EMERGENCY DEPARTMENT RESOURCE PREDICTION USING EXPLAINABLE DEEP LEARNING

DEEP LEARNING CLASSIFICATION AND MODEL EXPLAINABILITY FOR PREDICTION OF MENTAL HEALTH PATIENTS EMERGENCY DEPARTMENT VISIT

By SAJJAD RASHIDIANI, B.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Master of Applied Science

McMaster University
MASTER OF APPLIED SCIENCE (2022)
Hamilton, Ontario, Canada (Elec. & Comp. Engineering)

TITLE:	Deep Learning Classification and Model Explainability for Prediction of Mental Health Patients Emergency Department Visit
AUTHOR:	Sajjad Rashidiani BS (Electrical Engineering), Isfahan University of Technology, Isfahan, Iran
SUPERVISOR:	Dr. Thomas E. Doyle
NUMBER OF PAGES:	xvii, 150

Lay Abstract

In this document, an Artificial Intelligence (AI) approach for predicting 6-month Emergency Department (ED) visits is proposed. In this approach, the questionnaires gathered from children and youth admitted to an outpatient or inpatient clinic are converted to a text representation called Textionnaire. Next, AI is utilized to analyze the Textionnaire and predict the possibility of a future ED visit. This method was successful in about 75% of the time. In addition to the AI solution, an explainability component is introduced to explain how the natural language processing algorithm identifies the high risk patients.

Abstract

The rate of Emergency Department (ED) visits due to mental health and drug abuse among children and youth has been increasing for more than a decade and is projected to become the leading cause of ED visits. Identifying high-risk patients well before an ED visit will enable mental health care providers to better predict ED resource utilization, improve their service, and ultimately reduce the risk of a future ED visit. Many studies in the literature utilized medical history to predict future hospitalization. However, in mental health care, the medical history of new patients is not always available from the first visit and it is crucial to identify high risk patients from the beginning as the rate of drop-out is very high in mental health treatment. In this study, a new approach of creating a text representation of questionnaire data for deep learning analysis is proposed. Employing this new text representation has enabled us to use transfer learning and develop a deep Natural Language Processing (NLP) model that estimates the possibility of 6-month ED visit among children and youth using mental health patient reported outcome measures (PROM). The proposed method achieved an Area Under Receiver Operating Characteristic Curve of 0.75 for classification of 6-month ED visit. In addition, a novel method was proposed to identify the words that carry the highest amount of information related to the outcome of the deep NLP models. This measurement of word information using

Entropy Gain increases the explainability of the model by providing insight to the model attention. Finally, the results of this method were analyzed to explain how the deep NLP model achieved a high classification performance.

To hundreds of students who were imprisoned or were deprived from their rights to study during the Woman, Life, Freedom movement.

Acknowledgements

I would like to deeply appreciate the guidance and support of my supervisor, Dr. Thomas Doyle. Dr. Doyle has contributed greatly to my project through his expertise and knowledge in the field. I would also like to acknowledge supports of all the lab members of Biomedic.AI Lab, particularly my colleague and friend Mr. Md Asif Khan.

Next, I would like to thank the Dr. Sassi, Dr. Duncan, Dr. Pires, Dr. Samavi, and Dr. Mauluddin for their supports through out this project. They provided me with all the domain knowledge required for this project and trusted me with finding the solutions present in this thesis.

I would also like to recognize my family (Mom, Dad, Sahar, Shirin) who supported me in pursuing my academic interests, as well as friends (Alireza, Maral, Ghazal, and Kamran) who made my masters journey in Canada an amazing part of my life.

Lastly, I would like to acknowledge McMaster University, my home for the past two years, for funding my project and research.

Table of Contents

Lay Abstract **iii**

Abstract **iv**

Acknowledgements **vii**

Definitions, and Abbreviations **xvii**

Declaration of Academic Achievement **xviii**

1 Introduction **1**

1.1 Motivation 1

1.2 Overview of the Field 2

1.3 Thesis Questions 4

1.4 Evaluation Overview 6

1.5 Organization and Scope 7

2 Literature Review **9**

2.1 Feature Selection 10

2.2 Data Imbalanced Problem 11

2.3 Transfer Learning 14
2.4 Softmax 15
2.5 Transformer Models and BERT 16
2.6 Decision Tree Classifier 23
2.7 Medical Nomenclature 23
2.8 Evaluation Metrics 25
2.9 Systematic Review 26

3 Domain Data **46**
3.1 Data 46
3.2 Preprocessing 49

4 Machine Learning Model **51**
4.1 Question to Text Conversion (Textionnaire) 51
4.2 Artificial Intelligence Method 55
4.3 Baseline Model 58
4.4 Attention Information for Explainability 58

5 Methodology **65**
5.1 Machine Learning Pipeline 65
5.2 Experiments 71

6 Results **78**
6.1 Experiment 0: Base Line Performance 78
6.2 Experiments 1-4: Verification of the Efficacy of Textionnaire 79
6.3 Experiment 5: Evaluation of Wide and Deep Architecture 81

6.4 Verification and Interpretation of the Attention Information Method . 82

7 Discussion **99**

7.1 What Is the Rational of Using Textionnaire? 99

7.2 How Effective Is the Textionnaire? 100

7.3 How Well Does the Deep Model Perform Compared to the Shallow Model? 100

7.4 How Well Does the Wide and Deep Architecture Perform Compared to Other Models? 101

7.5 What Are the Implications of This Study? 101

7.6 What Does the Explainability Method Tell Us? 102

7.7 What Are the Limitations of This Study? 102

8 Conclusion **104**

A Literature Review Data **107**

A.1 Search Strings 107

B Additional Questionnaire Information **117**

B.1 Private Mental Health Dataset 117

B.2 NSDUH Dataset 117

C Sophistication Scores **118**

C.1 Sophistication Score Nomenclature 118

C.2 Scoring System 122

D Clustering of Words Based on Meaning **125**

E Attention Information Tables **127**

List of Figures

2.1 In the first iteration a base-learner (A Decision Tree in this case) splits the data to two groups, one includes only circles (circle class) the other one includes only squares (square class). The missclassified samples are displayed in red. In the second iteration an other base-learner splits the previous square class that has 3 miss-classifications in two new groups reducing the number of miss-classifications to only 1 (one square in new circle group). Finally in the third iteration the last base-learner splits the last group with miss-classification to two groups with perfect classification. 12

2.2 Distribution of a) a balanced, b) an imbalanced one-dimensional dataset, and c) the randomly undersampled version of (b). 13

2.3 How Transfer Learning Works. The similarity between the colours of Source and Target data blocks, and Source and Target Task blocks indicates the similarity between them. 15

2.4 Transformer Architecture 16

2.5 Embedding of a BERT model with a maximum length of 15. 19

2.6 An Encoder layer of BERT model. Note that LLT stands for Learned Linear Transformation where LLT(Q) generates Query, LLT(K) generates Key, and LLT(V) creates the Value matrix. 20
2.7 Multi-Head Attention architecture consist of several attention heads . 22
2.8 PRISMA chart of the review 32
2.9 Distribution of readmission rates in different datasets. 35
2.10 Number of papers focused on predicting readmission after different types of initial admission. 37
2.11 Number of studies focused on different outcomes of interest. 38
2.12 Number of studies used different categories of predictors. 40
2.13 Number of studies used different categories of machine machine algorithms. 42
2.14 Box plot of AUROC of machine machine algorithms category. 43
2.15 Scatter plot of the Quality Assessment Score of different studies and their AURCO labeled by their best performing models. 44
4.1 Generated Textionnaire for one of the NSDUH-B dataset's samples. . 55
4.2 Deep and shallow models. 56
4.3 The Wide and Deep architecture proposed for 6-month ED visit prediction. 57
4.4 An example of attention weight assignment. 60
4.5 The process of generating Attention Maps 61
4.6 The process of finding the most important pairs in each head of each encoder layer. 64

6.1 Attention Information aggregated across all layers and attention heads for a Textionnaire with a sentence replaced with irrelevant sentence. . 95
6.2 Attention Information aggregated across all layers and attention heads for a Textionnaire with original sentences. 96
6.3 Attention Information aggregated across all layers and attention heads for a Textionnaire with a sentence replaced with paraphrased sentence. 98
D.1 Clustering of words using their semantic similarity scores generated via Spacy library. 126

List of Tables

2.1 Data sources used in the papers reviewed in this study and the papers used those data sources 34

2.2 Papers stratified by the imbalanced data handling methods. 36

2.3 List of papers that used each one of the feature selection methods. . . 39

2.4 Papers stratified based on the their imputation method. 41

3.1 Private CYMHP and Public NSDUH Datasets Characteristics 48

4.1 The questionnaire options from "Because of a physical, mental or emotional condition, do you have serious difficulty concentrating, remembering, or making decisions?" are represented by these descriptive sentences in the generated text. Codes 1, 2, and 94 denote a positive, negative, and "Don't Know" answer to this question, respectively. . . 53

6.1 Performance metrics of the SVM and Shallow model tested on the tabular version of the Private dataset and OCHS-EBS. 78

6.2 AUROC of Deep and Shallow models across four experiments. 79

6.3 Specificity of deep and shallow models across four experiments. 79

6.4 Sensitivity of deep and shallow models across four experiments. . . . 80

6.5 Performance metrics of the Wide and Deep model on the Textionnaire and OCHS-EBS, age and sex and the baseline model on the raw data. 81

6.6 List of most informative pairs of words identified by the attention information algorithm and their normalized information values on the private dataset. Pairs are sorted by the encoder layer index and the head number where Layer 0 is the first encoder. 87

6.7 List of most informative pairs of words identified by the attention information algorithm and their normalized information values in experiment 6. Pairs are sorted by the encoder layer index and the head number where Layer 0 is the first encoder. 93

A.1 Total quality assessment scores of papers. 112

A.2 Sample Size of Datasets used in each study. 114

A.3 List of machine learning algorithms and the papers that used them. 116

E.1 List of most informative pairs of words identified by the attention information algorithm and their normalized information values in experiment 7. Pairs are sorted by the encoder layer index and the head number where Layer 0 is the first encoder. 132

Definitions, and Abbreviations

Definitions

All-Cause Readmission

An admission to an acute care hospital within specified period of discharge from the same or another acute care hospital.

Abbreviations

AI	Artificial intelligence
ED	Emergency Department
NLP	Natural Language Processing
SVM	Support Vector Machine
GBM	Gradient Boosting Machine

Declaration of Academic Achievement

The following is a declaration that the research represented in this thesis was completed by Mr. Sajjad Rashidiani and acknowledges the contributions of Dr. Thomas E. Doyle . Mr. Sajjad Rashidiani contributed to the inception of the study, study design, and was responsible for the data preprocessing, conducting analysis, experiments, and the writing of the manuscript. Dr. Doyle contributed to the inception of the study, study design, computation power needed to run the experiments, review of the manuscript, and has provided guidance and support at all stages of this thesis.

This thesis contains no material that has been submitted or published previously, in whole or in part, for the award of any other academic degree or diploma. Except where otherwise indicated, this thesis is Sajjad Rashidiani's work.

Chapter 1

Introduction

1.1 Motivation

Emergency Department (ED) visits due to mental health reasons across Ontario has demonstrated an increasing pattern from 2006 to 2011 [49]. This 32.5% increase led to listing Mental Health and Addiction, along with Respiratory and Abdominal/Pelvic complaints, as the top 3 common causes of ED visits in Northwest Ontario [49, 78]. A more recent study [30] indicates that this ascending pattern has persisted. The number of mental health or addiction-related ED visits in 2017 experienced an 89.1% hike since 2006. The trend suggests that mental health and addiction could become the most common reason for ED admission [78]. This deteriorating situation highlights the importance of ED visit prevention.

Furthermore, unplanned hospital admissions after an initial admission puts a massive burden on the healthcare systems [11, 122]. In an attempt to limit the costs and increase the quality of care, policy makers worldwide have been interested in reducing the hospital readmission rate in the last few decades [24, 66]. As as result, countries

like the United States, the United Kingdom, and Germany have introduced monetary incentives to encourage efforts toward reducing readmission rates [66]. Since ED visits after an initial outpatient mental health visit are also considered an unplanned hospital admission, this research is also aligned with this worldwide effort to improve care delivery.

Lavergne et al. [69] suggest that more psychiatric outpatient visits can decrease the odds of ED admission. Consequently, a better resource allocation in the mental health care system, which improves the availability of this service to those at high risk of ED visits, could lower the overall rate of ED visits. Identifying high-risk individuals is essential to any effort to improve mental health care delivery. For the reasons mentioned above, the focus of this research is on identifying patients with a high risk of ED visits within six months window using machine learning algorithms.

1.2 Overview of the Field

The Centre for Medicare and Medicaid Services (CMS) defines hospital readmission as "an admission to an acute care hospital within 30 days of discharge from the same or another acute care hospital" [1]. However, many researchers have considered different definitions for hospital readmission in the literature. Some researchers studied hospital readmission on different time intervals such as 7 days [99, 113], 60 days [96, 119], 90 days [54, 56, 70, 84, 91, 101, 121], 180 days [17, 84, 87], 1 year [68, 75, 119],and 2 years [25]. Other researchers considered *all-cause* readmission [22, 77, 127], whereas other studies examined readmission due to a specific disease [17, 29, 35, 79]. Finally, most studies considered any admission to hospital after the discharge, but a few studies focused on specific types of readmission, such as admission to ICU after initial

admission [71, 113].

Considering this broad definition of hospital readmission in the literature, the literature about readmission prediction using machine learning and deep learning algorithms were systematically reviewed to obtain a clear picture of the studies in the field of hospital readmission prediction. A detailed description of this review and its findings is provided in section 2.9. However, in the remainder of this section we discuss the gaps identified in the literature review.

Identified Gaps in the Literature

Gap 1 – Lack of Research on Mental Health Readmission Prediction: Cardiac disease-related, all-cause, and post ICU readmission were the most popular types of readmissions among the surveyed papers. Despite the concerning statistics about increasing number of mental health related ED visits that discussed in section 1.1, only a handful of studies focused on mental health-related readmission prediction. This highlights and important gap in readmission prediction studies.

Gap 2 – Reliance on Medical History Data: A vast majority of studies used medical history data to predict hospital readmission. However, the medical history data is not always available to mental health organizations from the first episode of care. Nonetheless, it is important to identify high risk patients in the early visits. That is because the rate of drop out from mental health treatments is high and an incomplete mental health treatment is a risk factor for adverse outcomes [124].

Gap 3 – Limited Research Using Deep Learning Methods: Although the inclusion and exclusion criterion of the systematic review was determined in a way

to keep the most sophisticated studies (Appendix C) only 12 papers (out of 51 papers) used deep learning algorithms. Since deep learning algorithms demonstrated a superior performance compared to machine learning algorithms in readmission prediction and other medical tasks, it is worthwhile to study application of deep learning algorithms in readmission prediction more extensively.

1.3 Thesis Questions

Based on the mental health team requirements and the identified gaps from the literature review, several research questions were defined. The research questions (RQ#) that are studied in this research are presented below:

RQ1: ***Using only questionnaire data, can a reliable deep learning model be developed to predict if a patient of the mental health clinic (inpatient or outpatient) is deemed high-risk with resulting Emergency Department visit within next 6 months?***

The importance of identifying high risk mental health patients from the early visits is discussed in section 1.1, and section 1.2. Furthermore, utilizing only the questionnaire data that is gathered in the first visit as the basis for classification, addresses the second gap (lack of medical history data) identified in the literature.

RQ2: ***Can pre-trained deep learning models be used on questionnaire data to improve classification performance?***

Questionnaires are frequently used to gather information in many medical and non-medical applications ranging from mental health, pain, and other health-related areas

to business management and politics. With the growing interest in using artificial intelligence in medical applications, many researchers have applied machine learning algorithms to questionnaire data in various contexts[42, 100, 111]. However, the sample size in many medical surveys is very small. For example, in seven systematic reviews of medical questionnaires[3, 28, 43, 48, 90, 106, 118] 128 surveys were studied and the average number of participants was 467 with the largest study population being 4451 participants. These numbers are sufficient for their particular application, but are too small for training a robust deep learning model. Deep learning algorithms outperformed the machine learning models in many of the surveyed studies on readmission prediction tasks. However, this lack of data limits the effectiveness of deep learning models. The mental health questionnaire dataset used in this study also suffers from the small sample size problem. A popular solution to the small size of the datasets is transfer learning from a similar, but larger dataset [110, 135]. This technique enables researchers to apply deep learning algorithms to datasets with small sample sizes. Nevertheless, the data used in the initial training must be similar to the data used in the new task [33]. A small change in the wording of the questions or the options in a fixed response questionnaire might entirely change the answering pattern. Thus a major challenge for transfer learning is to find similar questionnaire datasets in the domain of interest with a large number of records. Consequently, it is important to answer this question in order to gain a better classification performance in the task defined in RQ1.

RQ3: ***Could text representation of questionnaire data be used in classification instead of structured tabular representation?***

It is difficult to find a questionnaire with large sample size which has similar questions and similar patterns in the data to under study questionnaire. Questionnaire data is regularly presented in tabular structured format. With limited samples and data, the author hypothesized that a text representation of the questionnaire results could offer a unique approach to expanding the available data through a machine learning approach call Transfer Learning (section 2.3). In this study, a text representation of the questionnaire data is introduced and verified for its application in classification. This text representation is addressed as *Textionnaire* through out this document.

RQ4: ***How to make our method explainable?***

An undeniable advantage of the machine learning models over the deep learning algorithms is their explainability. When it comes to medical applications, explainability is an essential component to build trust in medical AI. As a result, in this research an approach is studied to explain the outcomes of the deep learning model.

1.4 Evaluation Overview

To answer the research questions listed in section 1.3 (RQ1-RQ4), different methods will be introduced in later chapters of this document. This section introduces the validation approaches utilized to verify the proposed methods.

Since readmission prediction is an imbalanced data problem, Specificity, Sensitivity, and Area Under Receiver Operating Curve (AUROC) were used as performance

metrics to evaluate the classification performance of the proposed methods. Since no other comparable study was found in the literature, different vanilla machine learning algorithms were studied to establish a baseline. The proposed deep learning architecture were compared with this baseline to measure the validity of the response provided to the first research question.

Since second and third research questions are tied to each others, they share the same evaluation processes. In three different experiments the AUROC, Specificity, and Sensitivity of the domain relevant pre-trained deep learning model fine tuned on the Textionnaires are compared with that same metrics for a neural network model (Baseline Model), trained and tested on the tabular questionnaire data. The experiments are designed to compare i) the effectiveness of the proposed method on two different datasets and classification tasks. ii) The effectiveness of proposed method against gathering more training data. iii) The impact of proposed method on the on the generalizability of classification. A detailed description of theses experiments is provided in chapter 5. Finally, the impact of using knowledge driven features (OCHS-EBS) in prediction was tested by comparing the performance of Wide and Deep model with the Baseline model and the Deep architecture.

Finally, four experiments were designed to verify the relevance of identified pairs and the respective calculated attention information.

1.5 Organization and Scope

In the following chapters a thorough review of required background knowledge, review of literature, description of utilized data, machine learning algorithms used in this study, and the methodology to test those algorithms to answer the research questions,

and the result of this study is provided. Finally the results are discussed and the finding of this study is concluded.The order of information in this document is as follows:

- *Chapter 2* provides the background knowledge required to understand content of this document. It also includes the details of literature review conducted as a part of this study.
- *Chapter 3* covers the required information about the datasets utilized in this study.
- *Chapter 4* describes the machine learning algorithms used in this project and the explainability component designed for the proposed method.
- *Chapter 5* explains the experiments conducted to investigate the effectiveness of the proposed methods and answer the identified research questions.
- *Chapter 6* lists the results of the experiments introduced in chapter 5.
- *Chapter 7* discusses how the results answer the research questions of this study and states the findings of this study. Furthermore this chapter describes the limitations of this study and propose new investigations for future steps.
- *Chapter 8* summarizes the information of this document to conclude the study.

Chapter 2

Literature Review

In this chapter, the terminology and techniques used in this thesis are explained, covering:

1. Feature Selection
2. Data Imbalanced Problem
3. Transfer Learning
4. Softmax
5. Transformer Model and BERT
6. Decision Tree Classifier
7. Medical Nomenclature
8. Evaluation Metrics

Finally, this section presents a systematic review of the literature on readmission prediction and analysis of the gaps in the literature.

2.1 Feature Selection

In a dataset with N features or attributes, feature selection is the act of selecting a subset of N' features ($N' < N$) in an attempt to improve the model performance [36, 65, 73, 85]. In other words, in the feature selection process, the irrelevant and redundant features are removed, and the machine learning model is trained on the most relevant features [64, 73, 85]. Feature selection is utilized to reduce overfitting, improve model performance and reduce training time. A *filter* is a feature selection method through which the relationship between the inputs and targets of the classification is determined outside of the final model [27, 67]. Features with the strongest relationships are included in the final subset and used to train and test the final model. As an example of a common filter method, Gradient Boosting Machine (GBM) is used as a feature filter. In this research we apply GBM to limit the length of Textionnaires to fit into the limitations of the gold standard transformer model (BERT) discussed in section 2.5 and subsection 4.2.1. In subsection 2.1.1, the machine learning algorithm that is used as a filter for feature selection is explained.

2.1.1 Gradient Boosting Machine

Gradient Boosting Machine (GBM) is a machine learning algorithm that generates its outcomes by aggregating the outcome of several base-learners trained on the same set of features [86]. In this algorithm, every base learner is trained to estimate the gradient of the classification made by the previous base learners. In this way a linear combination of these base learners leads to a high classification performance as they are compensating the remaining gradient. Zhang et al. [131] explains the process of error reduction using multiple base-learners with a simple example demonstrated

in Figure 2.1. The amount of increase in the quality of split after every split on each feature is calculated using Friedman Mean Squared Error, Mean Squared Error, or Squared Error. The amount of improvement after each split is calculated and normalized across all the calculated values. These normalized values are considered the feature importance[5]. A popular implementation of GBM is available in Scikit-Learn library [94] under the *"GradientBoostingClassifier"* class. In this implementation, the *"DecisionTreeRegressor"* class is the base learner and the Friedman MSE [5, 86] method is used to measure the quality of split and as a result the feature importance.

2.2 Data Imbalanced Problem

A dataset is imbalanced when the proportion of the number of samples in one or more classes is very small and majority of samples belong to another class. In this situation, machine learning algorithms are prone to bias toward the majority class. For example, Figure 2.2a depicts the distribution of a one-dimensional balanced dataset consisting of 2 classes where half of the samples belong to each class. Although the distribution of classes coincide, each distribution is still distinguishable and the number of samples in each is approximately equal. Figure 2.2b shows the distribution of another dataset, except one of its classes accounts of 95% of samples. Although the orange samples in both examples come from a normal distribution with the same characteristics and the source distribution blue samples is also the same in both cases, in Figure 2.2b the distinction between the two classes in difficult. In such a case, the easiest way to minimize the loss function is to classify all samples as the majority class. As a result, a model can easily ignore the minority class, estimate all samples with majority class

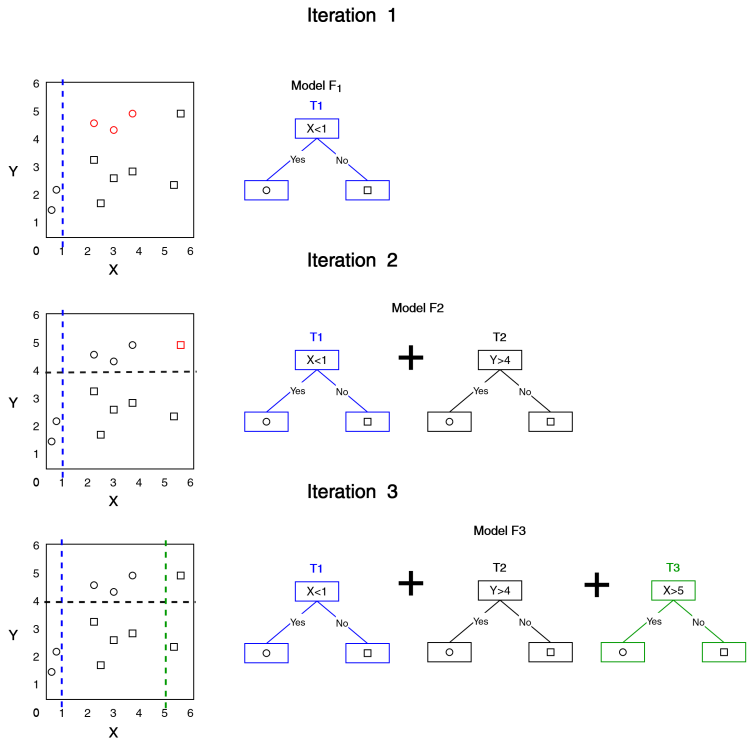

Figure 2.1: In the first iteration a base-learner (A Decision Tree in this case) splits the data to two groups, one includes only circles (circle class) the other one includes only squares (square class). The missclassified samples are displayed in red. In the second iteration an other base-learner splits the previous square class that has 3 miss-classifications in two new groups reducing the number of miss-classifications to only 1 (one square in new circle group). Finally in the third iteration the last base-learner splits the last group with miss-classification to two groups with perfect classification.

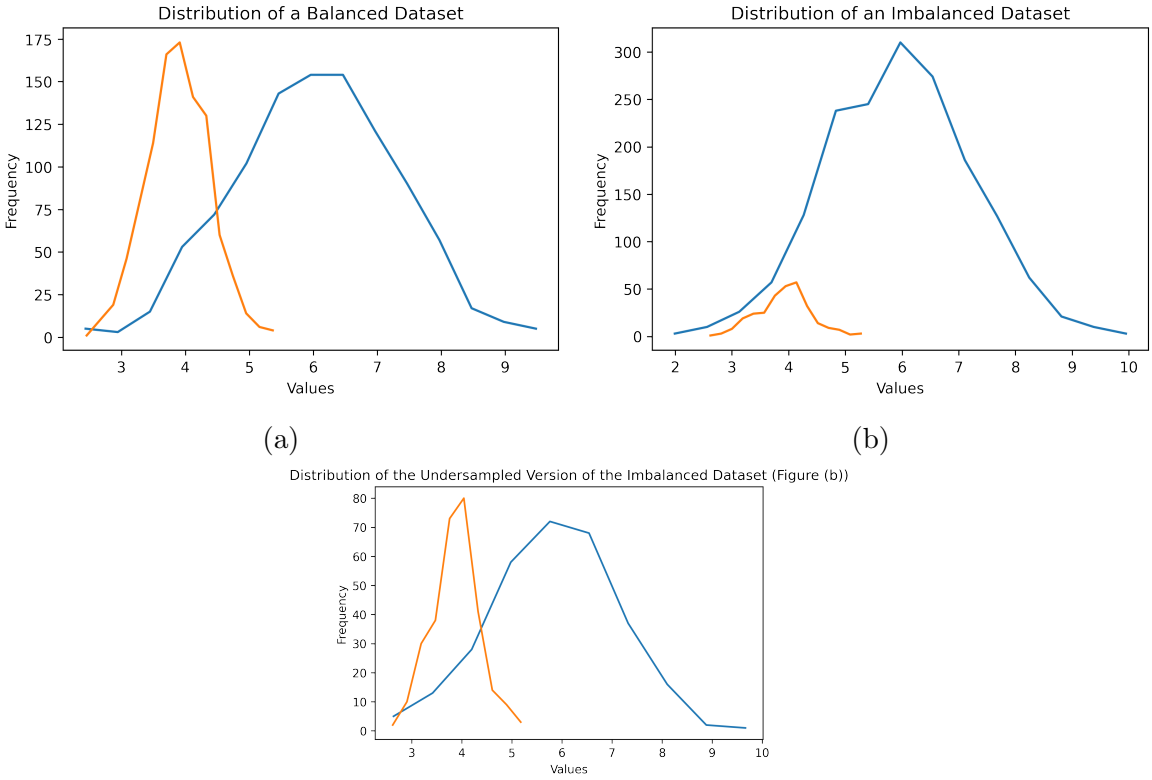

Figure 2.2: Distribution of a) a balanced, b) an imbalanced one-dimensional dataset, and c) the randomly undersampled version of (b).

distribution, and without careful inspection appear to achieve a high performance due to the calculated accuracy (95%). There are methods in literature to address and compensate for imbalance [29, 61, 81]. For this study, a random undersampling is used to counter the data imbalanced situation. Figure 2.2c depicts the distribution of dataset where the blue class is randomly undersampled.

2.2.1 Random Undersampling

In random undersampling, the majority class is randomly sampled to create a subset with the same number of samples as the minority class. This subset is used along

with the minority class for training the machine learning model.

2.3 Transfer Learning

Transfer Learning is a method utilized to improve the performance of machine learning algorithms using the knowledge learned from other similar datasets. Pan and Yang [89] defines Transfer Learning based on several components: i) a domain D consisting of a feature set χ, and a probability distribution $P(X)$, where X belongs to χ, ii) a task T including a label space y, which are the true labels, and a prediction function $f(x)$ which predicts the label for a sample X. For a target task T_T, transfer learning aims to improve the learning of the target predictive function $f_T(x)$ from the D_T using the knowledge learned in a sources task T_S from a source domain D_S [89]. In other words, when a model is trained on source data to predict source labels, the relation between the source domain and source target is transferable to a new model that is to be trained on similar data for a similar task. This knowledge accelerates the learning process of the new model and enables the model to learn complex relationships between the target domain and associated labels.[33]. Figure 2.3 depicts a block diagram that explains the concept of transfer learning. This technique enables researchers to train deep learning models on datasets with a small number of samples and gain higher performance while avoiding overfitting. For example, in this study, a BERT model was trained on a social media text dataset (D_S) to predict depression (T_S). We utilized the pre-trained Encoders and tokenizer of that architecture and used them to predict 6-month ED visits of mental health patients (T_T) from Textionnaire (D_T) generated from a mental health-related questionnaire.

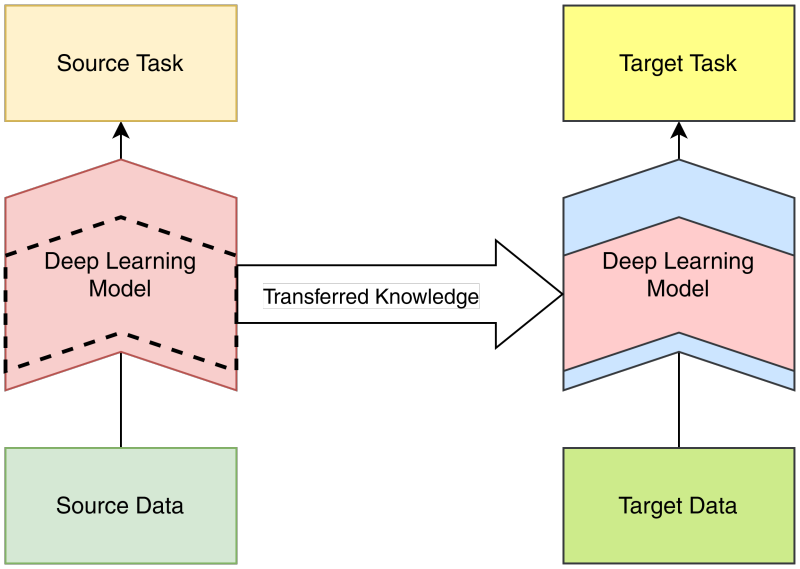

Figure 2.3: How Transfer Learning Works. The similarity between the colours of Source and Target data blocks, and Source and Target Task blocks indicates the similarity between them.

2.4 Softmax

The term *Softmax* refers to a soft argmax function or a normalized exponential function that calculates the probability distribution of K possible outcomes given a vector of K real numbers [15, 51]. The Softmax function is normally used as the activation function in multi-class classifications. In connection to Transformers, as defined in section 2.5, it is also utilized in the scaled dot product attention mechanism. Softmax of vector $\vec{X}$ is a vector with the same dimensions whose elements are calculated via Equation 2.4.1[16].

$$\sigma(\vec{X})_i = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}} \tag{2.4.1}$$

where x_i is the i^{th} element of $\vec{X}$ and K is the size of the vector.

2.5 Transformer Models and BERT

In 2017, Vaswani et al. [116] introduced a new architecture for machine translation called Transformer. The authors claimed that this architecture can be utilized in other Natural Language Processing (NLP) tasks and tested it on an English constituency parsing task. Transformer architecture consists of input and output embedding, a stack of encoders, and a stack of decoders. Figure 2.4 depicts the transformer architecture introduced in [116]. Researchers in [39] utilized only the input embedding and

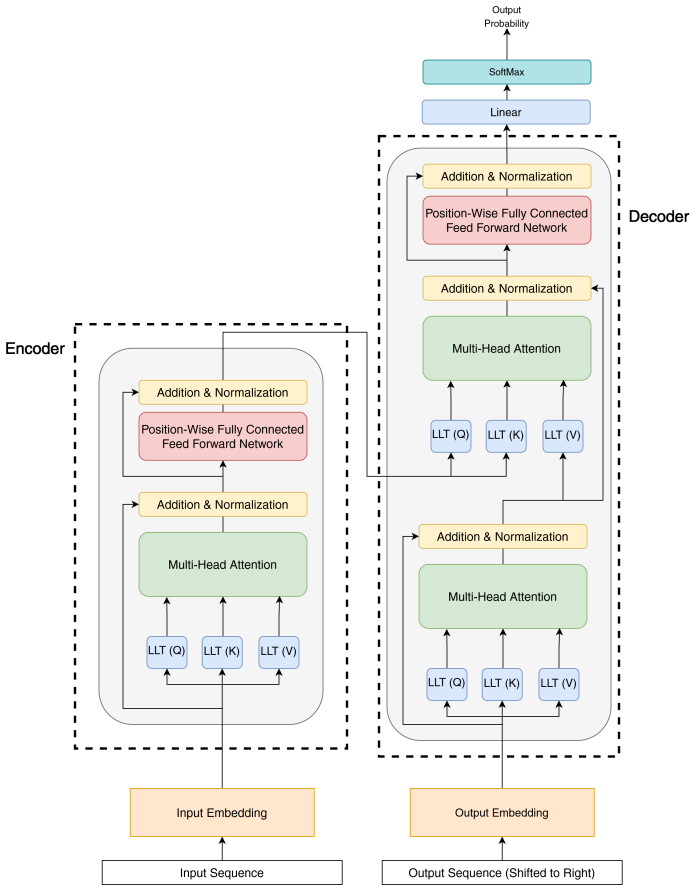

Figure 2.4: Transformer Architecture

stack of encoders from the Transformer model and proposed a Bidirectional Encoder

Representation of Transformer (BERT). This architecture is designed to pre-train from unlabelled data and be adopted for a wide range of NLP task by adding one task-relevant layer to BERT. In the BERT architecture the text is fed to a Tokenizer unit which converts the text to vector of numbers. This Tokenizer is demonstrated as the Input Embedding unit in Figure 2.4. This vector of embeddings is fed to the encoder layer and processed through the underlying attention heads and positional neural networks of the encoder blocks to generate machine driven features. In classification tasks, these features are then fed into several layers of neural network to perform the final classification. As in this study only BERT architecture is used, in the following sub-sections, different parts of BERT are explained in detail and the other components of the Transformer architecture are skipped.

2.5.1 Embedding

Devlin et al. [39] present the embedding used in BERT as a three-step embedding:

1. The input text is converted to a vector of tokens using WordPiece embedding[128]. The WordPiece embedding algorithm starts with the letters in the training datasets and puts together those letters that have the most cooccurrence probability and generates sub-words. Then, it considers the generated sub-words as one letter and repeats the previous steps. The process of generating new sub-words or expanding the existing sub-words continues until it reaches a certain vocabulary size. The first token of every input text is assigned to a classification-specific token ([CLS]) and sentences are separated using a separation token ([SEP]). If the length of the input sequence is smaller than the maximum length of input (a hyper-parameter), pad tokens ([PAD]) are added

to the end of the sequence to generate a sequence of size maximum length.

2. The tokens are replaced with a *Learned Token Embedding* also known as IDs.

3. A *Learned Segment Embedding* is added to tokens to indicate whether a token belongs to sentence A or sentence B.

4. A *Positional Embedding* is summed with the previous token embedding to indicate the position of the token within the sequence.Vaswani et al. [116] uses sine and cosine waves with various wave length in every dimension of the positional embedding. In other words, a position embedding is a matrix of size $Sequence\ Length \times Embedding\ Dimension$. Every column in this matrix contains a Sine (Even columns), or Cosine (Odd columns) wave calculated using Equation 2.5.1, and Equation 2.5.2, respectively.

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}}) \tag{2.5.1}$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}}) \tag{2.5.2}$$

where pos is the position of word in the sequence, i is the column index and d_{model} is the dimension of embedding.

At the end of embedding layer, every token is represented with a vector of size d_{model} that contains the token word embedding, sentence assignment, and information about the position of the token in the sequence. Figure 2.5 depicts these steps for a simple example.

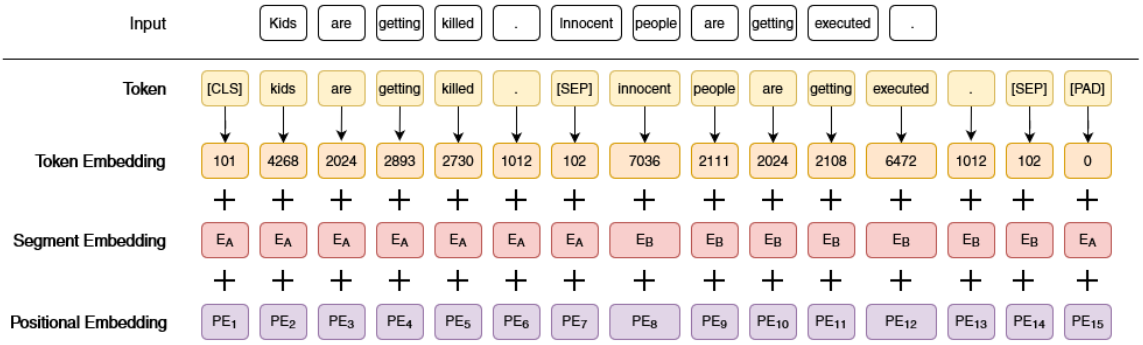

Figure 2.5: Embedding of a BERT model with a maximum length of 15.

2.5.2 Encoder

An Encoder is one of the building blocks of the BERT architecture which is designed for learning and extracting the information within a body of text. A BERT architecture may have one or more encoders cascaded where all of these encoders share the same hyper parameters. As every encoder is fed by the output of the previous encoder, the features extracted by deeper encoders are much more complex than the features extracted by the previous layers.

Figure 2.6 depicts an encoder block of BERT model. An encoder block has a multi-head attention sub-layer and a feed forward sub-layer which is a position-wise fully connected feed-forward neural network. The output of every sub-layer is summed with its input through a residual connection[52] and normalized using layer normalization[10, 116]. In the following, Multi-Head Attention and Feed Forwards sub layer are explained.

Multi-Head Attention

The *attention mechanism* is used to enable the model to perceive the relationship between words. For example, in the example provided in Figure 2.5 the words “are”

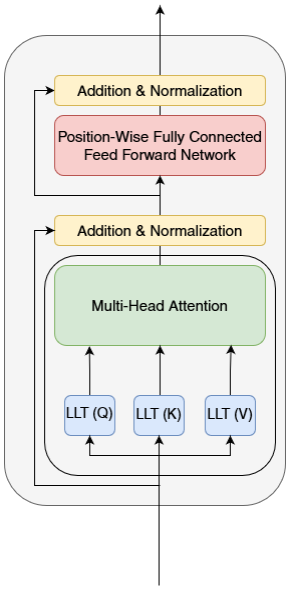

Figure 2.6: An Encoder layer of BERT model. Note that LLT stands for Learned Linear Transformation where LLT(Q) generates Query, LLT(K) generates Key, and LLT(V) creates the Value matrix.

and "getting" are used in both sentences. The model needs to understand that each one of them is used to describe the status of which subject. Attention heads enable the model to understand these relationships [6]. Multiple attention heads allow model to process more relationships between words in an input sequence. Attention weights are calculated and applied to the input sequence of embeddings using Equation 2.5.4

$$\text{Attention}(Q, K, V) = \text{softmax}\,(\frac{QK^T}{\sqrt{d_k}})V \tag{2.5.3}$$

where Key (K), Query (Q), and Value (V) are three linearly transformed copies of the input sequence with d_k, d_k, and d_v dimensions, respectively. This is also known as scaled dot product attention [116]. In multi-head attention mechanism, several linear transformations are utilized to create different queries, keys and values to generate

apply different sets of attention weights using Equation 2.5.3. This architecture enables the model to attend to various sub-spaces in every position in the sequence using only one attention head. Figure 2.7 depicts the proposed architecture for multi-head attention sub-layer. These parallel scaled dot-product attentions are also known as attention heads. The outputs of these heads are concatenated and projected to create the output of multi-head attention sub-layer [116] as it is described in Equation 2.5.4.

$$\text{MultiHead}(Q, K, V) = \text{Concat}\,(head_1, ..., head_h)W^O \tag{2.5.4}$$

where

$$\text{head}_\text{i}(Q, K, V) = \text{Attention}\,(QW_i^Q, KW_i^K, VW_i^V) \tag{2.5.5}$$

Where projections are learned parameters $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, and $W_i^O \in \mathbb{R}^{hd_{model} \times d_v}$, and h is the number of heads.

fix dimensions in the output ot attention head

Position-Wise Fully Connected Feed-Forward Network

The position-wise fully connected feed-forward network is a two-layer fully connected neural network with ReLU activation function for the hidden layer[116]. The same feed forward network is applied separately to every token in the input sequence. The hidden layer of this network is larger than the d_{model}, but the output layer has d_{model} neurons. For example, in BERT, $d_{model} = 512$ while $d_{hiddenlayer} = 2048$ [116].

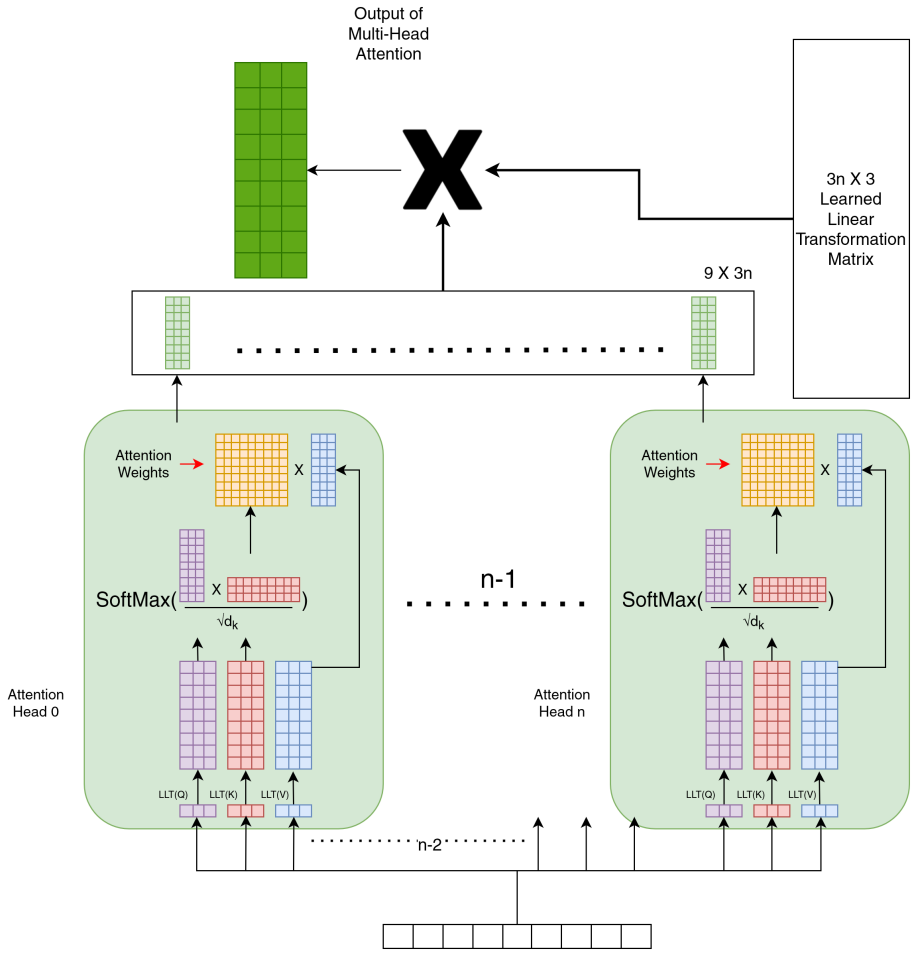

Figure 2.7: Multi-Head Attention architecture consist of several attention heads

2.5.3 Classifier

Since BERT is designed as a pretrained model for a wide variety of NLP tasks, a classification layer is not a part of the architecture introduced in [39]. However, as in this study, when BERT is utilized for a binary classification task the task-relevant layer must be a classifier layer. The classifier is applied to the "[CLS]" token of the last encoder layer of the BERT architecture to perform a classification task.

2.6 Decision Tree Classifier

Decision trees are a popular type of machine learning algorithm for performing classification based on multiple features, or for prediction of a target variable. This algorithm splits a population into branches that construct an inverted tree with a root node, internal nodes, and leaf nodes. This is an efficient non-parametric algorithm suitable for dealing with large, complicated datasets without imposing a complicated parametric structure [107].

To determine the most discriminating features in each step and to determine whether or not the splitting must continue, a decision tree must apply a criterion. There are several criterion measures, such as Gini Impurity, Log Loss, and Information Gain. The Information Gain measures the expected reduction in Shannon entropy caused by dividing the samples according to a feature [93]. Split Information Gain is calculated using Equation 2.6.1.

$$\text{SplitInfo}(A) = -\sum_{j=1}^{v} \frac{|Dj|}{|D|} \times \log_2\left(\frac{|Dj|}{|D|}\right) \tag{2.6.1}$$

2.7 Medical Nomenclature

Some mental health related terms are used frequently in this document. In the following subsections these terms are described in details.

2.7.1 Questionnaire

In this document "Questionnaire" refers to Patient Reported Outcome Measure (PROM) which is commonly used to determine mental health treatment plans and to monitor

the progress and clinical outcomes [63]. Several questionnaire are used in this study and a detailed description of these questionnaires is available in chapter 3.

2.7.2 Outpatient Clinic

Outpatient Clinic or Outpatient Department is a part of a hospital in which diagnosis and care are provided to those patients who do not need an overnight stay in the hospital [40].

2.7.3 Inpatient

An inpatient patient is admitted to a health care institution for receiving a treatment that requires at least one overnight stay [41].

2.7.4 Emergency Department

Emergency Department is a part of hospital in which patients with severe or urgent conditions are treated [37].

2.7.5 Mental Health Scores

In this study, mental health scores are used as hand crafted features extracted from standard questionnaires. These scores are calculated using Ontario Child Health Study Emotional Behavioural Scales(OCHS-EBS) introduced in [20, 47]. OCHS-EBS is a 52-item checklist reported by parent and youth which provides a trustworthy dimensional and categorical measurement of 7 DSM-5 disorders (attention deficit hyperactivity, oppositional defiant, conduct, generalized anxiety, separation anxiety,

major depressive and social phobia disorders). More details about calculation of OCHS-EBS is provided here.

2.8 Evaluation Metrics

Readmission or ED visit are rare events compared to total number of patients. In such cases, researchers utilize Area Under Receiver Operating Curve (AUROC), Specificity, and Sensitivity as performance metrics. In this study, these scores are calculated using Python and Sci-Kit Learn library [94]. In the following subsections the performance metrics utilized in this study are explained.

2.8.1 Area Under Receiver Operating Characteristic Curve

A Receiver Operating Characteristic (ROC) Curve plots the True Positive Rate (TPR) of a binary classifier against its False Positive Rate (FPR) for a range discrimination threshold varied from 1 to 0. Area under ROC curve (AUROC) is introduced as a performance metric for classification[21]. An ideal binary classifier has a 100% TPR and 0% FPR and in return an AUROC of 1, while a random classifier has an AUROC of 0.5.

2.8.2 Sensitivity

Sensitivity of the model measures the ability of the model for predicting the positive class. In other words, it shows how well the model does in identifying the patients at

risk of ED visit. Sensitivity is calculated using Equation 2.8.1.

$$Sensitivity = \frac{TruePositive}{TruePositive + FalseNegative} \tag{2.8.1}$$

2.8.3 Specificity

Specificity measures the ability of the model for identifying the negative class. In other words, this metric measures how well it identify patients at risk of ED visits. Specificity is calculated using Equation 2.8.2.

$$Specificity = \frac{TrueNegative}{TrueNegative + FalsePositive} \tag{2.8.2}$$

This metric is normally considered along with Sensitivity because a high specificity and low sensitivity indicates a bias towards the negative class while a high sensitivity and low specificity shows a bias towards the positive class.

2.9 Systematic Review

The focus of this review was to identify common machine learning and deep learning techniques used in predicting hospital readmission from the last ten years of research literature. This section reports on the methods and finding of the systematic review of literature that has been done as a part of this study.

To the best of our knowledge, there are only six systematic reviews in the literature on the application of artificial intelligence in readmission prediction [8, 31, 55, 76, 105, 133] in the last decade. Four of those studies, conducted by [8, 55, 76, 133], have provided good insight into the types of data, models, and performance of models

used in this area. Notable limitations among those studies: [55] focused only on USA patients, [76] was limited to 28 to 30-day readmission after 2014, and [8, 133] did not include studies after 2018 and 2016, respectively. The other two reviews, conducted by [31, 105], compared statistical methods such as LACE with machine learning algorithms. [105] provided a clear picture about the performance of machine learning models compared to many conventional statistical patient outcome measures focused on cardiac disease, but unfortunately excluded the studies that did not compare ML and statistical models. Christodoulou et al. [31] focused on comparing traditional ML methods with logistic regression, but excluded new prediction methods as out of its scope.

None of these studies sketched a comprehensive picture of the utilized datasets, pre-processing tools and steps, sophisticated machine learning models, and predictive features frequently used in readmission prediction. Our study, not only encompasses all of those, but also provides information on the sample sizes of the datasets used in this field as well as positive case rates among the datasets. It makes it an excellent starting point for new researchers in this field.

2.9.1 Review Methodology

Following the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement, published in 2020 [82], we conducted a search in Pub-Med, Web of Science, IEEE Xplore databases. Studies between January 2010 and February 2022 whose full text was published in English were included in this search. The search term consists of “machine learning,” “deep learning,” “artificial intelligence,” “readmission,” “Patient Outcome,” “Unscheduled Admission,” and Rehospitalization. No

limitation has been applied to the cause of initial admission in the search process. The full search string used in each database is available in Appendix A.

2.9.2 Inclusion and Exclusion Criterion

Search results exported from the databases were imported into the COVIDENCE platform for duplicate removal and title and abstract screening. Two independent reviewers screened the title and abstract of included articles to determine whether they contained research relevant to readmission prediction using artificial intelligence techniques. A third independent reviewer resolved the conflicts in the screening phase. One hundred seventy-eight articles were selected for the full-text review. Articles meeting all the following inclusion criteria were included in the final step:

1. study reported a novel AI approach for readmission prediction and its performance;
2. study was published in a peer-reviewed publication; and
3. article was published in the English Language.

Studies were excluded based on the following:

1. study did not report an original contribution to the applications of ML algorithms in readmission prediction (e.g., review papers, or descriptive reports on readmission reduction systems in hospitals without mentioning the AI model used and their performance);
2. study did not explicitly specify the AI model used in their experiments;

3. study did not explicitly quantify the performance of their model using a standard performance measure (e.g., only reported that a model outperformed the others);

4. study applied a set of machine learning and statistical models on a new data set with no preprocessing techniques, feature selection, or model design;

5. machine learning algorithms were applied to identify the risk factors, not readmission prediction; and

6. the full text of the article was not available in English.

2.9.3 Data Extraction

The information gathered in this review can be categorized in two groups with each consisting of multiple sub-groups. The first category is focused on the domain-related data: i) the dataset used, ii) cause of initial admission, iii) sample sizes, iv) rates of readmission, and v) outcomes of interest. The second category is technical and implementation-related information: i) features and predictors, ii) feature selection methods, iii) imputation methods, iv) data balancing techniques, v) machine learning and deep learning algorithms used in predictions, and vi) the performance metrics used to assess the prediction performance. This information was compiled by Sajjad Rashidiani from the identified papers and then verified by another researcher (Asif Khan). All of the extracted data is presented as several tables linked by paper IDs and is available in the Appendix A. The items are reported in this document following the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) [114] and **CH**ecklist for critical **A**ppraisal and data

extraction for systematic **R**eviews of prediction **M**odelling **S**tudies (CHARMS) [83] checklist.

2.9.4 Quality Assessment

The quality of research included in this study have been assessed using a rubric defined by the Sajjad Rashidiani (See Appendix A). This rubric includes 14 questions that assess quality and information richness of the papers and the quality of AI algorithm implementation in the eligible papers. The quality assessment rubric evaluates if the paper reported i) data source, ii) patient settings, iii) patients' characteristics, iv) outcome of interest v) distribution of studied outcome in the subjects, vi) features or predictors used in AI algorithm, vii) the AI algorithms(s), and viii) the hyper parameters of the AI algorithms listed in the paper to assess the information richness of the paper from a reproduciblity point of view. The rubric also assesses if the study incorporated i) knowledge-driven feature selection, ii) data-driven feature selection, iii) imputation technique, iv) data balancing method, v) validation method, or vi) proper performance metric for imbalanced cases such as readmission prediction to measure the quality the implementation of AI algorithm in the eligible papers. By combining the quality of report assessment and the quality assessment of AI algorithm implementation this rubric provides a good insight into overall quality of the papers.

2.9.5 Data Synthesis and Analysis

Firstly, eligible papers ([7, 9, 12, 13, 14, 17, 18, 22, 23, 25, 29, 35, 38, 44, 45, 50, 54, 56, 58, 58, 61, 68, 70, 71, 72, 75, 77, 79, 80, 81, 84, 87, 88, 91, 92, 98, 99, 101, 103, 104, 108, 109, 113, 119, 120, 121, 123, 127, 129, 130, 132, 134**?**])were reviewed

and the characteristics of interest listed in subsection 2.9.3 were extracted. Secondly, a qualitative analysis was performed to find unique values in each characteristic. Then, papers that included those unique values were listed in front of each value providing a clustered view of the literature based on each one of the 11 different types of information listed in subsection 2.9.6. If a paper used multiple items in each characteristics its ID is listed in front of all of those items. For example a paper might use several Machine Learning algorithm or datasets. In such cases, its ID will appear in front of all the algorithms studied in that paper and all the datasets used for the research in the AI algorithm and data source tables.

The Area Under Receiver Operating Characteristic Curve (AUROC) of the best performing method in each paper was reported as it was the most common metric used in the literature to report the AI models performance. If a study reported on training and validation AUROC the validation AUROC was reported in our review and if no training AUROC were reported in the paper the listed AUROC values were considered as validation metrics. Other types of performance metrics reported in some studies were also recorded and listed in a separate spread shit available in the supplementary materials.

2.9.6 Results

Initially, 518 papers were included in this scoping review. These papers were uploaded in COVIDENCE online software and 97 duplicate were identified and removed from the cohort. Then, title and abstracts of 421 studies were screened and 213 irrelevant papers were removed from the batch. The full-text of 208 papers were reviewed. 51 papers met the defined inclusion criteria and included in this scoping review. This

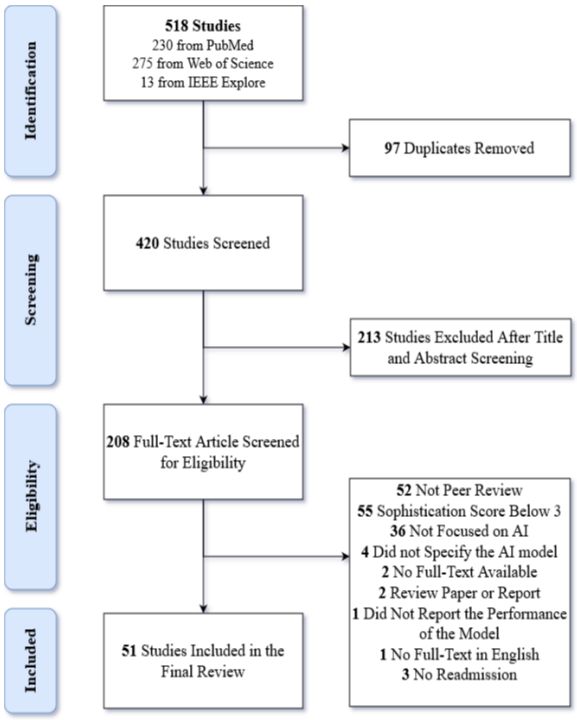

Figure 2.8: PRISMA chart of the review

process has been done by two researcher (Sajjad Rashidiani, Asif Khan) independently and the conflicts were resolved by consensus of the reviewers. Figure 2.8 depicts the PRISMA chart for this study.

Data Sources

A vast majority of studies (21) have used private datasets, gathered by the respective research team for their specific research. These private datasets include Electronic Health Records (EHR) of hospitals, medical centers, and other medical research institutes. In Qian et al. [96], the researchers gathered smartphone data for estimating 2-month readmission risk for patients undergoing surgery as part of their cancer treatment. None of these 21 studies have made their dataset publicly accessible. The

second most popular dataset for readmission prediction studies is MIMIC-III [60]. Table 2.1 provides ranked and detailed information about the data sources used in readmission prediction and the papers used each dataset.

Data Source	**Papers**
Private	Wolff et al. [127], Ashfaq et al. [9], Hung et al. [56], Golas et al. [50], Zhang et al. [132], Rodriguez et al. [99], Matheny et al. [77], Boag et al. [17], Brom et al. [23], Wang et al. [119], Madrid-García et al. [75], Cearns et al. [25], Desautels et al. [38], Xiao et al. [129], Qian et al. [96], Nguyen et al. [87], Park et al. [92],Miswan et al. [80], Baechle et al. [12], Landicho et al. [68]
MIMIC-III	Barbieri et al. [14],Lin et al. [71],Brom et al. [22],Zhang et al. [130],Wang et al. [120],Du et al. [44],Du et al. [45]
Cerner HealthFacts EMR database	Reddy and Delen [98],Shang et al. [104],Welchowski and Schmid [123],Miswan et al. [80]
National Readmission Dataset	Bolourani et al. [18],Li et al. [70],Allam et al. [7]
Sutter Community Connect's EpicCare	Sarijaloo et al. [101],Jamei et al. [58],Park et al. [91]
Geisinger Health System	Min et al. [79],Darabi et al. [35]
Northwestern Medicine Enterprise Data Warehouse (NMEDW)	Lineback et al. [72],Barber et al. [13]
Hospital Cost and Utilization Project (HCUP) State inpatient database	Symum and Zayas-Castro [109],Allam et al. [7]
Diabetic	Du et al. [44],Du et al. [45]
all-cause dataset	Du et al. [44],Du et al. [45]
National Health Insurance Research Database	Chi et al. [29]

American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) database	Kalagara et al. [61]
Tele-HF	Mortazavi et al. [84]
the Big Data Center of Taipei Veterans General Hospitals	Ou et al. [88]
AmsterdamUMCdb	Thoral et al. [113]
Mount Sinai Data Warehouse	Shameer et al. [103]
IBM® MarketScan® Commercial and Medicare Supplement Databases	Wang et al. [121]
WA Hospital Morbidity Data Collection (WAHMDC)	Zhou et al. [134]
Lace-score	Du et al. [45],
Readmission Analysis	Du et al. [45]
T-carer	Du et al. [45]
North American Consortium of the Study of End-Stage Liver Disease (NACSELD) cohort	Hu et al. [54]
Not Specified	Mohammadi et al. [81]

Table 2.1: Data sources used in the papers reviewed in this study and the papers used those data sources

Samples Sizes

Studies have used datasets with a wide range of sample sizes ranging from 49 samples in [96] to 335,815 samples in [58]. Out of 60 datasets (not necessarily from unique sources of data) utilized in these studies, 27 datasets had less than 5000 samples, and 10 datasets had more than 5000 samples and less than 10000 samples leading to a median value of 6581. Sample sizes of the datasets used in different papers is available in the Appendix A.

Readmission Rates

The median value of the readmission rates in the datasets used in studied papers is 13% indicating the imbalanced nature of the readmission prediction problem. Figure 2.9 depicts the distribution of the readmission rates and more information is available in the Appendix A.

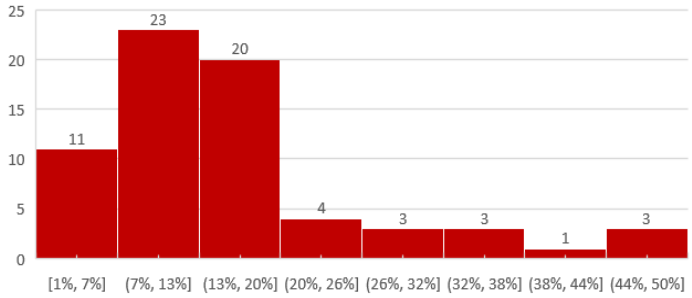

Figure 2.9: Distribution of readmission rates in different datasets.

Imbalance Data Handling

Although more than 80% of the datasets used in the reviewed papers were severely imbalanced (<20% positive cases), 25 studies did not use any method to compensate the imbalance rate of positive an negative cases in datasets. Table 2.2 lists the papers that used each of the imbalance data handling methods.

Imbalance Data Handling Method	**Papers**
Class Weighting	Chi et al. [29],Barbieri et al. [14],Ashfaq et al. [9],Mortazavi et al. [84],Wang et al. [119],Cearns et al. [25],Desautels et al. [38],Landicho et al. [68]
SMOTE	Kalagara et al. [61],Wolff et al. [127],Reddy and Delen [98],Shang et al. [104],Hung et al. [56],Symum and Zayas-Castro [109],Madrid-García et al. [75]
Random Undersampling	Mohammadi et al. [81],Zhou et al. [134],Nguyen et al. [87],Park et al. [92],Miswan et al. [80]
Random Upsampling	Chi et al. [29],Darabi et al. [35],Li et al. [70],Hung et al. [56]
Nearmiss Undersampling	Bolourani et al. [18]
KNIME Downsampling	Shang et al. [104]
SpreadSubsample	Hung et al. [56]
ClassBalancer	Hung et al. [56]
NCR Undersampling	Zhang et al. [132]
CIHL (Learns a weight for classes and a weight for every type of data)	Du et al. [44]
Graph based class imbalanced learning	Du et al. [45]

Table 2.2: Papers stratified by the imbalanced data handling methods.

Cause of Initial Admission

Predicting readmission after a cardiac treatment and the all-cause readmission prediction are the most frequent types of studies. Among the reviewed papers 13 were focused on cardiac disease patients and the same number of studies are focused on all-cause readmission prediction. Predicting readmission after receiving treatment for other types of disease is significantly less popular among the researchers in this area.

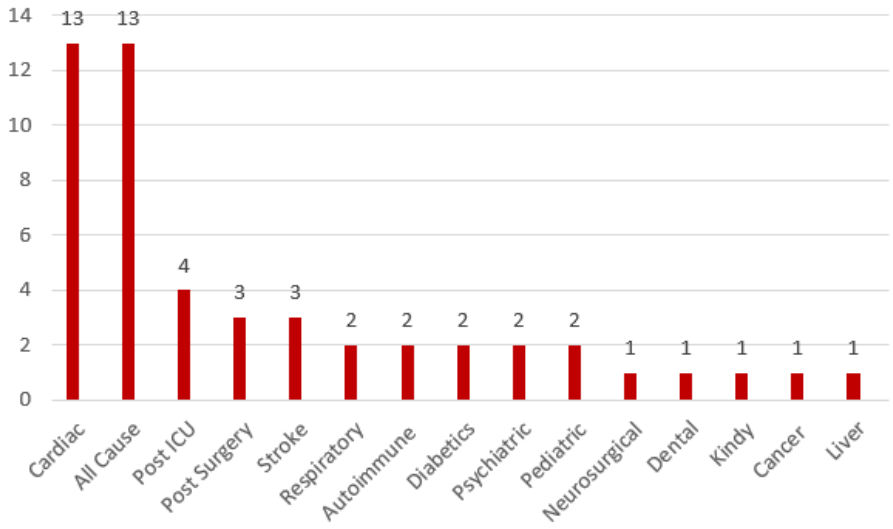

Figure 2.10: Number of papers focused on predicting readmission after different types of initial admission.

Figure 2.10 demonstrates the distribution of studies based on the cause of initial admission.

Outcomes of Interest

Prediction of 30-day all-cause hospital readmission is the most studied outcome among the papers. 31 papers addressed this specific out come while readmission prediction in an open ended window as the second most popular outcome of interest was studied by only 5 papers. Figure 2.12 demonstrates the number of studies that focused on each outcome of interest. Lists of papers studied each outcome of interest are available in Appendix A.

Predictors

In this literature review, predictors used in more than 10 papers were identified as a specific category and predictors used in 10 or less were listed under the "Others" category. Thus the 11 categories identified are: Demographics, Socioeconomic,

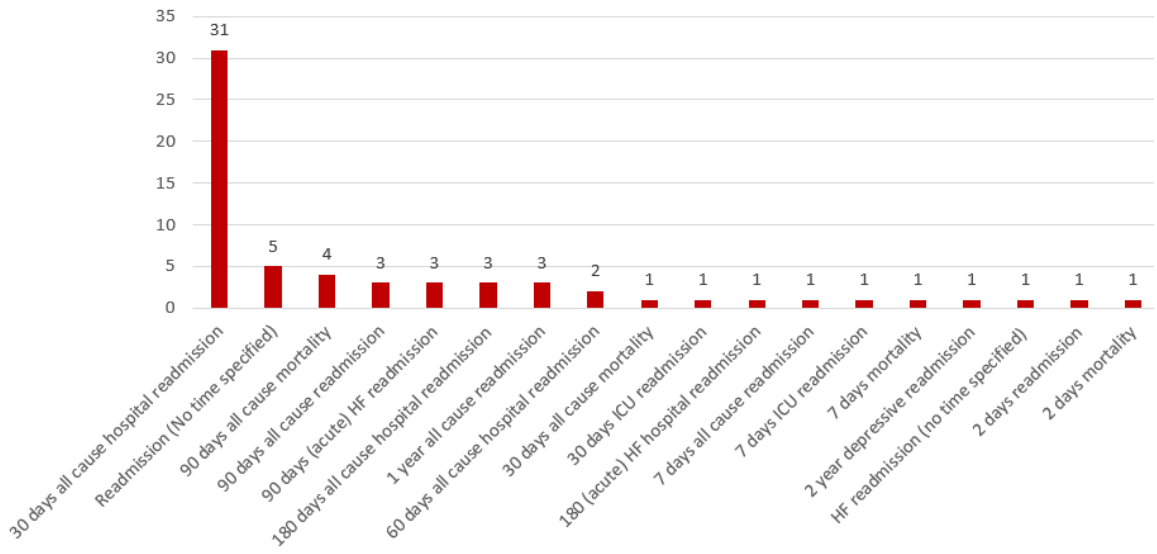

Figure 2.11: Number of studies focused on different outcomes of interest.

Medical History, Length of Stay, Diagnosis, Comorbidities, Medication, Procedure or Treatment, Vital Signs, Lab Results, and Others. Features listed under the "Others" category and the details about the predictors used in each paper can be found in Appendix A.

Feature Selection

From 51 papers reviewed in this study, 27 studies did not use any feature selection methods. 18 different methods used for feature selection in the papers. Some papers used several feature selection methods to select categorical and numerical variables or to test the effect of feature selection methods on the performance of classification. Table 2.3 lists the papers that used these feature selection methods in ranked order.

Feature Selection Technique	**Papers**
T-test	Bolourani et al. [18],Darabi et al. [35],Kalagara et al. [61],Li et al. [70],Sarijaloo et al. [101],Brom et al. [22],Hu et al. [54],Park et al. [91]

Chi-square	Bolourani et al. [18],Li et al. [70],Sarijaloo et al. [101],Symum and Zayas-Castro [109],Brom et al. [22],Hu et al. [54],Park et al. [91]
Linear Regression (LR)	Thoral et al. [113],Cearns et al. [25],Zhou et al. [134],Park et al. [91]
Correlation Based	Hung et al. [56],Shameer et al. [103],Landicho et al. [68]
Random Forest	Mortazavi et al. [84],Madrid-García et al. [75],Park et al. [92]
Varience Test	Bolourani et al. [18],Symum and Zayas-Castro [109]
Wilcoxon Rank-Sum	Bolourani et al. [18],Li et al. [70]
XGBoost	Lineback et al. [72],Zhou et al. [134]
LR+PCA	Lineback et al. [72],Barber et al. [13]
ML Based	Sarijaloo et al. [101],Zhang et al. [132]
Pearson's correlation coefficient	Darabi et al. [35]
Fisher's Test	Li et al. [70]
Knowledge Based	Shang et al. [104]
Kullback-Leibler (KL) divergence	Golas et al. [50]
Recursive Feature Elimination (RFE)	Symum and Zayas-Castro [109]
PCA	Madrid-García et al. [75]
LASSO+GRA (Grey Relational Analysis)	Miswan et al. [80]
Wraper	Landicho et al. [68]

Table 2.3: List of papers that used each one of the feature selection methods.

Missing Data Handling

Among those papers that specified measures for dealing with missing values, [25, 35, 80, 91, 99] used Multiple Imputation by Chained Equations (MICE), and [109,

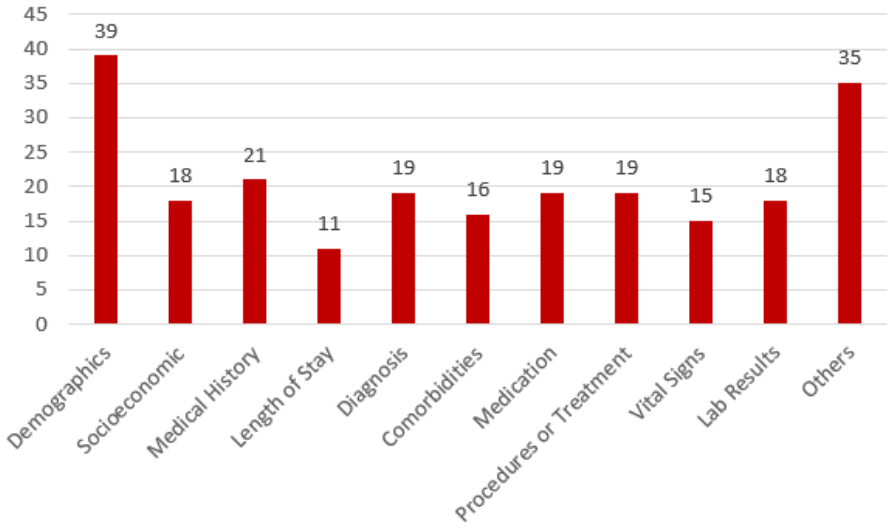

Figure 2.12: Number of studies used different categories of predictors.

123], utilized multiple regression chained equations. [61, 70, 92, 98] applied Listwise Deletion, and [101, 104, 132] removed features with missing value rate above a certain threshold. Boag et al. [17], Zhang et al. [132] replaced missing values in each predictor with the mean value of that feature. Also, [18] used SimpleImputer, but did not provide information about the parameters they used. With the default parameters SimpleImputer replaces the missing values with mean value of the feature. The details about other techniques that have been utilized is available in Table 2.4.

Imputation Technique	**Papers**
MICE	Darabi et al. [35],Rodriguez et al. [99],Cearns et al. [25],Miswan et al. [80],Park et al. [91]
Listwise Deletion	Kalagara et al. [61],Li et al. [70],Reddy and Delen [98],Park et al. [92]
Removing features with too many missing values	Shang et al. [104],Sarijaloo et al. [101],Zhang et al. [132]
Replaced with constant	Golas et al. [50],Zhang et al. [132]
Mean Replacement	Zhang et al. [132],Boag et al. [17]

multiple regression chained equations	Symum and Zayas-Castro [109],Welchowski and Schmid [123]
Simple Imputer	Bolourani et al. [18]
Mode Replacement	Shang et al. [104]
Last-Observation-Carried-Forward	Lin et al. [71]
kNN	Ou et al. [88]
Markov-chain Monte Carlo methods	Matheny et al. [77]
Median Replacement	Wang et al. [119]
naïve imputation	Mohammadi et al. [81]
RF	Zhou et al. [134]
Moving Average	Qian et al. [96]
predictive mean matching (PMM)	Landicho et al. [68]

Table 2.4: Papers stratified based on the their imputation method.

Machine Learning Models

Among the reviewed papers, 32 papers have implemented some variation of Linear Regression (LR) (e.g., LASSO, Elastic Net etc.). LR is either implemented as the main prediction tool or as a baseline model. 28 papers have implemented a tree-based classification such as Random Forest, Decision Tree, and Extra Tree. Boosting algorithms, including Gradient Boosting Machine, XGBoost, or AdaBoost are an other group of popular algorithms in readmission prediction implemented in 18 of the reviewed papers. Only 15 studies (29% of papers) proposed a new architecture or a new combination of methods for this specific task. Figure 2.13 depicts the usage frquency of each machine learning algorithm category in the studied cohort.

The relevant methods applied within this thesis are explained in the prior sections. Explanation of the additional methods listed here are beyond the scope of the thesis.

Detailed information about the papers that used each algorithm is available in Appendix A.

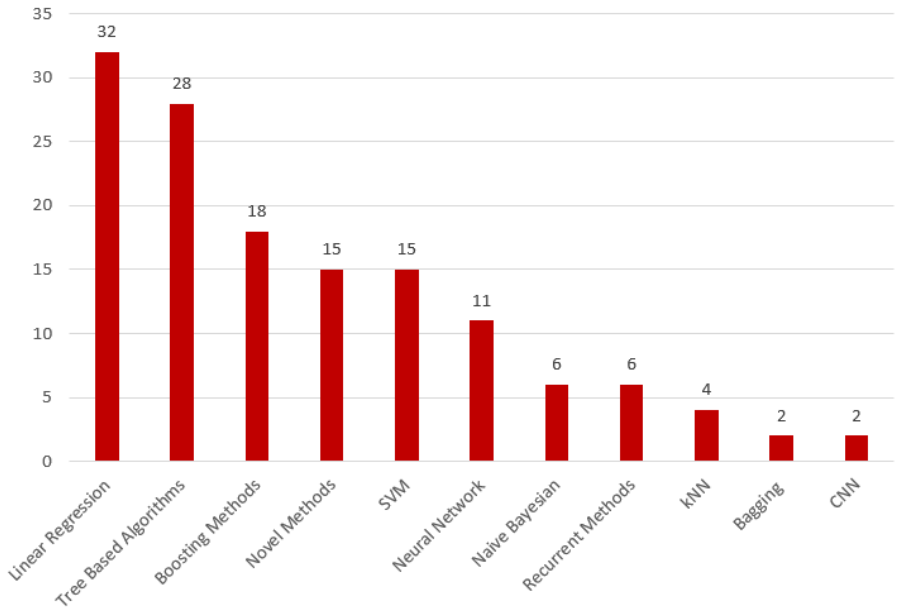

Figure 2.13: Number of studies used different categories of machine machine algorithms.

Model Performance

In this study the performance of the best performing algorithms in each paper has been extracted. The utilized models are presented in 7 categories. Boosting Methods, Linear Regression, Tree Based Algorithms, and Recurrent Methods categories are introduced in the previous section. Vanilla Algorithms category represents algorithms such as SVM, kNN or Naive Baysian while the Novel Algorithms category represents the papers that proposed a new method for readmission prediction. Figure 2.14 shows the box plot of the AUROC stratified by categories of artificial intelligence methods.

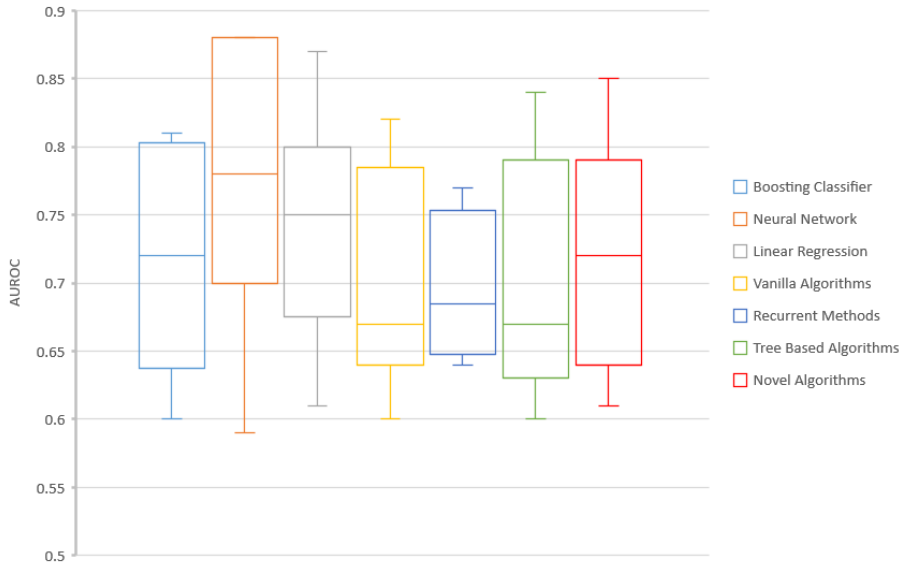

Figure 2.14: Box plot of AUROC of machine machine algorithms category.

Quality Assessment

Most papers identified in this review (41, 80%) scored 10 or higher indicating a high quality of implementation and reporting in those papers. Results of the quality assessment of the papers are available in the Appendix A. Figure 2.15 provides a combined view of the AUROC, quality assessment results, and the AI method categories. 2 papers with quality score less than 10 were not included in this plot as they did not report AUROC of their models.

2.9.7 Gap Analysis

In this study we identified 518 papers in readmission prediction. Of the 518 papers, only 51 were included after conducting a PRISMA review. These papers were selected were carefully reviewed to understand the state-of-the-art performance of sophisticated AI algorithms in predicting hospital readmission in various settings.

From the 21 accessible data sources for readmission prediction studies that were introduced, the top three popular datasets were MIMIC-III, Cerner HealthFacts EMR

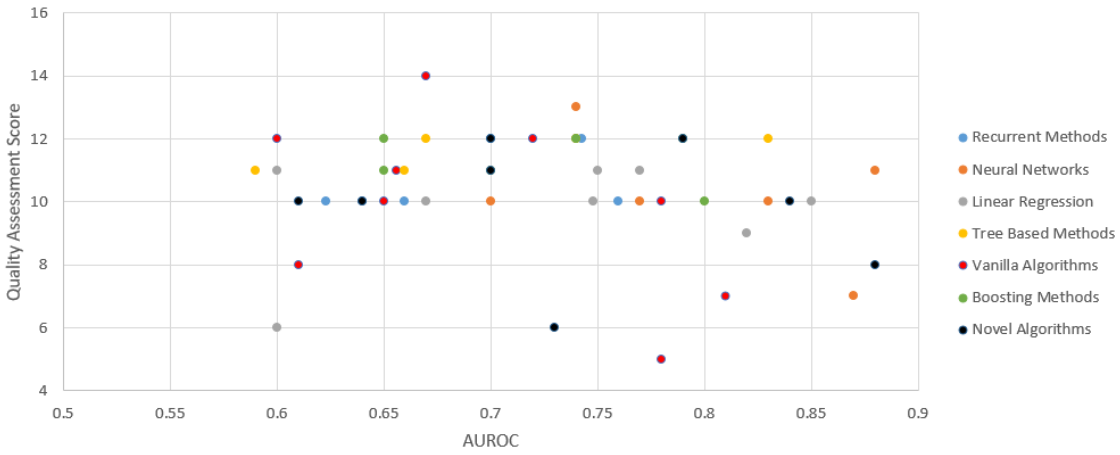

Figure 2.15: Scatter plot of the Quality Assessment Score of different studies and their AURCO labeled by their best performing models.

database, National Readmission Dataset. 18 feature selection techniques (Listed in Table 2.3) were used in readmission prediction studies where T-Test, Chi-Square methods, and using Linear Regression were more popular. 16 imputation method used in preprocessing the data were listed (Table 2.4) where MICE and Listwise Deletion were more frequently utilized. 11 types of predictors that used in at least 10 papers were listed and a list of papers that used these variable is provided in a supplementary table. 11 categories of frequently implemented AI algorithms in readmission prediction studies have been identified. Tree Based algorithms and Boosting methods were two most popular method after Linear Regression which was frequently used as baseline model. Finally, the papers were investigated for the methods they used for reducing the effect of data imbalance leading to listing 11 different methods of imbalance data handling. Class Weighting, SMOTE and Random Undersampling were the top three popular methods in this area.

The information presented provide a strong background for future studies in hospital readmission prediction. To the best of our knowledge there is no study that investigates the literature in such a granular level.

We found 5 gaps in the studies for hospital readmission prediction:

1. there is no dataset that is recognized as a benchmark dataset preventing researchers of comparing the performance of their methods against other studies in the literature;

2. private data sources are most common and this limits the repeatability of the published studies;

3. deep learning algorithms are less studied in this domain. Although the inclusion criteria of this review was determined to filter only the most sophisticated algorithms in the literature the number of algorithms that proposed novel deep learning algorithms are far less than those that used Tree-based and other Vanilla machine learning algorithms;

4. the outcomes of interest were mainly focused on post discharge readmission prediction while outcome prediction at the point of entry could also be helpful for shaping the course of treatment; and

5. studies were mainly focused on cardiac disease and "all-cause" readmission prediction. Hospital readmission after receiving treatments for other type of disease were less studied. For example, only 2 papers focused on readmission prediction for psychiatric disease.

6. A vast majority of the studies utilized medical history data. This approach limits the application of developed methods in areas such as mental where it is crucial to identify the high risk patients from early stages.

Chapter 3

Domain Data

This chapter provides a detailed description of datasets used in this study in the section 3.1. Furthermore, section 3.2 explains the preprocessing steps that has been taken to prepare them for respective experiments.

3.1 Data

In this section, Mental Health Questionnaire and text datasets used to test the proposed Textionnaire method, the deep architecture and the Attention Information algorithm are introduced. Then, the IMDB dataset utilized in testing the Attention Information explainability methods.

3.1.1 Mental Health Private Dataset

The Child and Youth Mental Health Program (CYMHP) is an outpatient mental health program offers mental care and services for children younger than 18 years of age at at McMaster University Medical Unit or Ron Joyce Children's Health Centre. Services include an inpatient unit, a day hospital program, outpatient services, emergency care, and outreach services for regions outside of Hamilton. The

Mental Health Questionnaire Dataset consists of a set of routinely collected parent- and youth-reported mental health intake assessments about children 18 years of age and younger who were registered in CYMHP as patient. At the intake to services, parent/caregivers of all children and youth aged 12 to 17 both completed a set of self-reported intake questionnaires that is also included in the private dataset. The questionnaire includes: 1) the Mental Health Questionnaire for Children and Youth (MHQ-CY), 2) measures of borderline personality features and temperament; and 3) a parent questionnaire about birth history, development, medication and family history of mental health problems. The mental health demands, functioning, and demographic characteristics of children and their caregivers is assessed using the MHQ-CY questionnaire. Characteristics such as risk factors, strengths, and modifiable behaviours of families seeking mental health care can be determined using this questionnaire. It also contains the Ontario Child Health Study Emotional Behavioural Scales [19, 46], which is a scale with verified validity and reliability in the evaluation of emotional and behavioural DSM-5 disorder symptoms. The MHQ-CY is completed by adolescents aged 12-17 (Youth version) and their caregivers (Caregiver version) either using iPads (pre-COVID, Before March 2020) or through Qualtrics (Post-COVID, April 2020 onwards) using email. Patient medical record numbers were utilized to link questionnaire responses to service utilization records extracted from the administrative hospital data. This included outpatient, inpatient, and emergency department visit data. Only the emergency department data was used to create 6-month emergency department visit labels. The administrative data used in this study were collected from October 2018 to April 2021. The questionnaires included in this research were acquired from September 2019 to April 2021 (n=1,393). The code book of this dataset including all the questions, response ranges and other information about this questionnaire is available in the Appendix B. We address this dataset as the *private dataset* throughout the paper. Only the questionnaires gathered from September 2019 to April 2021 (n=1393) were utilized in our study.

3.1.2 National Survey on Drug Usage and Health (NSDUH)

A publicly accessible dataset used in this research was the National Survey on Drug Usage and Health (NSDUH) [26], conducted by the United States federal government. This survey "provides nationally representative data on the use of tobacco, alcohol, and illicit drugs; substance use disorders; receipt of substance use treatment; mental health issues; and the use of mental health services among the civilian, non institutionalized population aged 12 or older in the United States" [2]. Responses from youth participants aged 12 to 17 surveyed from 2015 to 2019 (n=68,263) were used in this study. Due to lack of linked administrative data, the participants' response to a yes or no question about hospital admission in the past year (Question Code: $ANYSMH2$) was considered as the label. All other question indicating hospitalization history were removed from the datasets to avoid providing the answer to the model as a feature. The code book of this dataset including all the questions, response ranges and other information about this questionnaire is available in its public code book . Table 3.1 present additional information about NSDUH and Private datasets.

	Classes		**Age, years**		**Sex**	
Datasets	**Positive**	**Negative**	**Median**	**Range**	**Male**	**Female**
Private Data	635 (45.58%)	758 (54.42%)	14	4 to 17	616 (44.22%)	777 (55.78%)
NSDUH	10690 (15.94%)	56367 (84.06%)	-	12 - 17	34161 (50.94%)	32896 (49.06%)
NSDUH-B	9335 (50.00%)	9335 (50.00%)	-	12 - 17	8582 (45.97%)	10088 (54.03%)
NSDUH-SB	800 (50.00%)	800 (50.00%)	-	12 - 17	733 (45.81%)	867 (54.19%)
NSDUH-B (2015-2018)	7349 (50.00%)	7349 (50.00%)	-	12 - 17	7985 (54.33%)	6713 (45.67%)
NSDUH-B (2019)	515 (51.80%)	479 (48.20%)	-	12 - 17	497 (50.00%)	497 (50.00%)

Table 3.1: Private CYMHP and Public NSDUH Datasets Characteristics

3.1.3 Depression_Reddit

An additional public dataset utilized in this study is a text dataset consisting of depression-related Reddit and social medial posts published in [95] referred as *Depression_Reddit* by Ji et al. [59] (n=1,680).

3.1.4 Large Movie Review Dataset (IMDB)

A large movie review dataset [74], also known as IMDB, is a dataset for binary semantic text classification and widely used as a benchmark dataset. This dataset includes 50,000 movie reviews labeled with a binary label for positive and negative sentiments with an equal ratio. This dataset is used for evaluating our proposed method for increasing the interpretability of the BERT models.

3.2 Preprocessing

Several preprocessing steps were performed to prepare data for transformation and classification.

Mental Health Private Dataset

In the private dataset, three types of missing values were identified: i) The first category were the "Vague Missing Values." When a response to a check box was missing it was not clear if it is a negative response or a missing value. This type of missing values were replaced by 97. ii) According to the logic of questionnaire not all the participants had to answer all questions. The missing values associated with these "Valid Skips" were codded with 99. ii) All other missing values were replaced with 98. Because the focus of this study is to evaluate the effect of converting categorical questionnaire data to text, we also removed the free text responses from the dataset. Table 3.1 illustrates that the private dataset consists of 45.58% patients admitted to ED, which makes it a relatively balanced dataset. Thus, no balancing methods were required.

NSDUH

In the NSDUH dataset, we removed all the samples that had missing labels (i.e., the question . In this dataset the missing responses are reported as predetermined

constants. We removed questions with a missing response rate of 20% or more. To balance the data, we randomly down-sampled the majority class to have an equal number of samples as the minority class. After these preprocessing steps, we left with 18670 samples in the NSDUH dataset. We named this down-sampled version of the NSDUH dataset *Balanced NSDUH (NSDUH-B)*. Lastly, we selected 1,600 samples from the NSDUH-B to create a small balanced dataset that is comparable in size with the private dataset. We named this small sub-sample of NSDUH-B *NSDUH-SB* dataset.

Depression_Reddit

Random positive and negative labels were assigned to the samples creating a balanced dataset.

IMDB

No particular preprocessing were done on this dataset.

Chapter 4

Machine Learning Model

In this chapter the machine learning algorithm proposed to answer the research questions of this study are discussed. First, the questionnaire to text conversion method is explained and then the deep learning architectures used in this research are elaborated. Finally, the new method proposed to explain the outcomes of the deep learning models is explained.

4.1 Question to Text Conversion (Textionnaire)

In this section the process of generating Textionnaire representation for questionnaire data and its rational is explained.

4.1.1 Rational

Questionnaires are frequently used to gather information in many medical and non-medical applications ranging from mental health, pain, and other health-related areas to business management and politics. With the growing interest in using artificial intelligence in medical applications, many researchers have applied machine learning

algorithms to questionnaire data in various contexts [42, 100, 111]. However, the sample size in many medical surveys is very small. For example, in the 128 surveys studied in these seven systematic reviews of medical questionnaires [3, 28, 43, 48, 90, 106, 118], the average number of participants was 467, and the largest study population included 4451 participants. These numbers are sufficient for their particular application but are much smaller than the sample sizes required for training a deep learning model. This problem limits the effectiveness of deep learning models for many medical applications.

A popular solution to the small size of the datasets is transfer learning from a similar, but larger dataset [110, 135]. Transfer learning is defined as[110] improving the performance of a deep neural network on a target task by using the knowledge learned through training on a similar task (i.e. the source task). This technique enables researchers to apply deep learning algorithms to small datasets. Nevertheless, the data used in the initial training must be similar to the data used in the new task [33]. Since the questionnaire responses are gathered in a categorical format and a small change in the wording of the questions might entirely change the answering pattern, it is a challenge to find suitable questionnaire datasets in the domain of interests with a large number of records for transfer learning. It is the author's assertion that this problem can be addressed by transforming the questionnaire responses into a format that can leverage existing pre-trained models and transfer learning. Several studies [4, 62] have used pre-trained Natural Language Processing (NLP) models on questionnaire free text responses for classification tasks. However, analyzing the choice questions (as oppose to free text questions) of questionnaires using pre-trained deep learning models remains a less studied challenge. To solve this challenge, we need to understand how the information is stored in tabular format. Tabular representation of data usually has three parts: 1) The cell values, 2) the column labels, and 3) the definition of each row. For example, a table lacks meaningful information if we only know that the table has three columns and ten rows each filled with integers between zero to nine. However, we gain useful information, when the three columns

Option Code	Sentence Representation
1	I have serious difficulty concentrating remembering or making decisions because of a physical mental or emotional condition.
2	I do not have serious difficulty concentrating remembering or making decisions because of a physical mental or emotional condition.
94	I do not know if I have serious difficulty concentrating remembering or making decisions because of a physical mental or emotional condition.

Table 4.1: The questionnaire options from "Because of a physical, mental or emotional condition, do you have serious difficulty concentrating, remembering, or making decisions?" are represented by these descriptive sentences in the generated text. Codes 1, 2, and 94 denote a positive, negative, and "Don't Know" answer to this question, respectively.

in the same table are labeled as Hundreds, Tens, and Ones. In this case, we can infer that the table includes information of ten 3-digit numbers in an expanded form. In other words, if there is an 8 in the first column in one of the rows, we perceive it as eight hundred. In the same row, if there is a 9 in the second column, we consider that equal to ninety, and we read the 5 in the last column of this row as five. Still, we need a description to understand what those 3-digit numbers mean, but we can combine the information stored in column labels and the cell values and store the 3-digit positional notation of a base-10 number in a word format as "Eight hundred ninety five." Analogously, a text representation of the questionnaire data is conceptually similar to the word representation of a number stored in the expanded form. We introduce Textionnaire, a method that enables us to apply transfer learning to and utilize deep learning approaches on questionnaire datasets with small number of samples. Textionnaire transforms tabular questionnaire data into a text format. The text format representation enables us to use NLP for analyzing questionnaire data. In other words, instead of using tabular questionnaire data for training machine learning algorithms we generate representative text, for every questionnaire sample, that encompasses the information of that sample. Then, we use these Textionnaires to fine-tune a pre-trained NLP model for classification.

4.1.2 Feature Selection

Many NLP pre-trained models are suitable for a limited number of words from a specific domain. However, using all the questions of a questionnaire for generating Textionnaires may lead to very long text sequences. To address this challenge, we utilized a Gradient Boosting Machine (GBM) and feature importance scores calculated by GBM to rank the questions in the questionnaire based on their importance in GBM classifications. We utilize the highest ranking questions in creating Textionnaire to make sure it includes the most important information related to the classification problem and that our input data matches the boundaries of the pre-trained models. The number of questions included in the conversion can be determined by defining a cut-off threshold for the feature importance scores.

Algorithm 1 Questionnaire to Textionnaire Conversion

Input 1: Tabular questionnaire data
Input 2: Predefined sentences for every option of the selected questions
Output: A text representing the information within the questionnaire (Textionnaire)

```
for Every Participant do
    for Every Question Included do
        Load the selected option from the dataset (Cell Value)
        if The participant skipped this question or refused to answer then
            Skipp this question and do not add any sentence to Textionnaire.
        end if
        Load the predefined sentence representing this question and option
        Append the sentence to the Textionnaire
    end for
    return Textionnaire
end for
```

4.1.3 Questionnaire to Text Conversion

For every option of each selected question, we use the same words and grammatical structure as the question to generate a descriptive sentence that answers the question. Table 4.1 demonstrates an example of a question from NSDUH dataset, its options, and the generated sentences for each option. After selecting the most important

```
I have attended some type of school at some times during
the past 12 months. I would say none of the students
in my grade at school get drunk at least once a week.
During the past 12 months I have never gotten into a serious
fight at school or work. During the past 12 months I
have not participated in a problem solving communication
skills or selfesteem group. I believe that my religious
beliefs do not influence how I make decisions in my life.
I agree that my religious beliefs are a very important part
of my life. I have not moved in the past 12 months.
I do not have serious difficulty concentrating remembering
or making decisions because of a physical mental or
emotional condition.  In my  life I have had a period of time
lasting several days or longer when most of the day you
felt sad empty or depressed. In past year I have had
a major depressive episod with severe role impairment.
```

Figure 4.1: Generated Textionnaire for one of the NSDUH-B dataset's samples.

.

questions and manually creating descriptive sentences for every option, we follow the steps described in Algorithm 1 to construct Textionnaires. Figure 4.1 includes an example of Textionnaire generated using this method.

4.2 Artificial Intelligence Method

In this section, artificial intelligence algorithms utilized for classification of mental health patients with high risk of 6-month ED visit are discussed. In the following subsections, the pre-trained model used in this design is introduced and then the proposed wide and deep architecture is explained.

4.2.1 Deep Model

As explained in section 2.5, BERT models are frequently used in NLP tasks. Training these models takes a lot of time and resources and requires a large number of samples. However, there is a large inventory of pre-trained BERT models on variety of domains in the HuggingFace library [126]. In this study, the encoder layers of MentalBERT

are selected for our research because the generated Textionnaires are in the mental health domain. MentalBERT is a pretrained BERT model trained on mental health related social medial posts. This model has 12 encoders that each has 12 attention heads. The model dimension (d_{model}) is 796 and the maximum length of the input Textionnaires is set to be 256 tokens. In this study, a 3-layer neural network is applied to the [CLS] token of the sequence after 12 layers of encoders. The first, second, and output layers of the classification head consist of 128, 64 and 1 neurons, respectively. ReLU is utilized as activation function for the first and second layers while Sigmoid is the activation function for the output layer. This model is referenced as the *Deep Model* throughout this document. Figure 4.2 includes a block diagram of the deep model and the baseline model.

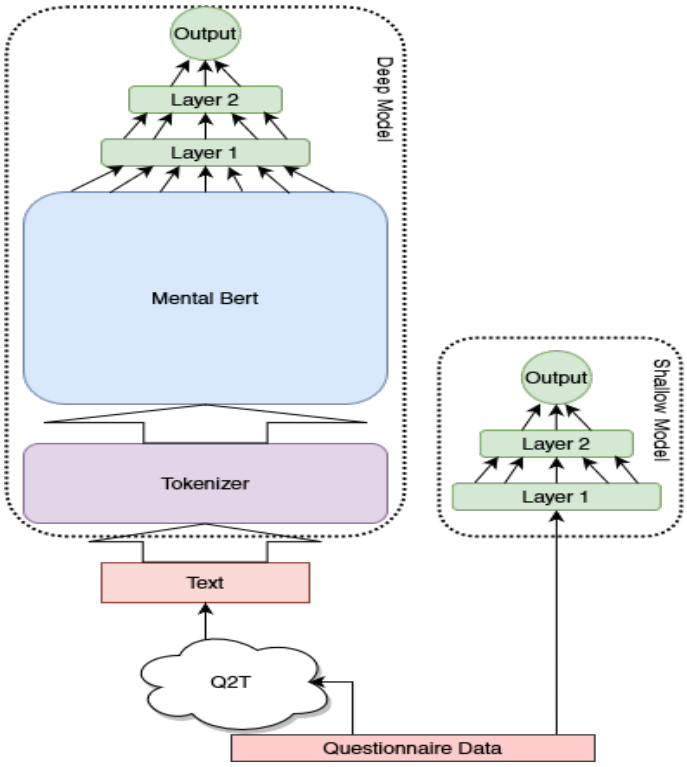

Figure 4.2: Deep and shallow models.

.

4.2.2 Wide and Deep Architecture

Using hand crafted features parallel to the machine generated features found to be useful in other complicated classification tasks [34]. This method combines the machine learned knowledge from the data with the features derived from the human experience in the field. In this study, we utilized Ontario Child Health Study Emotional Behavioural Scales (OCHS-EBS) [19, 46] as hand crafted or knowledge driven features along with the features generated by the MentalBERT deep encoder from the Textionnaires. In this architecture the output of the MentalBERT encoder along with the OCHS-EBS scores are fed into a three layer neural network similar to the one described in section 4.3. Figure 4.3 shows the wide and deep architecture.

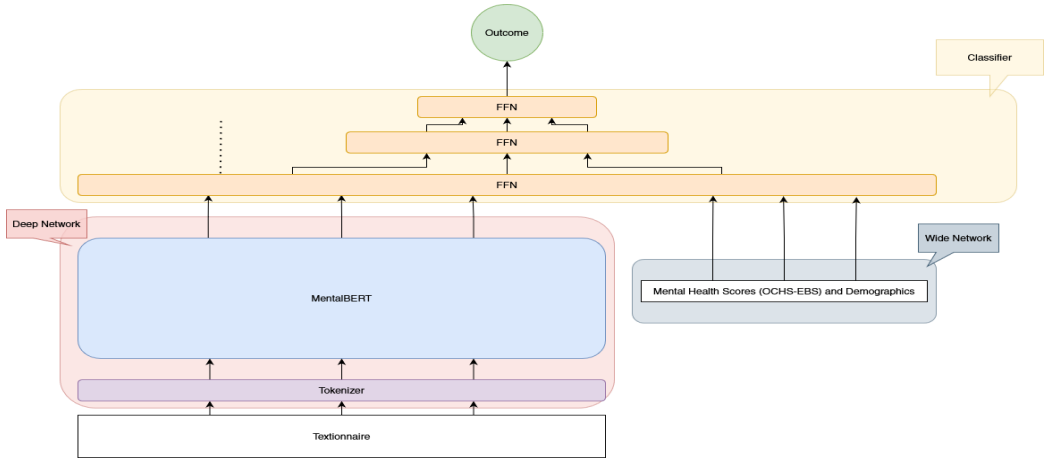

Figure 4.3: The Wide and Deep architecture proposed for 6-month ED visit prediction.

4.2.3 Train and Test Separation

In all classification experiments listed in chapter 5, 80% of the samples were used for the training, 10% for validation and 10% for testing. Samples in each set were randomly selected using the "train_test_split" function in the Sci-kit Learn library [94]

4.2.4 Training Details

The batch size used in this study is 16 and the model is trained for 5 epochs. The training time for the deep model on the Private dataset was 3 minutes and 51 seconds, and on the NSDUH dataset was 4 minutes and 5 seconds. Training the wide and deep network took similar time for training. The training and testing were performed on our lab GPU server (3 x RTX 2080 TI). The Adam optimizer is used with a ramp learning rate increase.

4.3 Baseline Model

The base line model is a three layer neural network. Similar to the classification head of the BERT model, the first and second layers have 128, and 64 neurons both with ReLU activation function, and its output layer has only one neuron with Sigmoid activation function. SVM, Linear Regression (LR), and GBM were also tested to be considered as baseline models. LR always resulted in lower training AUROC than testing AUROC, and GBM always over-fitted to the data. These models were tested with variety of hyper-parameters all resulting in similar outcomes as what listed before. SVM always had lower AUROC than the neural network which is consistent with the findings of the literature review. Furthermore, a neural network similar to the classification of the deep model provides more insight on the impact of the MentalBERT encoder on the performance improvement.

4.4 Attention Information for Explainability

Using BERT algorithms instead of other vanilla algorithms comes at a cost of losing explainability. Many of the tree based algorithms, SVM, LR, etc. can identify the most important features in the outcome of the classifier. However, deep models such as BERT are less explainable. As it is thoroughly discussed in section 2.5, the

transformers based models such as BERT architecture include multi-head attention mechanism. Some researchers [32, 97, 112, 117] have attempted to use attention weights to explain how BERT models work. On the other hand, in [57, 102] researchers found results that suggest that attention weights are not good means for interpreting BERT models. Vashishth et al. [115] tried to verify if attention weights are good means of explanation and concluded that both school of thoughts could be correct, stating "Attention weights are interpretable and are correlated with feature importance measures. However, this holds only for cases when attention weights are essential for model's prediction and cannot simply be reduced to a gating unit." Where the gating unit word used to suggest that amount of attention weights are not simply a direct indicator of the effect of the words in the model outcome. Whether the values of attention weights can be considered as explanations or not, their high predictive value in classifying the outcome of the BERT model indicates the importance of those weights in explaining the outcome of the BERT models. In this section, an entropy-based method is proposed to explain the outcomes of the pipeline proposed in this study. Although this method is applied to BERT component of the Deep model, suggesting this algorithm for explaining the BERT model in any pipeline requires further studies and developments on this method.

4.4.1 BERT Nomenclature

In the following, a few terms are introduced to support the explainability method in the following section.

Column or Row Words

Attention weights are stored in a square matrix and vary from every input sample to another. The dimension of the square matrix is the same as the maximum length of the input set in the model design process (256 in our case). In this square matrix each row is assigned to an input token in the same order as the input sequence. For

	[CLS]	It	is	cold	[PAD]
[CLS]	a_{00}	a_{01}	a_{02}	a_{03}	a_{04}
it	a_{10}	a_{11}	a_{12}	a_{13}	a_{14}
is	a_{20}	a_{21}	a_{22}	a_{23}	A_{24}
cold	a_{30}	a_{31}	a_{32}	a_{33}	a_{34}
[PAD]	a_{40}	a_{41}	a_{42}	a_{43}	a_{44}

Figure 4.4: An example of attention weight assignment.

example, in a sentence like "it is cold." after tokenization the first row is assigned to the $[CLS]$ token which is always the first token of each sequence, the second row belongs to word *it* and it continues to the last word and the rest of the rows are assigned to $[PAD]$ tokens that fill the remaining tokens to keep the input length fix and equal to maximum input length. A Similar assignment is done for the columns of the square matrix. In this way the element in the second row and third column of the matrix is recognized as the attention from the second token to the third token of the input sequence. The tokens assigned to each words are addressed as row-words, and the token tied to each column are called column-words. Figure 2.5 demonstrates an example of token assignment in BERT, and Figure 4.4 demonstrates the attention weights for the the "it is cold" sentence where maximum length of the model was 5.

Attention Map

As it is explained in the previous section the rows and columns are assigned to different words. This assignment varies from sample to sample which means a word could

appear as the fifth token in one sample and as the fourteenth token in another sample. In order to store the attention weights for all the pairs (pairs of row-words and column-words) in the same order the Attention Map matrix is introduced. This is a square matrix and its dimension is equal to the number of unique words in the group of input samples for which we are investigating the driving factors of the BERT model outcome. Similar to attention weights each row and column is assigned to a word.

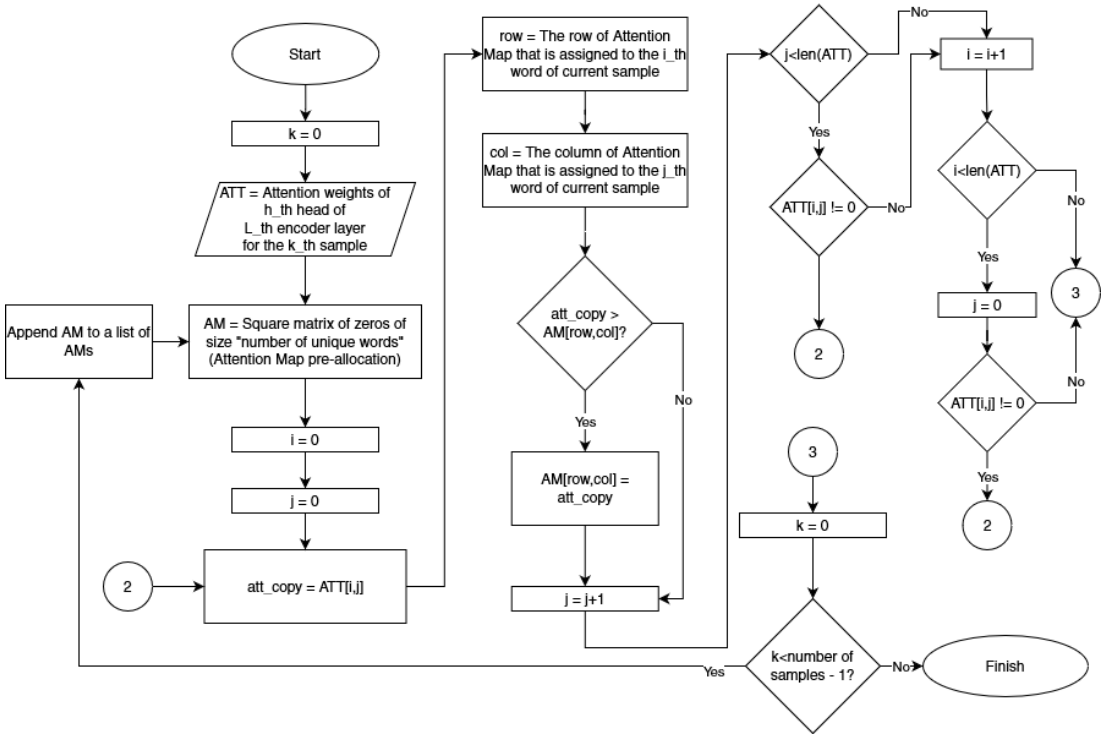

Figure 4.5: The process of generating Attention Maps

In this method, the most important attention weights are determined based on the amount of information gained about the BERT model outcome when splitting the sample on those attention weights. The feature importance of Decision Tree Classifier (DT) is used as a proxy to the information gains.

This matrix is created for each input sample and its cells are populated either with zero values (for those pairs that did not appear in the input sequence) or the attention weights for the pairs that appeared in the input sequence. If a word is repeated in

the input sequence several time and in return has several attention weights assigned to the pairs associated with it the largest attention weight is stored. The process of generating the Attention Maps for every sample is described in the flow chart depicted in Figure 4.5

Note: If a large number of unique words come with a frequency of 1, words can be clustered based on their meaning using the Spacy similarity score [53]. Spacy assigns a vector to each word that exists in its dictionary. These vectors are designed in a way that a Cosine distance of two vectors determines the similarity of the associated words. In such case, each row and column of the Attention Map is assigned to one cluster. Consequently, the dimension of the Attention Map will be equal to the number of identified meaning clusters. The process of clustering words based on their meaning similarity is described in a flowchart in the Appendix D as it is not required for the Textionnaire example, but is used in further tests on benchmark datasets such as IMDB.

4.4.2 Proposed Entropy Method for Explainability

All the extracted Attention Maps are flattened to a one-dimensional vector. The outputs of the BERT algorithm for every sample are extracted and utilized as the label. A DT is trained to use the flattened Attention Map and predict the outcome of the BERT model. The DT is set to use "Entropy" (An alias for Information Gain explained in section 2.6). After training, the feature importance scores are extracted and reshaped back to the original dimension of the Attention Map. The pairs associated with two largest feature importance scores are extracted and stored in a Data Frame. The block diagram in Figure 4.6 depicts the process of finding the most important pairs. The same process is done for all the layers and heads resulting in 288 (2×12 ($number\ of heads\ in\ each\ layer$) $\times$ 12 ($number\ of encoder\ layers$)) pairs of words that are not necessarily unique (Some words might be found important in multiple heads and layers).

Note: The DT is trained to classify the BERT outcomes in an attempt to extract

the information gain of attention weights. Consequently, it is not supposed to have a high performance on unseen samples. For this reason, the hyper-parameters of the DT are set to overfit to the training data.

4.4.3 Visualization

The extracted attention information (i.e. the feature importance score of the DT for each pair) is assigned to all the words involved in pairs. To this end, a Python script generates a Pandas [125] data frame where its columns are labeled with the input samples token and each row is assigned to a specific head in one of the encoder layers. Next, the extracted attention information values are stored in the cells associated with most important words identified using the attention map from the head and layer assigned to that row. To provide a sense of overall attention information, a column wise summation was performed and the outcome is presented in a bar chart indicating the amount of information carried by each weight across all attention heads in all layers Figure 6.2 demonstrate the outcome.

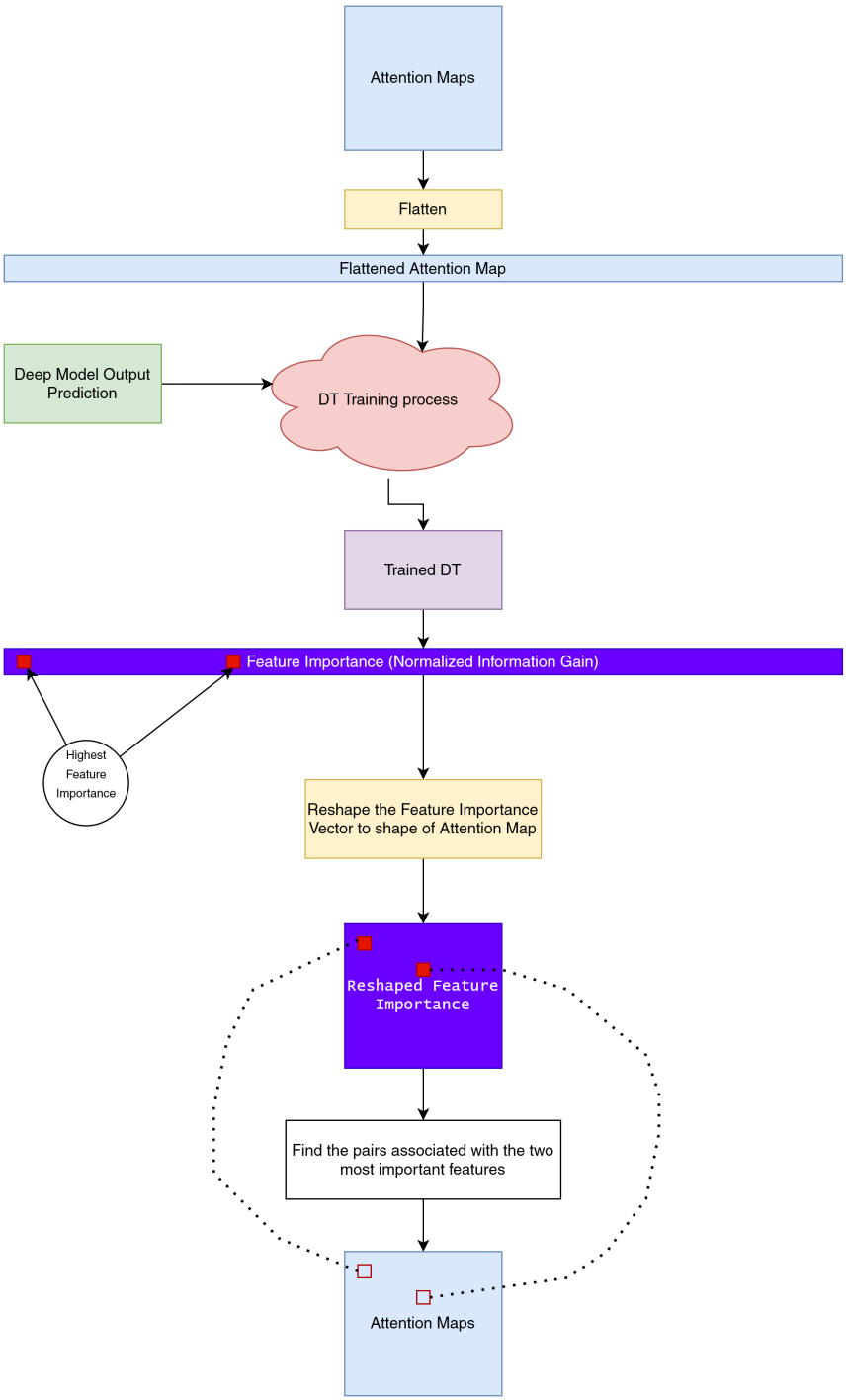

Figure 4.6: The process of finding the most important pairs in each head of each encoder layer.

.

Chapter 5

Methodology

In this chapter, the machine learning pipeline used to run the experiments is described. Furthermore, the experiments designed to verify the proposed algorithms in chapter 4 are explained.

5.1 Machine Learning Pipeline

First, the data described in chapter 3 is passed through the preprocessing steps described in section 3.2. Next, the processed data is converted to Textionnaire through the method described in section 4.1. The following questions were utilized in generating the Textionnaires from the Private dataset.

- Has this child/youth experienced any of the following? Please mark all that apply.
  - Been an overnight patient in a hospital or other setting for problems with emotions, attention or behaviours or use of drugs or alcohol
    - Yes, in the past 12 months.

- To the best of your knowledge did this child/youth ever seriously consider taking their own life or killing themselves?

    1. No, never
    2. Yes, but more than 12 months ago
    3. Yes, in the past 12 months
    4. Yes, in the past 4 weeks

- In the past 6 months, how would you rate this child/youth's mental health?

    1. Very poor
    2. poor
    3. moderate
    4. good
    5. very good

- During the past 12 months, did you, another family member or this child/youth visit an emergency room about concerns regarding this child/youth's mental health?

    1. Yes
    2. No

- To the best of your knowledge has this child/youth ever harmed themselves on purpose (e.g. cutting, scratching, hitting or burning themselves) but did not mean to take their life?

    1. No, never
    2. Yes, but more than 12 months ago
    3. Yes, in the past 12 months
    4. Yes, in the past 4 weeks

- To the best of your knowledge did this child/youth ever seriously consider taking their own life or killing themselves?

    1. No, never
    2. Yes, but more than 12 months ago
    3. Yes, in the past 12 months
    4. Yes, in the past 4 weeks

- Is this child/youth taking any medication or pills prescribed for mental health concerns?

    1. Yes
    2. No

- I rarely get enthusiastic about anything.

    1. Very False or Often False
    2. Sometimes or Somewhat False
    3. Sometimes or Somewhat True
    4. Very True or Often True

- Select the response that best represents how much you agree or disagree with each statement.

    – Activities like eating, dressing, washing and moving around are easy for my child.

        1. Strongly disagree
        2. Disagree
        3. Neither agree nor disagree
        4. Agree
        5. Strongly agree

Note: One of the 10 most important feature of the Private dataset was from the OCHS-EBS scores. Consequently, this score was ignored and only 9 questions were included in the process of generating Textionnaires. In experiments where NSDUH dataset was used the following questions were used to generate the Textionnaires:

- Because of a physical, mental or emotional condition, do you have serious difficulty concentrating, remembering, or making decisions?

    1. Yes
    2. No
    3. Don't KNOW

- Have you ever in your life had a period of time lasting several days or longer when most of the day you felt sad, empty, or depressed?

    1. Yes
    2. No
    3. Don't Know

- During the past 12 months have you participated in a problem solving, communication skills or self-esteem group?

    1. Yes
    2. No
    3. Don't Know

- Have you had a major depressive disorder in last year? (Generated variable from the response to two other questions)

    1. Yes
    2. No

- How many of the students in your grade at school would you say get drunk at least once a week?

    1. None of them
    2. A few of them
    3. Most of them
    4. All of them
    5. Don't Know

- During the past 12 months, how many times have you gotten into a serious fight at school or work?

    1. 0 times
    2. 1 or 2 times
    3. 3 to 5 times
    4. 6 to 9 times
    5. 10 or more times
    6. Don't Know

- These next questions are about the role that religious beliefs may play in your life. For each statement, please indicate whether you strongly disagree, disagree, agree, or strongly agree.

    – Your religious beliefs influence how you make decisions in your life.

        1. Strongly Disagree
        2. Disagree
        3. Agree
        4. Strongly Agree
        5. Don't Know

  - Your religious beliefs are a very important part of your life.
    1. Strongly Disagree
    2. Disagree
    3. Agree
    4. Strongly Agree
    5. Don't Know

- How many times in the past 12 months have you moved? Please include moves from one residence to another within the same city/town as well as those from one city/town to another.
  1. 0 time
  2. One time
  3. Two times
  4. Three or more times
  5. Don't Know

- Have you attended any type of school at any time during the past 12 months?
  1. Yes
  2. No
  3. Don't Know

Then the Textionnaires were fed into the Deep model described in subsection 4.2.1. When testing the wide and deep architecture, described in subsection 4.2.2, the OCHS-EBS are loaded parallel to the Textionnaires as it is depicted in Figure 4.3. Finally, the outcomes and the attention weights of the Deep model are investigated using the Attention Information method described in section 4.4 to explain the performance of the Deep model.

5.2 Experiments

In this section, we define the experiments conducted to answers the question listed in the section 1.3.

1. The following experiments were designed to answer RQ1: *Using only questionnaire data, can a reliable model be developed to predict if a patient of the mental health clinic (inpatient or outpatient) is deemed high-risk with resulting Emergency Department visit within next 6 months?*

    - Experiment 0 is designed to define a baseline for the described task as no similar study was found in the literature review.
    - Experiment 1 evaluates the classification performance when we have small sample size data.
    - Experiment 2 compares the classification performance of the Deep model trained on small number of samples with the shallow model trained on large number of samples to evaluate the efficacy of our method when a large number of samples are accessible.
    - Experiment 3 tests the effect of proposed method on generalizability of classification.
    - Experiment 4 ensures that the results the Deep mode are not that same given any random text.
    - Experiment 5 tests the performance of wide and deep architecture proposed to improve the classification task described in RQ1.

2. These experiments designed to answer RQ2: *Can pre-trained deep learning models be used on questionnaire data to improve classification performance?*

    - Experiment 1 tests the performance of pre-trained NLP models when fine-tuned on small samples size questionnaire data.

- Experiment 2 compares the performance of pre-trained NLP model fine-tuned on small sample size datasets with the Shallow model when trained on large dataset.
- Experiment 3 verifies the effect of pre-trained NLP model on generalizability of classification.

3. The following experiments were conducted to answer RQ3: *Could text representation of questionnaire data be used in classification instead of structured tabular representation?*
   - Experiment 1 evaluates the efficacy of the Textionnaire in improving the classification performance when we have small sample size data.
   - Experiment 2 evaluates the efficacy of Textionnaire when a large number of samples are accessible.
   - Experiment 3 tests the effect of Textionnaire on generalizability of classification.
4. The following experiments were used to test the method proposed to answer RQ4: *How to make our method explainable?* and interpret its outcomes:
   - Experiment 6 ensures that the extracted outcomes are not random.
   - Experiment 7 assesses if the order of features affects the DT algorithm and its feature importance.
   - Experiment 8 interprets the outcome of the attention information method. This experiment verifies if the presence of a word of a sentence in the input text is perceived by BERT as a flag for the presence of that sentence or the specific word matters.
   - Experiment 9 evaluates if the wording of the sentences in the text affects the outcome.

5.2.1 Experiment 0: Finding a Baseline Model

The first step to answer the first research question was to find out if machine learning algorithms can predict the 6-month emergency department visit using the questionnaire data and the OCHS-EBS. Based on the findings of the literature review on readmission prediction studies, several algorithms including SVM, LR, GBM, and neural network were tested on the tabular questionnaire data and OCHS-EBS to evaluate the predictive value of these features. LR could not fit to the data and consistently resulted in test AUROC higher than training AUROC. Also, GBM either overfit to the data or resulted in very poor AUROC. However, promising performance of the SVM and neural network models demonstrated the predictive value of the questionnaire and the OCHS-EBS for emergency department visit within 6 month of initial inpatient or outpatient mental health visit. Since the neural network model performed better on the task it was selected as the base line model.

5.2.2 Verification of the efficacy of Textionnaire

The following experiments have been conducted on the deep and shallow (Baseline) models to evaluate the impact of Textionnaire on the ED visit prediction (Private Dataset, and other classification tasks (NSDUH Dataset). In all the experiments the Shallow model was trained and tested on the tabular representation of the datasets and the Deep model was trained and tested on the Textionnaire representation of the Data.

Experiment 1 (Small Data vs. Small Data):

We fine-tuned our deep model and trained the shallow model on the private dataset. We compared the performance of these two models to see if using Textionnaires representation of the questionnaire data and pre-trained MentalBERT improved the classification performance.

Similarly, we fine-tuned the Deep model and trained the Shallow model on NSDUH-SB dataset to test if our method is applicable to other questionnaires.

Experiment 2 (Large Data vs. Small Data):

We trained the shallow model on the NSDUH-B dataset and fine-tuned the Deep model on the NSDUH-SB to compare the performance of the proposed solution (Textionnaire + Pre-trained NLP model) with the performance of the Shallow model trained on large number of data and assess how well this solution works compared to use of large number of samples.

Experiment 3 (Old Training vs. New Testing):

We used samples from 2015 to 2018 in the NSDUH-B dataset for training and 994 entries from 2019 NSDUH-B for testing. We trained the Shallow model on the tabular data of all samples between 2015 and 2018 and tested it on the tabular data of the selected entries from 2019. Also, we fine-tuned the Deep model on 1600 Textionnaire generated from 2015 to 2018 data samples and tested it on Textionnaires generated from samples from 2019 data. The results of this experiment, indicate which model has better performance on an unseen distribution (2019 samples) which is an indication of better generalizability on questionnaire data.

Experiment 4 (Deep vs. Random):

To ensure that the pre-trained model does not generate similar classification outcomes on any random mental health related text, we ran a dummy classification on Depression_Reddit dataset. In this dummy classification, we assigned random labels to Depression_Reddit dataset's texts to create a balanced classification dataset. The Deep model was fine-tuned on this dataset and tested on separate test set from the same dataset. We assume that a poor performance in this task would guarantees that the pre-trained model does not have same classification performance on any

mental health-related text and information in the questionnaire data drives the high classification performance.

5.2.3 Experiment 5: Evaluation of Wide and Deep Architecture

To test the impact of the wide and deep architecture on 6 month ED prediction it is tested on the private dataset. The model is trained with Textionnaires and OCHS-EBS and compared with the Deep model trained on only Textionnaire and shallow model trained on the tabular representation of the questionnaire. The AUROC is compared to evaluate the effectiveness of the wide and deep architecture.

5.2.4 Verification of the Attention Information Method

The attention information method is applied on the deep model to interpret the outcomes of the BERT model. It is not applied on the wide and deep model as we could not specify how the outcomes were affected by the OCHS-EBS. Two tests were performed to ensure that first, the same pairs and attention information values are not generated by any random labels and second the outcome of the algorithm is not affected by order of columns and rows in the attention maps. All of the experiments related to attention information technique used the Private dataset for the tests.

Experiment 6: Deep Model Outputs Vs. Random Labels

In this experiment, the attention maps were generated from the attention weight of the Deep model using the method described in section 4.4.1. However, the DT was trained to predict some random labels (Instead of BERT prediction outcomes as it is described in subsection 4.4.2). The hypothesis is that if the important pairs are a result of a relationship between the attention weights and the output the extracted important pairs must be different from the important pairs extracted from the DT

trained to predict the outcome of the deep model.

Experiment 7: Normal Order Vs. Shuffled Order

In this experiment, the DT was trained as described in subsection 4.4.2, but the flattened features were randomly shuffled before training of DT. The identified important pairs were compared with the important pairs extracted without shuffle. The hypothesis was that if the important pairs do not change with as a result of the shuffle the algorithm is not affected by the order of the features (i.e. the order of columns in the Attention Maps).

Two other tests are conducted to further analyze the outcome of this method.

Experiment 8: Does the Word Matter or Presence

In this experiment, one of the sentences which had the most important word or words, were replaced by some random irrelevant sentence in the process of generating the Textionnaire. In other words, instead of creating a descriptive sentence using the question and the options of the questionnaire a few random sentences were assigned to the options of associated question. In this way, when this question is answered in the questionnaire another sentence, which is not related to mental health, was added to the Textionnaire. All the steps including generating the Textionnaires, training the Deep model and those described in section 4.4 were repeated to extract the new attention information. The hypothesis is that if some words from these semantically irrelevant sentences are considered as the most important words, the presence of the sentence and not its wording matters to BERT algorithm.

Experiment 9: Is it Semantic or the Wording?

In this test, in the process of generating the Textionnaire the sentence containing the most important words, was replaced by a paraphrased sentence that carried the same semantic, but had different wording than the initial sentence. We hypothesize

that if the words that carry similar meaning as the initial most important words are considered as the most important words BERT is comprehending the semantic of the words. If other words in the same sentence get higher attention, it suggests that words are pointing to the semantic or presence of the sentence.

Chapter 6

Results

This chapter includes the result of experiments introduced in chapter 5.

6.1 Experiment 0: Base Line Performance

In Table 6.1, the performance metrics of the models tested to establish a baseline model are listed.

Metrics	SVM	Shallow Model
AUROC	0.48 ± 0.007	**0.61 ± 0.05**
Specificity	**0.8 ± 0.01**	0.73 ± 0.08
Sensitivity	**0.50 ± 0.003**	0.42 ± 0.08

Table 6.1: Performance metrics of the SVM and Shallow model tested on the tabular version of the Private dataset and OCHS-EBS.

Experiment 0 Analysis

The AUROC of the Shallow model suggest that the questionnaire and OCHS-EBS have predictive values for predicting 6-month ED visit among Mental Health patients.

Based on the findings of this experiment the Shallow model was determined as a baseline for this study.

6.2 Experiments 1-4: Verification of the Efficacy of Textionnaire

Table 6.2 includes the AUROC of the experiments 1 to 4 listed in subsection 5.2.2.

Experiments	**Training Data**	**Testing Data**	**Shallow Model**	**Deep Model**
#1	Private Dataset	Private Dataset	0.61 ± 0.05	**0.75 ± 0.01**
	NSDUH-SB	NSDUH-SB	0.66 ± 0.03	**0.77 ± 0.03**
#2	NSDUH-SB	NSDUH-SB	0.66 ± 0.03	**0.77 ± 0.03**
	NSDUH-B	NSDUH-B	0.72 ± 0.004	-
#3	2015-2018 NSDUH-SB	2019 NSDUH-SB	0.65 ± 0.02	**0.75 ± 0.005**
#4	Depression_Reddit	Depression_Reddit	-	0.53 ± 0.04

Table 6.2: AUROC of Deep and Shallow models across four experiments.

Table 6.3 presents the specificity of the models in experiments 1 to 4 explained in subsection 5.2.2.

Experiments	**Training Data**	**Testing Data**	**Shallow Model**	**Deep Model**
#1	Private Dataset	Private Dataset	0.73 ± 0.08	**0.84 ± 0.04**
	NSDUH-SB	NSDUH-SB	**0.82 ± 0.01**	0.78 ± 0.03
#2	NSDUH-SB	NSDUH-SB	**0.83 ± 0.02**	0.79 ± 0.04
	NSDUH-B	NSDUH-B	0.72 ± 0.01	-
#3	2015-2018 NSDUH-SB	2019 NSDUH-SB	**0.83 ± 0.00**	0.75 ± 0.00

Table 6.3: Specificity of deep and shallow models across four experiments.

To have a complete picture of the deep model and Textionnaire performance, Table 6.3

must be considered along with Table 6.4 that includes the Sensitivity of the model.

Experiments	**Training Data**	**Testing Data**	**Shallow Model**	**Deep Model**
#1	Private Dataset	Private Dataset	0.42 ± 0.08	**0.51 ± 0.02**
	NSDUH-SB	NSDUH-SB	0.35 ± 0.02	**0.67 ± 0.03**
#2	NSDUH-SB	NSDUH-SB	0.35 ± 0.02	**0.67 ± 0.01**
	NSDUH-B	NSDUH-B	0.66 ± 0.01	-
#3	2015-2018 NSDUH-SB	2019 NSDUH-SB	0.33 ± 0.00	**0.64 ± 0.00**

Table 6.4: Sensitivity of deep and shallow models across four experiments.

Experiment 1 Analysis

The superior performance of the Deep model on the Textionnaire representation of both of the questionnaire datasets in terms of AUROC indicates the effectiveness of this approach for utilizing transfer learning in analyzing questionnaire datasets with small number of samples. Consistency of the better performance across two different datasets and two different classification tasks (Explained in section 3.1) shows that the effectiveness of this approach is not limited to the Private dataset and 6-month ED visit.

Experiment 2 Analysis

Better performance of the Deep model on Textionnaire representation of a small dataset (NSDUH-SB) compared to the performance the Shallow model on large number of samples (NSDUH-B) in terms of AUROC highlights the usefulness of the proposed technique in improving the robustness of the model. This method improves the classification performance even compared to a Shallow model that is trained on a dataset which is more than 10 times larger in sample size and potentially includes a wider variety of possibilities.

Experiment 3 Analysis

Enhancement in performance in classification of samples from an unseen distribution (NSDUH - 2019) suggests an improvement in generalizability of the model.

Experiment 4 Analysis

Poor performance in classification of random labels suggests that the outcomes of previous experiments were not lucky coincidences and were derived by the information of the questionnaire data represented in Textionnaire format.

Overall, using of the proposed method improved robustness and generalizeablity of the model. According to Table 6.4 the Deep model has higher sensitivity which means it identifies the high risk patients with higher accuracy. It is an important qualification in medical classification.

6.3 Experiment 5: Evaluation of Wide and Deep Architecture

Table 6.5 includes the performance metrics of the wide and deep model.

Metrics	Shallow Model	Deep Model
AUROC	0.61 ± 0.05	**0.77 ± 0.01**
Specificity	0.73 ± 0.08	**0.89 ± 0.01**
Sensitivity	0.42 ± 0.08	**0.54 ± 0.02**

Table 6.5: Performance metrics of the Wide and Deep model on the Textionnaire and OCHS-EBS, age and sex and the baseline model on the raw data.

Experiment 5 Analysis

Improvement in terms of all three performance metrics when using the Wide an Deep architecture (Using Textionnaire and OCHS-ESB) suggests that using knowledge driven

and machine generated features can improve the classification performance for 6-month ED visit prediction.

6.4 Verification and Interpretation of the Attention Information Method

Recall experiments 6 and 7, where two tests were performed to ensure that: 1) the same pairs and attention information values are not generated by any random labels, and 2) the outcome of the algorithm is not affected by order of columns and rows in the attention maps. Table 6.6 lists the pairs of words identified as most informative by the Attention Information algorithm. In this table the normalized information value of each pair is listed.

Layer	Head	Row Words	Column Words	Normalized Information
0	0	child	member	0.771
0	1	child	mental	0.730
0	2	unfamiliar	himself	0.702
0	3	children	with	0.729
0	4	family	family	0.731
0	5	family	member	0.768
0	6	status	family	0.749
0	7	family	family	0.755
0	8	cutting	been	0.741
0	9	child	with	0.729
0	10	child	emergency	0.739
0	11	easy	dressing	0.723
1	0	purpose	month	0.713
1	1	family	emergency	0.695
1	2	emergency	family	0.678
1	3	with	wanted	0.685

Table 6.6 continued from previous page

Layer	Head	Row Words	Column Words	Feature Importance
1	4	family	visit	0.787
1	5	with	play	0.696
1	6	themselves	themselves	0.726
1	7	family	washing	0.736
1	8	month	parent	0.697
1	9	cutting	children	0.740
1	10	parents	member	0.657
1	11	easy	washing	0.683
2	0	month	member	0.703
2	1	medication	with	0.540
2	2	with	child	0.550
2	3	been	child	0.535
2	4	been	child	0.545
2	5	child	with	0.540
2	6	parent	been	0.550
2	7	months	emergency	0.538
2	8	prescribed	with	0.547
2	9	with	medication	0.552
2	10	been	overnight	0.540
2	11	family	member	0.645
3	0	parents	member	0.542
3	1	with	child	0.558
3	2	been	child	0.567
3	3	with	child	0.550
3	4	been	been	0.546
3	5	with	prescribed	0.544
3	6	been	been	0.550

Table 6.6 continued from previous page

Layer	Head	Row Words	Column Words	Feature Importance
3	7	months	emergency	0.544
3	8	with	child	0.540
3	9	room	taking	0.544
3	10	parents	parents	0.557
3	11	child	with	0.536
4	0	child	with	0.566
4	1	past	12	0.555
4	2	with	pills	0.554
4	3	regarding	medication	0.559
4	4	12	months	0.555
4	5	with	child	0.546
4	6	been	burning	0.548
4	7	been	overnight	0.553
4	8	months	during	0.543
4	9	12	emergency	0.546
4	10	with	been	0.550
4	11	regarding	months	0.555
5	0	child	been	0.564
5	1	been	overnight	0.553
5	2	youth	member	0.562
5	3	been	12	0.552
5	4	been	overnight	0.555
5	5	been	12	0.549
5	6	been	overnight	0.546
5	7	been	with	0.539
5	8	months	months	0.566
5	9	months	room	0.562

Table 6.6 continued from previous page

Layer	Head	Row Words	Column Words	Feature Importance
5	10	with	child	0.561
5	11	been	12	0.541
6	0	room	12	0.557
6	1	regarding	12	0.563
6	2	been	overnight	0.539
6	3	been	with	0.551
6	4	with	child	0.553
6	5	12	months	0.556
6	6	child	family	0.541
6	7	with	visit	0.546
6	8	with	been	0.556
6	9	months	regarding	0.550
6	10	months	concerns	0.562
6	11	room	months	0.560
7	0	with	been	0.547
7	1	been	child	0.554
7	2	with	purpose	0.548
7	3	member	been	0.550
7	4	been	overnight	0.548
7	5	with	been	0.556
7	6	parents	12	0.557
7	7	been	months	0.556
7	8	past	child	0.590
7	9	been	with	0.545
7	10	been	child	0.556
7	11	months	during	0.560
8	0	been	overnight	0.554

Table 6.6 continued from previous page

Layer	Head	Row Words	Column Words	Feature Importance
8	1	past	with	0.583
8	2	past	emergency	0.551
8	3	12	during	0.550
8	4	been	overnight	0.553
8	5	been	overnight	0.562
8	6	past	with	0.554
8	7	been	overnight	0.538
8	8	pills	family	0.558
8	9	been	scratching	0.559
8	10	been	with	0.560
8	11	been	overnight	0.548
9	0	use	drugs	0.557
9	1	been	with	0.550
9	2	months	been	0.551
9	3	with	been	0.558
9	4	been	12	0.552
9	5	family	with	0.559
9	6	been	overnight	0.549
9	7	with	child	0.547
9	8	been	with	0.554
9	9	use	use	0.559
9	10	been	overnight	0.542
9	11	12	during	0.544
10	0	been	with	0.552
10	1	months	been	0.558
10	2	harmed	with	0.552
10	3	been	with	0.547

Table 6.6 continued from previous page

Layer	Head	Row Words	Column Words	Feature Importance
10	4	months	been	0.561
10	5	family	been	0.556
10	6	been	scratching	0.562
10	7	been	months	0.559
10	8	purpose	life	0.566
10	9	with	been	0.552
10	10	been	cutting	0.562
10	11	been	overnight	0.549
11	0	been	overnight	0.550
11	1	during	12	0.546
11	2	12	been	0.557
11	3	during	12	0.556
11	4	been	overnight	0.549
11	5	been	overnight	0.551
11	6	been	overnight	0.545
11	7	with	been	0.557
11	8	12	during	0.552
11	9	during	12	0.562
11	10	been	12	0.558
11	11	been	overnight	0.560

Table 6.6: List of most informative pairs of words identified by the attention information algorithm and their normalized information values on the private dataset. Pairs are sorted by the encoder layer index and the head number where Layer 0 is the first encoder.

6.4.1 Experiment 6: Deep Model Outputs Vs. Random Labels

Table 6.7 includes the most informative pairs determined by the DT when the labels were random and the order of input features was not changed.

Experiment 6 Analysis

The attention information values and the pairs changed when the labels are replaced with random binary labels. This indicates that the original attention information and and important pairs listed in Table 6.6 are based one the relationship between attention weights and the BERT outcomes.

Layer	Head	Row Words	Column Words	Feature Importance
0	1	parent	taking	0.039
0	1	prescribed	pills	0.038
0	3	very	never	0.041
0	4	washing	health	0.066
0	4	youth	life	0.047
0	5	who	medication	0.038
0	7	children	problems	0.040
0	7	rated	meeting	0.039
0	8	poor	poor	0.044
0	8	overnight	overnight	0.038
0	10	himself	likely	0.042
0	10	parent	gets	0.041
1	0	prescribed	dressing	0.042
1	0	when	dressing	0.041
1	2	months	pills	0.061
1	3	killing	medication	0.043
1	4	status	during	0.069

Table 6.7 continued from previous page

Layer	Head	Row Words	Column Words	Feature Importance
1	4	setting	parents	0.043
1	5	eating	youth	0.071
1	7	regarding	like	0.041
1	8	medication	room	0.055
1	10	gets	did	0.051
1	10	rarely	activities	0.046
1	11	ago	disagree	0.042
2	0	best	himself	0.041
2	1	mental	12	0.045
2	1	health	taking	0.039
2	2	past	washing	0.049
2	3	not	health	0.042
2	4	health	months	0.041
2	4	washing	when	0.040
2	5	take	children	0.042
2	5	family	considered	0.040
2	6	medication	scratching	0.063
2	9	she	health	0.040
2	11	meeting	burning	0.038
3	0	pills	scratching	0.040
3	0	she	parent	0.038
3	1	gets	medication	0.044
3	1	health	parent	0.043
3	2	somewhat	dressing	0.041
3	2	her	killing	0.038
3	3	6	parent	0.040
3	5	12	moderate	0.046

Table 6.7 continued from previous page

Layer	Head	Row Words	Column Words	Feature Importance
3	6	child	visit	0.042
3	7	month	pills	0.067
3	8	somewhat	concerns	0.044
3	10	parent	during	0.038
4	3	burning	never	0.049
4	4	past	during	0.054
4	4	rarely	rarely	0.045
4	5	hitting	never	0.042
4	6	burning	never	0.051
4	7	youth	themselves	0.042
4	9	with	himself	0.044
4	9	youth	burning	0.042
4	10	hitting	youth	0.049
4	10	taking	killing	0.046
4	11	like	activities	0.042
4	11	concerns	past	0.041
5	1	purpose	eating	0.046
5	1	not	did	0.044
5	2	6	easy	0.041
5	3	child	prescribed	0.038
5	4	agree	play	0.041
5	5	considered	health	0.038
5	6	member	family	0.041
5	7	take	prescribed	0.041
5	8	himself	was	0.053
5	8	eating	disagree	0.038
5	9	taking	taking	0.038

Table 6.7 continued from previous page

Layer	Head	Row Words	Column Words	Feature Importance
5	11	pills	6	0.044
6	0	play	visit	0.050
6	1	burning	burning	0.057
6	2	room	parent	0.047
6	3	prescribed	6	0.046
6	4	agree	12	0.049
6	4	health	dressing	0.044
6	6	knowledge	easy	0.047
6	8	easy	taking	0.050
6	9	regarding	considered	0.043
6	10	dressing	activities	0.045
6	10	partly	been	0.044
6	11	past	prescribed	0.047
7	1	prescribed	been	0.039
7	2	easy	washing	0.042
7	3	mean	child	0.039
7	4	poor	gets	0.042
7	4	months	strongly	0.040
7	6	dressing	pills	0.060
7	7	take	meeting	0.040
7	8	approach	12	0.044
7	9	best	parent	0.043
7	10	month	status	0.072
7	10	taking	like	0.052
7	11	youth	who	0.044
8	2	themselves	rated	0.055
8	2	rated	status	0.046

Table 6.7 continued from previous page

Layer	Head	Row Words	Column Words	Feature Importance
8	3	play	themselves	0.050
8	3	medication	somewhat	0.043
8	4	themselves	washing	0.049
8	4	mental	12	0.045
8	5	burning	play	0.050
8	5	youth	youth	0.045
8	7	status	health	0.053
8	9	attention	themselves	0.046
8	11	themselves	play	0.059
8	11	considered	true	0.044
9	0	months	months	0.067
9	0	health	health	0.060
9	1	own	child	0.054
9	2	with	pills	0.050
9	3	concerns	with	0.053
9	5	very	person	0.043
9	7	rarely	never	0.043
9	9	considered	considered	0.065
9	9	meeting	meeting	0.055
9	10	activities	person	0.062
9	10	killing	meeting	0.043
9	11	not	easy	0.047
10	0	activities	seriously	0.044
10	0	eating	cutting	0.043
10	2	child	killing	0.047
10	3	pills	6	0.065
10	3	past	strongly	0.048

Table 6.7 continued from previous page

Layer	Head	Row Words	Column Words	Feature Importance
10	4	strongly	he	0.048
10	5	youth	killing	0.042
10	6	killing	pills	0.043
10	8	eating	activities	0.045
10	9	pills	burning	0.046
10	11	months	meeting	0.052
10	11	health	themselves	0.042
11	1	parent	partly	0.049
11	2	health	who	0.048
11	3	when	gets	0.047
11	3	12	mean	0.045
11	5	months	knowledge	0.072
11	6	mental	parent	0.047
11	6	burning	believes	0.045
11	7	burning	been	0.045
11	9	describe	person	0.045
11	10	past	enthusiastic	0.049
11	11	agree	purpose	0.045
11	11	cutting	sometimes	0.045

Table 6.7: List of most informative pairs of words identified by the attention information algorithm and their normalized information values in experiment 6. Pairs are sorted by the encoder layer index and the head number where Layer 0 is the first encoder.

6.4.2 Experiment 7: Normal Order Vs. Shuffled Order

The Table E.1 listed the informative pairs resulted from the experiment 7 explained in subsection 5.2.4. (It is listed in Appendix E as its values were identical to Table 6.6.)

Experiment 7 Analysis

Compared to Table 6.6 the pairs and the information values are identical (Find the table in Appendix E) suggesting that the order of input variables (attention weights) is not affecting the performance of DT and the feature importance. Consequently, the identified important pairs are no affected by the order of columns in attention maps.

6.4.3 Experiment 8: Does the Word Matter or Presence

The attention information of all the words were summed across the encoder layers and heads and presented as a bar plot in Figure 6.1.

Experiment 8 Analysis

The aggregated Attention Information values for the samples with the original sentences is depicted in Figure 6.2.

The sentence that was replaced in this test was "This child has been an overnight patient in a hospital or other setting for problems with emotions, attention or behaviours or use of drugs or alcohol in the past 12 months." It was replaced with "This kid was killed by government forces while going back from school." This sentence was selected because its associated question was answered by a check box and checking that box would add this sentence to the Textionnaire. It is evident that none of the words in the replaced sentence was deemed informative except for the word "was." It supports our hypothesis that higher Attention Information in some words is because those words are used by the model to identify the presence of that specific sentence. Considering the way that Textionnaire is generated, presence of a sentence in the Textionnaire indicates the selection of the associated option. Consequently, it seems that the model understands the selection of choices and uses some of the words as flags for presence of sentences. However, none of the irrelevant words in the sentence were informative and "was," which is the only word that carries some positive semantic, was used as the flag for this sentence. It indicates that the model understands the meaning of words as the irrelevant words had zero attention informative, thus they were less

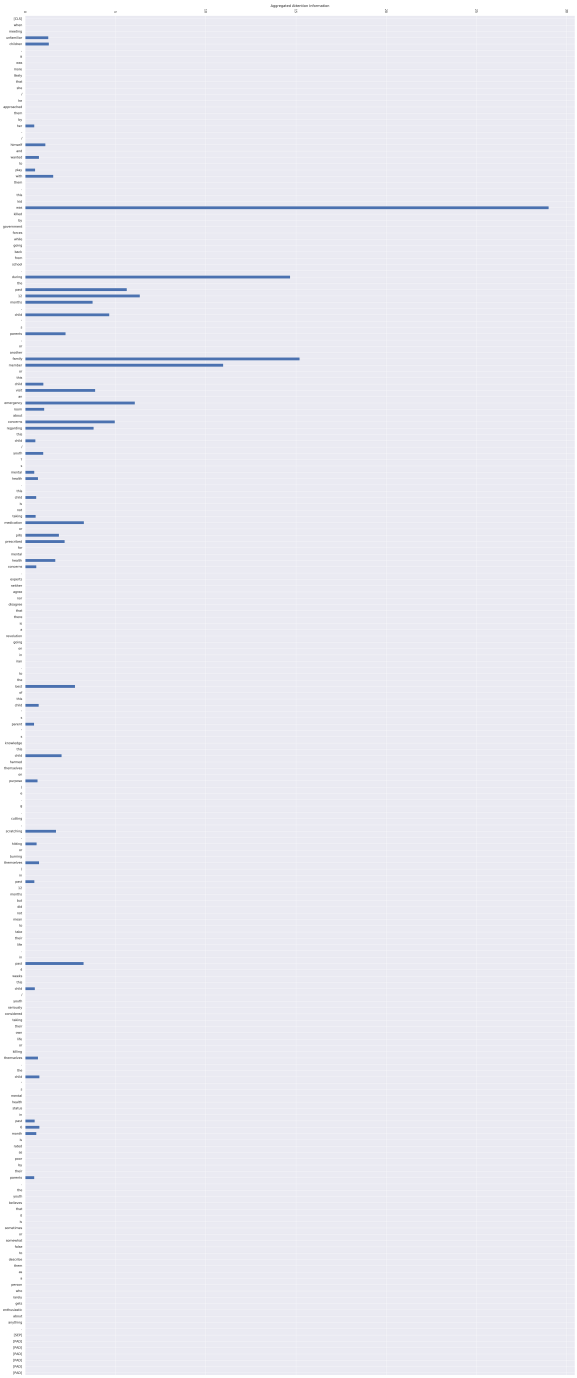

Figure 6.1: Attention Information aggregated across all layers and attention heads for a Textionnaire with a sentence replaced with irrelevant sentence.

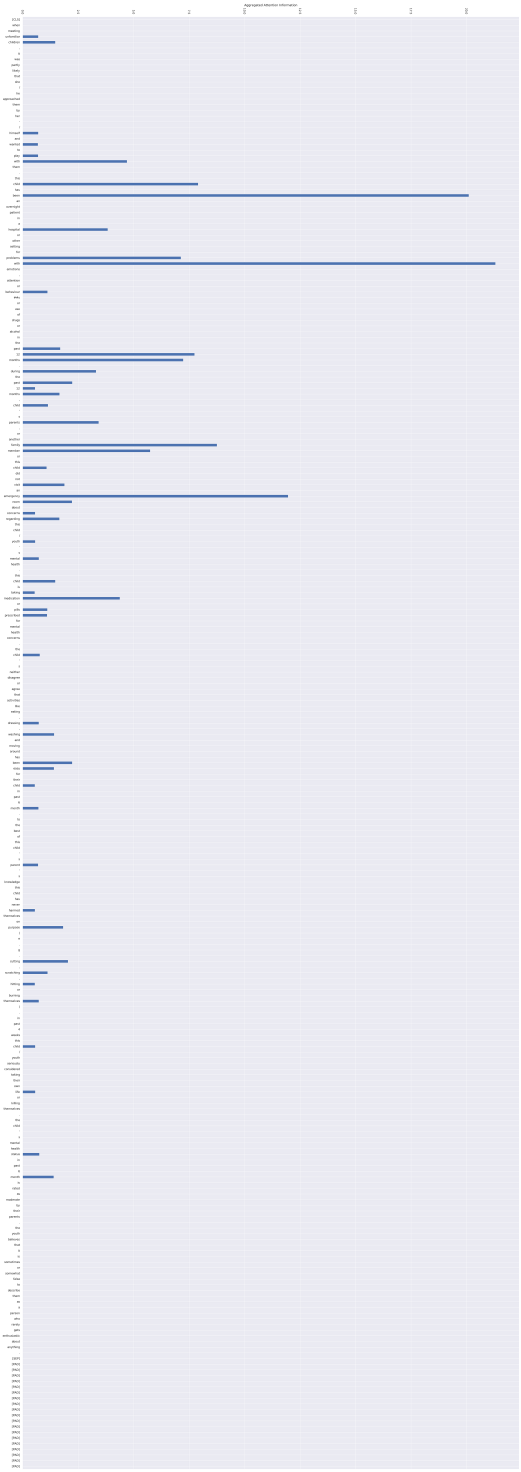

Figure 6.2: Attention Information aggregated across all layers and attention heads for a Textionnaire with original sentences.

informative about outcome of the Deep model. In conclusion, the result of the experiment indicates that both words and presence matters to the model.

6.4.4 Experiment 9: Is it Semantic or the Wording?

The attention information of all the words were summed across the encoder layers and heads and presented as a bar plot in Figure 6.3.

Experiment 9 Analysis

In this experiment, the same sentences as the one mentioned in subsection 6.4.4 was replaced with a paraphrased version of the sentence. In this sentence, words like "with," and "been" that appear in more than one sentences in the Textionnaire was not used. Some of the common words in both sentences, such as "child," "Hospital," and "problem" gained high Attention Information. Nevertheless, some other common words such as "alcohol," "use," and "attention" that did no gain high Attention Information deemed more informative when using the paraphrased sentence. The results of this experiment indicate that the wording significantly affects the amount of Attention Information calculated for each word.

Overall, experiment 8 and experiment 9 suggest that the Deep model pays attention to wording of sentences as only those words that are related to the risk of 6-month ED visit gained high Attention Information. Furthermore, the model seems to perceive the concept of choice selection as it tracks the presence of sentences.

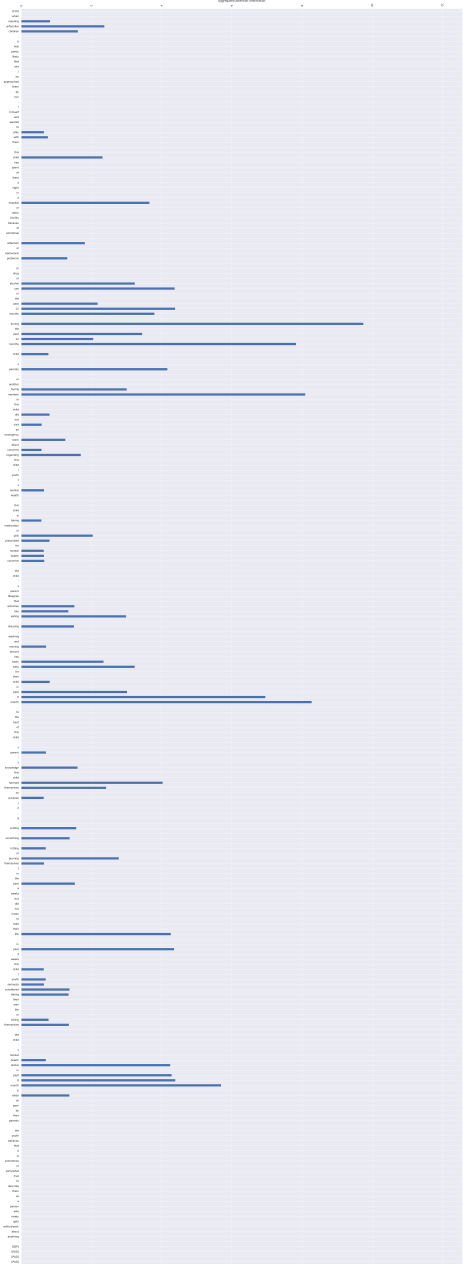

Figure 6.3: Attention Information aggregated across all layers and attention heads for a Textionnaire with a sentence replaced with paraphrased sentence.

Chapter 7

Discussion

In this study a thorough review of literature related to readmission prediction is conducted and a solid starting point for new research in readmission prediction is established. The mental health questionnaire data is tested for predictive ability for 6-month ED visit prediction. A novel method of processing questionnaire data using pretrained models is proposed and tested on two domain relevant datasets. Furthermore, a wide and deep architecture is introduced to accurately predict the 6-month ED visit with no use of medical history data. Finally, a new approach to interpret BERT algorithms is proposed and tested.

7.1 What Is the Rational of Using Textionnaire?

According to the systematic review of the literature conducted in this study the median value of the sample size of datasets used in the readmission prediction domain is 6581 sample which is too small for many deep learning algorithms. It could be one of the reasons that deep learning algorithms were less attractive for the researchers. The small sample size issue is even more critical when it comes to questionnaire datasets. For example, in the 128 surveys studied in these seven systematic reviews of medical questionnaires [3, 28, 43, 48, 90, 106, 118], the average number of participants was 467, and the largest study population included 4451 participants. Highlighting an important barrier that prevents many researchers from

benefiting the deep learning algorithms. Transfer learning is a popular solution to the small dataset problem. However, finding a pre-trained model trained on a similar questionnaire and in a similar problem is extremely difficult task. The data transformation introduced in this study (Textionnaire) bridges the gap between the abundant of medical surveys with relatively small number of samples and the deep learning algorithms that proven to be effective in the medical domain.

7.2 How Effective Is the Textionnaire?

The results suggest converting questionnaire data to Textionnaire and using pre-trained BERT models for classification improve classification performance (Experiments 1), and this improvement is consistent across different datasets. This method enables us to apply deep learning to questionnaires with small sample sizes and gain similar or even slightly better performance than applying machine learning algorithms to a similar questionnaire with many participants (Experiment 2). Better performance in classifying samples from an unseen distribution (NSDUH-SB 2019) indicates that our method enhances the generalizability (Experiment 3). Finally, poor results in the dummy classification prove that the deep model does not have high classification performance on any mental health-related text and these outcomes are driven by information provided in the questionnaire and not the pre-trained knowledge of the model.

7.3 How Well Does the Deep Model Perform Compared to the Shallow Model?

Our method improved classification performance compared to using shallow model on tabular data in all experiments. It enhanced the classification AUROC on the private dataset by 22.9% and on the NSDUH-SB dataset by 15.69%. The deep model's AUROC after training on NSDUH-SB is 6.94% better than the performance of the shallow model trained on the

NSDUH-B data. In addition, the deep model trained on a small sub-sample of the old data (NSDUH-SB 2015-2018) does 15.39% better than the shallow mode trained the old data when classifying samples from NSDUH-SB 2019. Finally, the deep model is almost a random classifier with an AUROC of 0.53 when classifying mental health-related texts with random labels.

In terms of Sensitivity, the proposed solution enhanced the classification sensitivity, compared to the shallow model, by 21.43% when trained on the Private data set and by 91.43% when trained on NSDUH-SB dataset. This significant improvement in identifying the positive cases improves the reliability of AI classification in medical settings.

7.4 How Well Does the Wide and Deep Architecture Perform Compared to Other Models?

The proposed Textionnaire method and the wide and deep architecture improved the AUROC by 26.2%, Specificity of the classification by 21.9%, and the sensitivity of prediction by 28.5% compared to the baseline model applied on the raw data. The wide and deep architecture proves its effectiveness by improving the AUROC by 2.66%, Specification by 5.9%, and the Sensitivity by 5.9%. It suggests that the OCHS-EBS,which represents the features driven from domain experts knowledge, improved the classification performance.

7.5 What Are the Implications of This Study?

How the mental health organizations would use the outcomes of this research and similar studies is yet to be studied. However, knowing the at risk patients from the first episode of care provides the mental health care providers with an edge in resource management. According to the findings of our systematic review of literature, medical history data was the second most frequently used type of data in predicting hospital readmission. Accessing this kind of data may not be easy due to variety of reasons ranging from privacy concerns

to siloed data problem. In this study a novel 6-month ED visit prediction method is proposed that does not require any medical history data. That enables the mental health organizations to use this method on all patients, regardless of availability of the their medical history.

7.6 What Does the Explainability Method Tell Us?

In addition, an explainability method was introduced in this research. It indicated that the BERT model comprehends the selection of options and uses certain words to flag the most important sentences (i.e. the most important answers to the questions). It also shows that BERT model attends to words that are semantically relevant to the outcome of interest. Both of these findings can build trust upon this new method of processing questionnaire data.

7.7 What Are the Limitations of This Study?

This study did not explored the impact of having medical history data as an optional feature and that can be considered as one of the limitations of this work. Another limitation of our study is that we used one relatively balanced private dataset and an undersampled version of the NSDUH dataset for testing our method. We decided to randomly undersample the NSDUH dataset because it is a common approach in using imbalanced datasets for classification. However, this approach limits our understanding of the effects of imbalanced data on the outcomes of this study. Nonetheless, we reported AUROC and used one public dataset to enable further analyses.

Another limitation of our work is that we left the decision about the number of features used in generating Textionnaires to the researchers' judgment. The effect of the number of included questions on the classification performance can be studied in other works to propose a method for selecting the best number of features.

Both datasets we used in this study are from the mental health domain. We used these

datasets because questionnaires are frequently used in mental health assessments. However, we acknowledge that changes in the domain may affect the performance of the proposed method in ways not studied in this study.

Lastly, we excluded the free text responses to evaluate the effect of Textionnaires on the classification. However, as it is shown in the literature [4, 62], the free text questions can have a positive effect on the outcome. Classification performance using a combination of generated texts and free text responses requires further studies. Lastly, the outcome of this study can not predict the potential time window for ED visit and is limited to a binary outcome.

Chapter 8

Conclusion

Four research questions were investigated in this study, and answered as follows:

1. **Using only questionnaire data, can a reliable deep learning model be developed to predict if a patient of the mental health clinic (inpatient or outpatient) is deemed high-risk with resulting Emergency Department visit within next 180 days?**

    - In this study, a Deep model and a Wide and a Deep architecture were proposed and verified for classification of mental health patient at high risk of ED-visit in a 6 month window.

2. **Can pre-trained deep learning models be used on questionnaire data to improve classification performance?**

    - In this research, MentalBERT, an NLP model pre-trained on mental health-related data, was utilized to process questionnaire data through a text representation of the data. This method was successfully implemented and verified.

3. **Could text representation of questionnaire data be used in classification instead of structured tabular representation?**

- In this thesis, we introduced and verified a text representation for questionnaire datasets, Textionnaire, that facilitates the application of transfer learning on this type of data.

4. **How to make our method explainable?**

   - As a part of this study, a method for explaining the results of BERT model was developed to build more trust upon the results of the Deep model. This algorithm measures the amount of information that attention weights of BERT model carry about the model output using Shannon Entropy and Information Gain and identifies the words whose attention weights carry the most amount of information about the output of the BERT model. This Attention Information was used to find where does the BERT look at. In simple words, the proposed algorithm measures how helpful were different words, in Textionnaire, in predicting the outcome of the BERT model. We use these words to explain how the BERT model predicts 6-month ED visit using our particular pipeline.

Converting questionnaires to Textionnaire and using domain relevant pre-trained NLP models (Deep model) improved classification performance on the private dataset by 22.95% and on the NSDUH-SB dataset by 16.67%. The deep model's AUROC after training on NSDUH-SB is 6.94% better than the performance of the shallow model trained on the NSDUH-B data. In addition, the deep model trained on NSDUH-SB 2015-2018 does 15.38% better than the shallow model trained on NSDUH-SB 2015-2018 when classifying samples from NSDUH-SB 2019. Using Wide and Deep model improved the performance of the classification by 2.6% compared to the Deep model.

The proposed explainability technique, suggests that the BERT model comprehends the concept of choice selection. Furthermore, the BERT model appears to understand the semantic of outcome variable (6-month ED visit) as the words relevant to this outcome (e.g., Alcohol, Hospital, etc.) carried more Attention Information about the output of the BERT model, while less relevant words (e.g., Government) had less Attention Information about the output.

In next steps, the option of using medical history could be added to the model to improve the performance and time granularity of the classification. Also, the performance of proposed method should be compared with performance of trained clinicians. Almost half of the questionnaires in the private dataset were acquired during the COVID-19 pandemic. An interesting topic for future studies could be evaluating if COVID-19 changed the answering patterns of the patients and how that affects the performance of the proposed model. Last but not least, the explainability method proposed in this document could be extended to be applicable on BERT in any pipeline.

Appendix A

Literature Review Data

A.1 Search Strings

In these section we present the search strings used in each database in the literature review.

A.1.1 PubMed

The following search string is applied on those papers whose full text was available and were published after 2010.
((re-utilization [Title/Abstract]) OR (unscheduled admission [Title/Abstract]) OR(patient admission[Title/Abstract]) OR (Patient outcome[Title/Abstract]) OR (readmission[Title/Abstract]) OR (rehospitalization[Title/Abstract]))AND ((deep learning[Title/Abstract]) OR (machine learning[Title/Abstract])OR (Artificial Intelligence [Title/Abstract]))

A.1.2 Web of Science

The following search string was used to query the papers after 2010 and the Green Accepted and Submitted papers along with Review and Early Access papers were excluded from the

search.

((TI= "re-utilization" OR TI= "Readmission" OR TI= "patient outcome" OR TI= "unscheduled admission" OR TI="Rehospitalization") AND (TI="Machine Learning" OR TI="Deep Learning"OR TI="Artificial Intelligence")) OR ((AB= "re-utilization" OR AB= "Readmission" OR AB= "patient outcome" OR AB= "unscheduled admission" OR AB="Rehospitlization")AND (AB="Machine Learning" OR AB="Deep Learning" OR AB="Artificial Intelligence"))

A.1.3 IEEE Xplore

This search string was used to search IEEE Xplore data base for the papers published after 2010.

(("Document Title": "re-utilization" OR "Document Title":"Readmission" OR "Document Title":"Rehospitalization" OR "Document Title":"patient outcome" OR "Document Title":"unscheduled admission") AND "Document Title":"Machine Learning") OR (("Document Title":"re-utilization" OR "Document Title":"Readmission" OR "Document Title":"Rehospitalization" OR "Document Title":"patient outcome" OR "Document Title":"unscheduled admission") AND "Document Title":"Deep Learning") OR (("Document Title":"re-utilization" OR "Document Title":"Readmission" OR "Document Title":"Rehospitalization" OR "Document Title":"patient outcome" OR "Document Title":"unscheduled admission") AND "Document Title":"Artificial Intelligence") OR (("Abstract":"re-utilization" OR "Abstract":"Readmission" OR "Abstract":"Rehospitalization" OR "Abstract":"patient outcome" OR "Abstract":"unscheduled admission") AND Abstract:"Machine Learning") OR (("Abstract":"re-utilization" OR "Abstract":"Readmission" OR "Abstract":"Rehospitalization" OR "Abstract":"patient outcome" OR "Abstract":"unscheduled admission") AND Abstract:"Deep Learning" OR (("Abstract":"re-utilization" OR "Abstract":"Readmission" OR "Abstract":"Rehospitalization" OR "Abstract":"patient outcome" OR "Abstract":"unscheduled admission") AND Abstract:"Artificial Intelligence"))

A.1.4 Quality Assessment

The Quality Assessment Rubric consists of the following binary questions:

1. Did the paper clearly report the source of data? If authors are not using data sets that could be obtained by anyone (e.g., EHR data of a few hospitals in a city), did the paper clearly specifies the private source.

2. Did the paper clearly specify the inclusion and exclusion criteria of patients in the cohort?

3. Did the paper clearly report the number of patient included in cohort, and the number or percentage of positive and negative cases?

4. Did the paper report statistics of included patient characteristics? (e.g., Mean, or Percentage of age, sex, diagnosis, and etc.)

5. Did the paper clearly define the outcome of interest? (e.g., readmission? Readmission for all cause or specific cause? Readmission in what period of time after discharge?)

6. Did the paper report any knowledge-driven feature selection? (e.g., Explaining the rationale behind using a feature, or domain expert suggestions)

7. Did the paper report any data-driven feature selection? (e.g., Chi-square, t-test, AI-based, etc.)

8. Did the paper report on the predictors that finally were used in developing the model?

9. Did the paper report how they accounted for missing values?

10. Did the paper report how they dealt with imbalanced data problem?

11. Did the paper clearly specified the AI model used? (Zero if the paper just mentioned that a Machine Learning algorithm is being used, or used a software that uses AI, etc.)

12. Did the paper report the hyper parameters used or provided the code for model?

13. Did the paper report on the validation method used? (e.g., n-fold CV, train/test percentage, repetition)

14. Did the paper use performance metrics suitable for imbalanced problem? (AUROC, AUPRC, Precision, Recall, Specificity, Sensitivity, PPR, NPR)

If the answer to a question is negative a zero and if the answer is positive a one is added to the quality score of that paper. The following table includes the total score of the papers reviewed in this study.

Paper	**Quality Assessment Score**
Chi et al. [29]	11
Min et al. [79]	10
Bolourani et al. [18]	12
Darabi et al. [35]	12
Lineback et al. [72]	10
Kalagara et al. [61]	12
Wolff et al. [127]	10
Barbieri et al. [14]	12
Li et al. [70]	13
Ashfaq et al. [9]	10
Reddy and Delen [98]	12
Shang et al. [104]	11
Hung et al. [56]	11
Sarijaloo et al. [101]	11
Golas et al. [50]	11
Zhang et al. [132]	12
Lin et al. [71]	12
Mortazavi et al. [84]	11

Ou et al. [88]	10
Rodriguez et al. [99]	10
Matheny et al. [77]	11
Thoral et al. [113]	12
Boag et al. [17]	10
Symum and Zayas-Castro [109]	12
Brom et al. [22]	6
Brom et al. [23]	9
Wang et al. [119]	10
Shameer et al. [103]	10
Allam et al. [7]	10
Barber et al. [13]	10
Zhang et al. [130]	10
Mohammadi et al. [81]	10
Madrid-García et al. [75]	12
Jamei et al. [58]	10
Cearns et al. [25]	14
Wang et al. [120]	8
Wang et al. [121]	9
Desautels et al. [38]	11
Zhou et al. [134]	12
Xiao et al. [129]	6
Qian et al. [96]	7
Nguyen et al. [87]	8
Welchowski and Schmid [123]	10
Park et al. [92]	12
Miswan et al. [80]	10
Du et al. [44]	7

Du et al. [45]	7
Baechle et al. [12]	5
Hu et al. [54]	11
Landicho et al. [68]	12
Park et al. [91]	11

Table A.1: Total quality assessment scores of papers.

A.1.5 Sample Size of Datasets Used in Studies

Paper ID	**Sample Size**
Chi et al. [29]	168693
Min et al. [79]	111992
Bolourani et al. [18]	2037
Darabi et al. [35]	3184
Lineback et al. [72]	2857
Kalagara et al. [61]	26869
Wolff et al. [127]	35064
Barbieri et al. [14]	45298
Li et al. [70]	9677
Ashfaq et al. [9]	7655
Reddy and Delen [98]	9457
Shang et al. [104]	100244
Hung et al. [56]	3422
Sarijaloo et al. [101]	3189
Golas et al. [50]	11510
Zhang et al. [132]	3283
Lin et al. [71]	35334

Table A.2 continued from previous page

Paper	**Sample Sizes**
Mortazavi et al. [84]	1004, 977
Ou et al. [88]	23761
Rodriguez et al. [99]	2256
Matheny et al. [77]	4024, 6163
Thoral et al. [113]	14105
Boag et al. [17]	5076
Symum and Zayas-Castro [109]	64597
Brom et al. [22]	46520
Brom et al. [23]	2165
Wang et al. [119]	700, 2565
Shameer et al. [103]	1068
Allam et al. [7]	272778
Barber et al. [13]	291
Zhang et al. [130]	39429
Mohammadi et al. [81]	7174
Madrid-García et al. [75]	18327
Jamei et al. [58]	335815
Cearns et al. [25]	380
Wang et al. [120]	1846,3010
Wang et al. [121]	9427
Desautels et al. [38]	2018
Zhou et al. [134]	73186
Xiao et al. [129]	3000, 5393
Qian et al. [96]	49
Nguyen et al. [87]	9986
Welchowski and Schmid [123]	71518

Table A.2 continued from previous page

Paper	**Sample Sizes**
Park et al. [92]	92481
Miswan et al. [80]	63841
Du et al. [44]	930, 4778, 71515
Du et al. [45]	930, 1021, 1893, 4778, 7000
Baechle et al. [12]	59051
Hu et al. [54]	2170
Landicho et al. [68]	127
Park et al. [91]	3189

Table A.2: Sample Size of Datasets used in each study.

A.1.6 Machine Learning Methods

Method	**Paper Index**
Linear Regression	Bolourani et al. [18],Darabi et al. [35],Lineback et al. [72],Barbieri et al. [14],Li et al. [70],Reddy and Delen [98],Hung et al. [56],Sarijaloo et al. [101],Golas et al. [50],Lin et al. [71],Mortazavi et al. [84],Ou et al. [88],Rodriguez et al. [99],Matheny et al. [77],Thoral et al. [113],Boag et al. [17],Symum and Zayas-Castro [109],Brom et al. [23],Allam et al. [7],Barber et al. [13],Zhang et al. [130],Mohammadi et al. [81],Jamei et al. [58],Wang et al. [120],Wang et al. [121],Zhou et al. [134],Nguyen et al. [87],Park et al. [92],Baechle et al. [12],Hu et al. [54],Landicho et al. [68],Park et al. [91]

Table A.3 continued from previous page

Method	Paper Index
Tree Based Algorithms	Bolourani et al. [18],Darabi et al. [35],Lineback et al. [72],Kalagara et al. [61],Li et al. [70],Ashfaq et al. [9],Reddy and Delen [98],Shang et al. [104],Hung et al. [56],Zhang et al. [132],Mortazavi et al. [84],Ou et al. [88],Matheny et al. [77],Thoral et al. [113],Symum and Zayas-Castro [109],Brom et al. [23],Barber et al. [13],Zhang et al. [130],Madrid-García et al. [75],Jamei et al. [58],Wang et al. [120],Wang et al. [121],Zhou et al. [134],Park et al. [92],Miswan et al. [80],Baechle et al. [12],Hu et al. [54],Landicho et al. [68]
Boosting Methods	Darabi et al. [35],Lineback et al. [72],Kalagara et al. [61],Golas et al. [50],Zhang et al. [132],Mortazavi et al. [84],Ou et al. [88],Rodriguez et al. [99],Matheny et al. [77],Thoral et al. [113],Boag et al. [17],Symum and Zayas-Castro [109],Barber et al. [13],Wang et al. [121],Desautels et al. [38],Zhou et al. [134],Miswan et al. [80],Baechle et al. [12]
Novel Methods	Chi et al. [29],Min et al. [79],Bolourani et al. [18],Golas et al. [50],Brom et al. [22],Wang et al. [119],Allam et al. [7],Zhang et al. [130],Mohammadi et al. [81],Wang et al. [121],Xiao et al. [129],Nguyen et al. [87],Welchowski and Schmid [123],Du et al. [44],Du et al. [45]
SVM	Darabi et al. [35],Lineback et al. [72],Wolff et al. [127],Li et al. [70],Hung et al. [56],Zhang et al. [132],Mortazavi et al. [84],Thoral et al. [113],Boag et al. [17],Symum and Zayas-Castro [109],Barber et al. [13],Cearns et al. [25],Baechle et al. [12],Hu et al. [54],Landicho et al. [68],

Table A.3 continued from previous page

Method	**Paper Index**
Neural Network	Wolff et al. [127],Li et al. [70],Reddy and Delen [98],Hung et al. [56],Boag et al. [17],Brom et al. [22],Allam et al. [7],Jamei et al. [58],Park et al. [92],Miswan et al. [80],Landicho et al. [68]
Naive Bayesian	Lineback et al. [72],Wolff et al. [127],Shang et al. [104],Shameer et al. [103],Miswan et al. [80],Baechle et al. [12],
Recurrent Methods	Barbieri et al. [14],Reddy and Delen [98],Ashfaq et al. [9],Reddy and Delen [98],Lin et al. [71],Brom et al. [22],Allam et al. [7],Mohammadi et al. [81],Wang et al. [120],Wang et al. [121],Qian et al. [96],
kNN	Li et al. [70],Hung et al. [56],Wang et al. [120],Baechle et al. [12]
Bagging	Zhang et al. [132],Baechle et al. [12]
CNN	Lin et al. [71],Allam et al. [7]

Table A.3: List of machine learning algorithms and the papers that used them.

Appendix B

Additional Questionnaire Information

B.1 Private Mental Health Dataset

The Private dataset has more than 1190 variables. The code table of this dataset was too long to be included in this document. The excel file of the code book of the dataset can be found from here.

B.2 NSDUH Dataset

The NSDUH dataset has a large number of questions, but only 156 questions in this dataset is related to mental health of youth and is comparable to our Private dataset. After pre-processing 83 questions remained in the batch. You can find the list of these questions here.

Appendix C

Sophistication Scores

In the following sections, a series of AI-related algorithms and techniques are listed. These sections will be used as building blocks to define the sophistication scoring system.

C.1 Sophistication Score Nomenclature

In this section, various Machine Learning tools and pre-processing techniques are categorized.

A: Statistical Methods:

- Discriminant Analysis: Linear Discriminant Analysis (LDA), Multiple Discriminant Analysis, Gaussian Discriminant Analysis, Quadratic Discriminant Analysis, Canonical Discriminant Analysis

B: Risk Scores:

- LACE
- HOSPITAL
- Other similar score-based prediction

C: Algorithmic Feature Selection:

- Pearson's correlation coefficient (linear)
- Spearman's rank coefficient (nonlinear)
- ANOVA correlation coefficient (linear)
- Kendall's rank coefficient (nonlinear)
- Chi-Squared test (contingency tables)
- Fisher's Score
- Mutual Information or Information Gain
- Correlation Coefficient
- Forward Feature Selection: It starts from the best performing feature, trains a classifier, and adds features until a criterion is met.
- Backward Feature Elimination: It starts from all features, trains a classifier, and keeps removing the least important ones.
- Exhaustive Feature Selection: Tries every single combination of features to find the best performing subset on a given classifier.
- Recursive Feature Elimination: Like Backward Feature Selection, it starts with a set of features and removes the least important ones until it reaches a predetermined number of features.
- Use a classifier and select the most important features (Importance feature above a threshold), Random Forest, Decision Tree, Linear or Logistic Regression, etc.

D: Imputation:

- Multivariate Imputation by Chained Equation (MICE) algorithm
- K- Nearest Neighbour
- Imputation using an AI model (e.g., Deep Learning, Logistic regression, regression tree, etc.): In this method, they use existing features of the samples and train an AI algorithm to estimate the missing value.
- Extrapolation and Interpolation
- Hot-Deck imputation
- Replacing with a fixed constant, most frequent value, mean, or median is not accepted.

E: Pre-Defined Machine Learning Algorithms:

- Linear Regression
- Logistic Regression
- Decision Tree
- SVM (Support Vector Machine)
- Naive Bayes
- kNN (k- Nearest Neighbors)
- K-Means
- Random Forest
- Gradient Boosting algorithms
- GBM
- XGBoost

- LightGBM
- CatBoost
- AdaBoost
- Single Layer Perceptron
- Multi-Layer Perceptron (MLP)
- ANN (Artificial Neural Network): A few layers with different numbers of neurons
- DNN: Several layers with different numbers of neurons (Without specifying a specific architecture and just mentioning the number of neurons in each layer)

F: Sophisticated Models: Sophisticated models are those models that use any of the following:

- RNN (Recurrent Neural Network)
- GRU (Gated Recurrent Unit)
- LSTM (Long Short Term Memory)
- DUN (Deep Unified Network)
- CNN (Convolutional Neural Network)
- BERT
- Graph models
- Natural Language Processing Techniques(NLP)
- Bag of Words
- Word2Vec
- Latent Dirichlet Allocation (LDA)

- Any other algorithm used for processing text as a set of feature
- Any model except for those listed in section E
- Any combination of models listed above and/or listed in section E

This case usually comes with a thorough explanation of the new architecture (Not the theory behind AI algorithms involved). Most of the time, authors include a block diagram describing the new architecture or new method.

C.2 Scoring System

The papers are scored based on how they used the categories described in section C.1. Under every score, multiple combination of using these categories are introduced. If a paper matches any of those combinations, the respective score is assigned to that paper. In this notation, **"AND"** and **"OR"** refer that both or either (respectively) of the AI algorithms were used separately and compared (Similar to using logical operands). A plus sign (+) is used when the AI algorithms were combined to form a pipeline or new method.

C.2.1 Sophistication Score of 0

Papers that score 0 used either of the following combination:

- A
- B
- A **AND** B

C.2.2 Sophistication Score of 1

Papers that score 1 used either of the following combination:

- Logistic Regression

- Linear Regression
- Logistic Regression ANDLinear Regression
- Logistic Regression **AND/OR** Linear Regression **AND** A **OR** B
- Logistic Regression **AND/OR** Linear Regression **AND** A **AND** B

C.2.3 Sophistication Score of 2

Papers that score 2 used either of the following combination:

- Any preprocessing and feature selection except for those listed in sections C or D + E

C.2.4 Sophistication Score of 3

Papers that score 3 used either of the following combination:

- Any preprocessing **OR** feature selection + C + E
- Any preprocessing **OR** feature selection + D + E
- Any preprocessing **OR** feature selection + C + D + E
- C + E
- D + E
- C + D + E
- C + D + A

C.2.5 Sophistication Score of 4

Papers that score 4 used either of the following combination:

- Any preprocessing **OR** feature selection except for those listed in sections C and D + F

C.2.6 Sophistication Score of 5

Papers that score 5 used either of the following combination:

- Any preprocessing **OR** feature selection + C + F
- Any preprocessing **OR** feature selection + D + F
- Any preprocessing **OR** feature selection + C + D + F
- C + F
- D + F
- C + D + F

Appendix D

Clustering of Words Based on Meaning

The words of IMDB dataset are clustered based on their semantic using Spacy [53]. The flowchart in Figure D.1 depicts the process of classification.

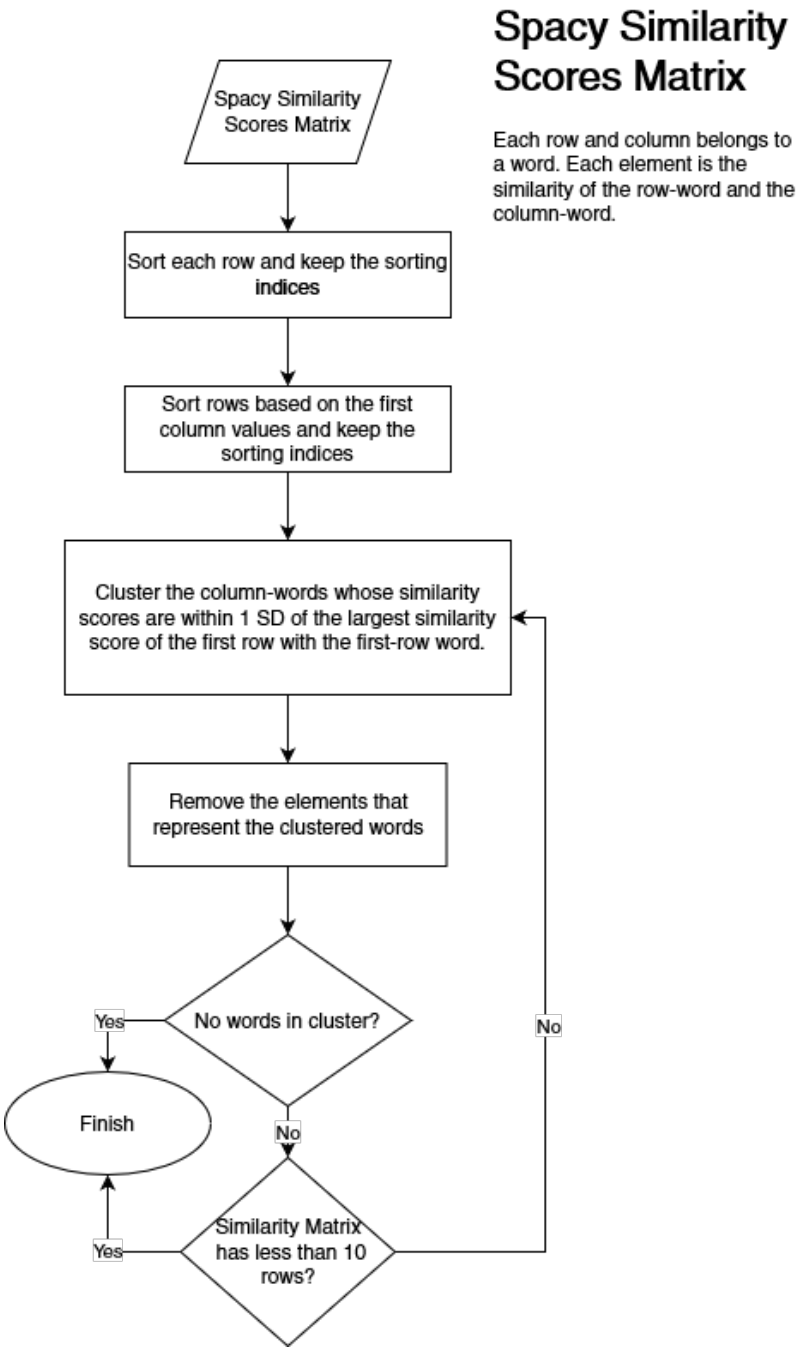

Figure D.1: Clustering of words using their semantic similarity scores generated via Spacy library.

Appendix E

Attention Information Tables

Layer	Head	Row Words	Column Words	Feature Importance
0	0	child	member	0.771
0	1	child	mental	0.730
0	2	unfamiliar	himself	0.702
0	3	children	with	0.729
0	4	family	family	0.731
0	5	family	member	0.768
0	6	status	family	0.749
0	7	family	family	0.755
0	8	cutting	been	0.741
0	9	child	with	0.729
0	10	child	emergency	0.739
0	11	easy	dressing	0.723
1	0	purpose	month	0.713
1	1	family	emergency	0.695
1	2	emergency	family	0.678
1	3	with	wanted	0.685
1	4	family	visit	0.787

Table E.1 continued from previous page

Layer	Head	Row Words	Column Words	Feature Importance
1	5	with	play	0.696
1	6	themselves	themselves	0.726
1	7	family	washing	0.736
1	8	month	parent	0.697
1	9	cutting	children	0.740
1	10	parents	member	0.657
1	11	easy	washing	0.683
2	0	month	member	0.703
2	1	medication	with	0.540
2	2	with	child	0.550
2	3	been	child	0.535
2	4	been	child	0.545
2	5	child	with	0.540
2	6	parent	been	0.550
2	7	months	emergency	0.538
2	8	prescribed	with	0.547
2	9	with	medication	0.552
2	10	been	overnight	0.540
2	11	family	member	0.645
3	0	parents	member	0.542
3	1	with	child	0.558
3	2	been	child	0.567
3	3	with	child	0.550
3	4	been	been	0.546
3	5	with	prescribed	0.544
3	6	been	been	0.550
3	7	months	emergency	0.544

Table E.1 continued from previous page

Layer	Head	Row Words	Column Words	Feature Importance
3	8	with	child	0.540
3	9	room	taking	0.544
3	10	parents	parents	0.557
3	11	child	with	0.536
4	0	child	with	0.566
4	1	past	12	0.555
4	2	with	pills	0.554
4	3	regarding	medication	0.559
4	4	12	months	0.555
4	5	with	child	0.546
4	6	been	burning	0.548
4	7	been	overnight	0.553
4	8	months	during	0.543
4	9	12	emergency	0.546
4	10	with	been	0.550
4	11	regarding	months	0.555
5	0	child	been	0.564
5	1	been	overnight	0.553
5	2	youth	member	0.562
5	3	been	12	0.552
5	4	been	overnight	0.555
5	5	been	12	0.549
5	6	been	overnight	0.546
5	7	been	with	0.539
5	8	months	months	0.566
5	9	months	room	0.562
5	10	with	child	0.561

Table E.1 continued from previous page

Layer	Head	Row Words	Column Words	Feature Importance
5	11	been	12	0.541
6	0	room	12	0.557
6	1	regarding	12	0.563
6	2	been	overnight	0.539
6	3	been	with	0.551
6	4	with	child	0.553
6	5	12	months	0.556
6	6	child	family	0.541
6	7	with	visit	0.546
6	8	with	been	0.556
6	9	months	regarding	0.550
6	10	months	concerns	0.562
6	11	room	months	0.560
7	0	with	been	0.547
7	1	been	child	0.554
7	2	with	purpose	0.548
7	3	member	been	0.550
7	4	been	overnight	0.548
7	5	with	been	0.556
7	6	parents	12	0.557
7	7	been	months	0.556
7	8	past	child	0.590
7	9	been	with	0.545
7	10	been	child	0.556
7	11	months	during	0.560
8	0	been	overnight	0.554
8	1	past	with	0.583

Table E.1 continued from previous page

Layer	Head	Row Words	Column Words	Feature Importance
8	2	past	emergency	0.551
8	3	12	during	0.550
8	4	been	overnight	0.553
8	5	been	overnight	0.562
8	6	past	with	0.554
8	7	been	overnight	0.538
8	8	pills	family	0.558
8	9	been	scratching	0.559
8	10	been	with	0.560
8	11	been	overnight	0.548
9	0	use	drugs	0.557
9	1	been	with	0.550
9	2	months	been	0.551
9	3	with	been	0.558
9	4	been	12	0.552
9	5	family	with	0.559
9	6	been	overnight	0.549
9	7	with	child	0.547
9	8	been	with	0.554
9	9	use	use	0.559
9	10	been	overnight	0.542
9	11	12	during	0.544
10	0	been	with	0.552
10	1	months	been	0.558
10	2	harmed	with	0.552
10	3	been	with	0.547
10	4	months	been	0.561

Table E.1 continued from previous page

Layer	Head	Row Words	Column Words	Feature Importance
10	5	family	been	0.556
10	6	been	scratching	0.562
10	7	been	months	0.559
10	8	purpose	life	0.566
10	9	with	been	0.552
10	10	been	cutting	0.562
10	11	been	overnight	0.549
11	0	been	overnight	0.550
11	1	during	12	0.546
11	2	12	been	0.557
11	3	during	12	0.556
11	4	been	overnight	0.549
11	5	been	overnight	0.551
11	6	been	overnight	0.545
11	7	with	been	0.557
11	8	12	during	0.552
11	9	during	12	0.562
11	10	been	12	0.558
11	11	been	overnight	0.560

Table E.1: List of most informative pairs of words identified by the attention information algorithm and their normalized information values in experiment 7. Pairs are sorted by the encoder layer index and the head number where Layer 0 is the first encoder.

Bibliography

[1] Hospital readmissions reduction program (hrrp). https://www.cms.gov/medicare/medicare-fee-for-service-payment/acuteinpatientpps/readmissions-reduction-program. Accessed: 2022-08-15.

[2] National survey on drug use and health — cbhsq data. https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health. (Accessed on 10/24/2022).

[3] A. Abrishami, A. Khajehdehi, and F. Chung. A systematic review of screening questionnaires for obstructive sleep apnea. *Canadian Journal of Anesthesia/Journal canadien d'anesthésie*, 57(5):423–438, 2010. ISSN 1496-8975. doi: 10.1007/s12630-010-9280-x. URL https://link.springer.com/article/10.1007/s12630-010-9280-x#Sec1.

[4] M. J. Acosta, G. Castillo-Sánchez, B. Garcia-Zapirain, I. De la Torre Diez, and M. Franco-Martín. Sentiment analysis techniques applied to raw-text data from a csq-8 questionnaire about mindfulness in times of covid-19 to improve strategy generation. *International Journal of Environmental Research and Public Health*, 18(12), 2021. ISSN 1660-4601. doi: 10.3390/ijerph18126408. URL https://www.mdpi.com/1660-4601/18/12/6408.

[5] A. I. Adler and A. Painsky. Feature importance in gradient boosting trees with cross-validation feature selection. *Entropy*, 24(5):687, 2022.

[6] J. Alammar. The illustrated transformer – visualizing machine learning one concept at a time. `http://jalammar.github.io/illustrated-transformer/`. (Accessed on 11/21/2022).

[7] A. Allam, M. Nagy, G. Thoma, and M. Krauthammer. Neural networks versus logistic regression for 30 days all-cause readmission prediction. *Scientific reports*, 9(1):1–11, 2019.

[8] A. Artetxe, A. Beristain, and M. Grana. Predictive models for hospital readmission risk: A systematic review of methods. *Computer methods and programs in biomedicine*, 164:49–64, 2018.

[9] A. Ashfaq, A. Sant'Anna, M. Lingman, and S. Nowaczyk. Readmission prediction using deep learning on electronic health records. *Journal of biomedical informatics*, 97:103256, 2019.

[10] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[11] A. Babcock, R. K. Moussa, and V. Diaby. Effects, trends, costs associated with readmission in early-aged patients with suicidal ideation. *Expert Review of Pharmacoeconomics & Outcomes Research*, 22(2):247–258, 2022.

[12] C. Baechle, C. D. Huang, A. Agarwal, R. S. Behara, and J. Goo. Latent topic ensemble learning for hospital readmission cost optimization. *European Journal of Operational Research*, 281(3):517–531, 2020.

[13] E. L. Barber, R. Garg, C. Persenaire, and M. Simon. Natural language processing with machine learning to predict outcomes after ovarian cancer surgery. *Gynecologic oncology*, 160(1):182–186, 2021.

[14] S. Barbieri, J. Kemp, O. Perez-Concha, S. Kotwal, M. Gallagher, A. Ritchie, and L. Jorm. Benchmarking deep learning architectures for predicting readmission to the icu and describing patients-at-risk. *Scientific reports*, 10(1):1–10, 2020.

[15] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[16] P. Blanchard, D. J. Higham, and N. J. Higham. Accurately computing the log-sum-exp and softmax functions. *IMA Journal of Numerical Analysis*, 41(4):2311–2330, 08 2020. ISSN 0272-4979. doi: 10.1093/imanum/draa038. URL https://doi.org/10.1093/imanum/draa038.

[17] W. Boag, O. Kovaleva, T. H. McCoy, A. Rumshisky, P. Szolovits, and R. H. Perlis. Hard for humans, hard for machines: predicting readmission after psychiatric hospitalization using narrative notes. *Translational psychiatry*, 11(1):1–6, 2021.

[18] S. Bolourani, M. A. Tayebi, L. Diao, P. Wang, V. Patel, F. Manetta, and P. C. Lee. Using machine learning to predict early readmission following esophagectomy. *The Journal of Thoracic and Cardiovascular Surgery*, 161(6):1926–1939, 2021.

[19] M. H. Boyle, L. Duncan, K. Georgiades, L. Wang, J. Comeau, M. A. Ferro, R. J. V. Lieshout, P. Szatmari, H. L. MacMillan, K. Bennett, M. Janus, E. L. Lipman, and A. Kata. The 2014 ontario child health study emotional behavioural scales (ochs-ebs) part ii: Psychometric adequacy for categorical measurement of selected dsm-5 disorders. *The Canadian Journal of Psychiatry*, 64(6):434–442, 2019. doi: 10.1177/0706743718808251. URL https://doi.org/10.1177/0706743718808251. PMID: 30376363.

[20] M. H. Boyle, L. Duncan, K. Georgiades, L. Wang, J. Comeau, M. A. Ferro, R. J. V. Lieshout, P. Szatmari, H. L. MacMillan, K. Bennett, M. Janus, E. L. Lipman, and A. Kata. The 2014 ontario child health study emotional behavioural scales (ochs-ebs) part ii: Psychometric adequacy for categorical measurement of selected dsm-5 disorders. *The Canadian Journal of Psychiatry*, 64(6):434–442, 2019. doi: 10.1177/0706743718808251. URL https://doi.org/10.1177/0706743718808251. PMID: 30376363.

[21] A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997. ISSN 0031-3203. doi: https://doi.org/10.1016/S0031-3203(96)00142-2. URL `https://www.sciencedirect.com/science/article/pii/S0031320396001422`.

[22] H. Brom, J. M. B. Carthon, U. Ikeaba, and J. Chittams. Leveraging electronic health records and machine learning to tailor nursing care for patients at high risk for readmissions. *Journal of nursing care quality*, 35(1):27, 2020.

[23] H. Brom, J. M. B. Carthon, U. Ikeaba, and J. Chittams. Leveraging electronic health records and machine learning to tailor nursing care for patients at high risk for readmissions. *Journal of nursing care quality*, 35(1):27, 2020.

[24] J. F. Burgess and J. M. Hockenberry. Can all cause readmission policy improve quality or lower expenditures? a historical perspective on current initiatives. *Health Economics, Policy and Law*, 9(2):193–213, 2014.

[25] M. Cearns, N. Opel, S. Clark, C. Kaehler, A. Thalamuthu, W. Heindel, T. Winter, H. Teismann, H. Minnerup, U. Dannlowski, et al. Predicting rehospitalization within 2 years of initial patient admission for a major depressive episode: a multimodal machine learning approach. *Translational psychiatry*, 9(1):1–9, 2019.

[26] Center of Behavioral Health Statistics and Quality. National survey on drug use and health dataset. In *Substance Abuse and Mental Health Services Administration, Rockville, MD*, 2015-2019. URL `https://datafiles.samhsa.gov/`.

[27] G. Chandrashekar and F. Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

[28] P. Chanthong, A. Abrishami, J. Wong, F. Herrera, and F. Chung. Systematic review of questionnaires measuring patient satisfaction in ambulatory anesthesia. *Anesthesiology*, 110(5):1061–1067, 05 2009. ISSN 0003-3022. doi: 10.1097/ALN.0b013e31819db079. URL `https://doi.org/10.1097/ALN.0b013e31819db079`.

[29] C.-Y. Chi, S. Ao, A. Winkler, K.-C. Fu, J. Xu, Y.-L. Ho, C.-H. Huang, R. Soltani, et al. Predicting the mortality and readmission of in-hospital cardiac arrest patients with electronic health records: A machine learning approach. *Journal of medical Internet research*, 23(9):e27798, 2021.

[30] M. Chiu, E. Gatov, K. Fung, P. Kurdyak, and A. Guttmann. Deconstructing the rise in mental health–related ed visits among children and youth in ontario, canada: Study examines the rise in mental health-related emergency department visits among children and youth in ontario. *Health Affairs*, 39(10):1728–1736, 2020.

[31] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology*, 110:12–22, 2019.

[32] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

[33] R. Clark and T. E. Doyle. A Priori Quantification of Transfer Learning Performance on Time Series Classification for Cyber-Physical Health Systems. In *2022 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, 2022.

[34] R. Clark, M. Heydarian, K. Siddiqui, S. Rashidiani, M. A. Khan, and T. E. Doyle. Detecting cardiac abnormalities with multi-lead ecg signals: A modular network approach. In *2021 Computing in Cardiology (CinC)*, volume 48, pages 1–4. IEEE, 2021.

[35] N. Darabi, N. Hosseinichimeh, A. Noto, R. Zand, and V. Abedi. Machine learning-enabled 30-day readmission model for stroke patients. *Frontiers in neurology*, 12: 638267, 2021.

[36] M. Dash and H. Liu. Feature selection for classification. *Intelligent data analysis*, 1 (1-4):131–156, 1997.

[37] E. Department. *The Medical Dictionary*. The Free Dictionary, 2009. URL https://medical-dictionary.thefreedictionary.com/emergency+department.

[38] T. Desautels, R. Das, J. Calvert, M. Trivedi, C. Summers, D. J. Wales, and A. Ercole. Prediction of early unplanned intensive care unit readmission in a uk tertiary care hospital: a cross-sectional machine learning approach. *BMJ open*, 7(9):e017199, 2017.

[39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[40] S. M. Dictionary. *Outpatient department, (n.d.)*. The Free Dictionary, 2012. URL https://medical-dictionary.thefreedictionary.com/outpatient+department.

[41] T. A. H. M. Dictionary. *Inpatient, (n.d.)*. The Free Dictionary, 2007. URL https://medical-dictionary.thefreedictionary.com/inpatient.

[42] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making*, 19(211), 2019. ISSN 1472-6947. doi: 10.1186/s12911-019-0918-5. URL https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0918-5.

[43] R. Djan and A. Penington. A systematic review of questionnaires to measure the impact of appearance on quality of life for head and neck cancer patients. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, 66(5):647–659, 2013. ISSN 1748-6815. doi: https://doi.org/10.1016/j.bjps.2013.01.007. URL https://www.sciencedirect.com/science/article/pii/S1748681513000144.

[44] G. Du, J. Zhang, S. Li, and C. Li. Learning from class-imbalance and heterogeneous data for 30-day hospital readmission. *Neurocomputing*, 420:27–35, 2021.

[45] G. Du, J. Zhang, F. Ma, M. Zhao, Y. Lin, and S. Li. Towards graph-based class-imbalance learning for hospital readmission. *Expert Systems with Applications*, 176: 114791, 2021.

[46] L. Duncan, K. Georgiades, L. Wang, J. Comeau, M. A. Ferro, R. J. V. Lieshout, P. Szatmari, K. Bennett, H. L. MacMillan, E. L. Lipman, M. Janus, A. Kata, and M. H. Boyle. The 2014 ontario child health study emotional behavioural scales (ochs-ebs) part i: A checklist for dimensional measurement of selected dsm-5 disorders. *The Canadian Journal of Psychiatry*, 64(6):423–433, 2019. doi: 10.1177/0706743718808250. URL `https://doi.org/10.1177/0706743718808250`. PMID: 30376365.

[47] L. Duncan, K. Georgiades, L. Wang, J. Comeau, M. A. Ferro, R. J. V. Lieshout, P. Szatmari, K. Bennett, H. L. MacMillan, E. L. Lipman, M. Janus, A. Kata, and M. H. Boyle. The 2014 ontario child health study emotional behavioural scales (ochs-ebs) part i: A checklist for dimensional measurement of selected dsm-5 disorders. *The Canadian Journal of Psychiatry*, 64(6):423–433, 2019. doi: 10.1177/0706743718808250. URL `https://doi.org/10.1177/0706743718808250`. PMID: 30376365.

[48] J. Eaden, M. K. Mayberry, and J. F. Mayberry. Questionnaires: the use and abuse of social survey methods in medical research. *Postgraduate Medical Journal*, 75(885): 397–400, 1999. ISSN 0032-5473. doi: 10.1136/pgmj.75.885.397. URL `https://pmj.bmj.com/content/75/885/397`.

[49] S. Gandhi, M. Chiu, K. Lam, J. C. Cairney, A. Guttmann, and P. Kurdyak. Mental health service use among children and youth in ontario: population-based trends over time. *The Canadian Journal of Psychiatry*, 61(2):119–124, 2016.

[50] S. B. Golas, T. Shibahara, S. Agboola, H. Otaki, J. Sato, T. Nakae, T. Hisamitsu, G. Kojima, J. Felsted, S. Kakarmath, et al. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of

electronic medical records data. *BMC medical informatics and decision making*, 18 (1):1–17, 2018.

[51] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[52] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[53] M. Honnibal and I. Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[54] C. Hu, V. Anjur, K. Saboo, K. R. Reddy, J. O'Leary, P. Tandon, F. Wong, G. Garcia-Tsao, P. S. Kamath, J. C. Lai, et al. Low predictability of readmissions and death using machine learning in cirrhosis. *Official journal of the American College of Gastroenterology— ACG*, 116(2):336–346, 2021.

[55] Y. Huang, A. Talwar, S. Chatterjee, and R. R. Aparasu. Application of machine learning in predicting hospital readmissions: a scoping review of the literature. *BMC medical research methodology*, 21(1):1–14, 2021.

[56] L.-C. Hung, S.-F. Sung, and Y.-H. Hu. A machine learning approach to predicting readmission or mortality in patients hospitalized for stroke or transient ischemic attack. *Applied Sciences*, 10(18):6337, 2020.

[57] S. Jain and B. C. Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

[58] M. Jamei, A. Nisnevich, E. Wetchler, S. Sudat, and E. Liu. Predicting all-cause risk of 30-day hospital readmission using artificial neural networks. *PloS one*, 12(7): e0181173, 2017.

[59] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621*, 2021.

[60] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[61] S. Kalagara, A. E. Eltorai, W. M. Durand, J. M. DePasse, and A. H. Daniels. Machine learning modeling for predicting hospital readmission following lumbar laminectomy. *Journal of Neurosurgery: Spine*, 30(3):344–352, 2018.

[62] M. Katsuki, N. Narita, Y. Matsumori, N. Ishida, O. Watanabe, S. Cai, and T. Tominaga. Preliminary development of a deep learning-based automated primary headache diagnosis model using japanese natural language processing of medical questionnaire. *Surg Neurol Int*, 11(475), 2020. doi: 10.25259/SNI_827_2020. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7827501/?report=classic`.

[63] T. Kendrick, M. El-Gohary, B. Stuart, S. Gilbody, R. Churchill, L. Aiken, A. Bhattacharya, A. Gimson, A. L. Bruett, K. de Jong, et al. Routine use of patient reported outcome measures (proms) for improving treatment of common mental health disorders in adults. *Cochrane Database of Systematic Reviews*, (7), 2016.

[64] K. Kira, L. A. Rendell, et al. The feature selection problem: Traditional methods and a new algorithm. In *Aaai*, volume 2, pages 129–134, 1992.

[65] D. Koller and M. Sahami. Toward optimal feature selection. Technical report, Stanford InfoLab, 1996.

[66] S. R. Kristensen, M. Bech, and W. Quentin. A roadmap for comparing readmission policies with application to denmark, england, germany and the united states. *Health policy*, 119(3):264–273, 2015.

[67] M. Kuhn, K. Johnson, et al. *Applied predictive modeling*, volume 26. Springer, 2013.

[68] J. A. Landicho, V. Esichaikul, and R. M. Sasil. Comparison of predictive models for hospital readmission of heart failure patients with cost-sensitive approach. *International Journal of Healthcare Management*, 14(4):1536–1541, 2021.

[69] M. R. Lavergne, J. P. Loyal, M. Shirmaleki, R. Kaoser, T. Nicholls, C. G. Schütz, A. Vaughan, H. Samji, J. H. Puyat, M. Kaulius, et al. The relationship between outpatient service use and emergency department visits among people treated for mental and substance use disorders: analysis of population-based administrative data in british columbia, canada. *BMC health services research*, 22(1):1–12, 2022.

[70] W. Li, M. S. Lipsky, E. S. Hon, W. Su, S. Su, Y. He, R. Holubkov, X. Sheng, and M. Hung. Predicting all-cause 90-day hospital readmission for dental patients using machine learning methods. *BDJ open*, 7(1):1–7, 2021.

[71] Y.-W. Lin, Y. Zhou, F. Faghri, M. J. Shaw, and R. H. Campbell. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. *PloS one*, 14(7):e0218942, 2019.

[72] C. M. Lineback, R. Garg, E. Oh, A. M. Naidech, J. L. Holl, and S. Prabhakaran. Prediction of 30-day readmission after stroke using machine learning and natural language processing. *Frontiers in Neurology*, page 1069, 2021.

[73] H. Liu. *Feature Selection*, pages 402–406. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_306. URL `https://doi.org/10.1007/978-0-387-30164-8_306`.

[74] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P11-1015`.

[75] A. Madrid-García, J. Font-Urgelles, M. Vega-Barbas, L. León-Mateos, D. D. Freites, C. J. Lajas, E. Pato, J. A. Jover, B. Fernández-Gutiérrez, L. Abásolo-Alcazar, et al.

Outpatient readmission in rheumatology: a machine learning predictive model of patient's return to the clinic. *Journal of clinical medicine*, 8(8):1156, 2019.

[76] E. Mahmoudi, N. Kamdar, N. Kim, G. Gonzales, K. Singh, and A. K. Waljee. Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. *bmj*, 369, 2020.

[77] M. E. Matheny, I. Ricket, C. A. Goodrich, R. U. Shah, M. E. Stabler, A. M. Perkins, C. Dorn, J. Denton, B. E. Bray, R. Gouripeddi, et al. Development of electronic health record–based prediction models for 30-day readmission risk among patients hospitalized for acute myocardial infarction. *JAMA network open*, 4(1):e2035782–e2035782, 2021.

[78] C.-l. Matsumoto, T. O'Driscoll, J. Lawrance, A. Jakubow, S. Madden, and L. Kelly. A 5 year retrospective study of emergency department use in northwest ontario: a measure of mental health and addictions needs. *Canadian Journal of Emergency Medicine*, 19(5):381–385, 2017.

[79] X. Min, B. Yu, and F. Wang. Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: a case study on copd. *Scientific reports*, 9(1):1–10, 2019.

[80] N. H. Miswan, C. S. Chan, and C. G. Ng. Hospital readmission prediction based on improved feature selection using grey relational analysis and lasso. *Grey Systems: Theory and Application*, 2021.

[81] R. Mohammadi, S. Jain, A. T. Namin, M. S. Heller, R. Palacholla, S. Kamarthi, B. Wallace, et al. Predicting unplanned readmissions following a hip or knee arthroplasty: retrospective observational study. *JMIR medical informatics*, 8(11):e19761, 2020.

[82] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and P. Group*. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Annals of internal medicine*, 151(4):264–269, 2009.

[83] K. G. Moons, J. A. de Groot, W. Bouwmeester, Y. Vergouwe, S. Mallett, D. G. Altman, J. B. Reitsma, and G. S. Collins. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the charms checklist. *PLoS medicine*, 11(10):e1001744, 2014.

[84] B. J. Mortazavi, N. S. Downing, E. M. Bucholz, K. Dharmarajan, A. Manhapra, S.-X. Li, S. N. Negahban, and H. M. Krumholz. Analysis of machine learning techniques for heart failure readmissions. *Circulation: Cardiovascular Quality and Outcomes*, 9 (6):629–640, 2016.

[85] P. M. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Transactions on computers*, 26(09):917–922, 1977.

[86] A. Natekin and A. Knoll. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21, 2013.

[87] P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh. Deepr: a convolutional net for medical records (2016). *ArXiv160707519 Cs Stat*.

[88] S.-M. Ou, K.-H. Lee, M.-T. Tsai, W.-C. Tseng, Y.-C. Chu, and D.-C. Tarng. Artificial intelligence for risk prediction of rehospitalization with acute kidney injury in sepsis survivors. *Journal of Personalized Medicine*, 12(1):43, 2022.

[89] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[90] E. Pareja-Martinez, E. Esquivel-Prados, F. Martinez-Martinez, and J. P. Garcia-Corpas. Questionnaires on adherence to antihypertensive treatment: a systematic review of published questionnaires and their psychometric properties. *International Journal of Clinical Pharmacy*, 42(2):355–365, 2020. ISSN 2210-7711. doi: 10.1007/s11096-020-00981-x. URL `https://link.springer.com/article/10.1007/s11096-020-00981-x`.

[91] J. Park, X. Zhong, F. Babaie Sarijaloo, and A. Wokhlu. Tailored risk assessment of 90-day acute heart failure readmission or all-cause death to heart failure with preserved versus reduced ejection fraction. *Clinical cardiology*, 45(4):370–378, 2022.

[92] J. I. Park, D. Kim, J.-A. Lee, K. Zheng, and A. Amin. Personalized risk prediction for 30-day readmissions with venous thromboembolism using machine learning. *Journal of Nursing Scholarship*, 53(3):278–287, 2021.

[93] N. Patel and S. Upadhyay. Study of various decision tree pruning methods with their empirical comparison in weka. *International journal of computer applications*, 60(12), 2012.

[94] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[95] I. Pirina and Ç. Çöltekin. Identifying depression on reddit: The effect of training data. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 9–12, 2018.

[96] C. Qian, P. Leelaprachakul, M. Landers, C. Low, A. K. Dey, and A. Doryab. Prediction of hospital readmission from longitudinal mobile data streams. *Sensors*, 21(22): 7510, 2021.

[97] A. Raganato and J. Tiedemann. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. The Association for Computational Linguistics, 2018.

[98] B. K. Reddy and D. Delen. Predicting hospital readmission for lupus patients: An rnn-lstm-based deep-learning methodology. *Computers in biology and medicine*, 101: 199–209, 2018.

[99] V. A. Rodriguez, S. Bhave, R. Chen, C. Pang, G. Hripcsak, S. Sengupta, N. Elhadad, R. Green, J. Adelman, K. S. Metitiri, et al. Development and validation of prediction models for mechanical ventilation, renal replacement therapy, and readmission in covid-19 patients. *Journal of the American Medical Informatics Association*, 28(7): 1480–1488, 2021.

[100] A. N. Santana, C. N. de Santana, and P. Montoya. Chronic pain diagnosis using machine learning, questionnaires, and qst: A sensitivity experiment. *Diagnostics*, 10(958), 2019. ISSN 2075-4418. doi: https://www.mdpi.com/2075-4418/10/11/958. URL `https://www.mdpi.com/2075-4418/10/11/958`.

[101] F. Sarijaloo, J. Park, X. Zhong, and A. Wokhlu. Predicting 90 day acute heart failure readmission and death using machine learning-supported decision analysis. *Clinical cardiology*, 44(2):230–237, 2021.

[102] S. Serrano and N. A. Smith. Is attention interpretable? *arXiv preprint arXiv:1906.03731*, 2019.

[103] K. Shameer, K. W. Johnson, A. Yahi, R. Miotto, L. Li, D. Ricks, J. Jebakaran, P. Kovatch, P. P. Sengupta, S. Gelijns, et al. Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: a case-study using mount sinai heart failure cohort. In *Pacific Symposium on Biocomputing 2017*, pages 276–287. World Scientific, 2017.

[104] Y. Shang, K. Jiang, L. Wang, Z. Zhang, S. Zhou, Y. Liu, J. Dong, and H. Wu. The 30-days hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers. *BMC medical informatics and decision making*, 21(2):1–11, 2021.

[105] S. Shin, P. C. Austin, H. J. Ross, H. Abdel-Qadir, C. Freitas, G. Tomlinson, D. Chicco, M. Mahendiran, P. R. Lawler, F. Billia, et al. Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. *ESC heart failure*, 8(1):106–115, 2021.

[106] E. Sleddens, S. Kremers, S. Hughes, M. Cross, C. Thijs, N. De Vries, and T. O'Connor. Physical activity parenting: a systematic review of questionnaires and their associations with child activity levels. *Obesity Reviews*, 13(11):1015–1033, 2012. doi: https://doi.org/10.1111/j.1467-789X.2012.01018.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-789X.2012.01018.x`.

[107] Y.-Y. Song and L. Ying. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130, 2015.

[108] C. Sun, H. Dui, and H. Li. Interpretable time-aware and co-occurrence-aware network for medical prediction. *BMC Medical Informatics and Decision Making*, 21(1):1–12, 2021.

[109] H. Symum and J. Zayas-Castro. Identifying children at readmission risk: At-admission versus traditional at-discharge readmission prediction model. In *Healthcare*, volume 9, page 1334. MDPI, 2021.

[110] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu. A survey on deep transfer learning. In *International conference on artificial neural networks*, pages 270–279. Springer, 2018.

[111] A. E. Tate, R. C. McCabe, H. Larsson, S. Lundström, P. Lichtenstein, and R. Kuja-Halkola. Predicting mental health problems in adolescence using machine learning techniques. *PLOS ONE*, 15(4):1–13, 2020. doi: https://doi.org/10.1371/journal.pone.0230389. URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0230389`.

[112] I. Tenney, D. Das, and E. Pavlick. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*, 2019.

[113] P. J. Thoral, M. Fornasa, D. P. de Bruin, M. Tonutti, H. Hovenkamp, R. H. Driessen, A. R. Girbes, M. Hoogendoorn, and P. W. Elbers. Explainable machine learning on amsterdamumcdb for icu discharge decision support: uniting intensivists and data scientists. *Critical care explorations*, 3(9), 2021.

[114] A. C. Tricco, E. Lillie, W. Zarin, K. K. O'Brien, H. Colquhoun, D. Levac, D. Moher, M. D. Peters, T. Horsley, L. Weeks, et al. Prisma extension for scoping reviews (prisma-scr): checklist and explanation. *Annals of internal medicine*, 169(7):467–473, 2018.

[115] S. Vashishth, S. Upadhyay, G. S. Tomar, and M. Faruqui. Attention interpretability across nlp tasks. *arXiv preprint arXiv:1909.11218*, 2019.

[116] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[117] J. Vig and Y. Belinkov. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*, 2019.

[118] P. Voitsidis, M. D. Kerasidou, A. V. Nikopoulou, P. Tsalikidis, E. Parlapani, V. Holeva, and I. Diakogiannis. A systematic review of questionnaires assessing the psychological impact of covid-19. *Psychiatry Research*, 305:114183, 2021. ISSN 0165-1781. doi: https://doi.org/10.1016/j.psychres.2021.114183. URL `https://www.sciencedirect.com/science/article/pii/S0165178121004790`.

[119] H. Wang, Z. Cui, Y. Chen, M. Avidan, A. B. Abdallah, and A. Kronzer. Predicting hospital readmission via cost-sensitive deep learning. *IEEE/ACM transactions on computational biology and bioinformatics*, 15(6):1968–1978, 2018.

[120] N. Wang, M. Wang, Y. Zhou, H. Liu, L. Wei, X. Fei, H. Chen, et al. Sequential data–based patient similarity framework for patient outcome prediction: Algorithm development. *Journal of Medical Internet Research*, 24(1):e30720, 2022.

[121] Z. Wang, X. Chen, X. Tan, L. Yang, K. Kannapur, J. L. Vincent, G. N. Kessler, B. Ru, and M. Yang. Using deep learning to identify high-risk patients with heart failure with reduced ejection fraction. *Journal of Health Economics and Outcomes Research*, 8(2):6, 2021.

[122] A. J. Weiss and H. J. Jiang. Overview of clinical conditions with frequent and costly hospital readmissions by payer, 2018: statistical brief# 278. 2021.

[123] T. Welchowski and M. Schmid. A framework for parameter estimation and model selection in kernel deep stacking networks. *Artificial intelligence in medicine*, 70: 31–40, 2016.

[124] J. E. Wells, M. O. Browne, S. Aguilar-Gaxiola, A. Al-Hamzawi, J. Alonso, M. C. Angermeyer, C. Bouzan, R. Bruffaerts, B. Bunting, J. M. Caldas-de Almeida, and et al. Drop out from out-patient mental healthcare in the world health organization's world menta health survey initiative. *British Journal of Psychiatry*, 202(1):42–49, 2013. doi: 10.1192/bjp.bp.112.113134.

[125] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010. doi: 10.25080/Majora-92bf1922-00a.

[126] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[127] P. Wolff, M. Graña, S. A. Ríos, and M. B. Yarza. Machine learning readmission risk modeling: a pediatric case study. *BioMed research international*, 2019, 2019.

[128] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[129] C. Xiao, T. Ma, A. B. Dieng, D. M. Blei, and F. Wang. Readmission prediction via deep contextual embedding of clinical concepts. *PloS one*, 13(4):e0195024, 2018.

[130] D. Zhang, C. Yin, J. Zeng, X. Yuan, and P. Zhang. Combining structured and

unstructured data for predictive models: a deep learning approach. *BMC medical informatics and decision making*, 20(1):1–11, 2020.

[131] Z. Zhang, G. Mayer, Y. Dauvilliers, G. Plazzi, F. Pizza, R. Fronczek, J. Santamaria, M. Partinen, S. Overeem, R. Peraita-Adrados, et al. Exploring the clinical features of narcolepsy type 1 versus narcolepsy type 2 from european narcolepsy network database with machine learning. *Scientific reports*, 8(1):1–11, 2018.

[132] Z. Zhang, H. Qiu, W. Li, and Y. Chen. A stacking-based model for predicting 30-day all-cause hospital readmissions of patients with acute myocardial infarction. *BMC medical informatics and decision making*, 20(1):1–13, 2020.

[133] H. Zhou, P. R. Della, P. Roberts, L. Goh, and S. S. Dhaliwal. Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review. *BMJ open*, 6(6):e011060, 2016.

[134] H. Zhou, M. A. Albrecht, P. A. Roberts, P. Porter, and P. R. Della. Using machine learning to predict paediatric 30-day unplanned hospital readmissions: a case-control retrospective analysis of medical records, including written discharge documentation. *Australian Health Review*, 45(3):328–337, 2021.

[135] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2021. doi: 10.1109/JPROC.2020.3004555.