RISK PREDICTION OF PATIENTS WITH SEVERE MENTAL DISORDER USING MACHINE LEARNING

A Thesis

submitted in partial fulfilment of the requirement for

the Degree of

Master of Technology in Intelligent Automation and Robotics

offered by

Department of Electronics and Telecommunication Engineering

Jadavpur University

May 2024

By

Sayantani Das

Registration No- **160223 of 2021-2022**

Examination Roll No- **M6IAR24006**

Under the Guidance of

Dr. Pratyusha Rakshit

Assistant Professor

Department of Electronics and Telecommunication Engineering

Jadavpur University, Kolkata- 700032

India

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

KOLKATA- 700032

CERTIFICATE OF RECOMMENDATION

This is to certify that Ms. Sayantani Das (002110704006) has completed her dissertation entitled "**Risk Prediction of Patients with Severe Mental Disorder using Machine Learning**", under the direct supervision of Dr. Pratyusha Rakshit, Department of Electronics and Telecommunication Engineering, Jadavpur University. We are satisfied with her work, which is being presented for the partial fulfilment of the degree of Master of Technology in Intelligent Automation and Robotics of Jadavpur University, Kolkata-700032

_____

Dr. Pratyusha Rakshit

Assistant Professor,

Department of Electronics and Telecommunication Engineering,

Jadavpur University,

Kolkata- 700032

_____

Dr. Manotosh Biswas

Head of the Department,

Department of Electronics and Telecommunication Engineering,

Jadavpur University,

Kolkata-700032

_____

Dr. Dipak Laha

Dean,

Faculty of Engineering and Technology,

Jadavpur University

Kolkata-700032

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

KOLKATA- 700032

CERTIFICATE FOR APPROVAL

The foregoing thesis entitled " **Risk Prediction of Patients with Severe Mental Disorder using Machine Learning**" is hereby approved as a creditable study of an Engineering subject carried out and presented in a manner satisfactory to warrant its subject carried out and presented in a manner satisfactory to warrant its acceptance as a pre-requisite to the degree of Master of Technology in Intelligent Automation and Robotics, offered by Department of Electronics and Telecommunication Engineering, for which it has been submitted. It is understood that, by this approval the undersigned does not necessarily endorse or approve any statement made, opinion expressed, or conclusion therein but approve this thesis only for the purpose for which it is submitted.

COMMITTEE ON FINAL EXAMINATION FOR EVALUATION OF THESIS

_____

_____

_____

Signature of Examiners

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

KOLKATA- 700032

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC THESIS

I hereby declare that this thesis entitled " **Risk Prediction of Patients with Severe Mental Disorder using Machine Learning** " contains literature survey and original research work by the undersigned candidate, as part of her Degree of Master of Technology in Intelligent Automation and Robotics, offered by Department of Electronics and Telecommunication Engineering.

All information have been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name : **SAYANTANI DAS**

Examination Roll No: **M6IAR24006**

Thesis title: **Risk Prediction of Patients with Severe Mental Disorder using Machine Learning**

_____

Signature of the candidate

ACKNOWLEDGEMENT

This thesis, entitled "**Risk Prediction of Patients with Severe Mental Disorder using Machine Learning**", is the culmination of work supported by many people. I am grateful for their assistance and encouragement.

First and foremost, I would like to express my sincere gratitude to my thesis supervisor, **Dr. Pratyusha Rakshit**, Assistant Professor in the Department of Electronics and Telecommunication Engineering at Jadavpur University. Her valuable guidance, insightful suggestions, and unwavering support have been instrumental in the successful completion of this thesis. Her positivity, confidence, and innovative ideas have been a constant source of inspiration.

I am also deeply grateful to our Head of Department, **Dr. Manotosh Biswas**, for his continuous help and support, and for allowing me to utilize all available resources for my thesis work.

I would like to acknowledge the assistance of **Mr. Shambhab Chaki**, a research scholar in our department, for his help with the idea, implementation, and analysis phases of this work. I am also thankful to my fellow project mates, friends, and the technical and non-technical staff of Jadavpur University, who have directly or indirectly contributed to the success of this thesis.

I want to express my heartfelt gratitude to my **parents** . Their invaluable love, encouragement, and support have been the foundation of all my achievements. Without their sacrifices, none of this would have been possible.

Sayantani Das

Jadavpur University

Kolkata- 700032

ABSTRACT

An Electronic Health Record is a digital database comprising of all the information related to a patient's medical history and treatment journey. It is a digital record of all the clinical activities, medications and lab tests that the patient has undergone. All the details regarding treatment costs are also available in an Electronic Health Record. They have immense advantage in the field of healthcare offering detailed insights into a patient's medical background enabling the healthcare advisors and physicians to deliver quality care to the patient.

However, healthcare costs are rising worldwide, including treatment costs , hospital and ICU stays, and medication expenses. In order to address this challenge it is essential to develop predictive models that can forecast the probable treatment path and also the future potential risk associated with a patient's disorder. This can also help to predict the future treatment expenses

In this thesis, we propose a machine learning approach to predict the risk of patients suffering from severe Mental Disorder with limited medical history available in Electronic Health Record. Here, we aim to predict the future risk and possible clinical activities of newly admitted Mental Disorder Patients, using the records of existing Mental Disorder patients available in Electronic Health Record.

We have made use of Hierarchical DBSCAN (Hierarchical Density Based Spatial Clustering of Applications with Noise) and Edit Distance to group the existing Mental Disorder Patients according to the similarity of their clinical activities. Markov Model have been used to test the interpretability of our approach and using Maximum Likelihood, we have assigned the new patient to the group with highest likelihood. Then, we have predicted future medications, diagnostics and lab tests for a certain number of days in future using Markov Model.

We have compared our risk-prediction model with other state-of-the-art models. This comparative analysis suggests that our proposed algorithm has outperformed other architectures.

CONTENTS

CHAPTER 1: INTRODUCTION 11

1.1 ELECTRONIC HEALTH RECORD (EHR) 11

1.1.1 HISTORY OF ELECTRONIC HEALTH RECORD 12

1.1.2 FUNCTIONALITIES OF ELECTRONIC HEALTH RECORD 13

1.1.3 BENEFITS OF ELECTRONIC HEALTH RECORD 13

1.1.4 LIMITATIONS OF ELECTRONIC HEALTH RECORD 14

1.2 RESEARCH SCOPE IN ELECTRONIC HEALTH RECORD 15

1.2.1 RESEARCH APPLICATIONS IN ELECTRONIC HEALTH RECORD 15

1.3 SCOPE OF AI IN ELECTRONIC HEALTH RECORD 16

1.3.1 PREDICTIVE ANALYSIS USING MACHINE LEARNING 16

1.3.2 TEXT EXTRACTION USING NATURAL LANGUAGE PROCESING 17

1.4 OVERVIEW OF PAST WORK USING ELECTRONIC HEALTH RECORD 19

1.5 RISK PREDICTION USING ELECTRONIC HEALTH RECORD 21

1.6 OBJECTIVE OF THE THESIS 21

1.7 ORGANIZATION OF THE THESIS 22

REFERENCES AND BIBLIOGRAPHY 23

CHAPTER 2 RISK PREDICTION OF PATIENTS WITH SEVERE MENTAL DISORDER USING MACHINE LEARNING 24

2.1 INTRODUCTION 24

2.1.1 PREVIEW 25

2.1.1.1 MARKOV CHAIN MODEL 25

2.1.1.2 HIERARCHICAL DBSCAN CLUSTERING ALGORITHM 26

2.1.1.3 EDIT DISTANCE 27

2.1.1.4 LIKELIHOOD PROBABILITY 28

2.1.1.5 MENTAL DISORDER 29

2.2 LITERATURE REVIEW 31

2.3 STRUCTURING OF DATA 34

2.3.1 IDENTIFYING UNIQUE COMBINATION OF SEQUENCE OF CLINICAL ACTIVITIES 36

2.4 GROUP FORMATION 38

2.5 REPRESENTING THE GROUPS USING MARKOV CHAIN RULE 39

2.6 FORECASTING PATIENT RISK WITH LIMITED CLINICAL ACTIVITY SEQUENCE 42

2.7 RESEARCH OUTCOMES 45

2.7.1 DATASET 45

2.7.2 MODEL EVALUATION 46

2.7.3 COMPARISON WITH STATE-OF-THE-ART MODELS 47

2.8 CONCLUSION 48

REFERENCE AND BIBLIOGRAPHY 52

CHAPTER 3 CONCLUSION AND FUTURE SCOPE 53

3.1 INTRODUCTION 53

3.2 SUMMARY 53

3.3 RESULTS AND DISCUSSION 54

3.4 FUTURE SCOPE 54

3.5 POTENTIAL IMPACT AND FUTURE DIRECTIONS 55

LIST OF FIGURES

Fig 1 BLOCK DIAGRAM OF AN EHR SYSTEM 10

Fig 2 SAMPLE ASSESSMENT OF $\lambda_{k,1}$ 44

Fig 3 DETERMIMING THE LIKELY SEQUENCE OF DIAGNOSES SETS FOR THE NEXT F=4 DAYS 45

Fig 4 REPRESENTING THE φ VALUES OBTAINED FROM OUR PROPOSED MODEL AND OTHER STATE-OF-THE-ART ARCHITECTURES 46

LIST OF TABLES

TABLE 1 REPRESENTATION OF THE ORGANIZATION OF RAW EHR DATA FOR N=2 PATIENTS 35

TABLE 2 ILLUSTRATIVE REPRESENTATION OF α_k 40

TABLE 3 ILLUSTRATION OF INITIAL PROBABILITIES OF MEMBERS OF D 42

TABLE 4 SHOWING OPTIMISED HYPER PARAMETERS FOR THREE STATE-OF-THE-ART ARCHITECTURES 47

Chapter 1

Introduction

1.1 Electronic Health Record

Electronic Health Records (EHRs) serve as a repository of patient's medical background. They contain a comprehensive record of a patient's medical history, including diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, and laboratory test results. EHRs are designed to facilitate smooth communication across various healthcare settings. They provide a wider perspective of the patients overall care. It serves as a centralized and accessible source of health related data that is securely stored and can be shared among healthcare providers and organizations involved in the care of a patient. EHRs enhances patient care by improving clinical decision-making, care coordination, efficiency, patient engagement while adhering to strict security and privacy standards.

An Electronic Health Record System can be defined as a structured database of health information of an individual. This information is designed to be shared over internet across various healthcare settings including hospitals and clinics. Three primary entities are involved within an EHR system:

i) Data owner: Data owner is typically the patient. They allow controlled external access to their health record.

ii) User: User can be an individual (a doctor, a nurse , healthcare administrator) or a healthcare organization who is requesting the data from EHR system

iii) Server: Server is responsible for receiving data from the data owner and make it available to the user who is requesting to access the EHR system. [1]

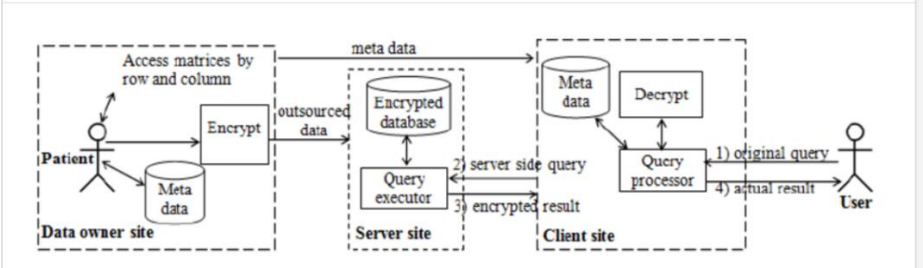


Fig 1: Block Diagram of an EHR System (from Pham *et al.* [1])

1.1.1 History of Electronic Health Record

Here, we are going to look back into the history of Electronic Health Record (EHR). We are going to focus on how EHRs have evolved over time. The origin of EHRs dates back in 1960's when the term "Electronic Health Record" was first introduced and electronic data processing began to gain popularity. The first major health systems, was The Mayo Clinic in Rochester, Minnesota. It marked a transition from paper medical records to adopting an Electronic Health Record (EHR) system. Regenstrief Medical Record System (RMRS) was one of the earliest implementation of EHR. It was used by the researchers of Regenstrief Institute in Indiana during 1970's. RMRS marked a significant step towards development of "Electronic Medical Record". They were the first systems to use computers for storing and retrieving data.

By 1990's, computers were used in medical offices for record keeping. EHRs were used to store data patient record, process claims and scan documents. The advent of internet marked a shift enabled recording and transfer of electronic medical records.

Digitizing health records is important as they allow the physicians and clinicians to easily access the patient's information, enhances coordination and collaboration among different provides. Digital health record systems also ensures better coordinated and effective healthcare treatment to patients. As technology advanced, digitizing health records became more accessible. Government initiatives were undertaken to encourage EHRs. The Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 in the United States, provided financial incentives to promote the adoption of EHRs, especially helping smaller providers integrate these systems.

EHRs have offered healthcare providers real-time access to patient records. They facilitate coordinated care across different providers and minimizes the risk of medical errors. They also help reduce costs associated with managing paper records, including printing, storage, and retrieval expenses. Presently, EHRs have become an integral part of healthcare systems. They have several advantages over the traditional paper based records. EHR software has become essential over the past decade since they facilitate an efficient communication between major medical systems. [2-3]

Currently further research is going on in order to increase the efficiency and functionality of EHRs by incorporating Artificial Intelligence and Machine Learning.

1.1.2 Functionalities of Electronic Health Record (EHRs)

1. Clinical Documentation: EHRs allow healthcare providers to create, update and access comprehensive clinical documentation, including progress notes, consultation reports, discharge summaries, and procedure notes. This functionality ensures that all the relevant patient information is accurately recorded and easily accessible to unauthorised users.

2. Order Entry and Management: EHR systems enable electronic order entry and management for medication, laboratory tests, imaging studies, referrals and other medical services. Healthcare providers can electronically submit orders, reducing dependency on paper based processes and minimizing errors associated with manual transcription.

3. Results Management: EHRs receive, store and display results from laboratory tests, imaging studies and other diagnostic procedures. This functionality allows healthcare providers to review and interpret results within the patient's electronic record, facilitating timely decision making and treatment planning.

4. Decision Support; EHR systems incorporate clinical decision support tools that provide alerts, reminders, and evidence based recommendations to assist healthcare providers in making informed clinical decisions. These tools help improve patient safety, adherence to best practices, adherence to clinical guidelines.

1.1.3 Benefits of Electronic Health Records (EHRs)

1. Enhanced Accessibility and Availability: EHRs provide healthcare providers with instant access to patient records from any location with internet connectivity. This enables timely decision making, especially in emergency situations, and facilitates coordinated care across multiple healthcare settings.

2. Improved Efficiency and Productivity: EHR systems streamline administrative tasks, automate workflows, and reduce paperwork, allowing healthcare providers to focus more on patient care. Tasks such as order entry, results management, and documentation can be completed more efficiently within the EHR system, leading to increased productivity.

3. Enhanced Patient Safety and Quality of Care: Decision support tools embedded within EHR systems help identify potential medication errors, adverse drug reactions, and other safety concerns. Alerts and remainders for drug interactions, allergies, and dosage adjustments help prevent errors and improve medication management, ultimately enhancing patient safety and quality of care delivered.

4. Facilitate Communication and Collaboration: EHRs enable seamless communication and information sharing among healthcare providers across different specialities, departments, and

care settings. This promotes collaboration and care coordination, reducing the risk of fragmented or duplicated care and improving continuity of care for patients.

5. Support for Evidence Based Practice and Research: EHRs provide a wealth of data that can be used for clinical research, quality improvement initiatives, and population health management. Data analytics tools within EHRs enable healthcare organizations to analyse trends, identify areas of improvement, and implement evidence-based practices to enhance patient outcomes.

1.1.4 Limitations of Electronic Health Record (EHRs)

Electronic Health Records have numerous advantages for patient health care and and medical research, yet they also present several challenges:

1. Privacy and Security Data : EHRs contain highly sensitive patient information and they are prime target for cyberattacks. Robust encryption techniques, secure access control are required to protect the privacy and security of these records.

2. Interoperability Challenges: Different healthcare providers often use EHR systems that are not compatible with each other. This lack of standardization can hinder the smooth exchange of information between institutions, resulting in fragmented patient care.

3. Accuracy of Data Entry: The quality of EHR data is dependent on the accuracy and completeness of the information input by healthcare providers. Errors, omissions, or inconsistencies in data entry can result in mistakes in patient care and research.

4. Impact on Workflow and Usability: Many EHR systems are not user-friendly and can be difficult to navigate, disrupting clinical workflows. Healthcare providers may spend a significant amount of time entering data, which can reduce the time available for patient interaction and care.

5. Cost and Implementation: Implementing and maintaining EHR systems can be costly. Smaller healthcare practices might find these expenses challenging, resulting in unequal adoption and use of EHR technology.

1.2 Research Scope in Electronic Health Record (EHRs)

There are immense research scopes in the field of Electronic Health Records. A wide range of topics can be explored which includes improving the quality, safety, and efficiency of healthcare through the effective use of digital health information systems. The research scope in EHRs encompasses various dimensions of technology, policy and practice like usability and user experience studies to clinical decision support interventions and data analytics. Investigations into patient engagement, health disparities, global perspectives, and emerging technologies offer insights in the field of healthcare information technology. Moreover, ethical considerations surrounding privacy, security, and data governance are integral to the research agenda, which ensures that the innovations in EHRs uphold principles of patient confidentiality. These researches can contribute to advancing knowledge, informing policy, and shaping the future of healthcare delivery, striving towards a more efficient, equitable and patient-centered healthcare system.

1.2.1 Research Applications in Electronic Health Records (EHRs)

Research applications using Electronic Health Records (EHRs) have significantly advanced medical studies by providing extensive and detailed patient data. EHRs enable researchers to access a wealth of information on patient histories, treatments, and outcomes, which can be utilized to identify trends, improve clinical practices, and develop new therapies. This integration of comprehensive health data into research facilitates more robust and accurate studies, ultimately enhancing the quality of healthcare. Now, we will be discussing on some research fields using EHRs.

1. Clinical Research: EHRs provide a rich source of data for clinical research, enabling researchers to conduct observational studies and outcome research. Researches can analyze EHR data to study disease patterns, treatment effectiveness and patient outcomes contributing to evidence based medicine and clinical decision making.

2. Population Health Management: EHRs support population health management initiatives by providing data analytics and reporting capabilities. EHR data can be used to identify at-risk populations, track health trends, and monitor outcomes at the population or community level.

3. Predictive Analytics: EHR data can be used for predictive analytics to forecast disease progression, identify high-risk patients, and anticipate healthcare needs. Researchers can develop predictive models and risk stratification algorithms using EHR data to support early intervention and targeted preventive care.

4. Health Services Research: EHR data can be used for health services research to study healthcare delivery processes, resource utilization patterns, and quality of care.

5. Translational Research: EHR facilitate translational research by bridging the gap between basic science and clinical practice. EHR can be used to study implementation and effectiveness of new therapies, interventions and technologies in real world clinical settings.

1.3 Scope of AI in Electronic Health Record (EHRs)

In the recent years, the healthcare industry has witnessed a transformative shift towards digitization, with Electronic Health Records (EHRs) becoming integral to modern healthcare delivery. As healthcare systems generate vast volumes of data, the need to harness the power of technology to effectively manage, analyse, and derive actionable insights from this data has become increasingly paramount. AI technologies, including machine learning, natural language processing, and image analysis offer opportunities to enhance functionalities and utility of EHR systems. From optimizing data management and clinical decision-making to enable personalized medicine and population health management, the scope of AI in EHR is vast. The unique features of machine learning and natural language processing (NLP) can help record the unique medical experiences of patients, search the large EHR data banks for important documents, gauge patient satisfaction, etc. The machine learning models combined with NLP can help healthcare providers in transcribing the speech from the voice recognition system into text. The algorithms can be trained to deal with huge amounts of patient data related to the treatment, equipment used for treatment, respective doctor consulted etc. Further, this data can be segmented based upon the individual patient, illness, treatment for illness, etc. This will enhance the document and information search from the large databases.

1.3.1 Predictive Analysis using Machine Learning

Predictive analysis in Electronic Health Records using Machine Learning techniques is transforming healthcare delivery. Machine learning algorithms can uncover hidden patterns and relationships, by leveraging vast amounts of patient data like demographics, medical history, diagnostic tests, and treatment outcomes. These insights helps to mitigate and anticipate health risks, identify early signs of diseases, design a treatment plan to individual patient needs, and optimize resource allocation. Machine learning driven predictive analytics helps clinicians with valuable decision support tools, leading to improved patient outcomes, reduced healthcare costs, and more efficient healthcare delivery systems. However, successful implementation requires addressing challenges related to data quality, privacy concerns, model

interpretability, and regulatory compliance to ensure the responsible and ethical utilization of predictive analytics in healthcare settings.

To implement predictive analysis using Machine Learning, following steps are implemented:

1. Data Collection: Gathering EHR data from electronic health record systems, including patient demographics, medical history, diagnoses, medications, lab results, and treatment plans.

2. Data Preprocessing: Cleaning and preprocessing the EHR data for analysis, which may involve handling of missing value, normalizing data, and encoding categorical variables.

3. Feature Engineering: Selecting and extracting relevant features from EHR data that are predictive of the outcome interest, such as disease onset or treatment response

4. Model Training: Training machine learning using historical EHR data and appropriate algorithms for the prediction task at hand, such as logistic regression, decision trees, or deep learning models.

5. Model Evaluation: Assessing the performance of the trained models using validation datasets and appropriate metrics, such as accuracy, precision, recall or area under the ROC curve

6. Deployment: Integrating the trained models into clinical workflows and EHR systems to provide real-time predictions and decision support for healthcare providers

Predictive analysis using machine learning in EHR has the potential to improve patient outcomes, optimize healthcare delivery, and reduce costs by enabling more personalized and proactive care. However, its essential to address the challenges related to data quality, privacy, and interpretability to ensure the responsible and ethical use of these technologies in healthcare.

1.3.2 Text Extraction using Natural Language Processing

In Electronic Health Records (EHRs), extracting text using Natural Language Processing (NLP) involves deciphering and interpreting unstructured clinical data. Initial step is to gain access to the EHR system, and identifying relevant text fields like clinical notes and diagnostic reports. Once identified, preprocessing steps include cleaning the text to remove noise and irrelevant information, ensuring a refined dataset for analysis. Tokenization breaks down the text into manageable units, while Named Entity Recognition (NER) techniques pinpoint and classify entities such as medical conditions, medications, and procedures. Following this, normalization links entities to standardized medical terminologies, enabling interoperability. Advanced NLP processes, including relationship extraction and sentiment analysis, extract insights from the text, aiding in understanding patient care pathways and sentiment analysis.

Quality assurance measures ensure the accuracy and reliability of the extracted information, refining NLP models iteratively. By leveraging NLP, healthcare organizations can unlock valuable insights from EHR text, enhancing clinical decision-making and patient care processes. One application of Text Extraction in EHR involves case detection which comprises identifying specific cases or instances of medical conditions, diseases or information stored in the EHR system. Case detection in EHRs is often facilitated through data analysis techniques, including machine learning, natural language processing (NLP), and statistical methods, applied to the structured and unstructured data stored in EHRs. The goal of case detection is to assist healthcare providers in identifying patients who may require targeted interventions, monitoring, or follow-up care, ultimately improving patient outcomes and healthcare delivery. Examples of case detection tasks include identifying patients with chronic diseases such as diabetes or hypertension, detecting adverse drug reactions, or identifying patients at risk of hospital readmission.

Steps to implement text extraction in EHR using Natural Language Processing:

1. Data Access and Preprocessing: It involves gaining access to the EHR system and identifying the relevant text fields containing clinical notes, discharge summaries, radiology reports, and other free-text data. Then the text is pre by removing noise, such as special characters, punctuation, and irrelevant information like headers and footers.

2. Tokenization and Sentence Segmentation: The pre processed text is tokenized into individual words or phrases, and segmented into sentences. This step breaks down the text into smaller units, facilitating further analysis.

3. Named Entity Recognition: Named Entity Recognition (NER) techniques are applied to identify and classify named entities in the text, such as patient names, healthcare providers, medical conditions, medications, procedures, and anatomical entities. NER models can be trained on annotated healthcare data to recognize domain-specific entities accurately.

4. Entity Linking and Normalisation: Identified entities are linked to standardized medical vocabularies or ontologies (e.g., SNOMED-CT, ICD-10, RxNorm) to normalize the extracted information and ensure interoperability across different healthcare systems. This step improves data consistency and facilitates data integration and analysis.

5. Relationship Extraction: The relationships between entities mentioned in the text are extracted, such as associations between medical conditions and treatments, patient-provider relationships, and temporal relationships (e.g., onset dates of symptoms, dates of procedures). This step helps in understanding the context of clinical events and patient care pathways.

6. Sentiment Analysis: The sentiment expressed in clinical text is analyzed to identify subjective information such as patient complaints, treatment satisfaction, or emotional states. Sentiment analysis techniques can help in monitoring patient experiences and sentiments over time and identifying areas for improvement in healthcare delivery.

7. Topic Modelling and Summarization: Topic modeling techniques are applied to identify latent topics or themes present in the clinical text data. Summarize large volumes of clinical

documents to extract key insights, trends, and patterns, facilitating data-driven decision-making and knowledge discovery.

8. Quality Assurance and Validation: The accuracy and reliability of the extracted information are validated through manual review by domain experts or comparison with gold standard datasets. Iteratively refine and improve the NLP models based on feedback to ensure high-quality text extraction results.

1.4 Overview of Past Work on Electronic Health Record

In this section we will be exploring the past research work done on EHRs .We will get a comprehensive understanding of the past studies, research works and the discoveries made in this field.

Electronic Health Records (EHRs) have transformed healthcare management by digitizing patient information. They contain everything from diagnoses to medication histories and test results. EHRs make it easy for healthcare providers to access this information quickly, improving coordination among different providers. They also come with helpful tools that give real-time guidance to healthcare professionals, making diagnoses and treatments more accurate. Plus, EHRs help identify patients who need preventive care, promoting wellness. Patients can also get involved in their care through EHRs by accessing their records online. Overall, EHRs are changing healthcare for the better by making care more coordinated, accurate, and patient-centered. Many research works have been implemented using EHRs. [4] Graña and Jackwoski (2015) examines the role of EHR in various aspects like data mining, security and privacy, developing decision support systems and so on. In data mining, EHRs are used in analysing vast amounts of patient data. It identifies trends and patterns which are useful for decision making and population health management. EHRs use encryption and access controls to protect sensitive patient data. Effective use of EHR can help in achieving a patient-focused healthcare like reduction in healthcare costs, improving the quality care. [5] Noteboom *et al* (2014) conducted research on how to effectively use EHRs in order to develop a patient-focused healthcare system. These include streamlining data entry processes to ensure accurate and comprehensive patient records, enhancing interoperability between EHR systems for seamless data exchange, empowering patients through access to their health records and online communication platforms, integrating clinical decision support tools to aid evidence-based decision-making, leveraging EHR data analytics for population health management and quality improvement initiatives, and providing continuous training and support to healthcare providers. Additionally, personalized care plans, remote monitoring, care coordination, patient safety initiatives, and patient education are essential components in optimizing EHR usage to prioritize patient engagement, clinical effectiveness, and overall healthcare quality.

Electronic Health Records began to gain popularity in 1990's. There was a shift from using traditional paper based records to the use of EHRs. Further with the advancements in technology, EHRs became the core of the healthcare system.[6] Evans (2016) showed how EHRs have evolved from 1992 till 2015. The paper also provides insights about the future aspects of EHR.

Recently, Artificial Intelligence and Machine Learning is being used in EHRs for enhancing its effectiveness and functionality. They help to uncover hidden patterns in the data, automate routine administrative tasks, forecast potential risks, enabling decision making. Several researches have been conducted on the integration of AI and ML into EHR systems [7]. EHR data can help in early forecasting of diseases. [8] Guru *et al* (2023) worked used machine learning algorithms for early prediction of diabetes. [9] Caroprese *et al* (2018) implemented deep learning tools for enhancing the efficiency of EHRs.

Deep Learning is a subset of machine learning which can automatically extract relevant feature from the raw data, they can model complex non-linear relationships in the data, handle large unstructured data. Deep Learning techniques can revolutionize EHRs further enhancing its capability and functionality.

Thus, we can get a summary of the research studies conducted using EHRs.

1.5 Risk Prediction using Electronic Health Record

Risk prediction using Electronic Health Records (EHR) uses patient's data stored in EHR systems to assess their likelihood of developing certain medical conditions or adverse health events in the future. This process typically utilizes machine learning and statistical modelling techniques to analyze a variety of patient factors, including demographic information, medical history, laboratory results, medication usage, and lifestyle factors. By identifying patterns and correlations within this data, predictive models can estimate an individual's risk of developing conditions such as diabetes, cardiovascular disease, or cancer, as well as predict the likelihood of experiencing events like hospital readmissions, complications from treatment, or adverse drug reactions. Risk prediction models can provide valuable decision support for healthcare providers by helping to identify high-risk patients who may benefit from targeted interventions, preventative measures, or closer monitoring, ultimately leading to more proactive and personalized patient care.

Risk prediction utilizes prior medical knowledge within Electronic Health Record (EHR), uses historical patient data and medical principles to predict future health risks. By integrating past medical history, laboratory results, medication records, and demographic information alongside medical guidelines and expert insights, predictive models can be developed which can identify patterns and indications of future potential risks. By incorporating medical knowledge directly into EHR systems, healthcare providers can gain a comprehensive understanding of each patient's unique risk profile, enabling them to implement proactive interventions, personalized treatment plans, and preventive measures. This approach not only enhances clinical decision-making but also facilitates early detection and intervention, ultimately improving patient outcomes and reducing healthcare costs.

1.6 Objective of the Thesis

This thesis is used to predict the risk of newly admitted patients suffering from severe mental disorder and having limited amount of medical history in the Electronic Health Record (EHR). Here, Edit distance and DBSCAN (Density Based Spatial Clustering of Applications with Noise) clustering algorithm have been used to categorize the mental disorder patients (whose records are already available in the EHR) into groups on the basis of similarity of the sequences of their clinical activities like diagnoses, lab tests and medications over their stays at Intensive Care Units (ICU). Next, Markov Chain Model has been used to mathematically represent each group. Risk prediction of the new patient belonging to each group is calculated using composite

likelihood. The patient is then assigned to the group with maximum likelihood. Markov Chain Model is then used to calculate the finite sequence of future activities of the unknown patient.

The main objective of this thesis is to propose a novel machine learning approach for early risk prediction of a patient suffering from severe mental disorder. Early and accurate risk prediction would help the medical practitioners design an effective treatment path and start with the required diagnoses at the earliest. It would also help to address the challenges associated with treatment costs, cost over hospital stays and the healthcare expenditures. Early risk prediction would help in estimating the healthcare and medication related expenditures, as it would help us the calculate the future set clinical activities and the treatment path of the patients. We have used MIMIC-III EHR database to study the effectiveness of the proposed method and the compared the proposed algorithm with other state-of-the art techniques.

1.7 Organization of the Thesis

Chapter 2 provides a detailed explanation of the thesis, presenting a new machine learning method for predicting the risk of severe mental disorders in patients. Chapter 3 offers a summary and conclusion of the thesis, along with insights into potential future research directions. References and bibliographies relevant to each chapter are included at the end of those chapters.

References & Bibliography

[1] Pham, Hue, Sven Wohlgemuth, I. Echizen, Dong Thuy, and Thuc Nguyen. "Fine-Grained Access Control for Electronic Health Record Systems." *Communications in Computer and Information Science* 124 (2010): 31-38. doi:10.1007/978-3-642-17644-9_4

[2] *A Brief History of Electronic Health Record.* (2023, July). Retrieved from https://www.eka.care/services/history-of-electronic-health-records

[3] *Electronic Health Records: A Comprehensive History of EHR Systems.* (2021, September 16). Retrieved from https://www.nethealth.com/blog/the-history-of-electronic-health-records-ehrs/

[4] Graña, M., & Jackowski, K. (2015). Electronic Health Record: A Review. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1375-1382). IEEE. https://doi.org/10.1109/BIBM.2015.7359879

[5] Noteboom, C., Motorny, S., Qureshi, S., & Sarnikar, S. (2014). Meaningful use of electronic health records for physician collaboration: A patient-centered health care perspective. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 656-666. https://doi.org/10.1109/HICSS.2014.87

[6] Evans, R. S. (2016). Electronic health records: Then, now, and in the future. *Yearbook of Medical Informatics, 1*(Suppl 1), S48-S61. https://doi.org/10.15265/IYS-2016-s006

[7] N'Gbesso, Y. (2020). Integration of Artificial Intelligence in Electronic Health Records: Impacts and Challenges.

[8] Guru, J., Shashidhar, R., H. L., Gururaj, R., Ravi, V., Almeshari, M., & Alzamil, Y. (2023). Electronic health record (EHR) system development for study on EHR databased early prediction of diabetes using machine learning algorithms. *The Open Bioinformatics Journal*.

[9] Caroprese, L., Veltri, P., Vocaturo, E., & Zumpano, E. (2018). Deep learning techniques for electronic health record analysis. *Proceedings of the 2018 IEEE International Conference on Information, Intelligence, Systems & Applications (IISA)* (pp. 1-4). https://doi.org/10.1109/IISA.2018.8633647

Chapter 2

Risk Prediction of Patients with Severe Mental Disorder using Machine Learning

2.1 Introduction

An Electronic Health Record (EHR) functions as a digital archive containing comprehensive medical information about a patient, including treatment histories, prescribed medications, diagnostic results, and incurred treatment costs. This extensive data collection has driven advancements in healthcare methodologies aimed at enhancing patient care and optimizing resource use. We have made a detailed discussion on EHR in chapter 1.[1]

EHR data often focus on immediate healthcare needs rather than long-term risk prediction, making it challenging to develop comprehensive predictive models. Previous studies mainly focused on predicting mortality or specific diseases rather than treatment paths. Additionally, while neural network models offer accuracy, they lack interpretability crucial for healthcare professionals.

With healthcare costs soaring globally, there is a need to manage expenses efficiently while ensuring that patients receive proper care, especially in the regions with limited medical resources. Predictive modelling using EHR data can help anticipate future risks for patients, even those with incomplete medical histories. However, creating accurate predictive models poses several challenges [2].

This thesis introduces a new machine learning approach to predict the early risk of severe mental disorders (MD) using EHR data. It divides MD patients into groups based on the similarity of their clinical activities during Intensive Critical Care Unit (ICU) stays, using DBSCAN clustering algorithm. Edit distance is used to assess the similarity between the sequence of their clinical activities. Each group is then represented by a Markov chain model, ensuring interpretability. In the prediction phase, the model identifies potential risks for new patients with limited medical histories by assigning them to the most similar group with highest likelihood and predicting their future clinical activities using the Markov chain model.

The proposed algorithm is evaluated using the MIMIC-III EHR database and compared with existing techniques, demonstrating superior performance. The paper is organized into sections discussing data organization, group formation based on clinical activities, Markov chain representation, risk prediction methodology, experimental results, and conclusion.

The following sections gives an overview of the methods that has been utilized in our work. Markov Chain Model will be explained first, followed by the examination of the Hierarchical Clustering DBSCAN algorithm and Edit Distance and finally, an overview of Likelihood

Probabilities will be provided. The application of the above methodologies in healthcare domain will also be highlighted.

2.1.1 Preview

In this section we are going to give an overview of the methods that have been utilized in our work. Markov Chain Model will be explained first, followed by the examination of the Hierarchical Clustering DBSCAN algorithm and Edit Distance (ED) and finally, an overview of Likelihood Probabilities will be provided. The application of the above methodologies in healthcare domain will also be highlighted.

2.1.1.1 Markov Chain Model

A Markov chain, credited to Andrey Markov, is a probabilistic framework used to forecast the probability of a series of events, relying on the state of the previous event. It's extensively employed in diverse fields, especially where sequential data analysis is crucial. This model provides a simple yet robust method to comprehend and evaluate processes over time. By applying mathematical concepts, it enables us to make reasoned predictions about forthcoming events, drawing from past observations.

The essence of the Markov process lies in figuring out how likely it is to move from one state to another. What sets the Markov model apart is its reliance solely on the current state to anticipate what comes next for a random variable. Put simply, a random variable is something whose values are shaped by chance outcomes. This aspect of the Markov model highlights its effectiveness and flexibility in understanding the workings of different systems. It gives us useful clues about the chances of moving between states, aiding in making well-informed decisions.

The Markov process is a key idea in stochastic modelling, featuring a unique trait called "memorylessness." In simpler terms, memorylessness refers to specific probability distributions where the likelihood of an event happening doesn't depend on how much time has passed since the last event occurred. This essentially means that the model doesn't remember its past state; it's as if it starts fresh with each event, making it impervious to previous happenings. As a result, the probabilities of future events aren't influenced by what's already occurred, underscoring the self-contained behavior of the model's dynamics.

The unique feature of a Markov process lies in its memoryless property. In essence, the predictions made within a Markov process solely rely on its present state, detached from any influence of past or future states. This means that the likelihood of future events occurring is

determined solely by the current state of the process, devoid of any dependence on its historical trajectory or future outcomes.

Markov Chain Models have various applications in the field of healthcare. They offer insights into disease dynamics, treatment efficacy, resource allocation and epidemiological trends. They are extensively used in disease progression modelling, where they simulate how diseases evolve by defining health states (like healthy, diseased, or recovered) and the transitions between them. This helps assess disease burden and evaluate treatment effectiveness. They can forecast the long-term consequences of different treatment approaches for chronic illnesses by analysing how patients move between health states based on treatment response and disease advancement. This aids clinicians in tailoring treatment plans for better patient outcomes.[3]

2.1.1.2 Hierarchical DBSCAN Clustering Algorithm

Hierarchical DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a variant of the traditional DBSCAN clustering algorithm that incorporates hierarchical clustering principles. This adaptation enhances DBSCAN's ability to detect clusters at different scales and provides a hierarchical structure of clusters within the dataset. In Hierarchical DBSCAN, the algorithm first identifies dense regions of data points similar to standard DBSCAN by considering two key parameters: epsilon (ε) and minPts. These parameters dictate the minimum distance and minimum number of points required to form a dense region or cluster, respectively. Once initial clusters are formed, Hierarchical DBSCAN employs hierarchical clustering techniques to organize these clusters into a hierarchical structure. This hierarchical structure represents relationships between clusters at different levels of granularity, allowing for the identification of clusters within clusters and capturing the hierarchical nature of the data. The hierarchical structure can be represented using dendrograms or trees, where clusters are nested within one another based on their similarities and spatial proximity. This hierarchical representation enables users to explore the dataset at various levels of detail, from overarching macro-clusters to finer-grained micro-clusters. One advantage of Hierarchical DBSCAN is its ability to handle datasets with varying densities and complex cluster structures. By incorporating hierarchical clustering principles, the algorithm can capture clusters of different sizes and shapes, making it more robust and versatile for clustering tasks. However, it's important to note that Hierarchical DBSCAN may come with increased computational complexity compared to standard DBSCAN, particularly when dealing with large datasets or high-dimensional data. Additionally, parameter selection, such as choosing appropriate values for epsilon (ε) and minPts, remains crucial for achieving meaningful clustering results. Hierarchical DBSCAN combines the strengths of both DBSCAN and

hierarchical clustering, offering a powerful approach for discovering and analyzing clusters in datasets with hierarchical structures and varying densities.

Hierarchical DBSCAN clustering is increasingly utilized in healthcare for patient stratification, disease subtyping, resource allocation, fraud detection, and clinical trial design. It aids in identifying patient subgroups based on health data, optimizing resource distribution, uncovering fraudulent activities in healthcare claims, and selecting homogeneous patient cohorts for clinical trials. By leveraging its ability to identify clusters in data with varying densities and hierarchical structures, hierarchical DBSCAN clustering facilitates personalized healthcare delivery, advances precision medicine approaches, and enhances healthcare system efficiency [4].

2.1.1.3 Edit Distance

Edit distance, also known as Levenshtein distance, measures how different two strings are by counting the minimum number of operations needed to change one string into the other. These operations usually involve inserting, deleting, or substituting individual characters.

For example, if there are two strings: "kitten" and "sitting." The edit distance between them represents the minimum number of edits needed to transform "kitten" into "sitting." Here is how this transformation can be achieved:

1. ‘k’ is replaced with ‘s’:

“kitten” → “sitten”

Number of Edits =1

2. There will be no change for ‘i’

“sitten” → “sitten”

Number of Edits =0

3. There will be no change for ‘t’

“sitten” → “sitten”

Number of Edits =0

4. Again, there will be no change for ‘t’

“sitten” → “sitten”

Number of Edits =0

5. ‘i’ is inserted after ‘t’, inplace of ‘e’

“sitten” → “sittin”

Number of Edits =1

6 ‘g’ is inserted after ‘n’

“sittin” → “sitting”

Number of Edits=1

Edit distance to transform the string “kitten” to “sitting” is 3

Edit distance plays a crucial role in various aspects of healthcare data management and decision-making [5]. In medical coding and billing, it ensures accuracy by comparing medical codes or descriptions, aiding in precise billing and record-keeping. Moreover, in clinical text mining, edit distance assists in analyzing unstructured clinical text from electronic health records, facilitating information retrieval and knowledge discovery. Patient matching across different sources is also enhanced through edit distance, ensuring comprehensive patient records by comparing demographics. Additionally, it identifies similar drug names to prevent medication errors and adverse reactions, while standardizing disease names improves interoperability and disease surveillance. Furthermore, in clinical decision support, edit distance matches patient data to clinical guidelines. Edit distance plays a crucial role in various aspects of healthcare data management and decision-making. In medical coding and billing, it ensures accuracy by comparing medical codes or descriptions, aiding in precise billing and record-keeping. Moreover, in clinical text mining, edit distance assists in analyzing unstructured clinical text from electronic health records, facilitating information retrieval and knowledge discovery. Patient matching across different sources is also enhanced through edit distance, ensuring comprehensive patient records by comparing demographics. Additionally, it identifies similar drug names to prevent medication errors and adverse reactions, while standardizing disease names improves interoperability and disease surveillance. Furthermore, in clinical decision support, edit distance matches patient data to clinical guidelines, providing personalized decision support for healthcare providers . Overall, edit distance contributes to accuracy, interoperability, and patient safety in healthcare data management and decision-making processes.

2.1.1.4 Likelihood Probability

Likelihood probability, essentially, is the measure of how probable an event is to occur given certain conditions or evidence [7]. It is comparable to assessing the chances of a particular outcome happening based on the information available. In healthcare, this concept is invaluable. For example, in diagnostic testing, likelihood probability aids clinicians in deciphering test results by evaluating the probability of specific outcomes given a patient's health status. This facilitates precise diagnoses and treatment plans. Furthermore, in risk assessment, likelihood probability helps gauge the probability of individuals developing certain health conditions, allowing healthcare providers to tailor preventive measures accordingly. Additionally, it plays a pivotal role in evaluating treatment efficacy, offering insights into the likelihood of treatment success based on various factors. By harnessing likelihood probability, healthcare professionals can make informed decisions, enhance patient care, and ultimately improve health outcomes.

In addition to diagnostic testing, likelihood probability is instrumental in clinical decision-making processes. Healthcare providers often encounter complex scenarios where they must weigh the risks and benefits of different treatment options. Likelihood probability assists in this decision-making by providing probabilistic estimates of various outcomes. For example, it helps predict the likelihood of a patient responding favourably to a particular treatment, developing complications, or experiencing disease progression. This information empowers clinicians to make evidence-based decisions tailored to each patient's unique circumstances, optimizing treatment efficacy and minimizing adverse events.

Moreover, likelihood probability contributes to patient-centered care by facilitating risk stratification and personalized medicine approaches. By assessing the probability of individuals developing certain health conditions based on their risk factors and genetic predisposition, healthcare providers can implement targeted preventive interventions and lifestyle modifications. This proactive approach not only improves patient outcomes but also reduces healthcare costs associated with managing preventable diseases.

Furthermore, likelihood probability is essential in healthcare quality improvement initiatives. Healthcare organizations utilize probabilistic models to analyse clinical data and identify areas for improvement. By quantifying the likelihood of adverse events, medical errors, or treatment complications, healthcare leaders can implement targeted interventions to enhance patient safety and care quality. Likelihood probability also supports outcome-based evaluation metrics, enabling healthcare organizations to measure performance, track progress, and drive continuous quality improvement efforts [8].

Thus, Likelihood probability serves as a cornerstone of evidence-based practice in healthcare, guiding diagnostic and treatment decisions, promoting patient-centered care, and driving quality improvement initiatives. By harnessing the power of probabilistic reasoning, healthcare professionals can navigate complex clinical scenarios with confidence, ultimately improving patient outcomes and advancing the delivery of high-quality healthcare services.

2.1.1.5 Mental Disorder

Mental disorders, which encompass a broad spectrum of conditions affecting mood, cognition, and behaviour, pose significant challenges globally. These disorders are a cause for concern due to their profound impact on individuals, families, and society as a whole. Mental disorders are a growing concern due to their widespread occurrence and rising rates globally. For example, depression affects millions worldwide, leading to substantial societal and economic challenges. This increase can be attributed to changes in society, lifestyle stressors, and biological factors. Modern lifestyles characterized by urbanization, economic pressures, and social isolation contribute to heightened levels of stress and mental health issues. Moreover, stigma and lack of access to mental healthcare hinder many people from seeking assistance promptly. Additionally, shortages of mental health professionals and inadequate funding worsen the problem.

One of the primary reasons for the heightened concern surrounding mental disorders is their widespread prevalence and increasing rates worldwide. Mental disorders are a growing concern due to their widespread occurrence and rising rates globally. According to the World Health Organization (WHO), conditions like depression affect millions of people worldwide, leading to significant societal and economic challenges. This increase can be attributed to changes in society, lifestyle stressors, and biological factors. According to the statistics published by WHO, in 2019, approximately 970 million people, or 1 in every 8 people worldwide, were living with a mental disorder, with anxiety and depression being the most common. The COVID-19 pandemic in 2020 led to a significant rise in these conditions, with initial estimates indicating a 26% increase in anxiety disorders and a 28% increase in major depressive disorders within a single year. Although there are effective prevention and treatment options available, many individuals with mental disorders do not receive adequate care. Additionally, they often face stigma, discrimination, and human rights violations [9]. Mental disorders globally contribute to 1 in 6 years lived with disability. Individuals with severe mental health conditions have a life expectancy that is 10 to 20 years shorter than the general population. Additionally, having a mental health condition elevates the risk of suicide [10].

Hence, Mental health is crucial for our overall well-being, affecting how we think, feel, and connect with others. However, mental health problems are becoming more common, impacting people from all walks of life. To tackle these challenges effectively, we need to take a comprehensive approach that includes both prevention and support. Preventive measures are especially important in stopping mental health issues before they start and helping people cope better. By spreading awareness, building supportive communities, and addressing social issues, we can create environments where mental well-being thrives, empowering everyone to lead fulfilling lives [11]. Here, are some statistics:

i) Raising Awareness about Mental Health: Educating people about mental health issues to reduce stigma and encourage early intervention.

ii) Enhancing Resilience: Teaching individuals for coping skills and techniques that better manage stress and overcome challenges.

iii) Providing supportive Environment : Creating accessible support systems, including social networks and community resources, to help individuals facing mental health challenges.

iv) Addressing Social Factors: Tackling social determinants of health such as poverty, inequality, and discrimination to prevent the onset of mental disorders.

With advancing technology, there has been a growing interest in utilizing innovative approaches to address mental health challenges. Machine learning can play a crucial role in detecting mental disorders by analysing patterns and data associated with symptoms and behaviours. This technology can sift through vast amounts of information, such as electronic health records, social media activity, and smartphone usage, to identify potential signs of mental health issues. By recognizing subtle patterns that may not be apparent to human observers, machine learning algorithms can provide valuable insights and assist healthcare professionals in early detection and intervention. Moreover, these algorithms can continuously learn and improve over time, enhancing their accuracy and effectiveness in identifying individuals at risk of mental disorders. Overall, machine learning holds great promise for revolutionizing mental health care by facilitating earlier detection and personalized treatment approaches [12].

This is the reason we have decided to concentrate on mental disorders. In our endeavour, we have innovatively introduced a new machine learning method specifically tailored for the early identification of these conditions.

2.2 Literature Review

Here, we will review some previous studies on Early Risk Prediction. Risk Prediction using EHR involves analysing the patient data from the digital health records and anticipate the potential health problems. Predictive analytics plays a crucial role in identifying the patterns and trends for forecasting future risk. Numerous research efforts have been undertaken on this topic. For example, [13] Wu *et al.* (2017) created a prediction model to detect the risk of Breast Cancer using the data from EHR. They developed two risk prediction models: one using Logistic Regression and the other Lasso Logistic Regression. EHRs consist of sequentially recorded information about a patient's medical history like clinical notes, medications, diagnoses, lab test results and so on. Hence EHRs can be considered as time series data. They are well suited for time series analysis where we need to understand the progression and trends in a patient's health which is crucial for effective prediction and intervention. Recurrent Neural Network (RNNs) is a type of artificial neural network designed to handle sequential data. They are ideal for time series analysis and sequence prediction. [14] Xu *et.al* (2019) used multitask Recurrent Neural Network to detect the risk associated with major adverse cardiovascular events. RNN is mainly used to capture the sequential occurences in the patient's medical history. [15] Yu *et.al* (2022) used EHR data along with genetic data for identifying the risk of heart failure for cancer patients.

[16] Cheng *et al*. (2016) proposed a deep learning methodology by phenotyping a patient's EHR followed by developing a four layer convolutional neural network. The proposed model was tested on a real-world EHR data warehouse under the scenario of predictive modeling of chronic diseases. In another study, [17] Che *et al*. (2017) proposed a deep learning framework using Generative Adversarial Network (GAN) and Convolutional Neural Network (CNN) to improve the prediction performance with limited EHR data.[18] Ma *et al*. (2018) proposed a framework PRIME for risk prediction tasks, which can incorporate prior medical knowledge into all state-of-the-art predictive models using posterior regularization technique.[19] Lastly, Cui *et al*. (2022) developed an automated risk prediction model, which can improve the performance of risk prediction models by automatically searching for an optimal model architecture for modeling complex EHR data." [20] Additionally, Grout et al. (2024) examined the effectiveness of utilizing deep learning techniques to predict future disease diagnoses through Electronic Health Records (EHRs), aiming to improve population health management.

Thus we can see that machine learning plays an important role for developing risk prediction models.

Several work has also been implemented to predict the healthcare treatment costs. [21] Panay *et.al* (2019) used Evidence Regression Model for forecasting treatment expenses. [22] Fauziah *et al.* (2022) used Decision Tree algorithm to forecast the duration of hospital stay and the associated treatment expense for diabetes patient.

Thus, the studies we examined highlight the importance of using EHRs to predict health problems early. Predictive analytics using Machine Learning helps us uncover hidden patterns and trends within EHR data to foresee future health risks. Additionally, we have seen how predictive analytics is crucial for estimating treatment and medication costs. We have also seen how Recurrent Neural Networks are crucial for modelling sequential time series EHR data. The above literatures highlights the importance of Machine Learning and Deep Learning in developing predictive models. The reviewed literature also indicates that machine learning and deep learning techniques are crucial for estimating healthcare costs.

However, the main challenge is creating predictive models that use patient records from EHRs to accurately predict future risks for new patients who do not have complete medical histories. These models must balance the timing of predictions with their accuracy. Additionally, the models should provide enough interpretability and stability to be useful across a variety of decision-making scenarios for a wide range of diseases, particularly severe ones.

Interpretability is crucial because it allows healthcare providers to understand how and why a prediction was made, which is essential for gaining trust and ensuring the model's adoption in clinical settings. When a model is interpretable, clinicians can trace the factors leading to a specific prediction, facilitating informed decision-making and enabling them to explain the reasoning to patients, thus improving patient engagement and compliance. Moreover, interpretability aids in identifying potential biases or errors within the model, allowing for adjustments and improvements. It also ensures that the model's predictions are not just accurate but also actionable and reliable across different patient populations. This stability across diverse patient groups and various types of decisions ensures that the model can be applied broadly and consistently, particularly in diagnosing and managing severe diseases where decision-making is critical.

By incorporating mechanisms for interpretability, such as transparent algorithms, feature importance metrics, or visualizations, predictive models can provide insights that align with clinical knowledge and expertise, ultimately enhancing the overall effectiveness of healthcare delivery. Keeping these considerations in mind, we employed the Markov chain rule in our work to test the interpretability of the model.

2.3 Structuring of Data

This section outlines the process of extracting and organizing raw data from electronic health records (EHR) containing information on N patients diagnosed with severe mental disorders (MD). The data is structured into sequences to align with the experimental framework.

Each patient's daily clinical activities are comprised of a finite set of diagnoses, lab tests, and medications. Let D denote the universal sets of diagnoses, L denote lab tests, and M denotes medications for all N MD patients in the EHR, the data is organized as follows:

For each MD patient Pi in the EHR:

S_i represents the total number of days spent in the Intensive Care Unit (ICU).

D_i^s denotes the set of diagnoses during the patient's s^{th} day of ICU stay ($s = 1, 2, \ldots, S_i$).

L_i^s represents the set of lab tests during the patient's s^{th} day of ICU stay ($s = 1, 2, \ldots, S_i$).

M_i^s represents the set of medications during the patient's s^{th} day of ICU stay ($s = 1, 2, \ldots S_i$).

Utilizing these definitions, sequences of clinical activities (SofCA) are constructed, consisting of three temporal sequences for i^{th} MD patient:

(i) Let D_i denote the sequence of sets of diagnoses

It is represented by the following equation:

$$D_i = \{ D_i^1, D_i^2, \ldots\ldots, D_i^3\} \quad (1)$$

(ii) Let Li denote the sequence of sets of lab tests

It is represented by the following equation:

$$L_i = \{ L_i^1, L_i^2, \ldots\ldots.., L_i^3\} \quad (2)$$

(iii) Let M_i denote the sequence of sets of medications

It is represented by the following equation:

$$M_i = \{M_i^1, M_i^2, \ldots\ldots.., M_i^3\} \quad (3)$$

From these definitions, we can write:

$D = \sum_{i=1}^{n} D_i$ (4)

$M = \sum_{i=1}^{n} M_i$ (5)

$L = \sum_{i=1}^{n} L_i$ (6)

The following table shows organization of raw EHR data with N=2 patients:

Patient P_i	No of days of stay at ICU S_i	S^{th} day at ICU	Set of diagnoses D_i^s of P_i on S^{th} day	Set of lab tests L_i^s of P_i on S^{th} day	Set of medications M_i^s of P_i on S^{th} day	Sequence of sets of diagnoses D_i of P_i	Sequence of sets of lab tests L_i of P_i	Sequence of sets of medications M_i of P_i
P_1	3	1	$D_1^1=\{d_2,d_4,d_7\}$	$L_1^1=\{l_2,l_4\}$	$M_1^1=\{m_2,m_4\}$	$D_1=\{D_1^1,D_1^2,D_1^3\}$ $=\{\{d_2,d_4,d_7\},\{d_1,d_3\}$ $\{d_1,d_2,d_7,,d_8\}\}$	$L_1=\{L_1^1,L_1^2,L_1^3\}$ $=\{\{l_2,l_4\},\{l_1\}$ $\{l_2,l_4,\}\}$	$M_1=\{M_1^1,$ $M_1^2,M_1^3\}$ $=\{\{m_2,m_4\},$ $\{m_1,m_3\}$ $m_3,m_5,\}\}$
		2	$D_1^2=\{d_1,d_3\}$	$L_1^2=\{l_1\}$	$M_1^2=\{m_1,m_3\}$			
		3	$D_1^3=\{d_1,d_2,d_7,,d_8\}$	$L_1^3=\{l_2,l_4\}$	$M_1^3=\{m_3,m_5\}$			
P_2	2	1	$D_2^1=\{d_4,d_7\}$	$L_2^1=\{l_1,l_3\}$	$M_2^1=\{m_5\}$	$D_2=\{D_2^1,D_2^2\}$ $=\{\{d_4,d_7\},\{d_1,d_2,d_7,,$ $d_8\}\}$	$L_2=\{L_2^1,L_2^2\}$ $=\{\{l_1,l_3\},\{l_2,l_4,\}$ $\}$	$M_2=\{M_2^1,$ $M_2^2\}$ $=\{\{m_5\},$ $\{m_2,m_4\}\}$
		2	$D_2^2=\{d_1,d_2,d_7\}$	$L_2^2=$ $\{l_2,l_3,l_4\}$	$M_2^2=\{m_2,m_4\}$			

Table:1: Representation of the organization of raw EHR data for the N=2 patient

In Table 1, The set of sequence of diagnoses D can be represented as :

$D= D_1 \cup D_2 = \{\{d_2, d_4, d_7\},\{d_1,d_3\},\{d_1,d_2,d_7,d_8\},\{d_4,d_7\},\{d_1,d_2,d_7\}\}$

$= \{ diag_1, diag_2, diag_3, diag_4, diag_5\}$ with $\alpha= |D| = 5$ (7)

The set of sequence of lab tests L can be represented as:

$L= L_1 \cup L_2 = \{\{l_2, l_4\},\{l_1\},\{l_1,l_3\},\{l_2,l_3,l_4\}\}$

$= \{ lab_1, lab_2, lab_3, lab_4\}$ with $\beta= |L| = 4$ (8)

The set of sequence of medications M can be represented as:

$M= M_1 \cup M_2 = \{\{m_2, m_4\},\{m_1,m_3\},\{m_1,m_2,,m_3,m_4\},\{m_5\}\{ m_3,m_5\}\}$

$= \{ med_1, med_2, med_3, med_4\}$ with $\gamma= |M| = 4$ (9)

2.3.1 Identifying Unique Combination of Sequence of Clinical Activities

Once we have extracted the raw EHR data and organized it into sequences of clinical activities (Diagnoses, Lab Tests, and Medications), the next step is to obtain the unique combination sequence of these clinical activities for each patient. This process allows us to group patients based on their unique combination sequences of clinical activities, which will be discussed in detail in the next section. Here, we will focus on the step-by-step process of obtaining these unique combinations from the sequences of clinical activities of existing MD patients.

Unique combination of diagnoses is obtained by:

1) Identification of Diagnoses Sequences: In Table 1, D_1 represents the diagnoses set for Patient 1 and D_2 represents the diagnoses set for Patient 2, each representing a distinct combination of medical diagnoses.

2) Compilation and Merging of Sequences: D_1 and D_2 are compiled into a unified set D, combining all the diagnoses sequences obtained from D_1 and D_2

3) Detection of Unique Combination: Through systematic analysis, unique combination within set D are isolated and identified. This involves iterating through each sequence and verifying its uniqueness in the context of the entire set. Unique sequences are retained while duplicates are excluded. The list of unique combinations gives us important insights into the various ways diagnoses can occur in our dataset. By counting these unique sequences, we gain a complete understanding of the different patterns of diagnoses found in our sample of patients.

Equation (7) shows the mathematical representation for obtaining the unique combination of diagnoses where ‘α’ denotes the number of unique combination of diagnoses obtained across D_1 and D_2

Unique combination of lab tests is obtained by:

1) Identification of Sequences of Lab Tests: In Table 1, L_1 represents the set for lab tests for Patient 1 and D_2 represents the set for lab tests for Patient 2, each representing a distinct combination of lab tests.

2) Compilation and Merging of Sequences: L_1 and L_2 are merged into a single set L, combining all the sequences obtained from L_1 and L_2

3) Detection of Unique Combination: Through systematic analysis, unique combination within set D are isolated and identified. This involves iterating through each sequence and verifying its uniqueness in the context of the entire set. Unique sequences are retained while duplicates are excluded. The list of unique combinations gives us important insights into the various ways diagnoses can occur in our dataset. By counting these unique sequences, we gain a complete understanding of the different patterns of diagnoses found in our sample of patients.

Equation (8) shows the mathematical representation for obtaining the unique combination of lab tests where ‘β’ denotes the number of unique combination of lab tests obtained across L_1 and L_2

Unique combination of medications is obtained by:

Equation (9) shows the mathematical representation for obtaining the unique combination of medications where ‘γ’ denotes the number of unique combination of medications obtained across M_1 and M_2

1) Identification of Sequences of Medications: In Table 1, M_1 represents the set of medications for Patient 1 and D_2 represents the set of medications for Patient 2, each representing a distinct combination of medications.

2) Compilation and Merging of Sequences: M_1 and M_2 are merged into a single set M, combining all the sequences obtained from M_1 and M_2

3) Detection of Unique Combination: Through systematic analysis, unique combination within set D are isolated and identified. This involves iterating through each sequence and verifying its uniqueness in the context of the entire set. Unique sequences are retained while duplicates are excluded. The list of unique combinations gives us important insights into the various ways diagnoses can occur in our dataset. By counting these unique sequences, we gain a complete understanding of the different patterns of diagnoses found in our sample of patients.

2.4 Group Formation

In the context of patients' diagnosed diseases, individual traits, and socio-economic backgrounds, their clinical activities can vary significantly. Clustering patients based on their Sequence of Clinical Activities (SofCA) can help identify groups of patients with similar medical profiles. This, in turn, simplifies predicting the risk for a new patient by comparing their SofCA to that of the group with the closest resemblance. By doing so, we reduce the computational burden of assessing the SofCA of all patients in the Electronic Health Record (EHR) dataset.

Creating groups with similar SofCA faces two main challenges: finding a suitable metric to measure SofCA similarity despite varying lengths, and selecting an efficient clustering algorithm. In this study, we utilize the edit distance (ED) between SofCAs to categorize them into K groups. The ED between SofCA sequences of patients consists of three components: ED between their sequences of diagnoses, lab tests, and medications.

i) ED between the sequence of set of diagnoses D_i and D_j , denoted by $\zeta_{i,j}$

ii) ED between the sequence of set of lab tests L_i and L_j , denoted by $\eta_{i,j}$

iii) ED between the sequence of set of medications M_i and M_j , denoted by $\theta_{i,j}$

The normalized ED between the sequences of diagnoses, lab tests, and medications is then calculated as:

iv) Normalized ED between the sequence of set of diagnoses of patients Pi and Pj are given as:

$$\bar{\zeta}_{i,j} = \zeta_{i,j} \; / \max_{p=1,p\neq q}^{N}(\max_{q=1}^{N}(\zeta_{p,q})) \qquad (10)$$

v) Normalized ED between the sequence of set of lab tests of patients Pi and Pj are given as:

$$\bar{\eta}_{i,j} = \eta_{i,j} \; / \max_{p=1,p\neq q}^{N}(\max_{q=1}^{N}(\eta_{p,q})) \qquad (11)$$

vi) Normalized ED between the sequence of set of medications of patients Pi and Pj are given as:

$$\bar{\theta}_{i,j} = \theta_{i,j} \; / \max_{p=1,p\neq q}^{N}(\max_{q=1}^{N}(\theta_{p,q})) \qquad (12)$$

The combined ED between SofCAs of P_i and P_j is calculated as:

$$ED_{i,j} = \max(\bar{\zeta}_{i,j}, \bar{\eta}_{i,j}, \bar{\theta}_{i,j}) \quad (13)$$

Finally, the combined estimation of ED between SofCAs is computed (as shown in equation 18). This distance measure is used to categorize N patients into K groups. Distance between two SofCAs denotes dissimilarity between them. The choice of algorithm for this categorization is hierarchical density-based spatial clustering of applications with noise (hierarchical DBSCAN).

Hierarchical DBSCAN is chosen for its ability to divide the dataset into groups based on the density of neighbouring patients, without requiring a predefined number of groups. Additionally, it is suitable for datasets with arbitrary shapes and noise. Considering the spatially distributed nature of EHR data and varying density levels, hierarchical DBSCAN fulfills these requirements effectively. The symbol 'K' represents the number of groups identified by the hierarchical DBSCAN method.

2.5 Representing the Groups using Markov Chain Rule

This step holds significant importance for a couple of fundamental reasons:

1) Firstly, it is vital to guarantee that the model we are putting forward can be easily understood and interpreted. This ensures that stakeholders, whether they are clinicians or decision-makers, can grasp how the model functions and the insights it provides.

2) Secondly, it is crucial to replicate real-world scenarios where decisions are made based on past clinical activities. By utilizing a first-order Markov chain model, we are essentially capturing the probabilistic nature of sequential events in the data, mirroring how decisions unfold in actual clinical settings. This approach helps us better understand and predict the sequential patterns in clinical data, which is essential for effective decision-making and improving patient outcomes.

For the k^{th} group G_k of MD patients ($k = \{1, 2, \ldots, K\}$),

Let the number of transitions from the i^{th} set of diagnoses $diag_i$ D to the j^{th} set of diagnoses $diag_j$ D for i, j = 1, 2, …, | D | be denoted as $\alpha_{i,j,k}$

Let the number of transitions from the i^{th} set of lab tests lab_i L to the j^{th} set of lab tests labj L for i, j = 1, 2, …, | L | be denoted as $\beta_{i,j,k}$

Let the number of transitions from the i^{th} set of medications med_i M to the j^{th} set of medications medj M for i, j = 1, 2, …, | M | be denoted as $\gamma_{i,j,k}$

Let the number of sequences of sets of diagnoses starting with $diag_i \in D$ for i = 1, 2, …, | D | be denoted as $\mu_k(diag_i)$

Let the number of sequences of sets of lab tests starting with $diag_i \in D$ for i = 1, 2, …, | D | be denoted as $\mu_k(diag_i)$

Let the number of sequences of sets of lab tests starting with $lab_i \in L$ for i = 1, 2, …, | L | be denoted as $\mu_k(lab_i)$ and

Let the number of sequences of sets of medications starting with $med_i \in M$ for i = 1, 2,.., | M | be denoted as $\mu_k(med_i)$

Each group G_k is now represented with three state transition probability matrices δ_k , ρ_k, ω_k

$$\delta_k = [\, pd_{i,j,k} \,]_{|D| \times |D|} \qquad (14)$$

$$\rho_k = [pl_{i,j,k}]_{|L| \times |L|} \qquad (15)$$

$$\omega_k = [pm_{i,j,k}]|_{M|^{\times} |M|} \qquad (16)$$

where δ_k denotes the state-transition probability representing all transitions within all the SofCAs related to the diagnoses set, ρ_k denotes the state-transition probability representing all transitions within all the SofCAs related to the lab tests set and ω_k denotes the state-transition probability representing all transitions within all the SofCAs related to medication set

In equation (14), $pd_{i,j,k}$ denotes the probability of transition from the current state x_t of diagnoses-set $diag_i$ to the next state x_{t+1} of diagnoses-set $diag_j$ and is obtained from the following equation consulting with previous definitions and sequence of sets of lab tests (SofCA for diagnoses) of patients of group G_k.

$$pd_{i,j,k} = prob_k (x_{t+1} = diag_j / x_t = diag_i) = \alpha_{i,j,k} / \sum_{i=1}^{|D|} \alpha_{i,j,k} \qquad (17)$$

An illustrative representation of α_k is provided in Figure 2 below. We have assumed that the total number of distinct daily diagnoses, represented by D, is 5, following the example depicted in Figure 1.

State Trans. Prob	$diag_1$	$diag_2$	$diag_3$	$diag_4$	$diag_5$
$diag_1$	$pd_{1,1,k} = 0.23$	$pd_{1,2,k} = 0.63$	$pd_{1,3,k} = 0.02$	$pd_{1,4,k} = 0.10$	$pd_{1,5,k} = 0.02$
$diag_2$	$pd_{2,1,k} = 0.28$	$pd_{2,2,k} = 0.17$	$pd_{2,3,k} = 0.02$	$pd_{2,4,k} = 0.52$	$pd_{2,5,k} = 0.01$
$diag_3$	$pd_{3,1,k} = 0.05$	$pd_{3,2,k} = 0.14$	$pd_{3,3,k} = 0.31$	$pd_{3,4,k} = 0.15$	$pd_{3,5,k} = 0.35$
$diag_4$	$pd_{4,1,k} = 0.12$	$pd_{4,2,k} = 0.06$	$pd_{4,3,k} = 0.39$	$pd_{4,4,k} = 0.26$	$pd_{4,5,k} = 0.17$
$diag_5$	$pd_{5,1,k} = 0.43$	$pd_{5,2,k} = 0.21$	$pd_{5,3,k} = 0.31$	$pd_{5,4,k} = 0.01$	$pd_{5,5,k} = 0.04$

Table 2. Illustrative representation of α_k

In Table 2, we illustrate the probability of transitions within the diagnoses set. For example, the probability of transitioning from the diagnoses-set $diag_2$ to $diag_4$ is denoted as $pd_{2,4,k} = 0.52$. The probability of transitioning from diagnoses-set $diag_1$ to $diag_2$ is 0.63. These probabilities are derived from the observed sequences of sets of diagnoses among MD patients within group G_k.

In equation (15), $pl_{i,j,k}$ denotes the probability of transition from the current state x_t of diagnoses-set $diag_i$ to the next state x_{t+1} of diagnoses-set $diag_j$ and is obtained from the following equation consulting with previous definitions and sequence of sets of diagnoses (SofCA for diagnoses) of patients of group G_k.

$$pl_{i,j,k} = prob_k (x_{t+1} = lab_j / x_t = lab_i) = \beta_{i,j,k} / \sum_{i=1}^{|D|} \beta_{i,j,k} \quad (18)$$

In equation (16), $pm_{i,j,k}$ denotes the probability of transition from the current state x_t of diagnoses-set $diag_i$ to the next state x_{t+1} of diagnoses-set $diag_j$ and is obtained from the following equation consulting with previous definitions and sequence of sets of diagnoses (SofCA for diagnoses) of patients of group G_k.

$$pm_{i,j,k} = prob_k (x_{t+1} = med_j / x_t = med_i) = \gamma_{i,j,k} / \sum_{i=1}^{|D|} \gamma_{i,j,k} \quad (19)$$

As $pd_{i,j,k}$, $pl_{i,j,k}$, $pm_{i,j,k}$ are probability distributions. They are distributed over the universal sets of daily diagnoses (D), lab tests (L), medications (M) respectively.

$$\sum_{i=1}^{|D|} pd_{i,j,k} = 1 \text{ for } i = 1,2,\ldots.|D| \quad (20)$$

$$\sum_{i=1}^{|L|} pl_{i,j,k} = 1 \text{ for } i = 1,2,\ldots.|L| \quad (21)$$

$$\sum_{i=1}^{|M|} pm_{i,j,k} = 1 \text{ for } i = 1,2,\ldots.|M| \quad (22)$$

We also calculate the initial probabilities of members of Diagnoses set, Lab test set and Medication set of MD patients in group G_k, which is shown below

$$init_pd_k(diag_i) = \mu_k(diag_i) / \sum_{i=1}^{|D|} \mu_k(diag_j) \quad (23)$$

$$init_pd_k(lab_i) = \mu_k(lab_i) / \sum_{i=1}^{|L|} \mu_k(lab_j) \quad (24)$$

$$init_pd_k(med_i) = \mu_k(med_i) / \sum_{i=1}^{|M|} \mu_k(med_j) \quad (25)$$

Calculating the initial probability for Diagnoses, Lab tests, and Medications helps to establish an understanding of the likelihood of encountering each specific element at the beginning of a patient's medical journey within a particular group. By determining these initial probabilities, we gain insights into which diagnoses, lab tests, and medications are most commonly encountered first by patients in group G_k

In equations (23),(24),(25) we calculate the initial probability for each member of the Diagnoses, Lab Test, and Medication sets for MD patients in group Gk, by dividing the count of sequences starting with each member by the total count of sequences starting with any member in the respective set.

In Table 3, we calculate the initial probabilities of members of Diagnoses set D using Table 2

Total no of sequences in G_k	Number of sequences starting with				
	$diag_1$	$diag_2$	$diag_3$	$diag_4$	$diag_5$
25	5	2	11	3	4
I $init_pd_k$	5/25	2/25	11/25	3/25	4/25

Table 3. Illustration of initial probabilities of members of D

D ={ $diag_1$, $diag_2$, $diag_3$, $diag_4$,$diag_5$ }

The process described above is repeated for all K groups of MD patients identified through the hierarchical DBSCAN algorithm in the preceding step.

2.6 Forecasting Patient Risk with Limited Clinical Activity Sequences

The objective of this stage is to predict the possible clinical actions of a new patient, referred to as P, during his/her stay at the ICU. This is based on their Sequence of Clinical Activities (SofCA), which includes limited sequences of diagnoses, lab tests, and medications over a period of S days. The SofCA for the patient is categorized into three types according to the principle mention in section 3.3. These categories represent the ordered sequences of sets of diagnoses, lab tests, and medications for the patient during their stay.

Let A denote the SofCA for the ordered sequence of sets of diagnoses of the new patient P for S days, B denote the SofCA for the ordered sequence of sets of lab tests and C denote the SofCA for the ordered sequence of medications. A, B, C can be represented as,

$A=\{ a_1, a_2, a_3,\ldots\ldots..,a_s\}$, where $A \in D$ (26)

$B=\{ b_1, b_2, b_3,\ldots\ldots\ldots,b_s\}$, where $B \in L$ (27)

$C=\{ c_1, c_2, c_3,\ldots\ldots\ldots..,c_s\}$, where $C \in M$ (28)

Here, D, L, M refers to the universal diagnoses set, lab test set and medication set respectively, which has been derived from SofCA of the existing MD patients in the EHR.

a_i denotes the set of diagnoses that the patient has undergone during his/her i^{th} day of stay at ICU for i= 1,2,…..S

b_i denotes the set of lab tests that the patient has undergone during his/her i^{th} day of stay at ICU for i= 1,2,…..S

c_i denotes the set of medications that the patient has undergone during his/her i^{th} day of stay at ICU for i= 1,2,…..S

Risk prediction of P comprises of predicting the clinical activities of P for i= S+1, S+2,…… This is done in two steps :

Step-1: Assessing the Probability of P among Groups of Patients with Mental Disorders:

In the initial phase of the study, we aim to evaluate the probability of a specific patient, denoted as P, being associated with various groups of patients with MD. This evaluation is performed using distinct Markov chain models tailored to each patient group. The process involves determining the likelihood of patient P belonging to a particular group, such as G_k, by examining three key aspects:

1) The likelihood $\lambda_{k,1}$ representing the sequence of set of diagnoses A to G_k
2) The likelihood $\lambda_{k,2}$ representing the sequence of set of lab tests B to G_k
3) The likelihood $\lambda_{k,3}$ representing the sequemce of set of lab tests C to G_k

We , calculate $\lambda_{k\,1}$, $\lambda_{k,2}$, $\lambda_{k,3}$ as follows:

$$\lambda_{k,1} = \text{init_pd}_k(a_1)\prod_{i=1}^{S-1} prob_k(a_{i+1}/a_i) \quad (29)$$

$$\lambda_{k,2} = \text{init_pl}_k(b_1)\prod_{i=1}^{S-1} prob_k(b_{i+1}/b_i) \quad (30)$$

$$\lambda_{k,3} = \text{init_pm}_k(c_1)\prod_{i=1}^{S-1} prob_k(c_{i+1}/c_i) \quad (31)$$

Here, a_1 denotes the initial set of diagnoses of the new patient P, b_1 denotes the initial set of lab tests of the new patient P, c_1 denotes the initial set of medical tests of the new patient P.

$\text{init_pd}_k(a_1)$, $\text{init_pl}_k(b_1)$, $\text{init_pm}_k(c_1)$ are the initial probabilities (as governed by the Group G_k), the initial probabilities are calculated using equations (23), (24) and (25)

$prob_k(a_{i+1}/a_i)$ denotes the probability of transition from the diagnoses state a_i to the state of diagnoses a_{i+1}

$prob_k(b_{i+1}/b_i)$ denotes the probability of transition from the lab tests state b_i to the state of diagnoses b_{i+1}

$prob_k(c_{i+1}/c_i)$ denotes the probability of transition from the medications state c_i to the state of medications c_{i+1}

We have calculated these probabilities using Markov Chain models of the group G_k. Next we evaluate the composite likelihood of the group G_k to the ordered SofCA of the new patient P, as shown in equation 32. P is assigned to the group with highest composite likelihood. We denote the group as G_l

$$\lambda_k = \lambda_{k,1}\, \lambda_{k,2} \lambda_{k,3} \quad (32)$$

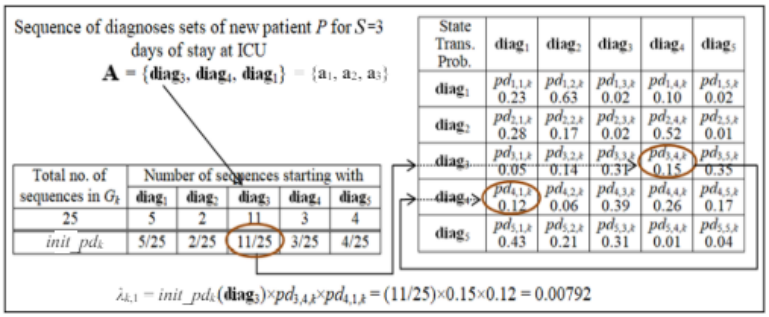


Fig 2: Sample assessment of $\lambda_{k\,1}$

In the above figure, we are assessing the probability $\lambda_{k,1}$ that the k^{th} group corresponds to the sequence of diagnosis sets for a new patient P

Step 2: Predicting Risk for P by Anticipating Future Clinical Activities within the Selected Group G_l

After the patient P is assigned to the group G_l with maximum likelihood probability to his/her ordered SofCA for S days of stay at ICU, next we have to identify the future set of activities of the patient from $(S+1)^{th}$ day onwards. We consult the state transition matrices (δ_l, ρ_l, ω_l) of G_l. Here δ_l, ρ_l, ω_l denotes the state transition probability matrices representing all the transitions within the diagnoses set, lab test set and the medication set respectively.

Considering the diagnoses set, the next most probable set of diagnoses, a^{S+1}, on the (S+1)th day, is evaluated by assessing the universal set D by consulting $prob_l(a_i|a_S)$ for each i from 1 to the total number of elements in D (for i=1,2,…..|D|). The set of diagnoses, denoted as $a_m \in D$, exhibiting the highest probability of transition $prob_l(a_m|a_S)$ from a_S is identified as a_{S+1}.

This indicates that the G_l group of MD patients most likely matching P has the highest frequency of transitions from the current diagnosis state $x_t = a_S$ to the next state $x_{t+1} = a_m$.

Mathematically, it is represented as:

$$prob_l\,(a_{s+1}/a_s\,) = max_{i=1}^{|D|}(prob_l(a_i/a_s)) \quad (33)$$

Similarly, set of lab tests lab $_{s+1}$ and medications med_{s+1} on $(S+1)^{th}$ ICU day is identified by:

$prob_l\ (b_{s+1}/b_s) = max_{i=1}^{|L|}(prob_l(b_i/b_s))$ (34)

and $prob_l\ (c_{s+1}/c_s) = max_{i=1}^{|M|}(prob_l(c_i/c_s))$ (35)

The same approach is applied to determine the finite sequence of F sets comprising diagnoses, lab tests, and medications throughout the patient's ICU stay on days (S+1), (S+2), ..., (S+F). An illustration of identifying the potential sequence of diagnosis sets for the next F = 4 days of a patient, who has already spent S = 3 days in the ICU, is presented in Figure 3. In this scenario, the SofCA of the k^{th} group G_k of MD patients, as depicted in Figure 2, is considered to be most analogous to patient P. Consequently, δ_k, as illustrated in the figures, is utilized for the risk prediction of patient P. Furthermore, P's SofCA, concerning their daily diagnosis set, is presumed to be identical to that depicted in Figure 3, with A = {diag3, diag4, diag1}.

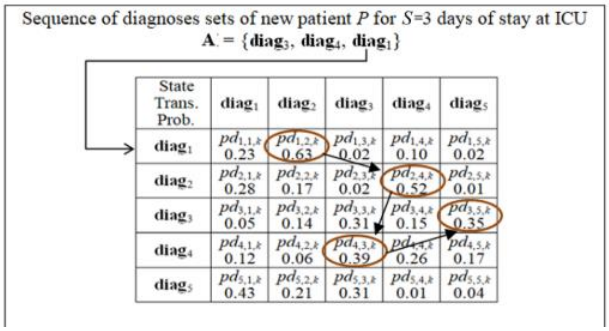


Fig 3: Determining the likely sequence of diagnosis sets for the next F = 4 days

2.7 Research Outcomes

In this section, we will be discussing about the results of our analysis. We will focus on the EHR database that we have used in our study, evaluate our study using a suitable performance metric and perform a comparative analysis with other state-of-art machine learning models

2.7.1 Dataset

In this study, we utilized the Medical Information Mart for Intensive Care (MIMIC)-III version 1.4 database, which is openly accessible via Physionet's official website. MIMIC contains anonymized health data from patients admitted to critical care units at Beth Israel Deaconess Medical Center between 2001 and 2012, totaling 46,520 patients. Among these, 19,868 patients were diagnosed with at least one mental disorder according to the International Classification of Diseases, Ninth Revision (ICD-9) codes falling within the range 290 to 319.

For our analysis, we have partitioned the data. 70% of the data is used for training, 10% of the data is used for validation and 20% data is used for testing. A 10- fold cross validation technique has been used for generalizing the state-of-art machine learning algorithms. All machine learning codes have been implemented using Python.

2.7.2 Model Evaluation

Edit Distance (ED) has been used to evaluate the performance of the proposed risk-prediction model. An overview of Edit Distance is provided in chapter 2.1.3. For the new patient P, who has already stayed in ICU for S days, the edit distance calculated between the predicted and true SofCA of P from $(S+1)^{th}$ day to $(S+F)^{th}$ day is used as the performance metric for evaluating the model.

Equations (26), (27), (28) represents the set of SofCAs of P during his/her stay in ICU for S days.

For the next F days of stay in ICU, let the true and predicted sequence of set of diagnoses be denoted by A_t and A_p respectively, B_t and B_p denote the true and predicted sequence of lab tests and C_t and C_p denote the true and predicted sequence of medications. They can be represented as:

$$A_p = \{a_{s+1}, a_{s+2}, \ldots\ldots, a_{s+F}\} \quad (36)$$

$$B_p = \{b_{s+1}, b_{s+2}, \ldots\ldots, b_{s+F}\} \quad (37)$$

$$Cp = \{c_{s+1}, c_{s+2}, \ldots., c_{s+F}\} \quad (38)$$

$$A_t = \{ {}^{t}a_{s+1}, {}^{t}a_{s+2}, \ldots.., {}^{t}a_{s+F}\} \quad (39)$$

$$B_t = \{ {}^{t}b_{s+1}, {}^{t}b_{s+2}, \ldots\ldots. {}^{t}b_{s+F}\} \quad (40)$$

$$Ct = \{{}^{t}c_{s+1}, {}^{t}c_{s+2}, \ldots\ldots {}^{t}c_{s+F}\} \quad (41)$$

Let the Edit distance (ED) between Ap and A_t be denoted by σ, ε denote the ED between B_p and B_t, ED between C_p and C_t be denoted by ϑ. Then we calculate the composite normalized Difference between the true and predicted SofCA of P, denoted by φ, which is shown below:

$$\varphi = \max(\sigma/F, \varepsilon/F, \vartheta/F) \quad (42)$$

It has been observed that the actual and predicted sequences are more similar for lower value of φ, which indicates better performance by the model.

S	*F*								
	10%	20%	30%	40%	50%	60%	70%	80%	90%
10%	0.0251	0.0581	0.0943	0.1423	0.2039	0.2403	0.3477	0.5247	0.6982
20%	0.0205	0.0456	0.0758	0.1245	0.1473	0.1972	0.2928	0.4014	NA
30%	0.0173	0.0354	0.0637	0.0877	0.1087	0.1581	0.2229	NA	NA
40%	0.0157	0.0275	0.0474	0.0631	0.0815	0.1134	NA	NA	NA
50%	0.0149	0.0237	0.0323	0.0459	0.0621	NA	NA	NA	NA
60%	0.0084	0.0133	0.0235	0.0292	NA	NA	NA	NA	NA
70%	0.0058	0.0078	0.0104	NA	NA	NA	NA	NA	NA
80%	0.0023	0.0042	NA	NA	NA	NA	NA	NA	NA
90%	0.0014	NA	NA	NA	NA	NA	NA	NA	NA

Fig 4. Table representing the values of φ obtained by varying S and F (calculated in terms of percentage of entire duration of stay of P in ICU)

Normalized Composite Difference or the error (φ) that we have obtained in equation (42) is dependent on S which is the total number of days for which P has stayed in ICU, and F which is the number of days for which our model has predicted the clinical activities of P. The table shown in Fig 6 depicts the value of φ obtained across different values obtained by varying S and F. For example, if P has stayed in ICU for 50 days, then S has been varied from 5 days (which is 10% of S) to 45 days (which is 90% of S). Similarly, F has also been varied from 5 days to 45 days which 10% and 90% of the entire duration, for which φ has been calculated. From the table shown in Fig 6, it can be observed that the prediction accuracy is more for smaller φ values, which is achieved for large S values and lower F values. We can say that our algorithm will be more effective in predicting the future SofCA of P of near future only with a longer duration of stay. However, from the table we can see that for S=10% and F=60%, we get a smaller φ value, which indicates that our proposed algorithm is robust irrespective of S and F values. Here "NA" stands for not applicable, used in situations where the combined total of S and F exceeds 100%.

2.7.3 Comparison with State-of-the-Art Models

Here, we have compared our proposed algorithm with other state-of-art models. State-of art models are the models which have achieved superior performance in a particular field or task. Researchers often validate their study against these models in order to evaluate the effectiveness of their approach. We have performed a comparative analysis of our proposed algorithm with three state-of-art machine learning techniques namely, extreme gradient descent boosting (XGB), bidirectional long short-term memory (BiLSTM) and minimal gated recurrent unit (MGRU). The above architectures are useful in predicting time series healthcare problems. LSTM-based architectures are efficient in learning from sequential data [23]. They are useful for tasks like disease monitoring and disease progression tasks. These models can identify the risk factors and the patterns associated with various diseases. We have used the above three benchmark architectures and validated our proposed model. In order to tune the hyperparameter of these machine learning algorithms we have used a 10-fold cross validation

The table below shows the optimized hyperparameters for XGB, BiLSTM and MGRU

Algorithm	XGBoost				BiLSTM				MinGRU	
Hyperparameters	No.of Estimators	Learning	Gamma	Max Tree Depth	Hidden Layers	Weight Regularization	Activation	Dropout	Learning Rate	Hidden Units
Value chosen	45	0.05	5	8	1	L1 =0, L2= 0.01	ReLU	0.1	0.0005	500

Table 4: Showing the optimised hyperparameters for three state-of-art models

S	Competing algorithms	F								
		10%	20%	30%	40%	50%	60%	70%	80%	90%
10%	Prop. Method	0.0251	0.0581	0.0943	0.1423	0.2039	0.2403	0.3477	0.5247	0.6982
	XGB	0.0609	0.0912	0.1203	0.1738	0.2235	0.2919	0.4192	0.5714	0.7716
	BiLSTM	0.0502	0.0823	0.1098	0.1492	0.2133	0.2323	0.3734	0.5603	0.7334
	MinGRU	0.0605	0.1107	0.1857	0.2727	0.3777	0.5044	0.6305	0.7428	0.8495
20%	Prop. Method	0.0205	0.0456	0.0758	0.1245	0.1473	0.1972	0.2928	0.4014	NA
	XGB	0.0347	0.0578	0.0890	0.1359	0.1724	0.2131	0.2981	0.3529	NA
	BiLSTM	0.0318	0.0489	0.0824	0.1414	0.1913	0.2442	0.3376	0.4991	NA
	MinGRU	0.0456	0.0907	0.1286	0.1933	0.2728	0.3763	0.5007	0.6072	NA
30%	Prop. Method	0.0173	0.0354	0.0637	0.0877	0.1087	0.1581	0.2229	NA	NA
	XGB	0.0254	0.0432	0.0755	0.1058	0.1339	0.1547	0.2434	NA	NA
	BiLSTM	0.0219	0.0381	0.0713	0.1042	0.1423	0.1891	0.2693	NA	NA
	MinGRU	0.0351	0.0693	0.1054	0.1413	0.2041	0.2849	0.3924	NA	NA
40%	Prop. Method	0.0157	0.0275	0.0474	0.0631	0.0815	0.1134	NA	NA	NA
	XGB	0.0188	0.0349	0.0583	0.0777	0.0927	0.1715	NA	NA	NA
	BiLSTM	0.0184	0.0327	0.0522	0.0785	0.1039	0.1521	NA	NA	NA
	MinGRU	0.0279	0.0494	0.0659	0.0764	0.1365	0.1873	NA	NA	NA
50%	Prop. Method	0.0149	0.0237	0.0323	0.0459	0.0621	NA	NA	NA	NA
	XGB	0.0157	0.0265	0.0378	0.0503	0.0745	NA	NA	NA	NA
	BiLSTM	0.0165	0.0278	0.0391	0.0494	0.0685	NA	NA	NA	NA
	MinGRU	0.0193	0.0286	0.0394	0.0515	0.0724	NA	NA	NA	NA
60%	Prop. Method	0.0084	0.0133	0.0235	0.0292	NA	NA	NA	NA	NA
	XGB	0.0094	0.0147	0.0226	0.0329	NA	NA	NA	NA	NA
	BiLSTM	0.0093	0.0134	0.0215	0.0323	NA	NA	NA	NA	NA
	MinGRU	0.0122	0.0171	0.0254	0.0371	NA	NA	NA	NA	NA
70%	Prop. Method	0.0058	0.0078	0.0104	NA	NA	NA	NA	NA	NA
	XGB	0.0068	0.0088	0.1098	NA	NA	NA	NA	NA	NA
	BiLSTM	0.0062	0.0094	0.0137	NA	NA	NA	NA	NA	NA
	MinGRU	0.0074	0.0092	0.0139	NA	NA	NA	NA	NA	NA
80%	Prop. Method	0.0023	0.0042	NA	NA	Na	NA	NA	NA	NA
	XGB	0.0028	0.0043	NA	NA	NA	NA	NA	NA	NA
	BiLSTM	0.0039	0.0054	NA	NA	NA	NA	NA	NA	NA
	MinGRU	0.0041	0.0075	NA	NA	NA	NA	NA	NA	NA
90%	Prop. Method	0.0014	NA	NA	NA	Na	NA	NA	NA	NA
	XGB	0.0015	NA	NA	NA	NA	NA	NA	NA	NA
	BiLSTM	0.0019	NA	NA	NA	NA	NA	NA	NA	NA
	MinGRU	0.0027	NA	NA	NA	NA	NA	NA	NA	NA

Fig 5: Comparing the φ values obtained from our proposed model and other state-of-the-art architectures

From the table shown in Fig. 5, the φ values obtained by the state-of-the -art models shows that they have achieved equal accuracy in predicting the future SofCA in almost all cases. But, when S is varied from 10% to 40% and F is varied from 60% to 90%, we can observe a significant difference in the performance of the models. Our proposed model has achieved better prediction accuracy (minimum φ) in most cases when compared to the other three architectures except in the following three cases:

i) For S =20% and F=80% , BiLSTM has achieved minimum φ value compared to our model

ii) XGB has achieved better accuracy for S = 30% and F=60%

iii) BiLSTM has outperformed our model when S=10% and F=60%

2.8 Conclusion

This thesis introduces a new machine learning method which is designed to forecast the likelihood of ICU patients developing severe mental disorders. Its innovation is highlighted by several factors. Primarily, in contrast to much of the existing research, this approach takes into account a variety of clinical activities to predict treatment outcomes, including diagnoses, lab tests, and medications. Secondly, the technique predicts the future sequence of clinical activities (SofCA) for a new ICU patient by focusing on a specific subset of mental disorder (MD) patients who have the highest likelihood of similarity, rather than using the electronic health records (EHRs) of all MD patients, as most existing studies do. This approach improves prediction accuracy. Thirdly, the hierarchical DBSCAN method is employed to cluster the complete SofCA of existing patients in the EHR, eliminating the need to predefine the number of clusters and allowing for clusters of any shape. Lastly, a Markov model is utilized to capture the transitions in clinical activities within each patient group, providing transition probabilities that medical practitioners can easily interpret to predict a new patient's treatment plan.

Future research will focus on applying the proposed method to various types of electronic health records (EHRs). This will involve refining the model to incorporate real-time human input and uncover hidden relationships among different features such as lab tests, medications, medical images, radiology reports, patient demographics, and environmental factors. Additionally, a significant area of interest will be implementing this method in real-time resource allocation scenarios.

References & Bibliography

[1] Seymour, Dr. Tom & Frantsvog, Dean & Graeber, Tod. (2014). Electronic Health Records (EHR). 10.19030/ajhs.v3i3.7139.

[2] Honeyford, K., Expert, P., Mendelsohn, E. E., Post, B., Faisal, A. A., Glampson, B., Mayer, E. K., & Costelloe, C. E. (2022, August 19). Challenges and recommendations for high quality research using electronic health records. Frontiers in Digital Health, 4, 940330. https://doi.org/10.3389/fdgth.2022.940330.

[3] McClean, S., Faddy, M., & Millard, P. (2006). Evaluating the Effectiveness of a Health and Community Care System Using Markov Models. In 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06) (pp. 777-782). Salt Lake City, UT, USA. https://doi.org/10.1109/CBMS.2006.164.

[4] Wang P, Li J. Implementation of Real-Time Medical and Health Data Mining System Based on Machine Learning. J Healthc Eng. 2021 Nov 19;2021:7011205. doi: 10.1155/2021/7011205. PMID: 34840702; PMCID: PMC8626197.

[5] Williams R, Buchan IE, Prosperi M, Ainsworth J. Using String Metrics to Identify Patient Journeys through Care Pathways. AMIA Annu Symp Proc. 2014 Nov 14;2014:1208-17. PMID: 25954432; PMCID: PMC4419997.

[6] Johny-Brooks-Barlett. (2018, January 4). Probability Concepts Explained: Maximum Likelihood Estimation. Towards Data Science. https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1.

[7] Zhang L, Ding X, Ma Y, Muthu N, Ajmal I, Moore JH, Herman DS, Chen J. A maximum likelihood approach to electronic health record phenotyping using positive and unlabeled patients. J Am Med Inform Assoc. 2020 Jan 1;27(1):119-126. doi: 10.1093/jamia/ocz170. PMID: 31722396; PMCID: PMC6913222.

[8] Smith, Matthew & Phillips, Rachael & Luque-Fernandez, Miguel & Maringe, Camille. (2023). Application of targeted maximum likelihood estimation in public health and epidemiological studies: a systematic review. Annals of Epidemiology. 86. 10.1016/j.annepidem.2023.06.004.

[9] World Health Organization. (2022, June 8). Mental disorders. https://www.who.int/news-room/fact-sheets/detail/mental-disorders.

[10] World Health Organization. Mental Health. https://www.who.int/health-topics/mental-health#tab=tab_1.

[11] Min JA, Lee CU, Lee C. Mental health promotion and illness prevention: a challenge for psychiatrists. Psychiatry Investig. 2013 Dec;10(4):307-16. doi: 10.4306/pi.2013.10.4.307. Epub 2013 Dec 16. PMID: 24474978; PMCID: PMC3902147.

[12] Jetli Chung, Jason Teo, "Mental Health Prediction Using Machine Learning: Taxonomy, Applications, and Challenges", *Applied Computational Intelligence and Soft Computing*, vol. 2022, Article ID 9970363, 19 pages, 2022. https://doi.org/10.1155/2022/9970363.

[13] Wu, Yirong, Burnside, Elizabeth, Cox, Jennifer, Fan, Jun, Yuan, Ming, Yin, Jie, Peissig, Peggy, Cobian, Alexander, Page, David, & Craven, Mark. (2017). Breast Cancer Risk Prediction Using Electronic Health Records. In Proceedings of the IEEE International Conference on Healthcare Informatics (ICHI) (pp. 224-228). DOI: 10.1109/ICHI.2017.62.

[14] E. Xu, S. Zhao, J. Mei, E. Xia, Y. Yu and S. Huang, "Multiple MACE Risk Prediction using Multi-Task Recurrent Neural Network with Attention," *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, Xi'an, China, 2019, pp. 1-2, doi: 10.1109/ICHI.2019.8904675.

[15] Yu, Zehao, Yang, Xi, Chen, Yiqing, Fang, Ruogu, Hogan, William, Gong, Yan, & Wu, Yonghui. (2022). Identify Cancer Patients at Risk for Heart Failure using Electronic Health Record and Genetic Data. In Proceedings of the IEEE International Conference on Healthcare Informatics (ICHI) (pp. 138-142). DOI: 10.1109/ICHI54592.2022.00032.

[16] Cheng, Y., Wang, F., Zhang, P., & Hu, J. (2016). Risk prediction with electronic health records: A deep learning approach. In *Proceedings of the 2016 SIAM International Conference on Data Mining* (pp. 432-440). https://doi.org/10.1137/1.9781611974348.49

[17] Che, Z., Cheng, Y., Zhai, S., Sun, Z., & Liu, Y. (2017). Boosting deep learning risk prediction with generative adversarial networks for electronic health records. In *2017 IEEE International Conference on Data Mining (ICDM)* (pp. 787-792). New Orleans, LA, USA. https://doi.org/10.1109/ICDM.2017.93

[18] Ma, F., Gao, J., Suo, Q., You, Q., Zhou, J., & Zhang, A. (2018). Risk prediction on electronic health records with prior medical knowledge. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1910-1919). https://doi.org/10.1145/3219819.3220020.

[19] Cui, S., Wang, J., Gui, X., Wang, T., & Ma, F. (2022, December). AutoMed: Automated medical risk predictive modeling on electronic health records. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 948-953). https://doi.org/10.1109/BIBM55620.2022.9995209. PMID: 37063974; PMCID: PMC10102833

[20] Grout, R., Gupta, R., Bryant, R., Elmahgoub, M. A., Li, Y., Irfanullah, K., Patel, R. F., Fawkes, J., & Inness, C. (2024). Predicting disease onset from electronic health records for population health management: a scalable and explainable deep learning approach. *Frontiers in Artificial Intelligence*, 6, 1287541. DOI: 10.3389/frai.2023.1287541. PMID: 38259826; PMCID: PMC10800619.

[21] Panay, B., Baloian, N., Pino, J. A., Peñafiel, S., Sanson, H., & Bersano, N. (2019). Predicting Health Care Costs Using Evidence Regression. *Proceedings*, 31(1), 74. DOI: 10.3390/proceedings2019031074.

[22] Fauziah, S., Renaldi, F., & Santikarama, I. (2022). Prediction of Length of Stay and Cost for Diabetes Inpatients using C4.5 Algorithm. In *2022 IEEE Creative Communication and Innovative Technology (ICCIT)* (pp. 1-5)

[23] Chen, T., & Guestrin, C. (2016, August). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672.2939785.

Chapter 3

Conclusion and Future Scope

3.1 Introduction

In this thesis we have introduced a new machine learning approach for predicting the risk of severe Mental Disorder (MD) patients. Early risk prediction of Mental Disorders(MD) is crucial as it can enable timely intervention and prevent other severe long term outcomes associated with these conditions. In the previous chapters, we have highlighted about the importance of early risk predictions using Electronic Health Records (EHRs). One of the advantages of early risk prediction is that it can help the clinical advisors and the doctors to plan a treatment path and start with the diagnosis as early as possible. Secondly, it can help to reduce the healthcare costs, as it can provide an overview of the future clinical expenses.

3.2 Summary

An overview of this work will be presented in this section. We developed a risk prediction model aimed at forecasting the risk of newly-admitted mental disorder patients, particularly those with limited medical records. Initially, we utilized the health records of existing MD patients available in the Electronic Health Record (EHR) system to analyze various clinical activities they underwent, including lab tests, diagnoses, and medications recorded on a day-to-day basis. This data organization involved structuring the sequences to obtain unique combinations for each patient, as detailed in Chapter 2.3. Subsequently, we categorized existing MD patients into groups based on the similarity of their clinical activities, utilizing Edit Distance as a measure, as explained in Chapter 2.4. After group formation, we employed a Markov Model to mathematically represent each group and simulate decision-making by constructing a state transition matrix, outlined in Chapter 2.5. To predict the risk of newly-admitted patients, we assigned them to the groups with the highest likelihood and then forecasted the sequence of their future clinical activities for a specified number of days, as described in Chapter 2.6. We have validated our model and compared it with the state-of-the art architectures. In majority of cases, our model depicted better prediction accuracy compared to these models.

3.3 Results and Discussion

In this study, we have used the Medical Information Mart for Intensive Care (MIMIC-III) version 1.4 database, specifically utilizing records pertaining to patients with mental disorders (MDs). Edit Distance (ED) serves as the performance metric for assessing our algorithm. We assess similarity by calculating the normalized composite difference between predicted and original sequences of clinical activities. A lower error value indicates greater similarity, reflecting superior performance of our model. Furthermore, our algorithm accurately predicts the risk and future clinical activities of new patients in the near future. It remains effective even with a longer patient history. Additionally, we conducted a comparative analysis with other state-of-the-art techniques such as BiLSTM, XGB, and MGRU. In most cases, our proposed risk prediction model surpasses these benchmark models. Further discussion on our research outcomes is provided in Chapter 2.8.

3.4 Future Scope

Now, we wish to focus on adapting the proposed method to accommodate diverse Electronic Health Record (EHR) systems. This involves refining the model by integrating real-time human input and uncovering hidden relationships among various patient features such as lab tests, medications, medical images, radiology reports, patient demographics, and environmental factors. Additionally, we also aim to explore the application of the developed method in real-time resource allocation scenarios within healthcare settings, addressing the challenges and opportunities presented by dynamic patient needs and resource constraints.

3.5 Potential Impact and Future Directions

The findings of this study demonstrate significant potential for enhancing patient care and management in the field of mental health. By leveraging the Medical Information Mart for Intensive Care (MIMIC-III) version 1.4 database, we specifically targeted records pertaining to patients with mental disorders (MDs), utilizing Edit Distance (ED) as a performance metric. This approach allows us to assess the similarity between predicted and actual sequences of clinical activities, where a lower error value indicates greater similarity and superior performance of our model.

Our algorithm not only accurately predicts the risk and future clinical activities of new patients in the near future but also maintains its effectiveness with longer patient histories. This capability is crucial for proactive healthcare management, allowing clinicians to anticipate and address potential mental health issues before they become critical.

Moreover, our comparative analysis with state-of-the-art techniques such as BiLSTM, XGB, and MGRU reveals that our proposed risk prediction model generally outperforms these benchmarks. This superior performance underscores the robustness and reliability of our approach, highlighting its potential for widespread application in clinical settings.

The implications of our research are far-reaching. With the ability to predict future clinical activities and assess risk with high accuracy, healthcare providers can make more informed decisions, leading to improved patient outcomes. This proactive approach can significantly reduce the burden on healthcare systems by preventing the escalation of mental health issues and optimizing resource allocation. In addition, the successful integration of our algorithm into diverse Electronic Health Record (EHR) systems can standardize and enhance predictive analytics across different healthcare environments. This could pave the way for more personalized and timely interventions, ultimately transforming the landscape of mental health care.

The potential impact of our work extends beyond individual patient care. By contributing to the development of more accurate and scalable predictive models, our research supports the broader goal of advancing healthcare analytics. This, in turn, fosters a more data-driven and evidence-based approach to managing mental health, benefiting both patients and healthcare providers alike.

In conclusion, our study not only advances the understanding and application of machine learning in mental health risk prediction but also sets the stage for future innovations in the field. As discussed in Chapter 2.8, our research outcomes highlight the promise of our approach and the opportunities for further refinement and application in real-world healthcare scenarios.