converted to article-level embeddings using Smooth Inverse Frequency (S. Arora, Liang, and Ma 2017). The generated pair of document level vectors are used for the calculation of distance metrics. Cosine similarity was calculated to find the similarity between two embedded vectors. Based on the users' reviews, the threshold of cosine similarity was set to determine the news articles that will be provided to users subsequently using this AI system (Pal et al. 2020). This pipeline served to complement the user's daily news consumption that suits their palette with an appropriate WHO guideline related to COVID-19 and WASH (Water Sanitation and Hygiene), thus augmenting healthcare awareness. In order to enhance engagement and provide increasingly relevant content, user feedback was sought at the end of each matching- the users marked each pair of WHO guidelines and news articles provided to them as either relevant or irrelevant. This active user feedback aided the machine learning backend in improving with each review by determining the type of news articles the user found relevant to a particular guideline. Further, any new article provided to the user took into account the previous learning, which enabled the deliverance of more relevant information with each feedback cycle.

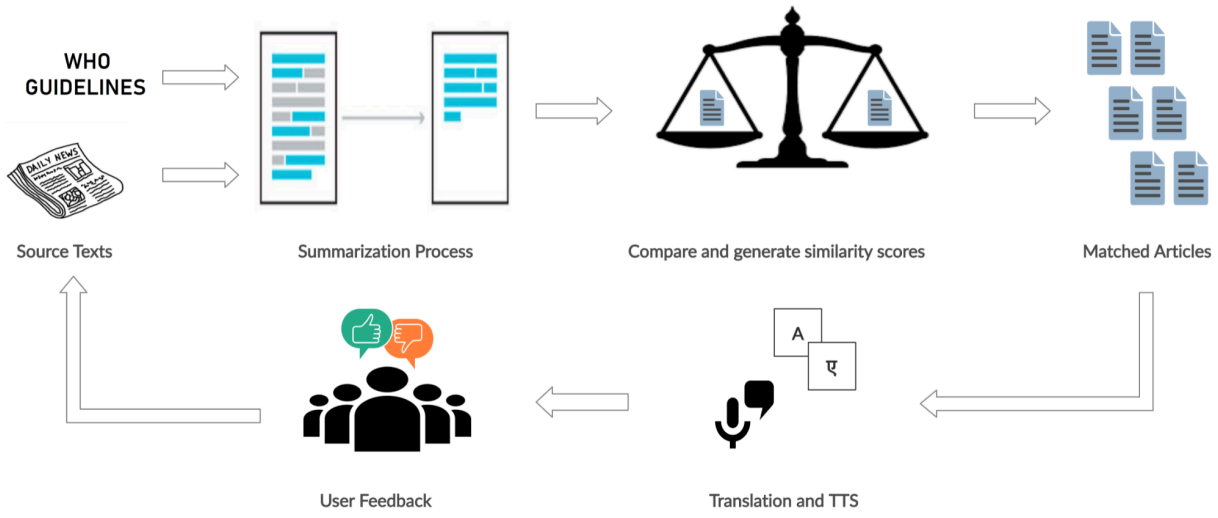


Figure 6.2: *NLP Pipeline. The pipeline takes in news articles and the World Health Organization (WHO) reports and constructs two-level sentence similarity between titles and the full-text to build a similarity score. Finally, the relevant texts are subject to translation and text to speech conversion for local language consumption (Hindi).*

6.2.3 Simplification

We made a deliberate attempt to convey context-specific and consumable information in a medium of the user's choice. Infographics based on WHO recommendations were used for effective presentation of preventive measures. Byte sized information packets were delivered in multiple local languages to ensure accessibility to various marginalized groups. Text-to-speech engines helped convert the information to an audio-visual format, thus reaching out to the less educated population. The application made use of the inbuilt Android text-to-speech model available in each android device. Mythbusters and government advisories, critical in countering misinformation and uncertainties surrounding the official guidelines, were credibly sourced and regularly updated. Mythbusters and government advisories, critical in countering misinformation and uncertainties surrounding the official guidelines, were credibly sourced from the official government website of the Ministry of Health and Family Welfare. The guidelines related to COVID-19 present in the MOHFW were updated on a daily basis, and the WashKaro application was simultaneously updated. Information on containment zones, hospitals, and hunger relief centers was provided in a geographical context, with directions imparted through Google Maps, a popular user-friendly interface. Regular notifications, worded positively to encourage participation, reminding the user to wash their hands and use masks in public places, were displayed.

6.2.4 Symptom Self Assessment

Low accessibility of the healthcare system, given the lockdown and social distancing measures in place, and a skewed ratio between the population who wish to get tested and medical professionals who can verify this need call for an effective alternative to screen patients (Emanuel et al. 2020). Thus, we devised a self-assessment tool for the symptoms of COVID-19, thereby enabling quicker identification of suspect cases who can then be guided to the Government helpline numbers and informed about proper self-quarantine protocols, nearby hospitals admitting COVID suspects, and testing centers. We defined the Suspect Case using the WHO Interim Guidance on Global surveillance for COVID-19 caused by human infection with the COVID-19 virus, and classified them further as a Suspect case (A), (B) or (C) (Interim guidance, 2023.). The 7-point questionnaire was designed using the case definitions from the WHO Interim Guidance verbatim. Based on the WHO criteria' application on the answers to the 7 questions, the user was notified about whether or not they were suspected of having COVID-19 (Fig. 6.3).

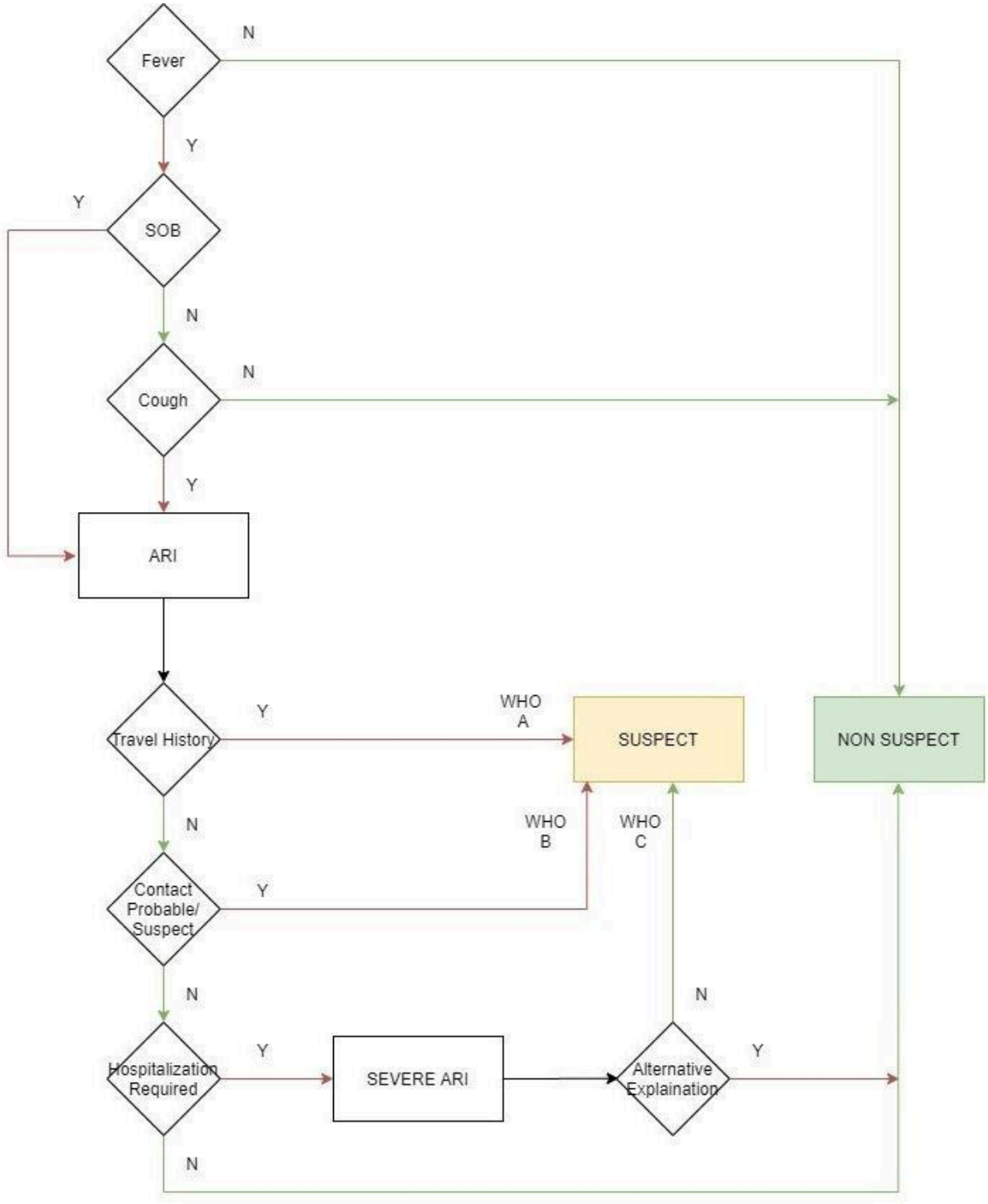


Figure 6.3. *Self Assessment Tool Flowchart. Based on the World Health Organization(WHO) Interim Guidance, a questionnaire and flowchart were developed to classify the responders as `Suspects' or `Non-suspects'. Here SOB refers to Shortness Of Breath, and ARI refers to Acute Respiratory Infection.*

6.2.5 Chatbot

Correct and officially verified information regarding the disease should be there at everyone's disposal. We have made a chatbot (Fig. 6.4) to serve this purpose, which has verified information from WHO, CDC, and additional government-approved sources. Existing Solution consists of an option Driven System ("WHO Health Alert Brings COVID-19 Facts to Billions via WhatsApp," 2023.) where a user needs to select through various lists of options to find answers to the Query. Thus, we devised a chatbot system designed to answer user queries using natural language. The current system consists of a Long Short Term Memory (LSTM) model fine-tuned on a Medical Question-Answers Dataset (MedQuAD) dataset (Ben Abacha and Demner-Fushman 2019). The Dataset was encoded using swivel embeddings generated on the Covid-19 open research dataset (L. L. Wang et al. 2020). All the data, including the training set, is incubated from Credible and Government controlled sources. Data for Answering the input query was taken from three Sources. Daily Statewise and India case count were extracted and updated on a real-time basis ("Covid19 Statewise Status" 2020). Data for Training the model was taken from the World Health Organisation's CoronaVirus Frequently Asked Questions and Centers for Disease Control and Prevention's CoronaVirus Frequently Asked Questions. The user Query is also passed through spelling correction using a symmetric delete spelling correction algorithm along with an artificial increasing frequency of words related to the disease, symptoms, etc ("Coronavirus (COVID-19) Outbreak Glossary" 2020) to increase the accuracy and effectiveness of the system.

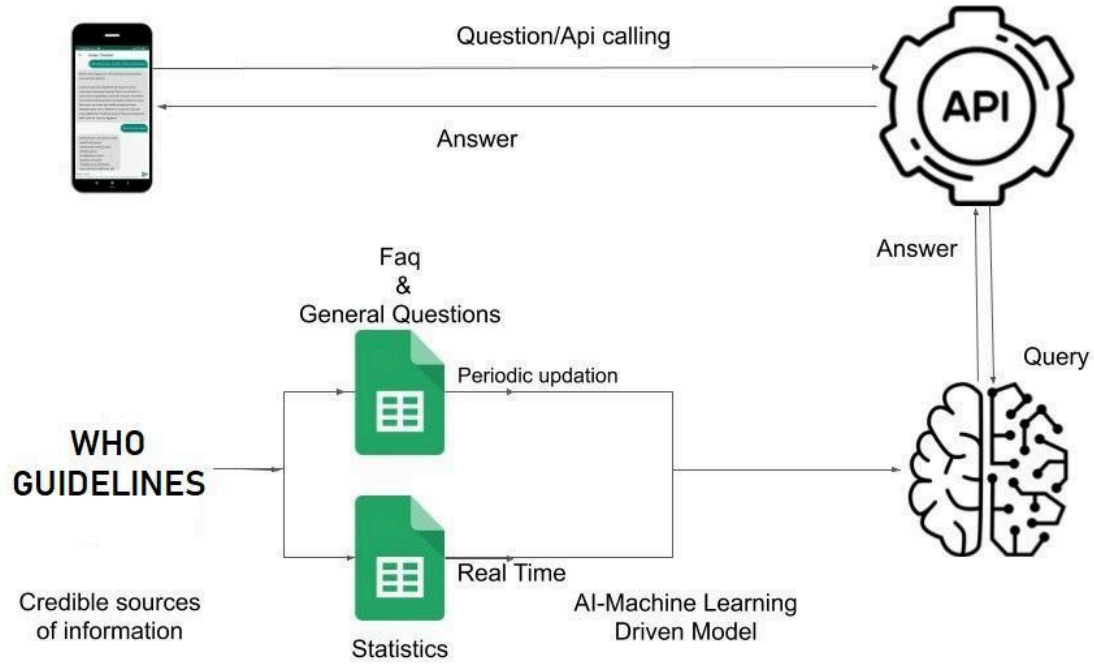


Figure 6.4. *Request-Response cycle in the chatbot. This is a schematic diagram depicting how the answer is displayed whenever a query is asked to the chatbot by a user.*

6.2.6 Active User Feedback

Anonymized data was collected through self-assessment usage analytics, Play Store managed user statistics, and easy to comprehend survey forms. Our suite deployed an anonymized public health survey that asked basic healthcare-related questions to understand the demographics and monitor the situation periodically. Audio-based feedback was taken from the user to ensure user convenience, establish a two-way dialogue, and prevent specific suggestions from being marginalized. Illustrations were used whenever possible to make the user aware of the collected data, hence protecting their right to information and privacy.

6.3 Results

The study aimed to improve article relevance using a machine learning approach and achieve information dissemination through better engagement in the local language. Our multi-pronged approach targeted to:

(i) Achieve a non-intrusive manner of healthcare knowledge dissemination.

(ii) Use local language to increase the participation of the target group (Female population).

(ii) Develop a self-assessment tool to identify the at-risk population at an early stage to mitigate the chances of community transmission.

6.3.1 Information Enrichment Over Time: The Number of `Relevant' Votes Increases

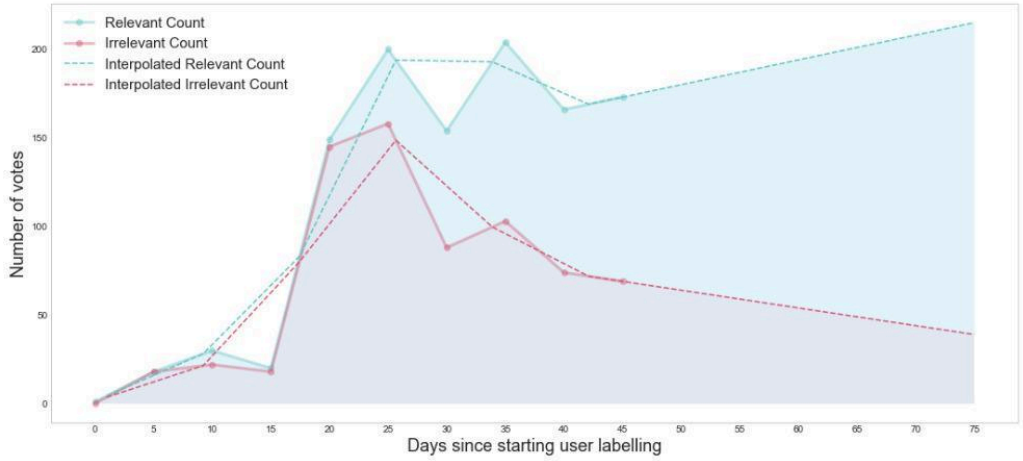


Figure 6.5. *Analysis of Natural Language Processing(NLP) pipeline. The graph shows relevance as a function of user feedback functionality in the app. Relevance is seen to increase with cumulative feedback over time. From day 0 onwards, the Relevant Count's angular coefficient is*

1.39 (±0.488), the angular coefficient of Irrelevant Count is -0.99 (±0.602), with an average slope difference of about 2.29.

With time and increased user feedback, the relevance of matching news articles with WHO reports increases, as seen in the rise in the number of relevant votes, owing to the constantly evolving machine learning model (Fig. 6.5). The number of irrelevant votes also decreases, validating our proposed methodology and providing increasingly relevant content from trusted sources to the user over time in the language of their preferred choice. At the beginning of the AI-based learning system on 15 March 2020, the number of `relevant' votes and `irrelevant' votes were both 18. On 25 March 2020, with increased user interaction and AI learning, the number of relevant votes was 173 and irrelevant votes was 69. The ratio of relevant votes to irrelevant votes rose from 1.0 to 2.5 over a period of one month.

6.3.2 Demographics- Females engaged more in Hindi

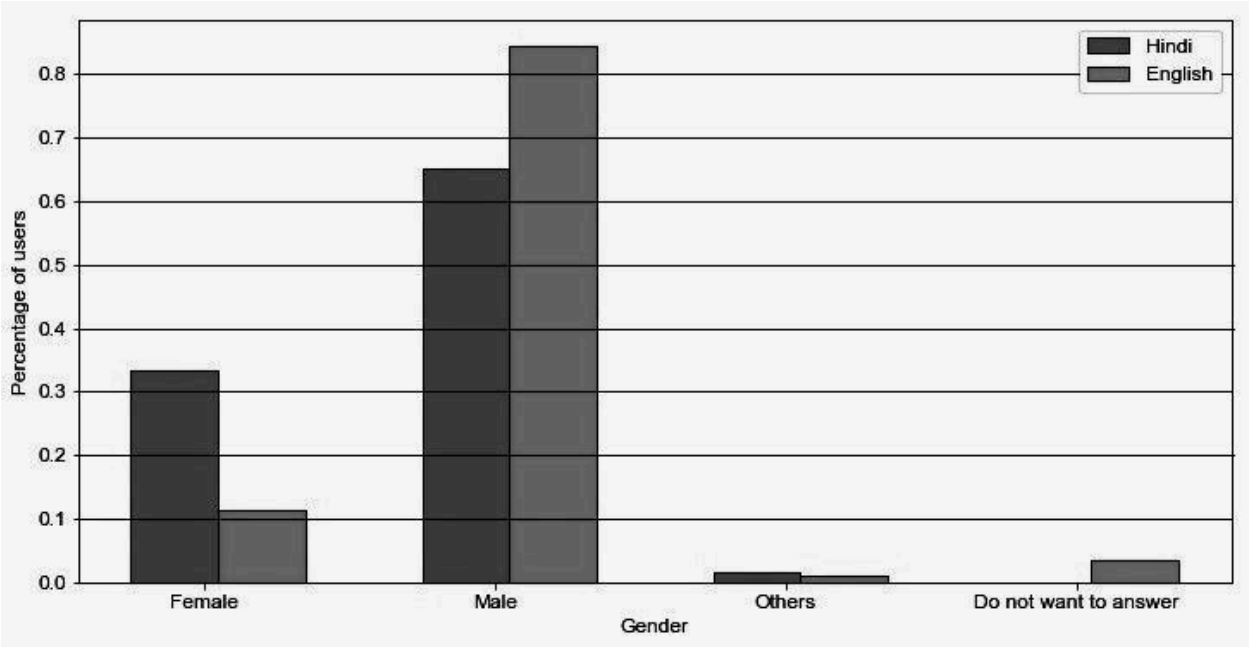


Figure 6.6. *Analysis of Public Health Survey. Distribution graphs showing the distribution of gender among Hindi and English Users. It clearly shows skewness in gender for English users whereas in the case of Hindi users it shows an approximate normalization among the genders.*

A total of 436 people took part in the English language based survey, and 126 took part in the Hindi language based survey. Fig. 6.6 was generated based on the survey conducted on the WashKaro app. The analysis of this plot suggests that the number of English users is more than the number of Hindi users. It also depicts that the overall number of male users is more than female users. A key insight observed from the data depicted that Hindi speaking female users (33% of total Hindi speaking users) were more than English speaking female users (11% of total

English speaking users). The census of India 2011 highlighted the disparity of literacy rates across genders, with 82.14% literacy rates amongst Indian males and 65.46% literacy rates in Indian females (Chandramouli and General 2011). This underscores the fact that using local languages empowers the sections of the population that might not have otherwise access to the information.

6.3.3 Target population: Users who reported higher than expected incidence

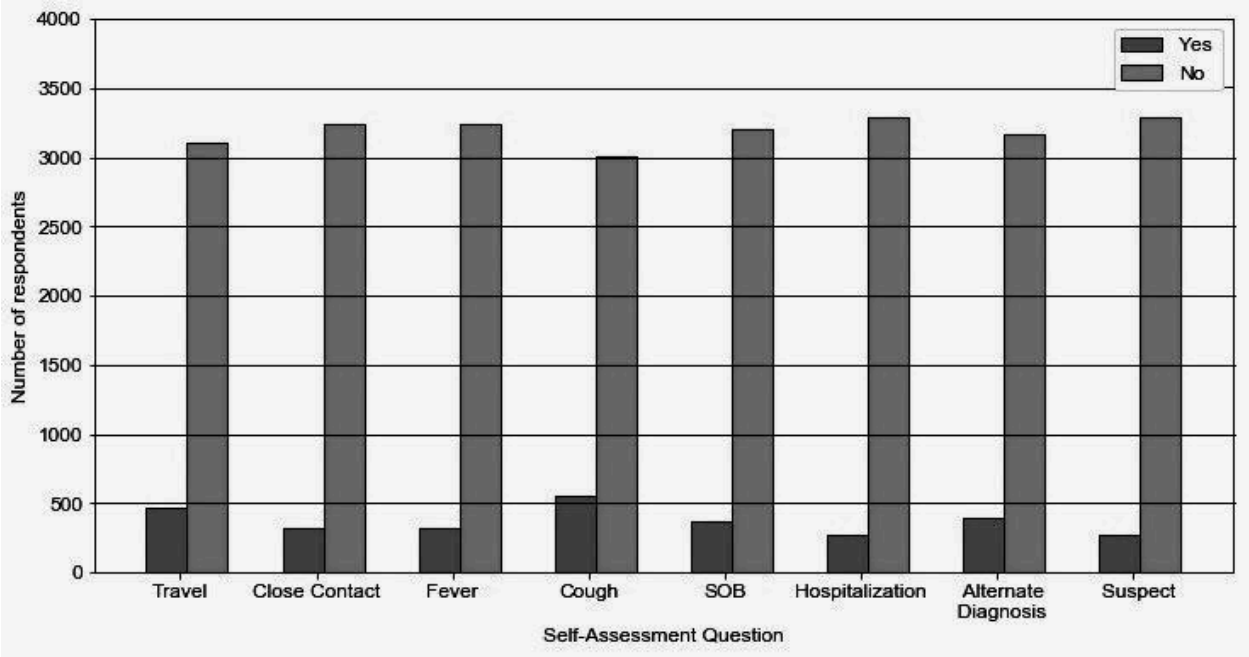


Figure 6.7. *Analysis of Self Assessment. A simple user-level self-assessment has been deployed to enable the general population to perform self-assessment and identify the population at risk, which can be used as an effective screening. A higher trend for a positive COVID-19 report in people who reported cough was observed. The symptoms of the disease have been known to change with strains, hence this approach of crowdsourcing information provides an agile approach to screen patients with specific symptoms.*

Based on the data collected from the symptom self-assessment 276 (7% of 3567 respondents) were found to be suspect cases according to their responses, while 3291 were non-suspect. 467 (13.09 %) of the respondents reported a travel history to locations reporting community transmission of COVID-19, 326 (9.13%) reported close contact with COVID-19 positive patient, 323 (9.05% of respondents) reported fever, 556(15.58%) reported cough, 367(10.28%) reported shortness of breath (SOB), 277 (7.77%) reported that they required Hospitalization and 395 (11.07%) respondents reported there was an alternate diagnosis for their condition (Fig. 6.7). A higher trend for a positive COVID-19 report in people who reported cough was observed. The symptoms of the disease have been known to change with strains. Hence this approach of

crowdsourcing information provides an agile approach to screen patients with specific symptoms.

6.4 Discussion

Our App was one of the first machine learning and natural language processing based approaches to tackle COVID-19 misinformation using machine learning. Over the years mHealth and machine learning have made significant contributions in the medical domain (Istepanian and Al-Anzi 2018). In the case of COVID-19, where reliable therapeutic strategies are still under experimentation, the role of such mHealth and machine learning interventions is critical (Alimadadi et al. 2020; P. Wang et al. 2020; Lv et al. 2021). Timely dissemination of trusted byte-sized information is extremely instrumental in mitigating the infodemic. With the majority of the world population staying at home due to lockdown, the increased amount of digital consumption opens up the scope to deploy these techniques as an effective social intervention to mitigate the infodemic by delivering the right information to the right people at the right time. When organizations across the globe are proactively testing various strategies to address the issue at hand, open-source software will play a vital role in such scenarios at a global stage to mitigate pandemics and infodemics at the root level.

Our study focuses on addressing these questions and has the following strengths. We provide an open-source template featuring functions like Symptom Self-Assessment, Notification amplifier that notifies the user for washing hands, which are required for fighting against epidemics and pandemics. This also helps in the propagation of the right information hence proper management of infodemic can be done without misleading the masses in crucial scenarios. The enhancement of a few features within the WashKaro application can help in serving as an effective intervention for the government and policymakers. Detailed questions can be formulated targeting the at-risk users identified using symptom self-assessment, which can be incorporated into the existing framework followed by the higher authorities and medical workers for predicting the suspects of COVID-19 at an early stage. After identification and testing of the at-risk population, our analysis can be extended to predict patients who have chances of being at risk in the near future from the definitive set of questions, based on the priority of each question. We can present this data to the suitable administration and decision-makers for taking effective measures against such individuals at an early stage. For the Suspect Cases, we can administer a second questionnaire to further stratify the risk of acute respiratory distress syndrome (ARDS) and septic shock by assessing the severity of symptoms and looking for identified risk factors like age and pre-existing comorbidities that are not included in the WHO Interim Guidance. This can aid in making decisions regarding home quarantine against hospital admission. The app can also be used to identify other suspect cases in the same household. Further, to assist the government authorities to identify those requiring testing for COVID-19, we can ask for contact details of the Suspect Cases with informed consent and relay them to the appropriate government authorities to enable targeted testing. A follow up of the suspect cases through push notifications,

advising testing and recording test results, can help ensure that complacence does not set in. Also, as highlighted by previous work, teaching interventions to women can be an important step for mitigating diseases (Caldwell and Caldwell 1993). We reason that the engagement in Hindi is due to the content rather than the nature of the App because the default language of the app content is English. Therefore, the users would have made a conscious effort to change the language to Hindi.

We believe that public health messaging is a key component in managing a health crisis and our approach is geared to make this more agile through machine learning and natural language models. The strength of our approach is in devising a real-world pipeline for local language based deployment of an infodemic mitigation solution. To the best of our knowledge, there are no Applications that use machine learning and natural language processing to provide the right information, to the right people, in the right format, and at the right time. Our pipeline summarized texts from newspapers and matched these to the official sources of information such as the WHO, before delivering these as bite-sized audios in Hindi. The second strength of our study is the online learning algorithm to optimize the threshold of relevance as a function of user feedback. The third strength of our study is in the need for minimum data to gain public health insights. Despite not collecting granular personal data, we were able to show trends such as the gender inequity in the usage of Hindi versus English on the App.

Finally, there are some limitations to the study that have been conducted prior to the revamp of the application. All applications with COVID-19 information were removed in association with the guidelines regarding COVID-19 related applications on the Google Play store. The time frame of the case study for COVID-19 was shortened due to this reason. Our application was the first application providing vetted information using machine learning. As the government rolled out the official apps, these were promoted on a larger scale to target the national audience. That may explain the drop in the number of users active on our app. Secondly, we did not incentivize the responses to the usage of our App. Hence the responses from more than 5000 people are less likely to harbor a systematic bias. The primary objective of this app was straightforward- “Does the user score for relevance increase over time as our natural language processing based filters improve over time with user feedback?” This objective reflects the quality of information and personal preferences such as the language, which were analyzed. This objective is less likely to have heterogeneous influences. Due to data privacy policies and our motivation to collect minimal data, we did not obtain granular information on individuals' locations and other important personal information. Hence the confounding factors may be limited to the number of Android smartphone users (the majority in India) and the rate of spread of disease.

The information from the survey in the result section conveys a potential gender bias, with 3.4 times more females engaging with the app in Hindi compared to males. Although the survey was conducted on a very small sample size of 426 users, which depicted a gender skewness impacting the generalizability of the findings. The study predominantly reflects the experiences

and preferences of female users in a specific language setting. Additionally, the study does not explicitly address the diversity of the user population in terms of age, socio-economic background, or geographical location. To enhance generalizability, future studies should consider a more diverse and representative user sample considered over a large group. Analyzing user experiences across various demographics can provide a more comprehensive understanding of the application's effectiveness in mitigating health misinformation. Moreover, addressing biases in user engagement is essential for ensuring equitable access and impact on a broader scale. Furthermore, the application's pipeline can be adapted for various other diseases, such as HIV and Tuberculosis, to counteract the impact of the misinformation.

The development of innovative approaches while protecting individual data yet gathering useful inference is an active area of research, and our further work will address this limitation in various public health scenarios. Therefore, we conclude that the role of digital health interventions in the form of systems articulating vetted messages needs to be explored effectively dealing with public health challenges, both during health emergencies and normal times addressing the Sustainable Development Goals (SDGs) put forward by WHO.

Since health information from credible sources is not necessarily prioritized for dissemination in conventional media, especially for less literate and non-English speaking sections of the society. The users were searching for such information at a time when little was available in a local language in India. Our App was the first such platform and >5000 people downloaded the app during the study window, among those 1545 were actively engaged users before Google pulled the plug on all COVID-19 apps. Going forward, we are continuing to develop natural language processing based services for extending this feasibility experiment for raising public health awareness about infectious diseases such as TB. The further scope and extension of this study involve quizzes, gamification, peer network building and social incentivization to engage users. Future work is planned to evaluate the human-centric design of the app collaboration with an NGO working on the ground. Our use of models to manage information through the use of machine learning optimizes human resources and is shown to be effective as per the measured parameters. We strongly believe that this approach may be relevant to other resource-constrained settings also.

In premise, Chapters 3, 4, and 6 explored the application and implementation of language models in addressing the COVID-19 pandemic from multiple perspectives. Each work addresses in tackling an unique aspect of the pandemic, hence providing measures to mitigate and diminish its effect. While *Strainflow* and *EvidenceFlow* offered approaches for predicting and analyzing COVID-19 cases and scientific research, respectively. WashKaro was aimed to increase awareness and abate the effect of infodemic. Notably, Washkaro raised awareness related to WASH practices as a means of curbing the spread of the virus. And restricted the spread of infodemic by providing “Right Information to the Right People in Right Format at Right Time”.

Abbreviations

WHO: World Health Organization

WASH: Water Sanitation Hygiene

EPI-WIN: WHO's Information Network for Epidemics

AI: Artificial Intelligence

NLP: Natural Language Processing

Chapter-7

Conclusion and Future work

Health datasets pertain information related to an individual's health and medical history, which can be collected, stored, analyzed and modeled for studying the spread and impact of a particular disease. Temporal decision is crucial for health datasets because it allows healthcare providers, medical practitioners and policy makers to make more informed decisions based on a data-driven modeling approach. Temporal decision support had a critical role in managing pandemics. It involves making decisions based on real-time data, identifying populations at high risk, and implementing public health interventions such as vaccination and social distancing measures to control and mitigate the spread of the disease. In this dissertation, I demonstrate the impact of language models for temporal decision in health datasets considering COVID-19 pandemic and Abnormal Shock Index as use cases.

7.1 Overview of the thesis

In the context of pandemics, genomic sequence analysis is essential because it can reveal important insights on evolution and transmission of the virus. By sequencing the virus's genome, researchers can keep track of its changes and spot any potential new variants. In the third chapter of this thesis, I tackled the problem of predicting pandemic progression using language models based on genomic dataset. Genomic sequences of COVID-19 were collated from the GISAID database. Spike regions were extracted from the whole genome of SARS-CoV-2 virus using the Covseq tool. Word2vec models were trained on the spike region of SARS-CoV-2 to capture the latent space of the virus, where codons were treated as words and sequences as paragraphs. Taking country as covariate, sample entropy was calculated using the monthly embedding of individual sequences for investigating the effect of mutation in viral genome. Analysis depicted that changes in codon level led to changes in the entropy of the underlying dimensional space. Boruta was then implemented to extract important dimensions (referred as DoCs) from entropy feature space. Utilizing these DoCs, we were able to predict the caseload of the country with a lead-time of two months. We developed *Strainflow*, an integrative framework built using language models, machine learning algorithms and mathematical algorithms for analyzing the temporal changes in latent space of viruses; while predicting the spread of viral strains. For the majority of countries, *StrainFlow* detected the increase in case loads 2 months before the Delta and Omicron surges. It was able to forecast the increase in COVID-19 cases in India due to the Omicron variant as early as the start of November 2021.

The second use case addresses the application of language models on COVID-19 literature and their use in mining evolving evidence. The COVID-19 pandemic led to an unprecedented

increase in scientific literature and presented challenges in terms of managing, synthesizing and distilling useful information. This exponential growth in scientific literature highlighted the need for rapid evidence mining using computational approaches. Language models can act as tools by assisting in knowledge discovery while analyzing real-time data to make informed decisions. In the fourth chapter, EvidenceFlow, was designed to track emerging themes in scientific literature based on the temporal drift within literature. The study aimed to exploit the temporal change in unsupervised word embeddings of the COVID-19 literature for capturing and tracking new knowledge. More than 150,000 COVID-19 articles published on the World Health Organization database were collated, from which medical entities were extracted using Named Entity Recognition (NER). For each month, separate Word2Vec models were trained to capture semantic drift in word similarities of published literature over time. Networks of most frequently occurring medical entities were constructed using cosine similarities as edge weights. On the basis of historical trends, topological characteristics of the upcoming month's network were forecasted using ARIMA model. While new links for upcoming months were predicted using supervised machine learning. Biomedical topics were tracked using community detection and alluvial diagrams as they changed throughout the course of the months. The inferences suggested that neurological problems and long COVID issues started to become more significant in March 2021 and June 2021. Prior analysis during August 2020 depicted thromboembolic complications as an emerging theme in the literature. The area under the receiver operating characteristic curve for the link prediction models was confirmed at 0.87. Based on the trends seen in previous months, the prediction modeling identified predisposing conditions, symptoms, cross-infection, and neurological consequences as the leading research issues in COVID-19 articles. The unified approach based on language models and machine learning-based allowed to capture patterns in semantic relationships over time. This method facilitated prediction of emerging links, which could contribute to steering research by capturing themes represented by clusters of medical entities.

The sixth chapter of this dissertation focused on real-time deployment of AI-suite for effective dissemination of accurate information to combat infodemic. With the rise in COVID-19 cases, chances of spreading misinformation went high as people were very less aware of the disease in context to its spread, symptoms and cure. Prompt solutions were required for effectively distilling relevant information from the internet in order to combat infodemic. To address this challenge, *WashKaro*, a multifaceted mobile application built using language models, machine translation and Artificial Intelligence to intervene the spread of misinformation. The study pipeline involved two datasets (i) WHO guidelines collated from WHO website; (ii) News articles scrapped from the Indian vernacular news source Dainik Jagran. The NLP pipeline exploits machine learning-based summarization techniques and language models to match COVID-19 related news articles with WHO guidelines. Word2Vec embedding has been used to generate embedding vectors for each word in both the documents, which were then converted to article-level embeddings using Smooth Inverse Frequency. To measure the similarity between

two embedded vectors, cosine similarity was calculated. In order to determine the relevance of news articles provided to users in context to WHO guidelines, a threshold cut-off for cosine similarity was set based on prior user feedback. To improve the machine learning backend, user feedback was collected at the end of each matching news article and WHO guideline. Additionally, it aided in continual learning and helped in identifying the type of news articles users found relevant to specific guidelines. This approach allowed enhancement of the machine learning system with each review, leading to a more personalized and effective user experience. The significance of the pandemic scenario has been validated by the number of downloads which amounts to a total of 5026 during the study window; among which 1545 were actively engaged users. The study also suggested that 3.4 times more females engaged with the App in Hindi as compared to males. Additionally, the relevance of AI-filtered news content doubled within 45 days of continual machine learning. Hence, our findings (i) demonstrated the potential of a mHealth platform during the advent of pandemic; (ii) highlighted the importance of providing accurate information in a user-friendly and accessible format (iii) and showed the importance of *WashKaro* which could serve as a template for other applications during future pandemics.

In critical care settings, temporal decisions assist medical practitioners in taking prompt, well-informed decisions that can enhance patient outcomes. Medical professionals can monitor patient progress, detect patterns and trends in patient data, and modify treatment regimens as necessary with the use of language models. This ultimately helps to reduce the length of stay in the ICU, lower mortality rates, and improve patient quality of life. In the fifth chapter, I explored the next use case which focused on temporal decision using language models in critical care settings. We aimed to create an explainable multi-modal early warning system (EWS) for 24-hour ahead prediction of abnormal shock index (SI) in critically ill patients, fusing routinely available vital signs and clinical notes. Unsupervised word embeddings of drugs and HOPI were generated using BERT-based architectures, which were integrated with vitals data. A total of 3117 features from vitals time-series combined with BERT-based features from clinical notes to train a series of machine learning models. Important predictive features were selected using Extra Tree Classifier, which were modeled using machine learning algorithms to predict abnormal shock index. The best multimodal pipeline (*ShockModes*) was then assessed for interpretability using SHAP features, which revealed Fourier-features of heart rate and heparin sodium prophylaxis as the most important predictor of outcome. On comparing the performance of vitals-based, notes-based, and multi-modal (vital+notes) classifiers, the best sensitivity and F1-score were achieved while using multi-modal classifiers (0.83 & 0.81). This demonstrates the potential of ShockModes over uni-modal pipelines, and the high sensitivity of models accounted for early warning systems. Clinical profiling across a range of critical care diseases, including hypertension, acute respiratory failure, chronic kidney disease, anemia, hyperlipidemia, atrial fibrillation, congestive heart failure and acute kidney failure, demonstrated that *ShockModes* can be applicable to assess and predict wide spectrum of critical care disorders. Overall, our study suggested that the multimodal, interpretable early warning system *ShockModes* can be used for

prognosticating SI based outcomes in ICU and emergency settings can assist clinicians in decision making, while saving multiple lives. *ShockModes* captured the semantic drift in the treatment provided to the patients over a time period for making predictions. Hence, this study depicted how language models unraveled the temporal changes occuring in the treatment plan of patients, while using this knowledge for enhanced patient outcomes.

7.2 Future Work

Application and implementation of language models have been studied since its inception in the late 2000's.With the advent of transformer architectures, language models have gained significant popularity in the recent decade and advanced the direction in research for NLP in healthcare. However, with the increased adoption of language models in healthcare, concerns have emerged regarding challenges such as data quality, model interpretability, and ethical concerns related to privacy and bias. Furthermore, addressing these challenges will assist in (i) Responsible deployment of language models in healthcare while minimizing potential harm; (ii) Development of language models that can handle the intricacies of healthcare data while upholding ethical and privacy considerations. Hence, a comprehensive approach must be taken to overcome the obstacles that impede the full utilization of language models in healthcare. In addition to addressing these challenges, language models can alo be utilized to mitigate the spread of misinformation related to healthcare on a real-time basis. The dissertation has a broad scope, and there are ample opportunities for future research to address the above challenges. The subsequent paragraphs highlight some specific areas where my research can be extended, and also provide a foundation for further investigation and advancement of the field.

7.2.1 Bias Amplification in Intersectional Subpopulations for Clinical Application by Large Language Models

Artificial Intelligence (AI) based algorithms are rapidly influencing decision making in diverse domains such as autonomous vehicle navigation, fraud detection, recommender systems with healthcare being no exception. Owing to the availability of large volumes of data in the form of images, electronic signals and electronic health records (EHR), a vast number of AI models have been trained and used to influence clinicians' decision-making process. In the present scenario, a mammoth amount of clinical notes are available containing densely rich information about a patient. However, capturing relevant information from such resources to assist clinicians in decision-making can be equally challenging. Language models have demonstrated a potential aid in addressing this challenge in a real-world setting. Language models can assist in identifying patterns in large amounts of patient data, which can help with early diagnosis and treatment of diseases. The advent of transformer architecture has caused a rampant surge in the development of sophisticated large language models (LLMs). Transformer architecture has the ability to capture vast amounts of semantic knowledge using self-attention mechanisms encoded as word embeddings. Hence, transformer-based language models achieved state-of-the-art results for a

number of clinical applications (Chetoui and Akhloufi 2022; Mayer, Cabrio, and Villata 2020; Shome et al. 2021; Shang et al. 2019). This has led to automating tasks, improved patient outcomes and reduced economic costs in clinical settings by assisting physicians in clinical thought processes.

However, despite its promise, there is growing concern regarding the fairness of these systems because AI algorithms have been shown to generate and amplify bias in a number of settings. Identifying bias in AI algorithms related to the healthcare domain is important because it can lead to inaccurate predictions, misdiagnosis, and potentially harmful treatment recommendations for patients. Such bias can be introduced at different stages of the algorithm development and deployment process, including data collection and preparation, algorithm design, and training. If not addressed, bias can lead to disparities in healthcare outcomes and exacerbate existing inequalities in the healthcare system. In the context of healthcare, bias is defined as the systematic and unintended favoring of individuals or groups on the basis of characteristics such as gender, age, ethnicity or other factors. Biases can be introduced in clinical notes through various ways, which include recruitment or collection strategy of data, analysis of data, and misinterpretation of data by AI systems. Additionally, these biases may arise due to various other factors, such as clinicians' implicit biases, differences in documentation practices or patient population demographics. This leads to skewed representations of the patient's medical history or health status. The presence of such biases can affect clinical decision-making, diagnosis, treatment and outcomes. Therefore bias analysis and mitigation become even more essential in a domain as critical as healthcare because unintended bias may hinder diagnosis leading to inequitable access to healthcare services for under-represented groups. A significant amount of work has already been done in the domain of algorithmic bias. For example, African Americans have been denied loans and given longer prison sentences compared to their Caucasian counterparts. In the healthcare domain, different studies (Angwin, Larson, and Mattu 2016) have come up with different notions of bias and demonstrated the same for various modalities such as clinical notes, medical scans and electronic signals. Seyyed-Kalantari et al. (Seyyed-Kalantari et al. 2021) illustrated underdiagnosis bias during triage in the diagnosis of patients, stating that underdiagnosis is potentially worse than misdiagnosis because the patient still receives medical care in the latter case. They analyzed under-served populations for chest radiographs and further reported that intersectional groups of under-served populations, such as Hispanic Female patients, are more prone to underdiagnosis bias. Further, Zhang et. al (H. Zhang et al. 2020) analyzed differences in the encoding of contextual embeddings for MIMIC-III dataset between marginalized and non-marginalized populations in terms of gender, ethnicity and insurance status. They showed that the majority group was always favored with regard to demographic denominators. Patient demographics such as gender, age, ethnicity and socio-economic status provide meaningful information used by clinicians. They can also lead to undesirable biases in AI predictions, which hinder access to healthcare services. Therefore, it is essential to design AI systems that capture the relevant information from demographic factors while minimizing the

effect of bias.

The study of biases in intersectional groups provides insight into the complexities and interactions of social determinants of health and the health disparities across population subgroups. Notably, there has been limited literature in the field of Artificial Intelligence (AI) backed healthcare that examines intersectional bias for multiple demographic dimensions in case of clinical notes. Ogungbe et. al. (Ogungbe, Mitra, and Roberts 2019) presented a survey of studies which illustrated the amplification of implicit bias for intersectional groups - the implicit bias of the participants was measured using IAT and the bias was shown to amplify for 2-fold (gender, age) and 3-fold demographics (gender, age, ethnicity). Another study (Kearns et al. 10--15 Jul 2018) examined the bias between demographic intersections such as young men and old women. Tan et al. (Y. C. Tan and Celis 2019) suggested methods to evaluate intersectional biases for contextual word embeddings and showed that biases for the intersection of two demographic dimensions were greater than the individual dimensions. Lalor et al. (Lalor et al. 2022) analyzed the intersectional bias in NLP related tasks across five text based datasets. It was reported that as the degree of intersection between groups increased, the Fairness Violation metric defined increased, indicating that the bias increases with degree of intersection. Given the staggering rate at which AI systems are being used by clinicians, the prevalence of biases in intersectional groups is an alarming cause of concern. The black box nature of Deep Learning models hinders explainability and affects decision making, thus preventing these intersectional groups from receiving timely and vital medical attention. Therefore, it is crucial to study bias in these models to ensure that decisions are not disproportionately influenced by factors such as race, gender, or socio-economic status.

7.2.2 Implementation of Language Models in Social Media

Another important area that remains unraveled in this broad spectrum of thesis was implementation of language models for mining and analyzing social media data. With the increasing popularity of social media platforms, there has been a surge of interest in using language models to analyze social media data for various applications, such as misinformation detection, sentiment analysis, mining social media trends related to vaccine beliefs, and emotion classification. Language models can effectively utilize vast amounts of unstructured text available on social media platforms to decipher trends on human behavior and characteristics. This can be utilized by policy makers to frame policy based on human sentiment and intervene at root level.

Social media plays a decisive role in propagating information, leading to the emergence of varying perceptions related to the pandemic (Cinelli et al. 2020). During the initial phase of national lockdown in several countries, Twitter had reported an increase of 24% in daily active users due to the increased usage of social media, the highest year-over-year growth rate reported

by the company to date. The rampant increase in social media engagement during the COVID-19 pandemic led to an alarming rise in the spread of misinformation. Social media platforms have been proven to be a breeding ground for misinformation and rumors, which can have serious consequences in terms of public health. This had far-reaching implications on public health, as misinformation often led to misguided actions and human behaviors. Healthcare organizations, policymakers, and researchers soon recognized the pressing need to address this issue. As a result, studies were conducted to investigate the extent of misinformation and its impact on public health. Prior studies have demonstrated the grim effects of misinformation related to COVID-19 on social media, and its impact in human behavior (Pian, Chi, and Ma 2021; Joseph et al. 2022; Loomba et al. 2021; Rocha et al. 2023). Misinformation has resulted in the proliferation of negative emotions such as confusion, fear, panic and anxiety, and posed challenges for health officials to disseminate accurate information to the public. This had led to reduced adherence to public health guidelines and mistrust of healthcare authorities. Additionally, misinformation has been linked to unnecessary hospitalizations and overuse of medical resources, which has put an immense strain on the healthcare system. Several events have been reported by studies such as "The 'Pandemic' of Disinformation in COVID-19" (Tagliabue, Galassi, and Mariani 2020), where mass media channels have shared incomplete or unverified updates on new treatments, myths about the use of masks, and errors by some hospital organizations that have resulted in higher reluctance from patients to seek medical attention. These findings highlight the need for effective strategies to limit the dissemination of false or misleading information and mitigate the impact of misinformation on social media for ensuring the well-being of individuals and the broader community. Misinformation on social media can mislead the public, affect their behavior, and lead to adverse health outcomes. Therefore, detecting and mitigating the spread of misinformation is crucial. Language models can act as a potent tool for mitigating misinformation, enabling means for detecting false information and providing users with informative content.

References

Abdalla, Mohamed, Moustafa Abdalla, Frank Rudzicz, and Graeme Hirst. 2020. “Using Word Embeddings to Improve the Privacy of Clinical Notes.” *Journal of the American Medical Informatics Association: JAMIA* 27 (6): 901–7.

Adamic, Lada A., and Eytan Adar. 2003. “Friends and Neighbors on the Web.” *Social Networks* 25 (3): 211–30.

Adams, Roy, Katharine E. Henry, Anirudh Sridharan, Hossein Soleimani, Andong Zhan, Nishi Rawat, Lauren Johnson, et al. 2022. “Prospective, Multi-Site Study of Patient Outcomes after Implementation of the TREWS Machine Learning-Based Early Warning System for Sepsis.” *Nature Medicine* 28 (7): 1455–60.

Adini, Bruria, Shepherd Roee Singer, Ronit Ringel, and Petra Dickmann. 2019. “Earlier Detection of Public Health Risks - Health Policy Lessons for Better Compliance with the International Health Regulations (IHR 2005): Insights from Low-, Mid- and High-Income Countries.” *Health Policy* 123 (10): 941–46.

Akkasi, Abbas, and Mari-Francine Moens. 2021. “Causal Relationship Extraction from Biomedical Text Using Deep Neural Models: A Comprehensive Survey.” *Journal of Biomedical Informatics* 119 (July): 103820.

Alanazi, Rayan. 2022. “Identification and Prediction of Chronic Diseases Using Machine Learning Approach.” *Journal of Healthcare Engineering* 2022 (February): 2826127.

Alimadadi, Ahmad, Sachin Aryal, Ishan Manandhar, Patricia B. Munroe, Bina Joe, and Xi Cheng. 2020. “Artificial Intelligence and Machine Learning to Fight COVID-19.” *Physiological Genomics* 52 (4): 200–202.

Almeida, Felipe, and Geraldo Xexéo. 2019. “Word Embeddings: A Survey.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1901.09069.

Alsentzer, Emily, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. “Publicly Available Clinical BERT Embeddings.” In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

Alsentzer, Emily, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. “Publicly Available Clinical BERT Embeddings.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1904.03323.

Amrollahi, Fatemeh, Supreeth P. Shashikumar, Fereshteh Razmi, and Shamim Nemati. 2020. “Contextual Embeddings from Clinical Notes Improves Prediction of Sepsis.” *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium* 2020: 197–202.

Angwin, J., J. Larson, and S. Mattu. 2016. “Machine Bias.” *Ethics of Data and*.

An, Weizhi, Yuzhi Guo, Yatao Bian, Hehuan Ma, Jinyu Yang, Chunyuan Li, and Junzhou Huang. 2022. “MoDNA: Motif-Oriented Pre-Training for DNA Language Model.” In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 1–5. BCB ’22 5. New York, NY, USA: Association for Computing Machinery.

Aouissi, Hani Amir, Mostefa Ababsa, and Aissam Gaagai. 2021. “Review of a Controversial Treatment Method in the Fight against COVID-19 with the Example of Algeria.” *Bulletin of the National Salmon Resources Center* 45 (1): 94.

Ariwala, Pinakin. 2019. “Top 14 Use Cases of Natural Language Processing in Healthcare.” Maruti Techlabs. March 19, 2019.

https://marutitech.com/use-cases-of-natural-language-processing-in-healthcare/.
Arora, Parul, Himanshu Kumar, and Bijaya Ketan Panigrahi. 2020. “Prediction and Analysis of COVID-19 Positive Cases Using Deep Learning Models: A Descriptive Case Study of India.” *Chaos, Solitons, and Fractals* 139 (October): 110017.
Arora, Sanjeev, Yingyu Liang, and Tengyu Ma. 2017. “A Simple but Tough-to-Beat Baseline for Sentence Embeddings.” *International Conference on Learning*. https://oar.princeton.edu/handle/88435/pr1rk2k.
Arslan, Hilal. 2021. “Machine Learning Methods for COVID-19 Prediction Using Human Genomic Data.” *Proceedings: A Conference of the American Medical Informatics Association / ... AMIA Annual Fall Symposium. AMIA Fall Symposium* 74 (1): 20.
Asgari, Ehsaneddin, and Mohammad R. K. Mofrad. 2015. “ProtVec: A Continuous Distributed Representation of Biological Sequences.” *arXiv [q-bio.QM]*. arXiv. http://arxiv.org/abs/1503.05140.
Avsec, Žiga, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. 2021. “Effective Gene Expression Prediction from Sequence by Integrating Long-Range Interactions.” *Nature Methods* 18 (10): 1196–1203.
Ayan, Necati, Sushil Chaskar, Anand Seetharam, Arti Ramesh, and Antonio A. de A. Rocha. 2021. “Mobility-Aware COVID-19 Case Prediction Using Cellular Network Logs.” In *2021 IEEE 46th Conference on Local Computer Networks (LCN)*, 479–86.
Bajwa, Junaid, Usman Munir, Aditya Nori, and Bryan Williams. 2021. “Artificial Intelligence in Healthcare: Transforming the Practice of Medicine.” *Future Healthcare Journal* 8 (2): e188–94.
Barabasi, A. L., and R. Albert. 1999. “Emergence of Scaling in Random Networks.” *Science* 286 (5439): 509–12.
Barros, Oscar, Richard Weber, and Carlos Reveco. 2021. “Demand Analysis and Capacity Management for Hospital Emergencies Using Advanced Forecasting Models and Stochastic Simulation.” *Operations Research Perspectives* 8 (January): 100208.
Beltagy, Iz, Kyle Lo, and Arman Cohan. 2019. “SciBERT: A Pretrained Language Model for Scientific Text.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1903.10676.
Beltagy, Iz, Matthew E. Peters, and Arman Cohan. 2020. “Longformer: The Long-Document Transformer.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2004.05150.
Ben Abacha, Asma, and Dina Demner-Fushman. 2019. “A Question-Entailment Approach to Question Answering.” *BMC Bioinformatics* 20 (1): 511.
Bennett, Casey C., and Thomas W. Doub. "Temporal modeling in clinical artificial intelligence decision-making and cognitive computing: Empirical exploration of practical challenges." In Proceedings of the 3rd SIAM Workshop on Data Mining for Medicine and Healthcare (DMMH). 2014.
Bepler, Tristan, and Bonnie Berger. 2019. “Learning Protein Sequence Embeddings Using Information from Structure.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1902.08661.
Bitnun, A. 2015. “Etymologia: Bonferroni Correction.” *Emerging Infectious Diseases* 21 (2): 289.
Blomberg, Bjørn, Kristin Greve-Isdahl Mohn, Karl Albert Brokstad, Fan Zhou, Dagrun Waag Linchausen, Bent-Are Hansen, Sarah Lartey, et al. 2021. “Long COVID in a Prospective Cohort of Home-Isolated Patients.” *Nature Medicine* 27 (9): 1607–13.
Boag, Willie, and Hassan Kané. 2017. “AWE-CM Vectors: Augmenting Word Embeddings with

a Clinical Metathesaurus." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1712.01460.
Bohlin, Ludvig, Daniel Edler, Andrea Lancichinetti, and Martin Rosvall. 2014. "Community Detection and Visualization of Networks with the Map Equation Framework." In *Measuring Scholarly Impact: Methods and Practice*, edited by Ying Ding, Ronald Rousseau, and Dietmar Wolfram, 3–34. Cham: Springer International Publishing.
Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. "Enriching Word Vectors with Subword Information." *Transactions of the Association for Computational Linguistics* 5: 135–46.
Bonanno, Fabrizio Giuseppe. 2011. "Clinical Pathology of the Shock Syndromes." *Journal of Emergencies, Trauma, and Shock* 4 (2): 233–43.
Brandes, Nadav, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. 2022. "ProteinBERT: A Universal Deep-Learning Model of Protein Sequence and Function." *Bioinformatics* 38 (8): 2102–10.
Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.
Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models Are Few-Shot Learners." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2005.14165.
Bussi, Yuval, Ruti Kapon, and Ziv Reich. 2021. "Large-Scale K-Mer-Based Analysis of the Informational Properties of Genomes, Comparative Genomics and Taxonomy." *PloS One* 16 (10): e0258693.
Butte, A. J. 2001. "Challenges in Bioinformatics: Infrastructure, Models and Analytics." *Trends in Biotechnology* 19 (5): 159–60.
Bu, Zhan, Yuyao Wang, Hui-Jia Li, Jiuchuan Jiang, Zhiang Wu, and Jie Cao. 2019. "Link Prediction in Temporal Networks: Integrating Survival Analysis and Game Theory." *Information Sciences* 498 (September): 41–61.
Caldwell, J. C., and P. Caldwell. 1993. "Roles of Women, Families, and Communities in Preventing Illness and Providing Health Services in Developing Countries." *For Developing Countries. Edited by JN …*. https://www.ncbi.nlm.nih.gov/books/NBK236447/pdf/Bookshelf_NBK236447.pdf#page=266.
Cao, Linan, Pei Liu, Jialong Chen, and Lei Deng. 2022. "Prediction of Transcription Factor Binding Sites Using a Combined Deep Learning Approach." *Frontiers in Oncology* 12 (June): 893520.
Cao, Yu, Wei Bi, Meng Fang, and Dacheng Tao. 2020. "Pretrained Language Models for Dialogue Generation with Multiple Input Sources." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2010.07576.
CDC. 2023. "SARS-CoV-2 Variant Classifications and Definitions." Centers for Disease Control and Prevention. March 28, 2023. http://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html.
Chandramouli, C., and Registrar General. 2011. "Census of India 2011." *Provisional Population Totals. New Delhi: Government of India*, 409–13.
Charoenkwan, Phasit, Chanin Nantasenamat, Md Mehedi Hasan, Balachandran Manavalan, and Watshara Shoombuatong. 2021. "BERT4Bitter: A Bidirectional Encoder Representations from Transformers (BERT)-Based Model for Improving the Prediction of Bitter Peptides." *Bioinformatics* 37 (17): 2556–62.
Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "SMOTE: Synthetic

Minority Over-Sampling Technique." *Journal of Artificial Intelligence Research* 16 (June): 321–57.
Cheng, Tzu-Heng, Yi-Da Sie, Kuang-Hung Hsu, Zhong Ning Leonard Goh, Cheng-Yu Chien, Hsien-Yi Chen, Chip-Jin Ng, et al. 2020. "Shock Index: A Simple and Effective Clinical Adjunct in Predicting 60-Day Mortality in Advanced Cancer Patients at the Emergency Department." *International Journal of Environmental Research and Public Health* 17 (13). https://doi.org/10.3390/ijerph17134904.
Chen, Mingda, and Kevin Gimpel. 2021. "TVStoryGen: A Dataset for Generating Stories with Character Descriptions." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2109.08833.
Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. KDD '16. New York, NY, USA: Association for Computing Machinery.
Chetoui, Mohamed, and Moulay A. Akhloufi. 2022. "Explainable Vision Transformers and Radiomics for COVID-19 Detection in Chest X-Rays." *Journal of Clinical Medicine Research* 11 (11). https://doi.org/10.3390/jcm11113013.
Cho, Hyejin, and Hyunju Lee. 2019. "Biomedical Named Entity Recognition Using Deep Neural Networks with Contextual Information." *BMC Bioinformatics* 20 (1): 735.
Choi, Jeong-Hyeon, and Hwan-Gue Cho. 2002. "Analysis of Common K-Mers for Whole Genome Sequences Using SSB-Tree." *Genome Informatics. International Conference on Genome Informatics* 13: 30–41.
Choi, Youngduck, Chill Yi-I Chiu, and David Sontag. 2016. "Learning Low-Dimensional Representations of Medical Concepts." *AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science* 2016 (July): 41–50.
Christ, Maximilian, Nils Braun, Julius Neuffer, and Andreas W. Kempa-Liehr. 2018. "Time Series FeatuRe Extraction on Basis of Scalable Hypothesis Tests (tsfresh – A Python Package)." *Neurocomputing* 307 (September): 72–77.
Cinelli, Matteo, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. "The COVID-19 Social Media Infodemic." *Scientific Reports* 10 (1): 16598.
Clusmann, Jan, Fiona R. Kolbinger, Hannah Sophie Muti, Zunamys I. Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, et al. 2023. "The Future Landscape of Large Language Models in Medicine." *Communication & Medicine* 3 (1): 141.
"Coronavirus (COVID-19) Outbreak Glossary." 2020. KFF. March 18, 2020. https://www.kff.org/glossary/covid-19-outbreak-glossary/.
Cosgriff, Christopher V., and Leo Anthony Celi. 2020. "Exploiting Temporal Relationships in the Prediction of Mortality." *The Lancet. Digital Health*.
"Covid19 Statewise Status." 2020. MyGov.in. March 28, 2020. https://www.mygov.in/corona-data/covid19-statewise-status/.
Craig, Erin, Carlos Arias, and David Gillman. 2017. "Predicting Readmission Risk from Doctors' Notes." *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1711.10663.
Dalla-Torre, Hugo, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, et al. 2023. "The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics." *bioRxiv*. https://doi.org/10.1101/2023.01.11.523679.

Daniels, Lori B., Junting Ren, Kris Kumar, Quan M. Bui, Jing Zhang, Xinlian Zhang, Mariem A. Sawan, Howard Eisen, Christopher A. Longhurst, and Karen Messer. 2021. “Relation of Prior Statin and Anti-Hypertensive Use to Severity of Disease among Patients Hospitalized with COVID-19: Findings from the American Heart Association’s COVID-19 Cardiovascular Disease Registry.” *PloS One* 16 (7): e0254635.

Davenport, Thomas, and Ravi Kalakota. 2019. “The Potential for Artificial Intelligence in Healthcare.” *Future Healthcare Journal* 6 (2): 94–98.

Desautels, Thomas, Jacob Calvert, Jana Hoffman, Melissa Jay, Yaniv Kerem, Lisa Shieh, David Shimabukuro, et al. 2016. “Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach.” *JMIR Medical Informatics* 4 (3): e5909.

“Detrended Fluctuation and Detrended Cross-Correlation Analysis [R Package DCCA Version 0.1.1].” 2020, January. https://CRAN.R-project.org/package=DCCA.

De Vine, Lance, Mahnoosh Kholghi, Guido Zuccon, Laurianne Sitbon, and Anthony Nguyen. 2015. “Analysis of Word Embeddings and Sequence Features for Clinical Information Extraction.” In , edited by B. Hachey and K. Webster, 21–30. Australia: Australasian Language Technology Association (ALTA).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1810.04805.

Doddapaneni, Sumanth, Gowtham Ramesh, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2021. “A Primer on Pretrained Multilingual Language Models.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2107.00676.

Dong, Wei. 2014. “The Detection of Fraudulent Financial Statements: An Integrated Language Model.” core.ac.uk. 2014. https://core.ac.uk/download/pdf/301362997.pdf.

Dubois, Sebastien, Nathanael Romano, David C. Kale, Nigam Shah, and Kenneth Jung. 2017. “Effective Representations of Clinical Notes.” *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1705.07025.

Du, Jingcheng, Peilin Jia, Yulin Dai, Cui Tao, Zhongming Zhao, and Degui Zhi. 2019. “Gene2vec: Distributed Representation of Genes Based on Co-Expression.” *BMC Genomics* 20 (Suppl 1): 82.

Ebadi, Ashkan, Pengcheng Xi, Stéphane Tremblay, Bruce Spencer, Raman Pall, and Alexander Wong. 2021. “Understanding the Temporal Evolution of COVID-19 Research through Machine Learning and Natural Language Processing.” *Scientometrics* 126 (1): 725–39.

Ehwerhemuepha, Louis, Theodore Heyming, Rachel Marano, Mary Jane Piroutek, Antonio C. Arrieta, Kent Lee, Jennifer Hayes, James Cappon, Kamila Hoenk, and William Feaster. 2021. “Development and Validation of an Early Warning Tool for Sepsis and Decompensation in Children during Emergency Department Triage.” *Scientific Reports* 11 (1): 8578.

Eklund, Anton, Mona Forsman, and Frank Drewes. 2022. “Dynamic Topic Modeling by Clustering Embeddings from Pretrained Language Models: A Research Proposal.” In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, 84–91. Online: Association for Computational Linguistics.

Elnaggar, Ahmed, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion

Jones, Tom Gibbs, et al. 2022. “ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (10): 7112–27.
Emanuel, Ezekiel J., Govind Persad, Ross Upshur, Beatriz Thome, Michael Parker, Aaron Glickman, Cathy Zhang, Connor Boyle, Maxwell Smith, and James P. Phillips. 2020. “Fair Allocation of Scarce Medical Resources in the Time of Covid-19.” *The New England Journal of Medicine* 382 (21): 2049–55.
“EPI-WIN Updates.” 2023. Accessed June 27, 2023. https://www.who.int/teams/risk-communication/epi-win-updates.
Esteva, Andre, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. 2021. “Deep Learning-Enabled Medical Computer Vision.” *NPJ Digital Medicine* 4 (1): 5.
Faruqui, Manaal, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2014. “Retrofitting Word Vectors to Semantic Lexicons.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1411.4166.
Freund, Yoav, Robert E. Schapire AT, and T. Labs. 1999. “A Short Introduction to Boosting.” yorku.ca. 1999. http://www.yorku.ca/gisweb/eats4400/boost.pdf.
Freund, Yoav, and Robert E. Schapire. 1997. “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting.” *Journal of Computer and System Sciences* 55 (1): 119–39.
Friedman, Jerome H. 2002. “Stochastic Gradient Boosting.” *Computational Statistics & Data Analysis* 38 (4): 367–78.
Garvin, Erica. 2019. “8 Use Cases for Natural Language Processing (NLP) Technology in Healthcare.” HIT Consultant Media. January 22, 2019. https://hitconsultant.net/2019/01/22/natural-language-processing-nlp-technology-use-cases/.
Gehrmann, Sebastian, Franck Dernoncourt, Yeran Li, Eric T. Carlson, Joy T. Wu, Jonathan Welt, John Foote Jr, et al. 2018. “Comparing Deep Learning and Concept Extraction Based Methods for Patient Phenotyping from Clinical Narratives.” *PloS One* 13 (2): e0192360.
“Genomic Data.” 2023. Snowflake. Accessed June 29, 2023. https://www.snowflake.com/trending/genomic-data.
“Gensim.” 2023. PyPI. Accessed June 28, 2023. https://pypi.org/project/gensim/.
Geurts, Pierre, Damien Ernst, and Louis Wehenkel. 2006. “Extremely Randomized Trees.” *Machine Learning* 63 (1): 3–42.
“Global Research on Coronavirus Disease (COVID-19).” 2023. Accessed June 28, 2023. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov.
Goh, Kim Huat, Le Wang, Adrian Yong Kwang Yeow, Hermione Poh, Ke Li, Joannas Jie Lin Yeow, and Gamaliel Yu Heng Tan. 2021. “Artificial Intelligence in Sepsis Early Prediction and Diagnosis Using Unstructured Data in Healthcare.” *Nature Communications* 12 (1): 711.
Gourdeau, Daniel, Olivier Potvin, Patrick Archambault, Carl Chartrand-Lefebvre, Louis Dieumegarde, Reza Forghani, Christian Gagné, et al. 2022. “Tracking and Predicting COVID-19 Radiological Trajectory on Chest X-Rays Using Deep Learning.” *Scientific Reports* 12 (1): 1–14.
Greff, Klaus, Rupesh K. Srivastava, Jan Koutnik, Bas R. Steunebrink, and Jurgen Schmidhuber. 2017. “LSTM: A Search Space Odyssey.” *IEEE Transactions on Neural Networks and*

Learning Systems 28 (10): 2222–32.
Gultepe, Eren, Jeffrey P. Green, Hien Nguyen, Jason Adams, Timothy Albertson, and Ilias Tagkopoulos. 2014. "From Vital Signs to Clinical Outcomes for Patients with Sepsis: A Machine Learning Basis for a Clinical Decision Support System." *Journal of the American Medical Informatics Association: JAMIA* 21 (2): 315–25.
Güneş, İsmail, Şule Gündüz-Öğüdücü, and Zehra Çataltepe. 2016. "Link Prediction Using Time Series of Neighborhood-Based Node Similarity Scores." *Data Mining and Knowledge Discovery* 30 (1): 147–80.
Gu, Yu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2007.15779.
Hadfield, James, Colin Megill, Sidney M. Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A. Neher. 2018. "Nextstrain: Real-Time Tracking of Pathogen Evolution." *Bioinformatics* 34 (23): 4121–23.
Hagberg, Aric, Pieter Swart, and Daniel S Chult. 2008. "Exploring Network Structure, Dynamics, and Function Using Networkx." LA-UR-08-05495; LA-UR-08-5495. Los Alamos National Lab. (LANL), Los Alamos, NM (United States). https://www.osti.gov/biblio/960616.
Hammoud, Ibrahim, I. V. Ramakrishnan, Mark Henry, and Eric Morley. 2020. "Multimodal Early Septic Shock Prediction Model Using Lasso Regression with Decaying Response." In *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, 1–3. ieeexplore.ieee.org.
Hand, David, and Peter Christen. 2018. "A Note on Using the F-Measure for Evaluating Record Linkage Algorithms." *Statistics and Computing* 28 (3): 539–47.
"Health Data." 2023. TheFreeDictionary.com. Accessed July 7, 2023. https://medical-dictionary.thefreedictionary.com/health+data.
Hearst, M. A., S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. 1998. "Support Vector Machines." *IEEE Intelligent Systems and Their Applications* 13 (4): 18–28.
He, Kai, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2023. "A Survey of Large Language Models for Healthcare: From Data, Technology, and Applications to Accountability and Ethics." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2310.05694.
He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. "Deep Residual Learning for Image Recognition." *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1512.03385.
Henry, Katharine E., Roy Adams, Cassandra Parent, Hossein Soleimani, Anirudh Sridharan, Lauren Johnson, David N. Hager, et al. 2022. "Factors Driving Provider Adoption of the TREWS Machine Learning-Based Early Warning System and Its Effects on Sepsis Treatment Timing." *Nature Medicine* 28 (7): 1447–54.
Hie, Brian, Ellen Zhong, Bryan Bryson, and Bonnie Berger. 2020. "Learning Mutational Semantics." *Advances in Neural Information Processing Systems* 33: 9109–21.
Hie, Brian, Ellen D. Zhong, Bonnie Berger, and Bryan Bryson. 2021. "Learning the Language of Viral Evolution and Escape." *Science* 371 (6526): 284–88.
Hochreiter, S., and J. Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9 (8): 1735–80.
Huang, Bi, Yanmin Yang, Jun Zhu, Yan Liang, Huiqiong Tan, Litian Yu, Xin Gao, and Jiandong

Li. 2014. “Usefulness of the Admission Shock Index for Predicting Short-Term Outcomes in Patients with ST-Segment Elevation Myocardial Infarction.” *The American Journal of Cardiology* 114 (9): 1315–21.
Huang, Jian, Keyang Xu, and V. G. Vinod Vydiswaran. 2016. “Analyzing Multiple Medical Corpora Using Word Embedding.” In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, 527–33.
Huang, Kexin, Jaan Altosaar, and Rajesh Ranganath. 2019. “ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1904.05342.
Hu, Ke, Tara N. Sainath, Bo Li, Nan Du, Yanping Huang, Andrew M. Dai, Yu Zhang, Rodrigo Cabrera, Zhifeng Chen, and Trevor Strohman. 2023. “Massively Multilingual Shallow Fusion with Large Language Models.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2302.08917.
Hu, Yibo, Mohammadsaleh Hosseini, Erick Skorupa Parolin, Javier Osorio, Latifur Khan, Patrick Brandt, and Vito D’Orazio. 2022. “COnfliBERT: A Pre-Trained Language Model for Political Conflict and Violence.” In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5469–82. Seattle, United States: Association for Computational Linguistics.
Interim guidance. 2023. “Global Surveillance for COVID-19 Caused by Human Infection with COVID-19 Virus.” Accessed June 27, 2023. https://apps.who.int/iris/bitstream/handle/10665/331506/WHO-2019-nCoV-SurveillanceGuidance-2020.6-eng.pdf.
Ioannidis, John P. A., Eran Bendavid, Maia Salholz-Hillel, Kevin W. Boyack, and Jeroen Baas. 2022. “Massive Covidization of Research Citations and the Citation Elite.” *Proceedings of the National Academy of Sciences* 119 (28): e2204074119.
Ioannidis, John P. A., Michael E. Stuart, Shannon Brownlee, and Sheri A. Strite. 2017. “How to Survive the Medical Misinformation Mess.” *European Journal of Clinical Investigation* 47 (11): 795–802.
Istepanian, Robert S. H., and Turki Al-Anzi. 2018. “M-Health 2.0: New Perspectives on Mobile Health, Machine Learning and Big Data Analytics.” *Methods* 151 (December): 34–40.
Ji, Yanrong, Zhihan Zhou, Han Liu, and Ramana V. Davuluri. 2021. “DNABERT: Pre-Trained Bidirectional Encoder Representations from Transformers Model for DNA-Language in Genome.” *Bioinformatics* 37 (15): 2112–20.
Johnson, Alistair E. W., Tom J. Pollard, Lu Shen, Li-Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. “MIMIC-III, a Freely Accessible Critical Care Database.” *Scientific Data* 3 (May): 160035.
Jonkmans, Nils, Valérie D’Acremont, and Antoine Flahault. 2021. “Scoping Future Outbreaks: A Scoping Review on the Outbreak Prediction of the WHO Blueprint List of Priority Diseases.” *BMJ Global Health* 6 (9). https://doi.org/10.1136/bmjgh-2021-006623.
Joseph, Andrew M., Virginia Fernandez, Sophia Kritzman, Isabel Eaddy, Olivia M. Cook, Sarah Lambros, Cesar E. Jara Silva, et al. 2022. “COVID-19 Misinformation on Social Media: A Scoping Review.” *Cureus* 14 (4): e24601.
Kaka, N. 2019. “Digital India: Technology to Transform a Connected Nation.”
Kapoor, Amol, Xue Ben, Luyang Liu, Bryan Perozzi, Matt Barnes, Martin Blais, and Shawn O’Banion. 2020. “Examining COVID-19 Forecasting Using Spatio-Temporal Graph Neural

Networks.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2007.03113.
Kearns, Michael, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 10--15 Jul 2018. “Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness.” In *Proceedings of the 35th International Conference on Machine Learning*, edited by Jennifer Dy and Andreas Krause, 80:2564–72. Proceedings of Machine Learning Research. PMLR.
Kennedy, Jason N., and Kristina E. Rudd. 2022. “A Sepsis Early Warning System Is Associated with Improved Patient Outcomes.” *Cell Reports. Medicine*. Elsevier.
Kholghi, Mahnoosh, Lance De Vine, Laurianne Sitbon, Guido Zuccon, and Anthony Nguyen. 2016. “The Benefits of Word Embeddings Features for Active Learning in Clinical Information Extraction.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1607.02810.
Koch, Erica, Shannon Lovett, Trac Nghiem, Robert A. Riggs, and Megan A. Rech. 2019. “Shock Index in the Emergency Department: Utility and Limitations.” *Open Access Emergency Medicine: OAEM* 11 (August): 179–99.
Kodali, Prakash Babu, Sibasis Hense, Swarajya Kopparty, Gangadhar Rao Kalapala, and Banashri Haloi. 2020. “How Indians Responded to the Arogya Setu App?” *Indian Journal of Public Health* 64 (Supplement): S228–30.
Kong, Hyoun-Joong. 2019. “Managing Unstructured Big Data in Healthcare System.” *Healthcare Informatics Research* 25 (1): 1–2.
Kormilitzin, Andrey, Nemanja Vaci, Qiang Liu, and Alejo Nevado-Holgado. 2021. “Med7: A Transferable Clinical Natural Language Processing Model for Electronic Health Records.” *Artificial Intelligence in Medicine* 118 (August): 102086.
Kouzy, Ramez, Joseph Abi Jaoude, Afif Kraitem, Molly B. El Alam, Basil Karam, Elio Adib, Jabra Zarka, Cindy Traboulsi, Elie W. Akl, and Khalil Baddour. 2020. “Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter.” *Cureus* 12 (3): e7255.
Kryściński, Wojciech, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. “Improving Abstraction in Text Summarization.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1808.07913.
Kumawat, Spraha, Inna Yadav, Nisha Pahal, and Deepti Goel. 2021. “Sentiment Analysis Using Language Models: A Study.” In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 984–88.
Kursa, Miron B., Aleksander Jankowski, and Witold R. Rudnicki. 2010. “Boruta – A System for Feature Selection.” *Fundamenta Informaticae* 101 (4): 271–85.
Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. “Diachronic Word Embeddings and Semantic Shifts: A Survey.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1806.03537.
Lalor, John, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. “Benchmarking Intersectional Biases in NLP.” In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3598–3609. Seattle, United States: Association for Computational Linguistics.
Lau, Ivan Shun, Zeljko Kraljevic, Mohammad Al-Agil, Shelley Charing, Alan Quarterman, Harold Parkes, Victoria Metaxa, et al. 2021. “Natural Language Word Embeddings as a Glimpse into Healthcare Language and Associated Mortality Surrounding End of Life.” *BMJ Health & Care Informatics* 28 (1). https://doi.org/10.1136/bmjhci-2021-100464.
LaValley, Michael P. 2008. “Logistic Regression.” *Circulation* 117 (18): 2395–99.
Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and

Jaewoo Kang. 2020. “BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining.” *Bioinformatics* 36 (4): 1234–40.
Lee, Kathy, Sadid A. Hasan, Oladimeji Farri, Alok Choudhary, and Ankit Agrawal. 2017. “Medical Concept Normalization for Online User-Generated Texts.” In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, 462–69. IEEE.
Le, Nguyen Quoc Khanh, Quang-Thai Ho, Trinh-Trung-Duong Nguyen, and Yu-Yen Ou. 2021. “A Transformer Architecture Based on BERT and 2D Convolutional Neural Network to Identify DNA Enhancers from Sequence Information.” *Briefings in Bioinformatics* 22 (5). https://doi.org/10.1093/bib/bbab005.
Le, Quoc, and Tomas Mikolov. 22--24 Jun 2014. “Distributed Representations of Sentences and Documents.” In *Proceedings of the 31st International Conference on Machine Learning*, edited by Eric P. Xing and Tony Jebara, 32:1188–96. Proceedings of Machine Learning Research. Bejing, China: PMLR.
Letunic, Ivica, and Peer Bork. 2021. “Interactive Tree Of Life (iTOL) v5: An Online Tool for Phylogenetic Tree Display and Annotation.” *Nucleic Acids Research* 49 (W1): W293–96.
Lever, Jake, and Russ B. Altman. 2021. “Analyzing the Vast Coronavirus Literature with CoronaCentral.” *Proceedings of the National Academy of Sciences of the United States of America* 118 (23). https://doi.org/10.1073/pnas.2100766118.
Liaw, Andy, and Matthew Wiener. 2002. “Classification and Regression by randomForest.” cogns.northwestern.edu. 2002. https://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf.
Li, Hong-Liang, Yi-He Pang, and Bin Liu. 2021. “BioSeq-BLM: A Platform for Analyzing DNA, RNA and Protein Sequences Based on Biological Language Models.” *Nucleic Acids Research* 49 (22): e129.
Linderman, George C., Manas Rachh, Jeremy G. Hoskins, Stefan Steinerberger, and Yuval Kluger. 2019. “Fast Interpolation-Based T-SNE for Improved Visualization of Single-Cell RNA-Seq Data.” *Nature Methods* 16 (3): 243–45.
Lin, Jianghao, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Xiangyang Li, Chenxu Zhu, et al. 2023. “How Can Recommender Systems Benefit from Large Language Models: A Survey.” *arXiv [cs.IR]*. arXiv. http://arxiv.org/abs/2306.05817.
Lin, Zeming, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, et al. 2022. “Language Models of Protein Sequences at the Scale of Evolution Enable Accurate Structure Prediction.” *bioRxiv*. https://doi.org/10.1101/2022.07.20.500902.
Li, Qianqian, Jiajing Wu, Jianhui Nie, Li Zhang, Huan Hao, Shuo Liu, Chenyan Zhao, et al. 2020. “The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity.” *Cell* 182 (5): 1284–94.e9.
Liu, Boxiang, Kaibo Liu, He Zhang, Liang Zhang, Yuchen Bian, and Liang Huang. 2020. “CoV-Seq, a New Tool for SARS-CoV-2 Genome Analysis and Visualization: Development and Usability Study.” *Journal of Medical Internet Research* 22 (10): e22299.
Liu, Yiheng, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, et al. 2023. “Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2304.01852.
Liu, Yue, Tao Ge, Kusum Mathews, Heng Ji, and Deborah McGuinness. 2015. “Exploiting Task-Oriented Resources to Learn Word Embeddings for Clinical Abbreviation Expansion.” In *Proceedings of BIoNLP 15*, edited by Kevin Bretonnel Cohen, Dina Demner-Fushman, Sophia Ananiadou, and Jun-Ichi Tsujii, 92–97. Beijing, China: Association for

Computational Linguistics.
Liu, Zhichao, Ruth A. Roberts, Madhu Lal-Nag, Xi Chen, Ruili Huang, and Weida Tong. 2021. "AI-Based Language Models Powering Drug Discovery and Development." *Drug Discovery Today* 26 (11): 2593–2607.
Loomba, Sahil, Alexandre de Figueiredo, Simon J. Piatek, Kristen de Graaf, and Heidi J. Larson. 2021. "Measuring the Impact of COVID-19 Vaccine Misinformation on Vaccination Intent in the UK and USA." *Nature Human Behaviour* 5 (3): 337–48.
Loper, Edward. 2023. *NLTK: The Natural Language Toolkit*.
Lopez-Leon, Sandra, Talia Wegman-Ostrosky, Carol Perelman, Rosalinda Sepulveda, Paulina A. Rebolledo, Angelica Cuapio, and Sonia Villapol. 2021. "More than 50 Long-Term Effects of COVID-19: A Systematic Review and Meta-Analysis." *Scientific Reports* 11 (1): 16144.
Łozinski, Tomasz, Wojciech T. Markiewicz, Tadeusz K. Wyrzykiewicz, and Kazimierz L. Wierzchowski. 1989. "Effect of the Sequence-Dependent Structure of the 17 Bp AT Spacer on the Strength of Consensus-like E.coli Promoters in Vivo." *Nucleic Acids Research* 17 (10): 3855–63.
Lundberg, Scott M., and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 30. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.
Luo, Xiao, Priyanka Gandhi, Susan Storey, and Kun Huang. 2022. "A Deep Language Model for Symptom Extraction From Clinical Text and Its Application to Extract COVID-19 Symptoms From Social Media." *IEEE Journal of Biomedical and Health Informatics* 26 (4): 1737–48.
Lv, Hao, Lei Shi, Joshua William Berkenpas, Fu-Ying Dao, Hasan Zulfiqar, Hui Ding, Yang Zhang, Liming Yang, and Renzhi Cao. 2021. "Application of Artificial Intelligence and Machine Learning for COVID-19 Drug Discovery and Vaccine Design." *Briefings in Bioinformatics* 22 (6). https://doi.org/10.1093/bib/bbab320.
Madani, Ali, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R. Eguchi, Po-Ssu Huang, and Richard Socher. 2020. "ProGen: Language Modeling for Protein Generation." *arXiv [q-bio.BM]*. arXiv. http://arxiv.org/abs/2004.03497.
Maher, M. Cyrus, Istvan Bartha, Steven Weaver, Julia di Iulio, Elena Ferri, Leah Soriaga, Florian A. Lempp, et al. 2022. "Predicting the Mutational Drivers of Future SARS-CoV-2 Variants of Concern." *Science Translational Medicine* 14 (633): eabk3445.
Maheshwari, Kamal, Brian H. Nathanson, Sibyl H. Munson, Seungyoung Hwang, Halit O. Yapici, Mitali Stevens, Carlos Ruiz, and Charles F. Hunley. 2020. "Abnormal Shock Index Exposure and Clinical Outcomes among Critically Ill Patients: A Retrospective Cohort Analysis." *Journal of Critical Care* 57 (June): 5–12.
Ma, Long, and Yanqing Zhang. 2015. "Using Word2Vec to Process Big Text Data." In *2015 IEEE International Conference on Big Data (Big Data)*, 2895–97. ieeexplore.ieee.org.
Mandal, Saurav, Tanmoy Roychowdhury, and Alok Bhattacharya. 2021. "Pattern of Genomic Variation in SARS-CoV-2 (COVID-19) Suggests Restricted Nonrandom Changes: Analysis Using Shewhart Control Charts." *Journal of Biosciences* 46 (1). https://doi.org/10.1007/s12038-020-00131-5.
Mayer, Tobias, Elena Cabrio, and Serena Villata. 2020. "Transformer-Based Argument Mining for Healthcare Applications." In *ECAI 2020*, 2108–15. Amsterdam, NY: IOS Press.
McGaughey, Jennifer, Dean A. Fergusson, Peter Van Bogaert, and Louise Rose. 2021. "Early

Warning Systems and Rapid Response Systems for the Prevention of Patient Deterioration on Acute Adult Hospital Wards." *Cochrane Database of Systematic Reviews* 11 (11): CD005529.

Melin, Patricia, Julio Cesar Monica, Daniela Sanchez, and Oscar Castillo. 2020. "Multiple Ensemble Neural Network Models with Fuzzy Response Aggregation for Predicting COVID-19 Time Series: The Case of Mexico." *Healthcare (Basel, Switzerland)* 8 (2). https://doi.org/10.3390/healthcare8020181.

Mencía, Eneldo Loza, Gerard de Melo, and Jinseok Nam. 2016. "Medical Concept Embeddings via Labeled Background Corpora." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4629–36. Portorož, Slovenia: European Language Resources Association (ELRA).

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1301.3781.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. "Distributed Representations of Words and Phrases and Their Compositionality." *Advances in Neural Information Processing Systems* 26. https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

"Modeling COVID-19 Scenarios for the United States." 2020. *Nature Medicine* 27 (1): 94–105.

Mo, Shentong, Xi Fu, Chenyang Hong, Yizhen Chen, Yuxuan Zheng, Xiangru Tang, Zhiqiang Shen, Eric P. Xing, and Yanyan Lan. 2021. "Multi-Modal Self-Supervised Pre-Training for Regulatory Genome Across Cell Types." *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/2110.05231.

Mugnai, Mauro L., Clark Templeton, Ron Elber, and D. Thirumalai. 2020. "Role of Long-Range Allosteric Communication in Determining the Stability and Disassembly of SARS-COV-2 in Complex with ACE2." *bioRxiv : The Preprint Server for Biology*, December. https://doi.org/10.1101/2020.11.30.405340.

Mulyar, Andriy, Elliot Schumacher, Masoud Rouhizadeh, and Mark Dredze. 2019. "Phenotyping of Clinical Notes with Improved Document Classification Models Using Contextualized Neural Language Models." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1910.13664.

Mushtaq, Rizwan. 2011. *Augmented Dickey Fuller Test*. SSRN.

Nagori, Aditya, Lovedeep Singh Dhingra, Ambika Bhatnagar, Rakesh Lodha, and Tavpritesh Sethi. 2019. "Predicting Hemodynamic Shock from Thermal Images Using Machine Learning." *Scientific Reports* 9 (1): 91.

Nagori, Aditya, Pradeep Singh, Sameena Firdos, Vanshika Vats, Arushi Gupta, Harsh Bandhey, Anushtha Kalia, et al. 2021. "Generalized Prediction of Hemodynamic Shock in Intensive Care Units." *bioRxiv*. medRxiv. https://doi.org/10.1101/2021.01.07.21249121.

"Natural Language Processing 101: A Guide to NLP in Clinical Documentation." 2022. *IMO* (blog). April 6, 2022. https://www.imohealth.com/ideas/article/natural-language-processing-101-a-guide-to-nlp-in-clinical-documentation/.

"Natural Language Processing in Healthcare Medical Records." 2023.ForeSee Medical. Accessed June 27, 2023. https://www.foreseemed.com/natural-language-processing-in-healthcare.

Neumann, Mark, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. "ScispaCy: Fast and

Robust Models for Biomedical Natural Language Processing.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1902.07669.
Ng, Patrick. 2017. “dna2vec: Consistent Vector Representations of Variable-Length K-Mers.” *arXiv [q-bio.QM]*. arXiv. http://arxiv.org/abs/1701.06279.
Nguyen, Phuoc, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. 2017. “$\mathtt {Deepr}$: A Convolutional Net for Medical Records.” *IEEE Journal of Biomedical and Health Informatics* 21 (1): 22–30.
Ofstad, Eirik H., Jan C. Frich, Edvin Schei, Richard M. Frankel, and Pål Gulbrandsen. 2014. “Temporal Characteristics of Decisions in Hospital Encounters: A Threshold for Shared Decision Making? A Qualitative Study.” *Patient Education and Counseling* 97 (2): 216–22.
Ogungbe, Oluwabunmi, Amal K. Mitra, and Joni K. Roberts. 2019. “A Systematic Review of Implicit Bias in Health Care: A Call for Intersectionality.” *IMC Journal of Medical Science* 13 (1): 005.
Özcan, Alper, and Şule Gündüz Öğüdücü. 2017. “Supervised Temporal Link Prediction Using Time Series of Similarity Measures.” In *2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN)*, 519–21. ieeexplore.ieee.org.
Pal, Ridam, Rohan Pandey, Vaibhav Gautam, Kanav Bhagat, and Tavpritesh Sethi. 2020. “A Cross-Lingual Natural Language Processing Framework for Infodemic Management.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2010.16357.
Panhuis, Willem G. van, Proma Paul, Claudia Emerson, John Grefenstette, Richard Wilder, Abraham J. Herbst, David Heymann, and Donald S. Burke. 2014. “A Systematic Review of Barriers to Data Sharing in Public Health.” *BMC Public Health* 14 (November): 1144.
Pan, Yu-Hsiang, Yung-Hung Wang, Sheng-Fu Liang, and Kuo-Tien Lee. 2011. “Fast Computation of Sample Entropy and Approximate Entropy in Biomedicine.” *Computer Methods and Programs in Biomedicine* 104 (3): 382–96.
Pappagari, Raghavendra, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. “Hierarchical Transformers for Long Document Classification.” In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 838–44. ieeexplore.ieee.org.
Park, Minjun, Seung-Woo Seo, Eunyoung Park, and Jinhan Kim. 2022. “EpiBERTope: A Sequence-Based Pre-Trained BERT Model Improves Linear and Structural Epitope Prediction by Learning Long-Distance Protein Interactions Effectively.” *bioRxiv*. https://doi.org/10.1101/2022.02.27.481241.
Patel, Kevin, Divya Patel, Mansi Golakiya, Pushpak Bhattacharyya, and Nilesh Birari. 2017. “Adapting Pre-Trained Word Embeddings For Use In Medical Coding.” In *BIoNLP 2017*, 302–6. Vancouver, Canada,: Association for Computational Linguistics.
Patel, Mohan P., Vivek B. Kute, Sanjay K. Agarwal, and COVID-19 Working Group of Indian Society of Nephrology. 2020. “‘Infodemic’ COVID 19: More Pandemic than the Virus.” *Indian Journal of Nephrology* 30 (3): 188–91.
Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2012. “Scikit-Learn: Machine Learning in Python.” *arXiv [cs.LG]*. arXiv. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https:/.
Pendergrass, Sarah A., and Dana C. Crawford. 2019. “Using Electronic Health Records To Generate Phenotypes For Research.” *Current Protocols in Human Genetics / Editorial Board, Jonathan L. Haines ... [et Al.]* 100 (1): e80.
Peng, Baolin, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. “Instruction

Tuning with GPT-4.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2304.03277.
Peng, Yifan, Shankai Yan, and Zhiyong Lu. 2019. “Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1906.05474.
Peymani, Payam, Tania Dehesh, Farnaz Aligolighasemabadi, Mohammadamin Sadeghdoust, Katarzyna Kotfis, Mazaher Ahmadi, Parvaneh Mehrbod, et al. 2021. “Statins in Patients with COVID-19: A Retrospective Cohort Study in Iranian COVID-19 Patients.” *Translational Medicine Communications* 6 (1): 1–14.
Pham, Trang, Truyen Tran, Dinh Phung, and Svetha Venkatesh. 2016. “DeepCare: A Deep Dynamic Memory Model for Predictive Medicine.” In *Advances in Knowledge Discovery and Data Mining*, 30–41. Springer International Publishing.
Pian, Wenjing, Jianxing Chi, and Feicheng Ma. 2021. “The Causes, Impacts and Countermeasures of COVID-19 ‘Infodemic’: A Systematic Review Using Narrative Synthesis.” *Information Processing & Management* 58 (6): 102713.
Polnaszek, Brock, Andrea Gilmore-Bykovskyi, Melissa Hovanes, Rachel Roiland, Patrick Ferguson, Roger Brown, and Amy J. H. Kind. 2016. “Overcoming the Challenges of Unstructured Data in Multisite, Electronic Medical Record-Based Abstraction.” *Medical Care* 54 (10): e65–72.
Potes, Cristhian, Bryan Conroy, Minnan Xu-Wilson, Christopher Newth, David Inwald, and Joseph Frassica. 2017. “A Clinical Prediction Model to Identify Patients at High Risk of Hemodynamic Instability in the Pediatric Intensive Care Unit.” *Critical Care / the Society of Critical Care Medicine* 21 (1): 282.
Powers, David M. W. 2020. “Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2010.16061.
Prass, Taiane Schaedler, and Guilherme Pumi. 2019. “On the Behavior of the DFA and DCCA in Trend-Stationary Processes.” *arXiv [math.ST]*. arXiv. http://arxiv.org/abs/1910.10589.
Qin, Lei, Qiang Sun, Yidan Wang, Ke-Fei Wu, Mingchih Chen, Ben-Chang Shia, and Szu-Yuan Wu. 2020. “Prediction of Number of Cases of 2019 Novel Coronavirus (COVID-19) Using Social Media Search Index.” *International Journal of Environmental Research and Public Health* 17 (7). https://doi.org/10.3390/ijerph17072365.
Rabiner, L. R. 1989. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.” *Proceedings of the IEEE* 77 (2): 257–86.
Rahman, Asif, Yale Chang, Junzi Dong, Bryan Conroy, Annamalai Natarajan, Takahiro Kinoshita, Francesco Vicario, Joseph Frassica, and Minnan Xu-Wilson. 2021. “Early Prediction of Hemodynamic Interventions in the Intensive Care Unit Using Machine Learning.” *Critical Care / the Society of Critical Care Medicine* 25 (1): 388.
Rajpurkar, Pranav, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, et al. 2017. “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning.” *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1711.05225.
Ramesh, Sidharth, Abhiraj Tiwari, Parthivi Choubey, Saisha Kashyap, Sahil Khose, Kumud Lakara, Nishesh Singh, and Ujjwal Verma. 2021. “BERT Based Transformers Lead the Way in Extraction of Health Information from Social Media.” In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, 33–38. Mexico City, Mexico: Association for Computational Linguistics.
Rana, Abhijeet, Aman Garg, Rishabh Jain, and Aditi Sharma. 2023. “Audio Examination –

Mental Health Diagnosis in Healthcare through Audio Analytics." In *Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing*, 24–29. IC3-2023. New York, NY, USA: Association for Computing Machinery.

Řehůřek, Radim, and Petr Sojka. 2010. "Software Framework for Topic Modelling with Large Corpora," January. https://paperswithcode.com/paper/software-framework-for-topic-modelling-with.

Rezaei, Nima. 2021. *Coronavirus Disease - COVID-19*. Springer Nature.

Rocha, Yasmim Mendes, Gabriel Acácio de Moura, Gabriel Alves Desidério, Carlos Henrique de Oliveira, Francisco Dantas Lourenço, and Larissa Deadame de Figueiredo Nicolete. 2023. "The Impact of Fake News on Social Media and Its Influence on Health during the COVID-19 Pandemic: A Systematic Review." *Journal of Public Health* 31 (7): 1007–16.

Rodríguez, Alexander, Anika Tabassum, Jiaming Cui, Jiajia Xie, Javen Ho, Pulak Agarwal, Bijaya Adhikari, and B. Aditya Prakash. 2021. "DeepCOVID: An Operational Deep Learning-Driven Framework for Explainable Real-Time COVID-19 Forecasting." *Proceedings of the AAAI Conference on Artificial Intelligence* 35 (17): 15393–400.

Rosvall, Martin, and Carl T. Bergstrom. 2010. "Mapping Change in Large Networks." *PloS One* 5 (1): e8694.

Rouchka, Eric C., Julia H. Chariker, and Donghoon Chung. 2020. "Variant Analysis of 1,040 SARS-CoV-2 Genomes." *PloS One* 15 (11): e0241535.

Sahlgren, Magnus. 2008. "The Distributional Hypothesis." *Italian Journal of Disability Studies* 20: 33–53.

Samyoun, Sirat, Sudipta Saha Shubha, Md Abu Sayeed Mondol, and John A. Stankovic. 2021. "iWash: A Smartwatch Handwashing Quality Assessment and Reminder System with Real-Time Feedback in the Context of Infectious Disease." *Smart Health (Amsterdam, Netherlands)* 19 (March): 100171.

Sanchez-Martinez, Sergio, Oscar Camara, Gemma Piella, Maja Cikes, Miguel Ángel González-Ballester, Marius Miron, Alfredo Vellido, Emilia Gómez, Alan G. Fraser, and Bart Bijnens. 2021. "Machine Learning for Clinical Decision-Making: Challenges and Opportunities in Cardiovascular Imaging." *Frontiers in Cardiovascular Medicine* 8: 765693.

Sariyar, Murat, Stephanie Suhr, and Irene Schlünder. 2017. "How Sensitive Is Genetic Data?" *Biopreservation and Biobanking* 15 (6): 494–501.

"Scispacy." 2023. Scispacy. Accessed June 28, 2023. https://allenai.github.io/scispacy/.

Seymour, Christopher W., Vincent X. Liu, Theodore J. Iwashyna, Frank M. Brunkhorst, Thomas D. Rea, André Scherag, Gordon Rubenfeld, et al. 2016. "Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)." *JAMA: The Journal of the American Medical Association* 315 (8): 762–74.

Seyyed-Kalantari, Laleh, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. 2021. "Underdiagnosis Bias of Artificial Intelligence Algorithms Applied to Chest Radiographs in under-Served Patient Populations." *Nature Medicine* 27 (12): 2176–82.

Shams, Abdullah Bin, Ehsanul Hoque Apu, Ashiqur Rahman, Md Mohsin Sarker Raihan, Nazeeba Siddika, Rahat Bin Preo, Molla Rashied Hussein, Shabnam Mostari, and Russell Kabir. 2021. "Web Search Engine Misinformation Notifier Extension (SEMiNExt): A Machine Learning Based Approach during COVID-19 Pandemic." *Healthcare (Basel, Switzerland)* 9 (2). https://doi.org/10.3390/healthcare9020156.

Shang, Junyuan, Tengfei Ma, Cao Xiao, and Jimeng Sun. 2019. “Pre-Training of Graph Augmented Transformers for Medication Recommendation.” *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1906.00346.

Shao, Zhenwei, Zhou Yu, Meng Wang, and Jun Yu. 2023. “Prompting Large Language Models with Answer Heuristics for Knowledge-Based Visual Question Answering.” *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/2303.01903.

Sharma, Avinash, K. D. V. Prasad, Sadashiva V. Chakrasali, Dankan Gowda V, Chanakya Kumar, Abhay Chaturvedi, and A. Azhagu Jaisudhan Pazhani. 2023. “Computer Vision Based Healthcare System for Identification of Diabetes & Its Types Using AI.” *Measurement: Sensors* 27 (June): 100751.

Sharma, Richa, Sudha Morwal, and Basant Agarwal. 2022. “Named Entity Recognition Using Neural Language Model and CRF for Hindi Language.” *Computer Speech & Language* 74 (July): 101356.

Sherstinsky, Alex. 2020. “Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network.” *Physica D. Nonlinear Phenomena* 404 (March): 132306.

Shishir, Tushar Ahmed, Iftekhar Bin Naser, and Shah M. Faruque. 2021. “In Silico Comparative Genomics of SARS-CoV-2 to Determine the Source and Diversity of the Pathogen in Bangladesh.” *PloS One* 16 (1): e0245584.

Shoemaker, W. C. 1996. “Temporal Physiologic Patterns of Shock and Circulatory Dysfunction Based on Early Descriptions by Invasive and Noninvasive Monitoring.” *New Horizons* 4 (2): 300–318.

Shome, Debaditya, T. Kar, Sachi Nandan Mohanty, Prayag Tiwari, Khan Muhammad, Abdullah AlTameem, Yazhou Zhang, and Abdul Khader Jilani Saudagar. 2021. “COVID-Transformer: Interpretable COVID-19 Detection Using Vision Transformer for Healthcare.” *International Journal of Environmental Research and Public Health* 18 (21). https://doi.org/10.3390/ijerph182111086.

Shu, Yuelong, and John McCauley. 2017. “GISAID: Global Initiative on Sharing All Influenza Data - from Vision to Reality.” *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* 22 (13). https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.

Singh, Romil, Likhita Shaik, Ishita Mehra, Rahul Kashyap, and Salim Surani. 2020. “Novel and Controversial Therapies in COVID-19.” *The Open Respiratory Medicine Journal* 14 (December): 79–86.

Si, Yuqi, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. “Enhancing Clinical Concept Extraction with Contextual Embeddings.” *Journal of the American Medical Informatics Association: JAMIA* 26 (11): 1297–1304.

Smilkov, Daniel, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B. Viégas, and Martin Wattenberg. 2016. “Embedding Projector: Interactive Visualization and Interpretation of Embeddings.” *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1611.05469.

Smith, M. E. Beth, Joseph C. Chiovaro, Maya O’Neil, Devan Kansagara, Ana R. Quiñones, Michele Freeman, Makalapua L. Motu’apuaka, and Christopher G. Slatore. 2014. “Early Warning System Scores for Clinical Deterioration in Hospitalized Patients: A Systematic Review.” *Annals of the American Thoracic Society* 11 (9): 1454–65.

Song, Xinying, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2020. “Fast WordPiece Tokenization.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2012.15524.

Steinberg, Ethan, Ken Jung, Jason A. Fries, Conor K. Corbin, Stephen R. Pfohl, and Nigam H. Shah. 2021. “Language Models Are an Effective Representation Learning Technique for Electronic Health Record Data.” *Journal of Biomedical Informatics* 113 (January): 103637.

Sun, Mengxin, Jason Baron, Anand Dighe, Peter Szolovits, Richard G. Wunderink, Tamara Isakova, and Yuan Luo. 2019. “Early Prediction of Acute Kidney Injury in Critical Care Setting Using Clinical Notes and Structured Multivariate Physiological Measurements.” *Studies in Health Technology and Informatics* 264 (August): 368–72.

Taboubi, Bilel -., Bechir Brahem, and Hatem Haddad. 2022. “ICompass at WANLP 2022 Shared Task: ARBERT And MARBERT for Multilabel Propaganda Classification of ARabic Tweets.” In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, 511–14. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics.

Tagliabue, Fabio, Luca Galassi, and Pierpaolo Mariani. 2020. “The ‘Pandemic’ of Disinformation in COVID-19.” *SN Comprehensive Clinical Medicine* 2 (9): 1287–89.

Tang, Gongbo, Mathias Müller, Annette Rios, and Rico Sennrich. 2018. “Why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1808.08946.

Tan, Yi Chern, and L. Elisa Celis. 2019. “Assessing Social and Intersectional Biases in Contextualized Word Representations.” *Advances in Neural Information Processing Systems* 32. https://proceedings.neurips.cc/paper/2019/hash/201d546992726352471cfea6b0df0a48-Abstract.html.

Tan, Yiming, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. “Evaluation of ChatGPT as a Question Answering System for Answering Complex Questions.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2303.07992.

Taquet, Maxime, Quentin Dercon, Sierra Luciano, John R. Geddes, Masud Husain, and Paul J. Harrison. 2021. “Incidence, Co-Occurrence, and Evolution of Long-COVID Features: A 6-Month Retrospective Cohort Study of 273,618 Survivors of COVID-19.” *PLoS Medicine* 18 (9): e1003773.

Tarcar, Amogh Kamat, Aashis Tiwari, Vineet Naique Dhaimodker, Penjo Rebelo, Rahul Desai, and Dattaraj Rao. 2019. “Healthcare NER Models Using Language Model Pretraining.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1910.11241.

Tasnim, Samia, Md Mahbub Hossain, and Hoimonty Mazumder. 2020. “Impact of Rumors and Misinformation on COVID-19 in Social Media.” *Journal of Preventive Medicine and Public Health = Yebang Uihakhoe Chi* 53 (3): 171–74.

Tekiroğlu, Serra Sinem, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. “Using Pre-Trained Language Models for Producing Counter Narratives Against Hate Speech: A Comparative Study.” In *Findings of the Association for Computational Linguistics: ACL 2022*, 3099–3114. Dublin, Ireland: Association for Computational Linguistics.

Tello, Manuel, Eric S. Reich, Jason Puckey, Rebecca Maff, Andres Garcia-Arce, Biplab Sudhin Bhattacharya, and Felipe Feijoo. 2022. “Machine Learning Based Forecast for the Prediction of Inpatient Bed Demand.” *BMC Medical Informatics and Decision Making* 22 (1): 55.

“Time Series Analysis and Computational Finance [R Package Tseries Version 0.10-54].” 2023, May. https://CRAN.R-project.org/package=tseries.

Tomcala, Jiri. 2018. “Time Series Entropies [R Package TSEntropies Version 0.9],” October.

https://CRAN.R-project.org/package=TSEntropies.
Tshitoyan, Vahe, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. 2019. “Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature.” *Nature* 571 (7763): 95–98.
Ullah, Zahid, Muhammad Usman, Siddique Latif, and Jeonghwan Gwak. 2023. “Densely Attention Mechanism Based Network for COVID-19 Detection in Chest X-Rays.” *Scientific Reports* 13 (1): 1–14.
Uludoğan, Gökçe, Elif Ozkirimli, Kutlu O. Ulgen, Nilgün Karalı, and Arzucan Özgür. 2022. “Exploiting Pretrained Biochemical Language Models for Targeted Drug Design.” *Bioinformatics* 38 (Suppl_2): ii155–61.
Varoquaux, Gaël, and Veronika Cheplygina. 2022. “Machine Learning for Medical Imaging: Methodological Failures and Recommendations for the Future.” *Npj Digital Medicine* 5 (1): 1–8.
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1706.03762.
Vats, Vanshika, Aditya Nagori, Pradeep Singh, Raman Dutt, Harsh Bandhey, Mahika Wason, Rakesh Lodha, and Tavpritesh Sethi. 2022. “Early Prediction of Hemodynamic Shock in Pediatric Intensive Care Units With Deep Learning on Thermal Videos.” *Frontiers in Physiology* 13 (July): 862411.
Wang, Longyue, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. “Document-Level Machine Translation with Large Language Models.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2304.02210.
Wang, Lucy Lu, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, et al. 2020. “CORD-19: The COVID-19 Open Research Dataset.” *ArXiv*, April. https://www.ncbi.nlm.nih.gov/pubmed/32510522.
Wang, Peipei, Xinqi Zheng, Jiayang Li, and Bangren Zhu. 2020. “Prediction of Epidemic Trends in COVID-19 with Logistic Model and Machine Learning Technics.” *Chaos, Solitons, and Fractals* 139 (October): 110058.
Wang, Ryan Yixiang, Tim Qinsong Guo, Leo Guanhua Li, Julia Yutian Jiao, and Lena Yiqi Wang. 2020. “Predictions of COVID-19 Infection Severity Based on Co-Associations between the SNPs of Co-Morbid Diseases and COVID-19 through Machine Learning of Genetic Data.” In *2020 IEEE 8th International Conference on Computer Science and Network Technology (ICCSNT)*, 92–96.
Wang, Yanshan, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. 2018. “A Comparison of Word Embeddings for the Biomedical Natural Language Processing.” *Journal of Biomedical Informatics* 87 (November): 12–20.
Wardi, Gabriel, Morgan Carlile, Andre Holder, Supreeth Shashikumar, Stephen R. Hayden, and Shamim Nemati. 2021. “Predicting Progression to Septic Shock in the Emergency Department Using an Externally Generalizable Machine-Learning Algorithm.” *Annals of Emergency Medicine* 77 (4): 395–406.
White, Jacob. 2020. “PubMed 2.0.” *Medical Reference Services Quarterly* 39 (4): 382–87.
“WHO Coronavirus (COVID-19) Dashboard.” 2023. Accessed June 27, 2023. https://covid19.who.int.

"WHO Health Alert Brings COVID-19 Facts to Billions via WhatsApp." 2023. Accessed June 27, 2023. http://www.who.int/news-room/feature-stories/detail/who-health-alert-brings-covid-19-facts-to-billions-via-whatsapp.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, et al. 2019. "HuggingFace's Transformers: State-of-the-Art Natural Language Processing." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1910.03771. 2020. "Transformers: State-of-the-Art Natural Language Processing." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.

World Health Organization. 2020. "Novel Coronavirus (2019-nCoV) Technical Guidance." https://pesquisa.bvsalud.org/portal/resource/in/lis-LISBR1.1-46991.

"Wrapper Algorithm for All Relevant Feature Selection [R Package Boruta Version 8.0.0]." 2022, November. https://CRAN.R-project.org/package=Boruta.

Wu, Feng, Runtao Yang, Chengjin Zhang, and Lina Zhang. 2021. "A Deep Learning Framework Combined with Word Embedding to Identify DNA Replication Origins." *Scientific Reports* 11 (1): 844.

Wu, Honghan, Minhong Wang, Jinge Wu, Farah Francis, Yun-Hsuan Chang, Alex Shavick, Hang Dong, et al. 2022. "A Survey on Clinical Natural Language Processing in the United Kingdom from 2007 to 2022." *Npj Digital Medicine* 5 (1): 1–15.

"XGBoost Documentation — Xgboost 1.7.6 Documentation." 2023. Accessed June 28, 2023. https://xgboost.readthedocs.io/en/stable/.

Xie, Feng, Han Yuan, Yilin Ning, Marcus Eng Hock Ong, Mengling Feng, Wynne Hsu, Bibhas Chakraborty, and Nan Liu. 2022. "Deep Learning for Temporal Data Representation in Electronic Health Records: A Systematic Review of Challenges and Methodologies." *Journal of Biomedical Informatics* 126 (February): 103980.

Yadav, Subhash Kumar, and Yusuf Akhter. 2021. "Statistical Modeling for the Prediction of Infectious Disease Dissemination With Special Reference to COVID-19 Spread." *Frontiers in Public Health* 9 (June): 645405.

Yang, Xi, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G. Flores, et al. 2022. "GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2203.03540.

Yang, Xi, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, et al. 2022. "A Large Language Model for Electronic Health Records." *Npj Digital Medicine* 5 (1): 1–9.

Yilmaz, Alper. 2020. "Assessment of Mutation Susceptibility in DNA Sequences with Word Vectors." *Journal of Intelligent Systems: Theory and Applications* 3 (1): 1–6.

Yin, Zi, and Yuanyuan Shen. 2018. "On the Dimensionality of Word Embedding." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1812.04224.

Yu, Zhiguo, Trevor Cohen, Byron Wallace, Elmer Bernstam, and Todd Johnson. 2016. "Retrofitting Word Vectors of MESH Terms to Improve Semantic Similarity Measures." In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, 43–51. Auxtin, TX: Association for Computational Linguistics.

Zaib, Munazza, Wei Emma Zhang, Quan Z. Sheng, Adnan Mahmood, and Yang Zhang. 2022. "Conversational Question Answering: A Survey." *Knowledge and Information Systems* 64

(12): 3151–95.
Zarocostas, John. 2020. “How to Fight an Infodemic.” *The Lancet* 395 (10225): 676.
Zhang, G. Peter. 2003. “Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model.” *Neurocomputing* 50 (January): 159–75.
Zhang, Haoran, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. “Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings.” In *Proceedings of the ACM Conference on Health, Inference, and Learning*, 110–20. CHIL ’20. New York, NY, USA: Association for Computing Machinery.
Zhang, Lu, Xinyi Qin, Min Liu, Guangzhong Liu, and Yuxiao Ren. 2021. “BERT-m7G: A Transformer Architecture Based on BERT and Stacking Ensemble to Identify RNA N7-Methylguanosine Sites from Sequence Information.” *Computational and Mathematical Methods in Medicine* 2021 (August): 7764764.
Zhang, Zhenwei, and Ervin Sejdić. 2019. “Radiological Images and Machine Learning: Trends, Perspectives, and Prospects.” *Computers in Biology and Medicine* 108 (May): 354–70.
Zhao, Mengnan, Aaron J. Masino, and Christopher C. Yang. 2018. “A Framework for Developing and Evaluating Word Embeddings of Drug-Named Entity.” In *Proceedings of the BIoNLP 2018 Workshop*, 156–60. Melbourne, Australia: Association for Computational Linguistics.
Zhou, Binggui, Guanghua Yang, Zheng Shi, and Shaodan Ma. 2024. “Natural Language Processing for Smart Healthcare.” *IEEE Reviews in Biomedical Engineering* 17 (January): 4–18.
Zhou, Liyuan, Hanna Suominen, and Tom Gedeon. 2019. “Adapting State-of-the-Art Deep Language Models to Clinical Information Extraction Systems: Potentials, Challenges, and Solutions.” *JMIR Medical Informatics* 7 (2): e11499.
Zhou, Yichu, and Vivek Srikumar. 2022. “A Closer Look at How Fine-Tuning Changes BERT.” In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1046–61. Dublin, Ireland: Association for Computational Linguistics.
Zhu, Henghui, Ioannis Ch Paschalidis, and Amir Tahmasebi. 2018. “Clinical Concept Extraction with Contextual Word Embedding.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1810.10566.
Zhu, Yongjun, Erjia Yan, and Fei Wang. 2017. “Semantic Relatedness and Similarity of Biomedical Terms: Examining the Effects of Recency, Size, and Section of Biomedical Publications on the Performance of word2vec.” *BMC Medical Informatics and Decision Making* 17 (1): 95.
Ziems, Noah, Wenhao Yu, Zhihan Zhang, and Meng Jiang. 2023. “Large Language Models Are Built-in Autoregressive Search Engines.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2305.09612.