Department of Computer Science

MSc Data Science and Analytics

Academic Year 2022-2023

Predicting Hospital Readmission using Clinical BERT

TANVI MATHUR
2218834

A report submitted in partial fulfilment of the requirement for the degree of Master of Science

Brunel University
Department of Computer Science
Uxbridge, Middlesex UB8 3PH
United Kingdom
Tel: +44 (0) 1895 203397
Fax: +44 (0) 1895 251686

ABSTRACT

Hospital readmission within a short span of discharge is one of the major concerns among patients as well as healthcare professionals. This issue not only impacts the finances of individuals but also showcase how reliable a hospital is in terms of their treatments and diagnosis. The goal of this study is to use machine learning and deep learning methods to predict hospital readmission within 30 days of discharge among patients purely based on the clinical notes provided by doctors. We applied deep learning method **Clinical BERT** to predict hospital readmission using the latest **MIMIC-IV dataset**, further the results were compared with another deep learning model called long short-term memory. The study also implemented two machine learning models (SVM and logistic regression) for better result comparison. Based on the implementation of these models on a small set of MIMIC-IV dataset it was highlighted that machine learning models were efficient in producing accurate results between 80% to 83% while the deep learning model were not giving impressive results due to small size of data. The clinical BERT model was still competent with the other studies and achieved an accuracy of 67.16% while the LSTM model performance was extremely low with just 50.54% accuracy indicating that the complex LSTM model require more training time and data to perform efficiently.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr Alaa Marshan for bringing the weight of his considerable experience and knowledge to this project. The completion of this project within the specified time would not have been possible without his guidance at every step.

I would also like to thank our course leader Dr Alessandro Pandini and professor Dr Sarath Dantu, for constantly reminding about the guidelines to while conducting the research.

Sign in the box below to certify that the work carried out is your own. By signing this box, you are certifying that your dissertation is free from plagiarism. Make sure that you are fully aware of the Department guidelines on plagiarism (see the student handbook). The penalties if you are caught are severe. All material from other sources must be properly referenced and direct quotes must appear in quotation marks.

I certify that the work presented in the dissertation is my own unless referenced.

Signature: TANVI MATHUR

Date: 13-09-2023

TOTAL NUMBER OF WORDS: 12892

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

This study focuses on enhancing healthcare analytics by using vast clinical datasets and fine-tuning Clinical BERT. The study will also implement other machine learning and deep learning models for a better comparison of their performance. It aims to help practitioners manage patient care and reduce hospital readmissions.

1.1 Background

Hospital readmission within 30 days of release is widely used as a factor to determine the quality of treatment for inpatients and is a substantial contributor to growing healthcare expenditures. Almost one-fifth of Medicare users discharged from acute care hospitals are readmitted within 30 days, resulting in extra yearly expenses of several billion dollars (Michailidis et al., 2022). In recent times, researchers have paid special attention to the readmission rate, as it is regarded as a universally appealing and dependable indication of the effectiveness and quality of hospital treatment offered to patients. Readmission rates are a publicly disclosed indicator that may be used to compare hospitals and determine hospital service reimbursement (Hasan et al., 2010). Studies highlight that the trends in readmission rates within 30 days of discharge have remained similar over the years, ranging from 8% to 16%, readmission is most likely to happen within 10 days after discharge. The factors that lead to readmission were categorized into patient-related, disease-related, and system-related factors. Patient-related factors that influence readmission include age, gender, functional score, public financial assistance, and social support, while chronic illness is a major factor in readmission. Additionally, system-related factors such as hospital practice, discharge destination, and post-discharge support are also major causes of readmission (Wong et al., 2010).

Attempts were made to predict hospital readmission using standard methods that typically depend on a limited set of clinical data using simple calculations such as a modified LACE score that examines the length of hospital stay, acuity on admission, comorbidity, and emergency department visits. On the other hand, machine learning and artificial intelligence provide efficient means to predict readmissions, and when researchers compare traditional methods to machine learning approaches, the results are better for later (Hung et al., 2020). The study by Huang et al., (2021) reviewed 43 studies, each using various machine learning algorithms to predict readmission. The most popular algorithms were tree-based methods, including decision trees, random forests, and boosted tree methods. The second most popular algorithm was a neural network using either one hidden layer or multiple hidden layers (Deep Learning). Elastic-net was the third most used ML algorithm, followed by Support Vector Machine (SVM). The study by Huang et al., (2019) used Clinical BERT, which is a transformer-based deep learning model that specializes in clinical notes. It uses the MIMIC-III dataset with clinical notes, which is smaller than the hospital’s internal data. The Clinical BERT model has outperformed deep learning models in predicting readmission and yields a relative increase in recall at a fixed rate of false alarms (Huang et al., 2019). This dissertation predicts hospital readmissions using Clinical BERT and other models. The study sets itself apart from previous research since it utilizes the latest MIMIC-IV dataset.

1.2 Research aim and objectives

Aim

The aim of this research project is to develop a predictive model that utilizes a transformer- based model (Clinical BERT) to analyse hospital stay data and doctor notes to assist healthcare professionals to predict hospital readmission for patients after they are discharged. Further, the developed model will be compared to traditional machine learning model to judge its accuracy considering the limitations of data and computational resources on the new MIMIC-IV dataset.

Objectives

The following are our objectives:

- Review the literature on the use of different machine learning techniques to predict hospital readmission and how Clinical BERT can be a better approach for this prediction.
- Explore different data science methodologies and select the one appropriate for achieving the aim of this project.
- Explore and select the appropriate sub-datasets from the huge MIMIC-IV dataset that can be used to develop the required prediction algorithm.
- Build the model using Clinical BERT to predict hospital readmission for patients and to develop machine learning models like Support Vector Machine, logistic regression, and deep learning models like LSTM for a clear comparison of results.
- Evaluate the model and discuss the research findings.

1.3 Research approach

Over the last several years, the field of data mining has become increasingly popular. Different standard models such as KDD, SEMMA, and CRISP-DM have been defined for data mining, which provide a step-by-step approach to carrying out data mining tasks. By following these models, it becomes easier to implement data mining projects. The methodology used for this research is CRISP-DM (Cross-Industry Standard Process for Data Mining). This section of the paper will highlight CRISP-DM and how it is the best-suited model for achieving the objectives of the project.

CRISP-DM stands for CRoss-Industry Standard Process for Data Mining and comprises a cycle of six stages. It is a structured and comprehensive approach for predictive models as it deals with business understanding, data preparation, modelling, evaluation, and deployment, which are required for validation and accuracy of prediction (Koh and Tan, 2011; Azevedo and Santos, 2008).

- **Business Understanding**: The first step is Business understanding, which aims at identifying the objective of the business or research, assessing its needs, and determining how it can improve the current situation.
- **Data Understanding**: Data is a crucial component of any research, so the next step is to understand the data, which includes exploring the data and verifying its quality to ensure that it’s fit for the use of research.

- **Data Preparation**: Based on the understanding of the data and the research, the data needs to be prepared by performing cleaning and transformation on the data according to the requirements to implement the research.

- **Modelling:** This is the actual data analysis stage where different modelling techniques are used, like online analytics process, traditional methods such as regression, cluster, or discriminant analysis, and non-traditional statistical analysis like a neural network and regression, among others. This range of analysis is not surprising, as data mining has three different disciplines namely database management, statistics, and computer science, it includes artificial intelligence and machine learning.

- **Evaluation:** This is the stage where the results of different models are compared using a common yardstick, such as lift charts, profit charts, or diagnostic classification charts.

- **Deployment:** It relates to the actual implementation of the data mining models.

1.4 Dissertation outline

- This research paper consists of six chapters. The first chapter is an introduction giving an overview of the field of study and its importance.

- The second chapter is a literature review, which sets out to create a basic understanding of hospital readmission and discuss the results of different algorithms already implemented to predict hospital readmission and how the intended research is different from the existing research. It further discusses the challenges that could be faced while developing and implementing the algorithm.

- The third Chapter briefly discusses different methodologies and explains the one most suitable for the prediction algorithm in detail.

- Data Analysis is done in the fourth chapter, which explains the entire life cycle of the methodology used for research.

- In the fifth chapter, we discuss the results obtained using the methodology and modelling technique.

- The final chapter brings the previous analyses together to fulfil the overall purpose of the paper and add a conclusion.

CHAPTER 2: LITERATURE REVIEW

2.1 Overview

This chapter outlines hospital readmission's importance, reviews machine and deep learning approaches, and explores transformer models like BERT and Clinical BERT in readmission prediction.

2.2 Machine Learning for Hospital Readmission Prediction

A hospital readmission refers to when a person is admitted to a hospital again shortly (usually within 30 days) after their initial admission. This readmission can happen for various reasons, whether planned or unplanned, and can occur at the same hospital where they were initially admitted or a different one (Futoma et al., 2015). It is a major concern in the healthcare field as it brings in financial burden for public and private stakeholders and can lead to negative publicity of the hospitals (Jiang et al., 2018). According to the study by (Jiang et al., 2018) approximately 35% of unplanned readmissions in the United Kingdom result in a financial burden of £11 billion annually and lead to heightened risks of patient mortality or extended hospital stays. Another study by Medicare Payment Advisory Committee reports that around 17.6% of hospital admissions resulted in readmissions within 30 days of discharge, and out of these, 76% were potentially preventable. Thus, Readmission rate is used as a quality indicator for the services provided by health institutes because as per Affordable Care Act, hospitals with higher-than-average readmission rates are penalized (Futoma et al., 2015).

Machine learning is a field of artificial intelligence that replicates how people learn through data and algorithms. Over time, it enhances its accuracy as more data passes through it. Machine learning in the field of healthcare utilizes data to improve the effectiveness and overall quality of treatment by reducing the need for human involvement. Several research studies have already been conducted using machine learning to predict hospital readmissions, which will be further analysed in this study. For example, the dataset used for study was collected from multiple sources and patients were referred using anonymised codes, it performed various algorithms on the dataset like Support Vector Machine (SVM) with linear kernel, SVM with Radial Basis Function (RBF) kernel, balanced and weighted random forest were implemented to predict readmission and analyse which model is more accurate. The results indicated that balanced random forest model outperforms others with a sensitivity of 0.70% and an Area Under the Curve (AUC) value of 0.78% (Michailidis et al., 2022). While the study by (Li et al., 2020) categorized the prediction of readmissions based on the timing of the prediction, either at the time of hospital admission or during the discharge process. Thus, data was randomly split into training and testing datasets. The target variable is binary that is 0 & 1, where 1 indicates the patient is readmitted and 0 indicates no readmission (Li et al., 2020). Further Li et al., (2020) compared naive bayes, deep learning, extreme gradient boosting, random forest, and logistic regression with penalization. The results reflect that random forest is most predictive with an AUC greater than 0.79% at the time of hospital admission while during discharge extreme gradient boosting (XGBoost) performs the best with AUC of 0.882% (Li et al., 2020). Datasets mostly have a class imbalance, which was handled by Huang et al., (2022) in his research where under sampling technique was implemented to make the data balanced as it results in less bias and allows computational efficiency. The models were compared using area under the receiver operating curves (AUROC) and area under precision-recall curves (AUPRC) which highlighted that rule-based model outperformed all the other models like decision tree, random forest with AUROC of 0.659% and only XGBoost had a greater AUROC of 0.6606% but the AUPRC was similar for both 0.2147 (Huang et al., 2022). These previous studies highlight how machine learning has the potential to perform readmission prediction, but (Futoma et al., 2015) highlighted that machine learning models can be difficult to tune and are sometimes more challenging to interpret.

2.3 Deep Learning for Hospital Readmission Prediction

In Machine learning algorithms, feature selection is a manual process, which increases the likelihood of overlooking significant features or information. On the other hand, deep learning automatically extracts key features through its algorithm's unique structure. Deep learning is a more efficient approach as it has a layered architecture of artificial neural networks that functions as a human brain. Researchers previously utilized deep learning for hospital readmission, and it will be examined in this section.

Deep learning provides extremely flexible and robust approach for learning complicated, nonlinear decision boundaries between classes and aims at modelling data using nonlinear transformations (Futoma et al., 2015). The building block for these models is a feedforward neural network having three layers, an input layer of observed variables, a hidden layer of unobserved variables, and an output layer consisting of a prediction (Futoma et al., 2015). The studies by Bikram et al., (2022) and Sushrith et al., (2021) used Electronic Health Record (EHR) data from MIMIC-III dataset for majority of studies depending on the specific need the features from the dataset are utilized, readmission is the binary target variable, where 0 indicates negative while 1 represents positive cases (Ashfaq et al., 2019). The dataset is split into training, validation, and testing datasets by using cross validation which leads to different number of items in each fold as one patient may have multiple records (Lin et al., 2019).
The study by Li et al., (2020) where dataset is split based on period includes two sets one during hospital admission and other during discharge. The deep learning model for prediction results for admission and discharge were 0.7898 and 0.88 area under the curve (AUC) respectively and it also highlighted that 'destination post discharge' was not an effective feature in case of deep learning like it was for the tree-based algorithms (Li et al., 2020). Recurrent Neural Network (RNN) is another popular approach that incorporated patient representations (Bikram et al., 2022) while capturing the sequence of visits and is capable of handling temporal data much more efficiently (Sushrith et al., 2021). RNN utilizes previous hidden state (the learned representation or memory from the previous time step) and current input at a certain time to sequentially update the hidden state (Ashfaq et al., 2019) and it's an excellent choice for solving problems where the order of the data matters and is qualified as a choice designed for time series data and NLP (Sushrith et al., 2021). The variants of RNN like long short-term memory (LSTM) and gated recurrent unit (GRU) contain an internal recurrence loop along with three and two gates respectively to control information flow (Ashfaq et al., 2019). The long short-term memory (LSTM) is a sequence-to-sequence prediction which implies that for each network there is an input and output (Ashfaq et al., 2019). These are well suited for generating forecasts using time series data, particularly in the case of clinical measurements, presents challenges due to potential delays of uncertain duration and the presence of missing values within the time series (Sushrith et al., 2021). The LSTM model uses a bidirectional LSTM combined with an additional LSTM layer (Lin et al., 2019), followed by a dense decision layer with one output neuron activated by a

sigmoid function (Sushrith et al., 2021). Another deep learning modelling approach is Convolutional neural network (CNN), which is beneficial for analysing Electronic Health Record (EHR) time series data, that trains the data on both comprehensive and longitudinal. Convolution is applied with respect to time axis and dimensions (Sushrith et al., 2021). The computed feature maps are combined and connected to a densely connected decision layer, consisting of a single output neuron (Lin et al., 2019).

Deep learning models use Area Under the Curve (AUC) and F1 scores as the measure to evaluate the prediction quality of different models, the deep neural networks consistently had 3/5 times better AUC and outperformed the penalized regressions on different conditions with highest overall AUCs (Futoma et al., 2015). The LSTM model had two cases one where data was split into train and test while other case where its split based on time (Ashfaq et al., 2019). The later has a strong discriminating ability, with an AUC of 0.77 and an F1-score of 0.51. We show that combining human and machine derived EHR data with an LSTM network beats models that disregard either of these properties (Ashfaq et al., 2019). The research by (Lin et al., 2019) presented that the basic LSTM model clearly outperforms the conventional machine learning approaches and best baseline regression with L1 regularisation, as the sensitivity increases from 0.525 to 0.540 and 0.596 to 0.611. Next, using the operating cut points with high specificity of 0.85 and 0.8, we find that LSTM+CNN produces the greatest sensitivities of 0.548 (95% CI, 0.522-0.575) and 0.619 (95% CI, 0.597-0.642) both of which were higher than standard machine learning models (Lin et al., 2019).

Based on above discussed research potential of deep learning models can be judged as it boosted the prediction accuracy in statistical approaches, but they are most difficult to manage and interpret due to large numbers of model parameters and complex results to evaluate and high degree of flexibility exhibited by deep learning models like neural network can lead to overfitting.

2.4 Transformers-based Deep Learning Models

The Transformer, a notable deep learning model widely embraced across multiple domains, including natural language processing (NLP), computer vision (CV), and speech processing (Lin et al., 2021), relies on attention mechanisms while eliminating the need for recurrence and convolution to draw global dependencies between input and output (Vaswani et al., 2017). The transformers have three types of attention, these are self-attention, masked self-attention and cross attention. Self-attention, also known as intra-attention, refers to an attention mechanism that connects various positions within a single sequence to generate a representation of the sequence. This technique of self-attention has proven effective in various tasks such as understanding written passages, creating concise summaries, determining textual implications, and developing sentence representations that are not task specific (Cheng et al., 2016). Masked self-attention also known as autoregressive or causal attention is a type of self-attention where the transformer decoder has restricted queries to only attend to key-value pairs up to and including their own position, enabling parallel training. It applies a mask function to the attention matrix, setting to ensuring illegal positions are masked out. On the other hand, cross-attention queries are projected from the outputs of the previous decoder layer while the keys and values are projected using the outputs of the encoder (Lin et al., 2021).

The transformer-based model follows an encoder-decoder architecture, where the encoder takes the vector representations of tokens (words) from an input text and converts them into an internal representation. This internal representation is then utilized by a decoder to generate output sequences, such as translating the text into a different language (Chandra et al., 2023). Figure 2.1 represents the architecture of transformer-based model which comprises of multi-head attention (highlighted in orange) and fully connected feed-forward networks which are multilayer perceptron MLP models (highlighted in blue) used as intermediate components of the model in six layers of both the encoder and decoder blocks (Liu et al., 2021). To generate the input embeddings to the model, two schemes can be used depending on the training dataset, encoding the sentences using byte-pair encoding (Britz et al., 2017) or splitting tokens into word component vocabulary (Wu et al., 2016).

The training procedure is known as self-supervision and two common approaches used, which are autoregressive language modelling, where the model predicts the next token based on previous tokens (Peters et al., 2018) and masked language modelling (MLM), where a portion of tokens are masked, and the model predicts them using information from the unmasked tokens (Devlin et al., 2019).

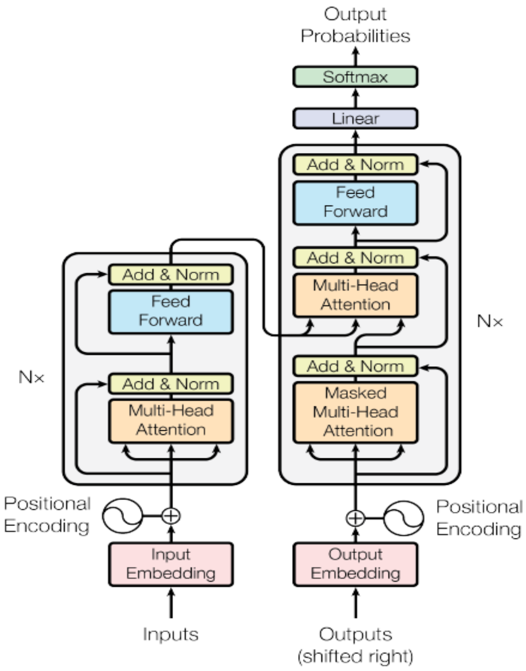


Figure 2.1: The Transformer - model architecture (Vaswani et al., 2017)

Following the initial introduction of the Transformer model, numerous other models have been proposed, utilizing the Transformer Architecture as a foundation. These models share a common characteristic of incorporating attention modules as essential components (Chandra et al., 2023).
Some of these models are:

1. Generative Pre-trained Transformer (GPT)
2. Transformer-XL
3. Text-To-Text Transfer Transformer (T5)
4. Bidirectional Encoder Representations from Transformers (BERT)

The generative pre-trained transformer (GPT) is a groundbreaking advancement in natural language processing, enabling machines to comprehend and communicate in a more human-like manner. It is a deep learning model that undergoes pre-training on extensive collections of text data using unsupervised learning techniques. It can then be fine-tuned for specific tasks such as generating text, analysing sentiment, language modelling, translating text, and classifying text (Yenduri et al., 2023). Through the pre-training phase, the model acquires knowledge and constructs representations of natural language, which can later be adjusted and fine-tuned to suit specific tasks in subsequent stages (Hou and Ji, 2023). GPTs have achieved substantial advancements, with their influence being evident in numerous industries including education, healthcare, manufacturing, agriculture, travel and transportation, e-commerce, entertainment, lifestyle, gaming, marketing, and finance (Yenduri et al., 2023). While GPT models generate coherent text, their contextual understanding is limited due to challenges like semantic comprehension, bias, handling subtleties, and figurative language. This can lead to errors despite grammatical correctness. Researchers are actively exploring ways to enhance contextual understanding for more reliable and accurate results, attracting a larger user base (Liu et al., 2021).

Next is transformer-XL, which enables learning dependency beyond a fixed length without disrupting temporal coherence. This self-attention model stands out as the first to significantly outperform RNNs in both character-level and word-level language modelling tasks, yielding superior results (Dai et al., 2019). Different studies have shown that it outperforms conventional RNN-based models (Al-Rfou et al., 2018). It has applications in various domains such as text generation, unsupervised feature learning, as well as image and speech modelling (Dai et al., 2019).

T5 or Text-To-Text Transfer Transformer is another example that utilizes a unified approach to transfer learning in natural language processing. This framework considers each text processing problem as a "text-to-text" problem, where input text is transformed into output text. By adopting this approach, the same model, objective, training procedure, and decoding process can be applied to all tasks (Raffel et al., 2020; Williams and Zipser, 1989). The main advantage of the text-to-text framework lies in its capacity to train a single model on a wide range of textual tasks using a consistent loss function and decoding procedure. This approach has proven successful in various tasks, including generative tasks such as abstractive summarization, classification tasks like natural language inference, and even regression tasks like STS-B (Raffel et al., 2019). Despite its simplicity, the text-to-text framework outperforms task-specific designs and produces innovative outcomes, especially when paired with volume. Within the study, the Text-to-Text Transfer Transformer (T5) is introduced as the employed model and framework. T5 leverages the text-to-text framework and showcases impressive outcomes in a variety of English-based NLP problems. In conclusion, the text-to-text transformer framework offers a simple yet powerful solution for transfer learning in natural language processing tasks. It enables the training of a single model across multiple tasks, resulting in state-of-the-art performance (Raffel et al., 2019).

Next variant is one of the most popular and latest transformers and it’s called Bidirectional Encoder Representations from Transformers (BERT). It is an encoder based on the Transformer architecture that is trained bidirectionally for language modelling. It utilizes a self-attention mechanism to learn representations of text, offering an alternative to traditional recurrent neural networks (RNNs). BERT-based architectures have emerged as the widely adopted standard for achieving top-notch performance on various natural language processing (NLP) tasks (Kovaleva et al., 2019) which is performed using BERT and it will be discussed in more detail in the coming sections.

While Transformer models have achieved impressive success in natural language processing, researchers are actively addressing several challenges. One challenge involves extending Transformers to handle diverse input and output modalities like images, audio, and video. Adapting the Transformer architecture to effectively process these modalities is an ongoing research area. Another challenge is exploring local attention mechanisms to efficiently handle large inputs and outputs by focusing on relevant information. Making generation less sequential is also important for faster inference and parallel processing. Additionally, applying attention-based models to tasks beyond machine translation, such as speech recognition and sentiment analysis, is an exciting direction of exploration. Overcoming these challenges will enhance the versatility and effectiveness of Transformers in tackling a wider range of problems (Vaswani et al., 2017).

2.4.1 BERT

BERT, which is short for Bidirectional Encoder Representation from Transformers, is an advanced model for language representation that has achieved remarkable success in complex tasks like classification, Natural Language Processing (NLP), and prediction. This deep learning model analyses text bidirectionally, considering both the left and right sides, to gain a thorough understanding of word context within sentences. Its ability to comprehend context makes it highly valuable for tasks such as sentiment analysis, question answering, and language translation, positioning BERT as a revolutionary model in the field of NLP. While BERT has demonstrated its effectiveness, its resource-intensive nature, characterized by billions of parameters, presents challenges for low-capability devices and applications with strict latency requirements (Aftan and Shah, 2023). In response, researchers have developed methods to compress large-scale Transformer-based models, including BERT (Ganesh et al., 2021). Moreover, BERT's versatility extends to domains like medical code assignment from clinical notes, where its effectiveness has been investigated through studies on knowledge transfer, hierarchical architectures, and label attention mechanisms (Ji et al., 2021). Additionally, BERT's impact on text summarization has been emphasized, demonstrating its application in both extractive and abstractive models through a comprehensive framework (Liu and Lapata, 2019).

The BERT model utilizes the Transformer architecture, which consists of multiple encoded layers and key components such as the encoder, attention mechanism, and self-attention layers. The encoder processes the input sequence, segmenting it into sub words, and passing it through multiple layers with self-attention mechanisms and feed-forward neural networks. This allows the attention mechanism to calculate attention scores for capturing sentence relationships, while the self-attention layers focus on different segments of the input sequence (Aftan and Shah, 2023). BERT undergoes pre-training by using a large amount of unlabelled text and employs two objectives: masked language modelling and next sentence prediction (Devlin et al., 2019). Masked language modelling involves randomly masking tokens and predicting them based on contextual understanding, enhancing the model's comprehension of word relationships. The next sentence prediction objective determines the consecutiveness of sentences, improving the model's understanding of sentence relationships. BERT effectively captures context and complex relationships through bidirectional processing and a multi-layer Transformer architecture (Ganesh et al., 2021). The pre-training process is detailed in the paper "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" by (Devlin et al., 2019), which provides insights into the specific objectives and techniques employed (Ji et al., 2021). In summary, BERT's pre-training is a vital step in developing a robust language representation model that achieves exceptional performance across various natural language processing tasks (Liu & Lapata, 2019).

BERT has demonstrated its effectiveness in a range of natural language processing (NLP) tasks, including sentiment analysis, named entity recognition, and question-answering (Aftan & Shah, 2023). It has achieved state-of-the-art results on benchmarks such as the Stanford Question Answering Dataset (SQuAD) and the GLUE benchmark (Aftan & Shah, 2023). BERT's applications have extended to commercial and medical domains, although research in these areas is limited (Aftan & Shah, 2023; Ji et al., 2021). While BERT has exhibited impressive performance and surpassed previous models in many NLP tasks (Devlin et al., 2019), its large size and computational demands pose challenges for deployment on resource-constrained devices or applications with strict latency requirements (Ganesh et al., 2021). Additionally, fine-tuning BERT for specific tasks necessitates a significant amount of labelled data, which may not always be readily available (Devlin et al., 2019). To address these limitations, future research aims to enhance BERT's efficiency and interpretability and explore its applications in new domains and languages (Ji et al., 2021). Overall, BERT has emerged as a potent language representation model, exhibiting remarkable performance across various NLP tasks. However, its size, computational requirements, and data availability should be taken into consideration. Ongoing research efforts strive to overcome these limitations and uncover new avenues for BERT's utilization in diverse domains and languages (Liu & Lapata, 2019).

2.4.2 Clinical BERT

Clinical BERT is a language model developed exclusively for the clinical domain, with the goal of comprehending clinical narratives such as physician notes (Lee et al., 2020). It is built on the BERT (Bidirectional Encoder Representations from Transformers) architecture, which is a well-known natural language processing system (Lamproudis et al., 2022). What sets Clinical BERT apart is its training methodology. It is pretrained on clinical data from the same domain, allowing it to excel in domain-specific reasoning for clinical language processing tasks, outperforming BERT models trained on general-domain datasets (Sushil et al., 2021). Clinical BERT develops a better knowledge of the specific language patterns, terminology, and context found in clinical narratives by exploiting in-domain clinical data during the pretraining phase. This specialised training gives Clinical BERT the capacity to successfully process and comprehend clinical text, making it a powerful tool for a variety of clinical language processing applications (Alsentzer et al., 2019).

The architecture of Clinical BERT is specifically designed for pretraining language models using large amounts of text data. During the pretraining process, the model is trained to predict missing words within a sentence based on the context provided by the surrounding words. In the case of Clinical BERT, it is pretrained using in-domain clinical data, including electronic health records, clinical notes, and other clinical text. This initial training is followed by fine-tuning on downstream clinical NLP tasks such as named entity recognition and classification (Lewis et al., 2020). During fine-tuning, the model is trained to predict accurate labels for the given task using a combination of supervised and unsupervised learning techniques. This approach allows Clinical BERT to learn from both labelled and unlabelled data, enhancing its performance on a wide range of clinical NLP tasks, including extracting information from electronic health records, and identifying medical concepts in clinical text (Lamproudis et al., 2022).

The performance of Clinical BERT has demonstrated immense potential for clinical language processing tasks, making it a popular choice for various applications such as literature search, question answering, and patient outcome prediction (Lamproudis et al., 2022). By employing clinical-specific contextual embeddings, Clinical BERT has outperformed BioBERT, another specialized BERT model, on multiple clinical tasks including named entity recognition and medical natural language inference (Lee et al., 2020). In fact, Clinical BERT achieved a new state-of-the-art performance of 82.7% accuracy on the MedNLI task, surpassing the previous state-of-the-art accuracy of 73.5% (Romanov and Shivade, 2018).

While BERT implementations like BioBERT exhibit superior domain-based reasoning compared to models trained on general-domain corpora, there still exists a considerable gap in performance when compared to human capabilities on these tasks. Alternative methods should be incorporated to enhance textual domain knowledge in medical natural language inference (NLI) tasks. These methods include additional language model pretraining on medical domain corpora, incorporating lexical match algorithms, dependency relations, or utilizing a trained retriever module (Sushil et al., 2021). By incorporating such approaches, the performance of BERT models can be improved, including Clinical BERT, on domain-specific inference tasks within the medical field. The evaluation solely focuses on English-language clinical text, leaving the performance of the models on clinical text in other languages unclear. Consequently, there is room for improvement in the performance of BERT models for clinical language processing tasks and to assess the generalizability and efficacy of clinical BERT across a wider range of clinical NLP tasks and linguistic contexts (Lamproudis et al., 2022).

2.5 Transformers Deep Learning Models for Hospital Readmission Prediction

One of the popular transformers deep learning models, Clinical BERT, which is a specialized version of the BERT model that is designed to model clinical notes and is used to predict hospital readmission. The dataset used by different studies to predict hospital readmission using clinical BERT is MIMIC-III and it includes clinical notes and other patient details. It contains electronic health records of 58,976 unique hospital admissions from 38,597 patients, the target variable set for the prediction is binary (0&1) thus the approach used is binary classification (Johnson et al., 2016; Huang et al., 2019).

First study conducted by (Huang et al., 2019) splits the data into five folds for independent runs, with four folds for pre-training (and training during fine-tuning) and the fifth for testing during fine-tuning to implement BERT pre training this resulted in Clinical BERT outperforming BERT model in terms of Area Under the Receiver Operating Characteristic Curve (AUROC) with a score of 0.714 ± 0.018 for Clinical BERT. This demonstrates enhanced capability Of Clinical BERT in capturing the subtleties of clinical terminology and comprehending the interconnections among medical concepts (Huang et al., 2019).

The other study, conducted by (Bikram et al., 2022) simply uses the admission data to fine tune the pre-existing BERT model. Further, clinical BERT was pre trained using a masked language modelling objective and then fine- tuned on a dataset of discharge summaries and the first few days of notes in the intensive care unit to predict 30-day hospital readmission (Banerjee et al., 2017; Huang et al., 2019). The evaluation of this models was based on a pre-defined metric motivated by a clinical challenge: useful classification rules for medicine need to have high positive predictive value (precision). The study by (Bikram et al., 2022) highlighted that clinical BERT performs better than traditional machine learning algorithms and other deep contextual representation techniques like Clinical XLNet. It also highlights that pre-training and fine-tuning Clinical BERT for the task makes prediction more efficient as the dataset with 1000 discharge summaries had Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC) of 0.721 and 0.70 respectively while the dataset with 10,000 had AUROC and AUPRC of 0.718 and 0.68 respectively (Bikram et al., 2022).

Based on all the previously discussed studies clinical BERT is computationally efficient and capable of capturing long-term dependencies and contextual information in unstructured text, making it suitable for representing clinical notes. BERT is more accurate in capturing clinical word similarities as compared to other models and it uses Local Interpretable Model Agnostic explanation (LIME) which provides an accurate interpretation of predictions made by any classifier or regressor. It achieves this by locally approximating the model using an interpretable model (Ribeiro et al., 2016). The interpretations, allows easy visual understanding of the contribution made by different terms for clinicians. Overall, clinical BERT is flexible framework and can readily be applied on other tasks (Vig, 2019).

The literature review emphasizes Clinical BERT's potential in hospital readmission prediction. Prior studies used limited MIMIC-III data, but we'll employ the comprehensive **MIMIC-IV** dataset, integrating Clinical BERT, LSTM, SVM, and logistic regression to assess their effectiveness.

2.6 Summary

This chapter introduces hospital readmission, its impact on healthcare, and the necessity of prediction. It reviews various approaches like machine learning, deep learning, and transformers. Each method, its predictive work, and limitations are emphasized, especially the use of BERT and Clinical BERT in transformer-based approaches. The chapter previews how this project differs from existing research. The next chapter will focus into prediction algorithm methodologies and efficiency comparisons to select the most suitable approach.

CHAPTER 3: RESEARCH APPROACH

This chapter focuses on the data used for this study and on different research approaches, aiming to select the best approach. To carry out the research, an ethical permission has been received.

3.1 Data Information

Data collecting is an important activity in every project. In this project, the dataset used is MIMIC-IV obtained from PhysioNet site. It is a large freely available credential-based dataset including health-related data of approximately 40,000 patients. The future readmission is a binary target variable calculated based on admission and discharge time features. If a patient is readmitted within 30 days of discharge, it is assigned 1 otherwise, 0 (Blinder, 2017).

3.2 Evaluation of Research Methodologies

Data mining comprises a series of systematically executed procedures. These procedures encompass various tasks that are commonly carried out as integral components of data mining (Yu Lin and McClean, 2001). Three popular approaches, CRISP-DM, SEMMA, and KDD will be compared and analysed to determine the best fit for this project's purpose.

3.2.1 SEMMA Model

SEMMA created by the SAS Institute stands for Sample, Explore, Modify, Model, and Access (Yu Lin and McClean, 2001). SEMMA is used to understand, organize, develop, and maintain data mining projects.

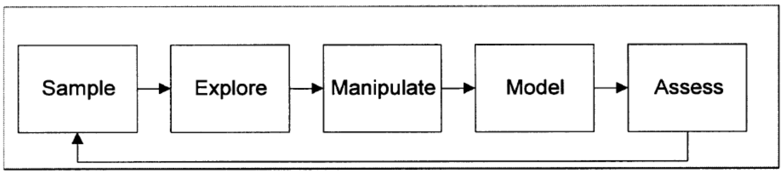


Figure 3.1: The SEMMA data mining process (Yu Lin & McClean, 2001)

Figure 3.1 presents the five stages of SEMMA model (Shafique and Qaiser, 2014):

1. **Sample:** This is an optional stage that involves selecting a subset of data from a larger dataset. The sample size should be appropriate to extract meaningful information and manipulate quickly.

2. **Explore:** This stage focused data understanding by searching for patterns, trends, and anomalies to refine the data discovery process.

3. **Manipulate:** The data undergoes changes using techniques like variable creation, selection, and transformation. The goal is to obtain dataset which helps in model to run efficiently.

4. **Model:** This stage constructs models using prepared data, employing various techniques based on their strengths and suitability to explore different data

combinations automatically, yielding results

5. Assess: In the final stage, we evaluate the reliability and utility of modelling findings by estimating model performance and assessing their practicality for decision-making.

3.2.2 KDD Model

KDD describes the comprehensive process of gathering knowledge from data, and data mining is a step in this process that involves using algorithms to extract meaningful patterns from the data. It also encompasses additional steps such as data preparation, selection, cleaning, integration of relevant knowledge, and accurate interpretation of results (Rahman et al., 2014), all of which contribute to the overall success of knowledge discovery (Fayyad et al., 1996).

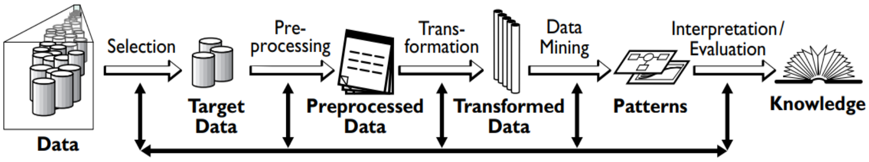


Figure 3.2: The KDD data mining process (Fayyad et al., 1996)

1. Selection: This stage focuses on selecting a specific set of variables or a sample of data that requires investigation or the creation of a target dataset.

2. Pre-processing: It involves in preparing the target dataset by eliminating any noise or inconsistencies that might hinder the accuracy of subsequent analyses.

3. Transformation: This stage focuses on converting the data from one format to another, enabling a smoother execution of the data mining process.

4. Data Mining: This stage aims at identifying hidden patterns from the dataset and modify them to achieve desired outcome.

5. Evaluation: The discovered patterns are evaluated to have better understanding of their significance and usefulness.

3.2.3 CRISP-DM Model

CRISP-DM is a structured and comprehensive approach for predictive models developed by Daimler Chrysler, SPSS, and NCR in 1999 (Shafique and Qaiser, 2014). It deals with business understanding, data preparation, modelling, evaluation, and deployment (Azevedo and Santos, 2008).

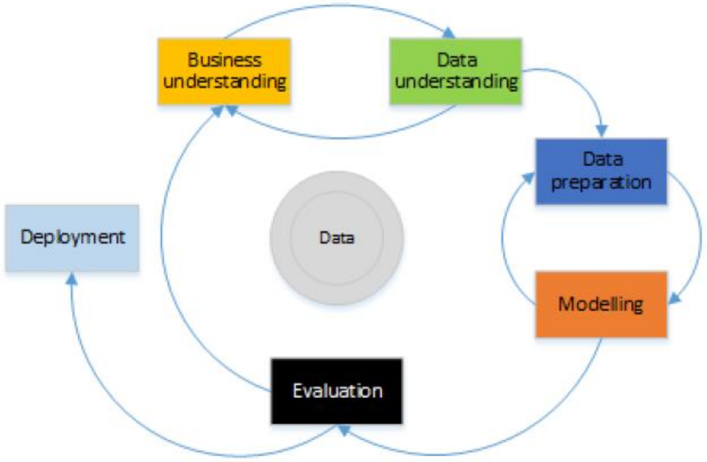


Figure 3.3: The CRISP-DM data mining process (Palacios et al., 2017)

The five stages of CRISP DM are as follows (Shafique & Qaiser, 2014):

1. **Business Understanding:** It involves understanding the objectives of the business or research. Key elements such as success criteria, business and data mining objectives, requirements, and technical and business terminologies are identified.

2. **Data Understanding**: This step involves assessing the gathered data by developing hypotheses to uncover hidden information ensuring that the data fits the research criteria.

3. **Data Preparation**: Based on the understanding of the data and research requirements, the data is prepared through cleaning and transformation processes like selecting records, tables, and attributes to meet the research objectives.

4. **Modelling**: This stage involves the actual analysis of the data utilizing modelling techniques such as online analytics, traditional methods like regression and clustering, and non-traditional statistical analysis like neural networks, are applied.

5. **Evaluation**: This stage focuses on comparing the results of different models using common evaluation measures such as accuracy, ROC curve or classification charts to assess the model’s performance.

6. **Deployment:** This final stage is concerned with the implementation of the data mining models. It involves organizing, reporting, and deciding how to utilize the information and findings effectively.

3.3 Comparative study of the Research Methodologies

SEMMA, KDD, and CRISP-DM are distinct methodologies in data analysis and mining. SEMMA emphasizes data preparation, modelling, and assessment but lacks capabilities for complex deep learning tasks. KDD excels with large datasets and pattern discovery but is less efficient with predictive models due to the absence of techniques like cross-validation and optimization. In contrast, CRISP-DM is a comprehensive, structured approach encompassing business understanding, data preparation, modelling, evaluation, and deployment, ensuring validation and prediction accuracy (Azevedo and Santos, 2008).

Despite differences, these methodologies share similarities. SEMMA, especially when using SAS Miner Software, aligns with the five KDD stages (Azevedo and Santos, 2008). For example, Selection in KDD corresponds to Sample in SEMMA, Pre-processing to Explore, Transformation to Modify, Data Mining to Model, and Interpretation/Evaluation to Assess. Selection to Pre-processing (KDD) or Sample to Explore (SEMMA) phases combine in the Data Understanding stage of CRISP-DM. Additionally, CRISP-DM's Business Understanding phase resembles initial steps before KDD processes, and its Deployment phase corresponds to steps after KDD completion (Palacios et al., 2017).

3.4 Implementation using CRISP-DM Algorithm

Based on the comparative study, we would use CRISP-DM to implement a prediction algorithm for hospital readmissions. The steps involved are as follows:

1. **Business Understanding**: Define organizational objectives and align them with data mining goals.

2. **Data Understanding:** Extract data from the PhysioNet site (MIMIC-IV dataset) with health-related data from around 40,000 patients. Perform exploratory data analysis.

3. **Data Preparation:** Extract relevant data, handle NULL values, transform the text data, and prepare the dataset.

4. **Modelling:** Train different models to predict hospital readmission based on various features.

5. **Evaluation:** Compare and assess the accuracy of different models for predicting hospital readmission.

6. **Deployment:** If all business objectives are met, plan for project deployment.

3.5 Ethics

Approval Status: This study received its ethical approval on **16-05-2023** from the Ethics Team to begin the research under the supervision of Prof. Alaa Marshan. For further reference please find the attached approval letter in the appendix section of this report.

CHAPTER 4: DATA ANALYSIS

This chapter explains how the algorithm was created using the CRISP-DM research approach and explains each of the six stages in detail. The code implementation involved in the different sections of this chapter will be attached in the appendix B for reference.

4.1 Business Understanding

This marks the project's outset, involving the comprehension of business objectives, the imperative for action, defining technical goals aligned with these objectives, and agreeing on the planned approach to achieve them (Clinton et al., 2000).

Business goals express project objectives using industry-specific terminology. In this project, the business goal is to reduce readmission costs for both patients and hospitals, as hospitals with higher-than-average readmission rates face penalties under the Affordable Care Act (Futoma et al., 2015; Johnson et al., 2023). Data mining goals, on the other hand, frame objectives in technical terms, transforming business issues into data mining goals. They specify different type of challenge like classification, prediction, or clustering and outline project requirements for successful completion. For the business objective, the technical goal is to identify patients at risk of readmission based on their medical history to prevent future readmissions.

The project aims to predict patient readmission using clinical notes, including radiological reports and discharge summaries within 30 days of discharge. The execution plan involves dataset selection, understanding, preprocessing for modelling, implementing prediction models, and finally evaluating the results to make informed predictions.

4.2 Data Understanding

To anticipate readmission, it is necessary to comprehend prior research initiatives carried out to choose the appropriate dataset. As a result, the initial stage of data collection is to look for scientific papers and evaluate the literature. After reading the publications and considering past research on predicting readmission, dataset from PhysioNet (Johnson et al., 2023) has been taken.

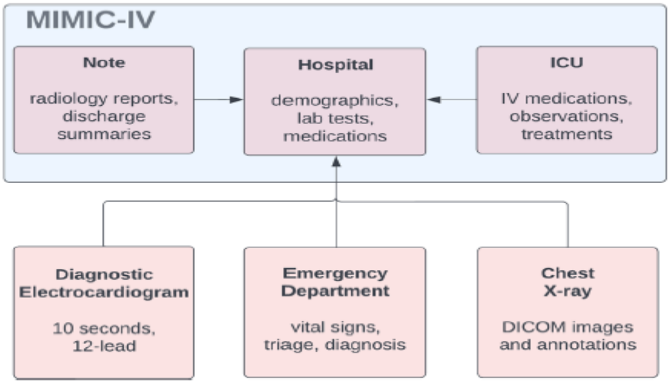

MIMIC-IV
Note
radiology reports, discharge summaries
Hospital
demographics, lab tests, medications
ICU
IV medications, observations, treatments
Diagnostic Electrocardiogram
10 seconds, 12-lead
Emergency Department
vital signs, triage, diagnosis
Chest X-ray
DICOM images and annotations

Figure 4.1: Modular structure of MIMIC-IV (Johnson et al., 2023)

The MIMIC-IV dataset, which is a credentialed access dataset, is an updated version of the MIMIC-III dataset and has been used in this project. The dataset consists of critical care data for over 40,000 patients and has adopted a modular approach to data organization. Figure 4.1 shows the structure of MIMIC-IV dataset for this study three specific files are used admissions from hospital module, radiology reports and discharge summaries from notes module.

4.3 Data Preparation

In this third phase, we analyse and process the acquired data to suit the modelling requirements. This comprises identifying pertinent data chunks, cleaning data, transforming data, and integrating numerous files into a uniform structure. Furthermore, syntactic data alterations are handled during this step (Clinton et al., 2000). In this section, we will go through all the actions that were performed to prepare the data for this study.

4.3.1 Selecting the Data

The three dataset's admissions, radiology report and discharge summary are processed to form a single dataset.

Admissions: Table 4.1 shows the 16 features of admissions table, and this dataset has 431,231 rows.

Column Names	Description
subject_id	Represents each unique patient.
hadm_id	Represents a single patient's unique admission to the hospital.
admittime	Date and time the patient was admitted to hospital.
dischtime	Date and time the patient was discharged from hospital.
deathtime	The time of in-hospital death for the patient.
admission_type	Represents the urgency of admission to hospital.
admit_provider_id	Anonymous identifier for the provider who admitted the patient.
admission_location	Location of the patient prior to arriving at the hospital.
discharge_location	Location of the patient after they are discharged from the hospital.
insurance	Patient has insurance or not.
language	Language spoken by patient.
marital_status	Marital status of patient.
race	Patient demographics
edregtime	Date and time of registration into hospital
edouttime	Date and time of discharge from hospital
hospital_expire_flag	Indicates if patient died within hospital or not.

Table 4.1 - Description Admissions dataset (Johnson et al., 2022)

Discharge: Table 4.2 shows the 8 features of discharge table, and this dataset has 431,231 rows.

Column Names	Description
note_id	Unique identifier for the given note.
subject_id	Unique identifier for specific an individual patient.
hadm_id	Represents a single patient's admission to the hospital.
note_type	Identify the type of notes: "DS" - discharge summary "AD" - discharge summary addendum
note_seq	Monotonically increasing integer which chronologically sorts the notes within note_type categories.
charttime	The time at which the note was charted.
storetime	Time at which the note was stored in the database.
text	The clinical text notes and reports.

Table 4.2 - Description Discharge dataset (Johnson et al., 2022)

Radiology: Table 4.3 shows the 8 features of radiology table, and this dataset has 2,321,355 rows.

Column Names	Description
note_id	Unique identifier for the given note.
subject_id	Unique identifier for specific an individual patient.
hadm_id	Represents a single patient's admission to the hospital.
note_type	Identify the type of notes: "RR" - radiology report "AR" - radiology report addendum
note_seq	Monotonically increasing integer which chronologically sorts the notes within note_type categories.
charttime	The time at which the note was charted.
storetime	Time at which the note was stored in the database.
text	The clinical text notes and reports.

Table 4.3 - Description Radiology dataset (Johnson et al., 2022)

These three datasets are prepared to form a single dataset using following steps:

- For all the three datasets the dead patients are removed from the dataset using deathtime and hadm_id.
- For the admissions table 4.1, all rows with admission type "surgical same day admission" are replaced with elective according to most healthcare system and experts' opinion (Johnson et al., 2023). The "elective" type is then replaced with NaN as it won't be considered for the study.
- Further two new columns next_admittime & next_admission_type are formed using group by function on subject_id and filled using admit time.
- Another new column is formed days_til_next_admit for which value is calculated using next_admittime & discharge time.
- For the discharge table 4.2 and radiology table 4.3, the subject_id and text column are extracted as a new dataset and then all the text is grouped based on subject_id.
- New dataset shown in table 4.4 is formed taking subject_id, hadm_id, next_admittime, days_til_next_admit, admission_type, text, admit, discharge

and death time.

Column Name	Description
subject_id	The unique id associated with each patient
hadm_id	The unique id is associated with each hospital admission for that patient
admittime	The date and time when patient got admitted to the hospital
dischtime	The date and time when patient got discharged from the hospital
days_til_next_admit	The number of days between two admissions
next_admittime	The date and time next admission is done
admission_type	The classification of admission type is stored here (like urgent, elective, emergency)
deathtime	The date and time when patient died
text	The clinical notes combined from discharge and radiology report are stored here

Table 4.4 - Dataset by combining the admission, discharge, and radiology tables.

- The target variable is formed based on days_til_next_admit column from table 4.4, if the value is less than 30 then value assigned is 1 else its 0.
- The final dataset for modelling is formed as shown in table 4.5, taking only the text column and readmit_lable column which is target variable. The dataset is huge with 167258 rows.
- The study has limited time and resources thus to handle the lack of computational power the dataset is randomly sampled to just 15000 rows, keeping random_state to 42 ensures that every time the code is executed, we get the same set of data.

Column Name	Description
Text	The clinical notes including discharge and radiology reports
label	Target variable representing readmission within 30 days

Table 4.5 – Dataset with the clinical notes and target variable.

4.3.2 Preparing the Data

1. Data Balancing

This is a crucial step to ensure modelling is implemented on balanced dataset and have a moderate prevalence which is a statistical concept to present the proportion of individuals in a population having disease (Youbi Idrissi et al., 2022). The data was having high prevalence difference where class 0 is **77.69%** and class 1 is **22.31%**, thus subsampling is applied to balance the classes as it is an efficient technique when there are computational barriers (Youbi Idrissi et al., 2022). Finally, the data is split into train, test and validation sets in **70%, 15%, 15%** ratio respectively.

2. Data Transformation

This is a major step as the dependent feature is in text format which is not accepted by models thus different transformations are applied to ensure the format of input data is accepted by models. First is **lowercasing** to ensure that no two same words if present in both upper and lower case are treated differently by the model while training (Lomet et al., 2000). Another important step is to remove any **punctuations, numbers, words with numbers, underscores, extra spacings** and **newline characters** to ensure that the dataset has only words which help model to perform efficiently.

4.4 Modelling & Evaluation

Followed by the previous section now this section will deal with discussing the approach of different models on the transformed dataset and then evaluating their results. As highlighted previously the study focuses on predicting future readmission of patients based on clinical notes provided and is performed using two machine learning and two deep learning models.

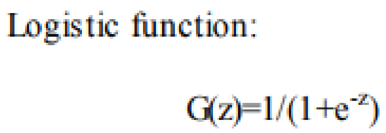
4.4.1 Machine Learning Models

The two machine learning models used in this study are **Support Vector Machine (SVM)** and **Logistic Regression** both have been used in a lot of studies before however there is not a lot of analysis done using the MIMIC-IV version 2.2. There are certain steps before supplying the model with data.

- First tokenization is applied to divide long text data into smaller chunks.
- Next is removing stop words and lemmatization where one removes the basic english words which does not hold any meaning to enhance the results while the later ensures that the words with similar meaning, but different grammatical format are not repeated as this might reduce the accuracy of model.
- Finally, vectorization is applied on the dataset to convert text data into numeric data as machine learning models accept only numeric data as input.

Model 1 – Logistic Regression

Logistic regression is a supervised machine learning model that is used for classification tasks, where goal is to predict the probability of weather an instance belongs to a particular class or not. It is like linear regression and uses logistic/ sigmoid function which is an S shaped curve, it takes the real value number and maps into 0 and 1(Sunitha et al., 2020). The model was implemented in this study inspired by the results of previous study by Rajkomar et al., (2018) which showed that logistic regression gave great results on MIMIC-III dataset, thus its worth exploring on MIMIC-IV dataset.



Logistic function:

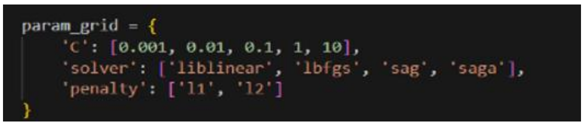
$$G(z)=1/(1+e^{-z})$$

Equation 4.1 Equation of Logistic function used in LR model.

First the model is trained on the data prepared by implementing data balancing and transformation, thus the training data is feed to the model and trained after which the model is tested on the test and validation datasets. Next, after evaluating data on the default parameters of the model next model is formed by hyperparameter tunning of logistic regression model.

Model 2 - Hyperparameter Tunned Logistic Regression

After training the data with default logistic regression parameters, the model was trained on parameters which are fixed before training process known as hyperparameters. The **Grid Search library** is used for setting up parameters which searches the best model from the grid as shown in figure 4.2 and for this study the parameters which were tuned to fixed values before training are solver is the specific optimization algorithm, penalty specifies the type of regularization applied on model while C represents regularization strength.



```
param_grid = {
    'C': [0.001, 0.01, 0.1, 1, 10],
    'solver': ['liblinear', 'lbfgs', 'sag', 'saga'],
    'penalty': ['l1', 'l2']
}
```

Figure 4.2 Hyperparameters used in LR model.

Evaluation of Logistic regression and Hyperparameter tunned LR model

The results of logistic regression model on validation and test dataset are shown below in table 4.6.

Model	Dataset	Accuracy	Precision	Recall	F1 score	Support
Logistic Regression	Validation dataset	80.77%	85% [0] 76% [1]	78% [0] 84% [1]	82% [0] 80% [1]	549 [0] 455 [1]
Logistic Regression	Test dataset	82.58%	84% [0] 81% [1]	83% [0] 83% [1]	83% [0] 83% [1]	497 [0] 508 [1]

Table 4.6 – Results of LR model on validation & test datasets.

The best model using the pre-defined parameters for hyperparameter tunning of logistic regression model is shown in table 4.7. Further the results of hyperparameter tunned LR model on test dataset is shown in table 4.8.

Selected Parameters	Best parameters
C	10
Penalty	L2
Solver	lbfgs

Table 4.7 – Best Model achieved by hyperparameter tunning.

Model	Dataset	Accuracy	Precision	Recall	F1 score	Support
Hyperparameter tunned LR	Test dataset	81.79%	84% [0] 80% [1]	79% [0] 85% [1]	81% [0] 82% [1]	497 [0] 508 [1]

Table 4.8 – Results of LR model on test dataset.

The confusion matrix and ROC curve are presented in figure 4.3 highlights that there are 430 correct predictions for class 0 and 380 for class 1 while the ROC curve has an area of 89% thus it is performing average in categorizing the data.

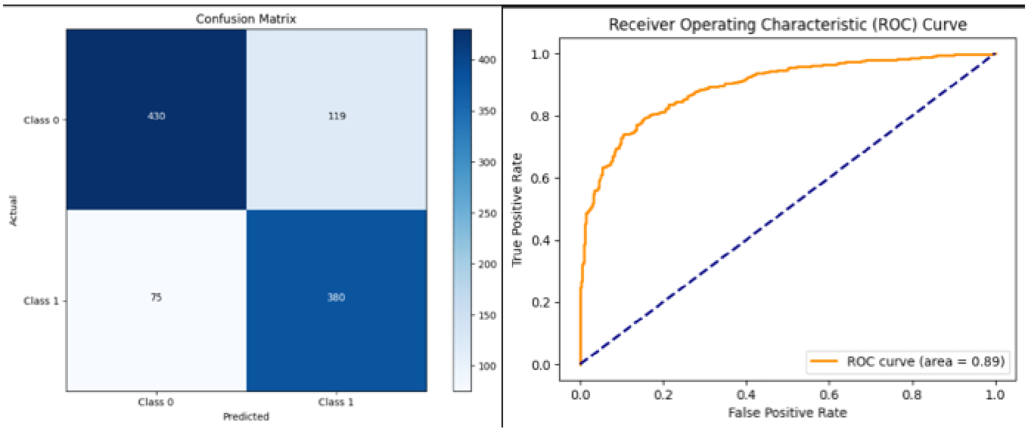


Figure 4.3 – Confusion Matrix & ROC curve for Logistic Regression

Comparison of the two models

The hyperparameter tunned model did not result in an improved model, the accuracy and F1 score is slightly low but the recall for class 1 is better than logistic regression model. This could be because of inadequate hyperparameter search space or overfitting of the data.

Model 3 – Support Vector Machine

Support Vector Machine is a supervised machine learning model that is used for classification as well as regression tasks. The objective of this model is to find optimal hyperplane to separate data points into classes. Figure 4.4 presents how the classification of data is done into 1 and 0 mathematically (Jegan et al., 2013). The literature review highlighted that SVM model performed well on MIMIC-III dataset and therefore exploring the same model on MIMIC-IV was necessary.

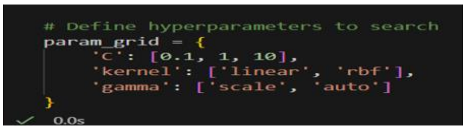
$$\hat{y} = \begin{cases} 1 & : \ w^T x + b \geq 0 \\ 0 & : \ w^T x + b < 0 \end{cases}$$

Figure 4.4 Linear classifier for SVM model.

The model is trained on the data prepared by implementing data balancing and transformation, thus the training data is feed to the model and trained after which the model is tested on the test and validation datasets. After evaluating data on the default parameters of the model next model is formed by hyperparameter tunning of support vector machine model.

Model 4 - Hyperparameter Tunned SVM

After training the data with default SVM parameters, the model was trained on parameters which are fixed before training process known as hyperparameters. The **Grid Search** library is used for setting up parameters which searches the best model from the grid and for this study the parameters which were tuned to fixed values before training are **C** which controls trade-off between maximizing margin and minimizing classification error, **Kernel** this decide the function used to transform input data into higher dimensional space and last is **Gamma** it is for certain kernels and control shape of decision boundary as shown in figure 4.5.



```
# Define hyperparameters to search
param_grid = {
    'C': [0.1, 1, 10],
    'kernel': ['linear', 'rbf'],
    'gamma': ['scale', 'auto']
}
```

Figure 4.5 Hyperparameters used in SVM model.

Evaluating Support Vector Machine and Hyperparameter tunned SVM model

The results of SVM model on validation and test dataset are shown below in table 4.9.

Model	Dataset	Accuracy	Precision	Recall	F1 score	Support
SVM	Validation dataset	80.97%	85% [0] 77% [1]	79% [0] 84% [1]	82% [0] 80% [1]	549 [0] 455 [1]
SVM	Test dataset	82.98%	83% [0] 83% [1]	82% [0] 84% [1]	83% [0] 83% [1]	497 [0] 508 [1]

Table 4.9 – Results of SVM model on validation & test datasets.

The best model using the pre-defined parameters for hyperparameter tunning of SVM model is shown in table 4.10. Further the results of hyperparameter tunned SVM model on test dataset is shown in table 4.11.

Selected Parameters	Best parameters
C	1
Gamma	Scale
Kernel	rbf

Table 4.10 – Best Model achieved by hyperparameter tunning.

Model	Dataset	Accuracy	Precision	Recall	F1 score	Support
Hyperparameter tunned SVM	Test dataset	82.68%	82% [0] 83% [1]	82% [0] 83% [1]	82% [0] 83% [1]	497 [0] 508 [1]

Table 4.11 – Results of best SVM model on test dataset.

The confusion matrix and ROC curve are presented in figure 4.6 highlights that there are 432 correct predictions for class 0 and 381 for class 1 while the ROC curve has an area of 90% thus it is performing average in categorizing the data.

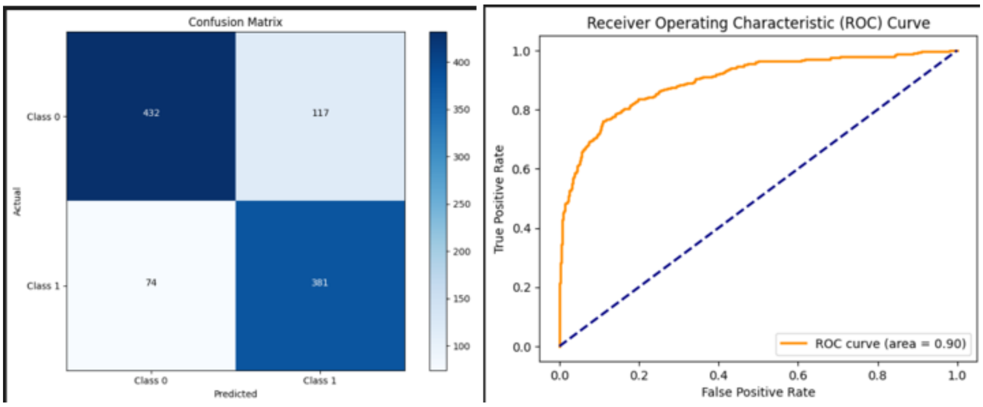


Figure 4.6 – Confusion Matrix & ROC curve for SVM

Comparison of the two models

The hyperparameter tunned model and standard SVM model almost similar results and thus the choice of which model to use depends on the specific requirements like if the goal is to simplify the model than standard SVM model is preferred choice.

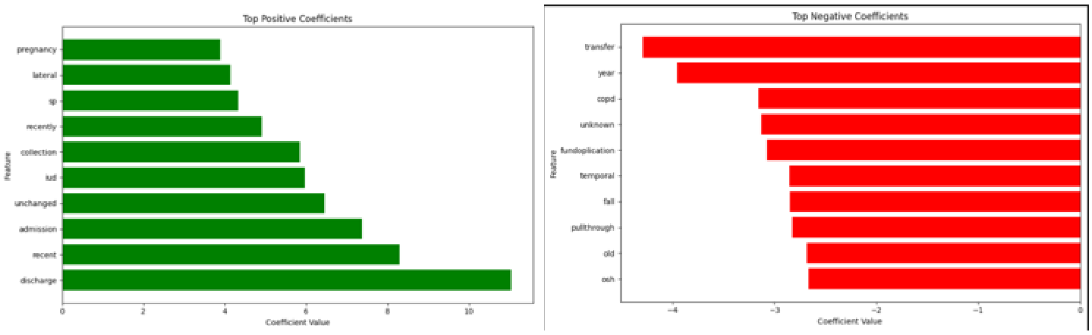


Figure 4.7 – Showing Positive & Negative effects of terms on models.

The figure 4.7 presents the terms in the text column that affect the accuracy of model in positive and negative terms, the green bar chart presents positive effect while red presents negative.

4.4.2 Deep Learning Models

The two deep learning models used in this study are **Clinical BERT** and **Long Short-Term Memory (LSTM)**. The LSTM model operates on the data that was transformed during the machine learning modelling stage, so the same vectorized data is used for this modelling while the Clinical BERT uses the pre-processed data, but the tokenization is performed separately using the Clinical BERT tokenizer.

Model 5 – Clinical BERT model

Clinical BERT is a BERT model which specializes in the clinical domain. It is a pretrained model on clinical data and can be fine tunned on the specific clinical notes depending on the goal of study (Lamproudis et al., 2022). There have not been any studies implemented using MIMIC-IV dataset for this model, the little implementation of this study is only on MIMIC-III dataset and the results were promising this inspired me to explore Clinical BERT model for this study. The pretrained model and the clinical BERT tokenizer are loaded, and the data is tokenized, further the dataset is converted into the hugging face and torch format to ensure it fits the model requirements as shown in figure 4.8.

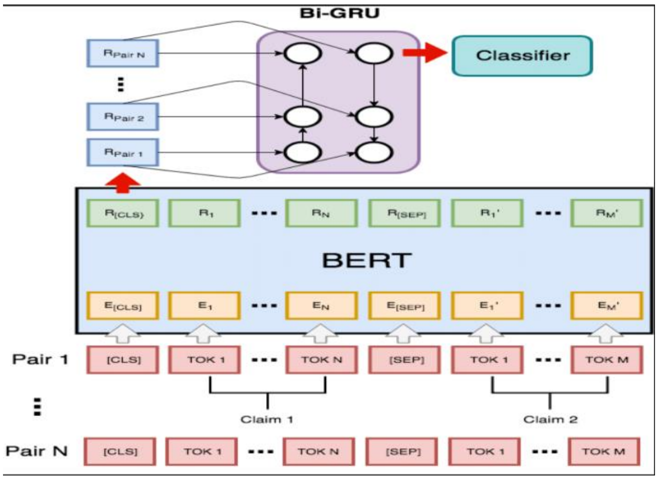
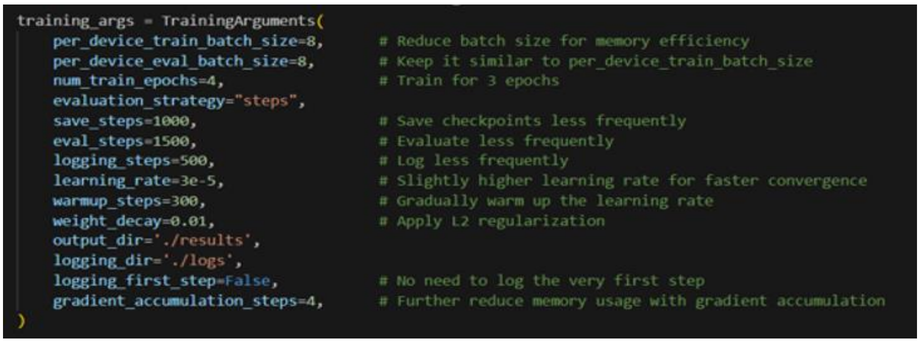


Figure 4.8 Pretrained BERT model (Durmus et al., 2019)**.**

The model is trained on different training arguments like training and evaluation batch size, number of times the model runs through entire dataset (epoch = 4) and others which can be seen in figure 4.9.



```
training_args = TrainingArguments(
    per_device_train_batch_size=8,      # Reduce batch size for memory efficiency
    per_device_eval_batch_size=8,       # Keep it similar to per_device_train_batch_size
    num_train_epochs=4,                 # Train for 3 epochs
    evaluation_strategy="steps",
    save_steps=1000,                    # Save checkpoints less frequently
    eval_steps=1500,                    # Evaluate less frequently
    logging_steps=500,                  # Log less frequently
    learning_rate=3e-5,                 # Slightly higher learning rate for faster convergence
    warmup_steps=300,                   # Gradually warm up the learning rate
    weight_decay=0.01,                  # Apply L2 regularization
    output_dir='./results',
    logging_dir='./logs',
    logging_first_step=False,           # No need to log the very first step
    gradient_accumulation_steps=4,      # Further reduce memory usage with gradient accumulation
)
```

Figure 4.9 Training Parameters of Clinical BERT model.

Evaluation of Clinical BERT model

The results of Clinical BERT model on validation and test dataset are shown below in table 4.12.

Model	Dataset	Accuracy	Precision	Recall	F1 score
Clinical BERT	Validation dataset	64.94%	61.07%	62%	61.73%
Clinical BERT	Test dataset	67.16%	68.54%	65%	66.59%

Table 4.12 – Results of Clinical BERT on validation & test datasets.

The confusion matrix and ROC curve are presented in figure 4.10 highlights that there are 346 correct predictions for class 0 and 329 for class 1 while the ROC curve has an area of 74% thus it is performing average in categorizing the data.

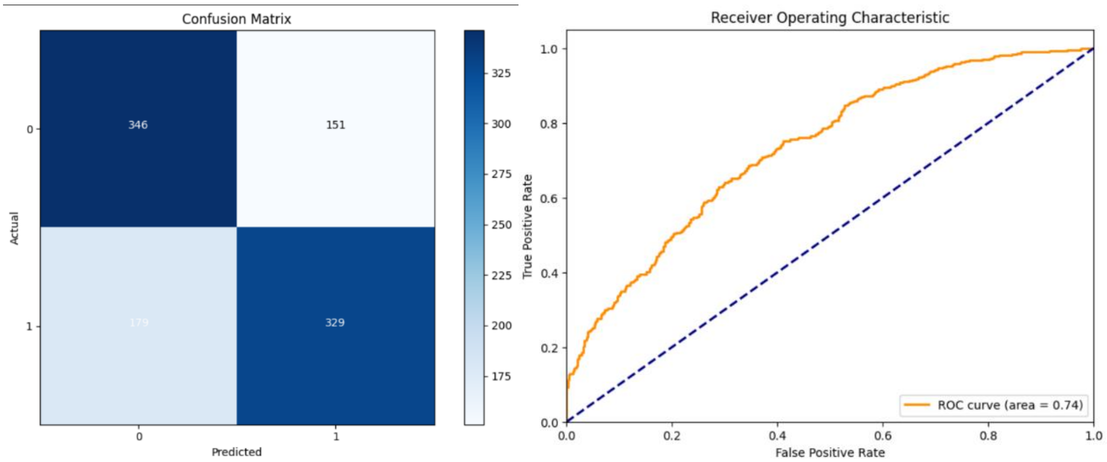
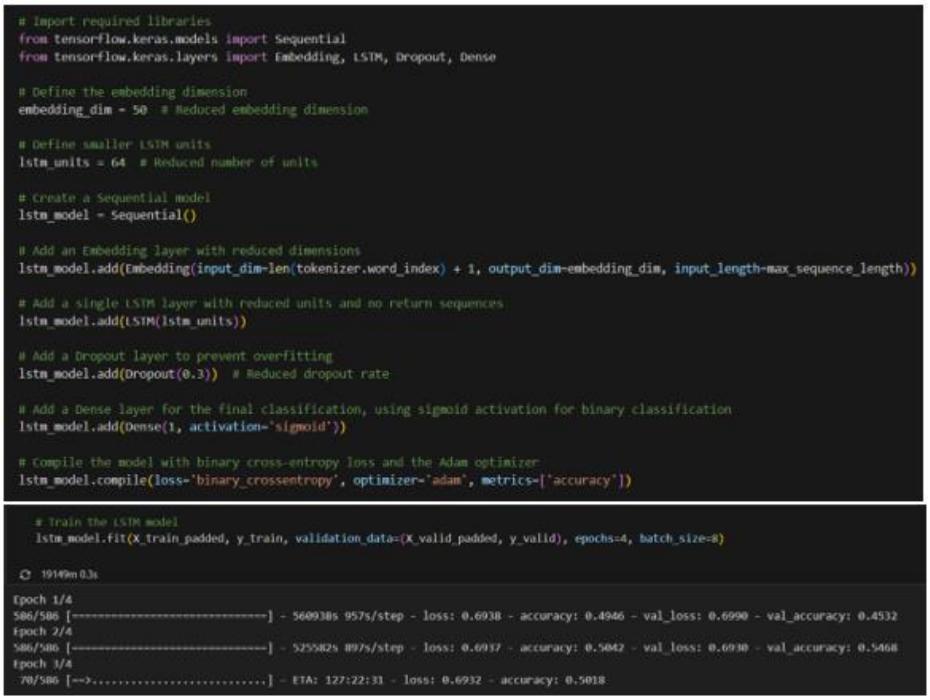


Figure 4.10 – Confusion Matrix & ROC curve for Clinical BERT

Model 6 – Long Short-Term Memory (LSTM) model

LSTM model is extension of recurrent neural networks (RNN) to remove the vanishing gradient problem of RNN. The model consists of subnetworks known as memory blocks that are used to remember input for long time (Reddy and Delen, 2018). After implementation of Clinical BERT model, based on the great results of LSTM model in the previous study by Lin et al., (2019) it was worth exploring this model.

The implementation of this model was done twice, first it was implemented on a high RAM of 64gb, parameters set to epoch=4 and batch size=8 with other parameters as shown in figure 4.11. This implementation was not successful and is still running. Figure 4.11 presents that epoch 1 & 2 completed while epoch = 3 is still running. This is due to the large dataset and the complex LSTM model which requires high computational resources and time. Due to both these constraints the model didn't finish within time with just 64GB RAM. Second implementation was done using the GPU, this implementation took a long time but was successful for epoch size 5 and batch size 16 the results for the same are presented in table 4.14.



```
# Import required libraries
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Embedding, LSTM, Dropout, Dense

# Define the embedding dimension
embedding_dim = 50  # Reduced embedding dimension

# Define smaller LSTM units
lstm_units = 64  # Reduced number of units

# Create a Sequential model
lstm_model = Sequential()

# Add an Embedding layer with reduced dimensions
lstm_model.add(Embedding(input_dim=len(tokenizer.word_index) + 1, output_dim=embedding_dim, input_length=max_sequence_length))

# Add a single LSTM layer with reduced units and no return sequences
lstm_model.add(LSTM(lstm_units))

# Add a Dropout layer to prevent overfitting
lstm_model.add(Dropout(0.3))  # Reduced dropout rate

# Add a Dense layer for the final classification, using sigmoid activation for binary classification
lstm_model.add(Dense(1, activation='sigmoid'))

# Compile the model with binary cross-entropy loss and the Adam optimizer
lstm_model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
```

```
  # Train the LSTM model
  lstm_model.fit(X_train_padded, y_train, validation_data=(X_valid_padded, y_valid), epochs=4, batch_size=8)

19149m 0.3s

Epoch 1/4
586/586 [==============================] - 560938s 957s/step - loss: 0.6938 - accuracy: 0.4946 - val_loss: 0.6990 - val_accuracy: 0.4532
Epoch 2/4
586/586 [==============================] - 525582s 897s/step - loss: 0.6937 - accuracy: 0.5042 - val_loss: 0.6930 - val_accuracy: 0.5468
Epoch 3/4
 70/586 [==>...........................] - ETA: 127:22:31 - loss: 0.6932 - accuracy: 0.5018
```

Figure 4.11 Training Parameters of LSTM model.

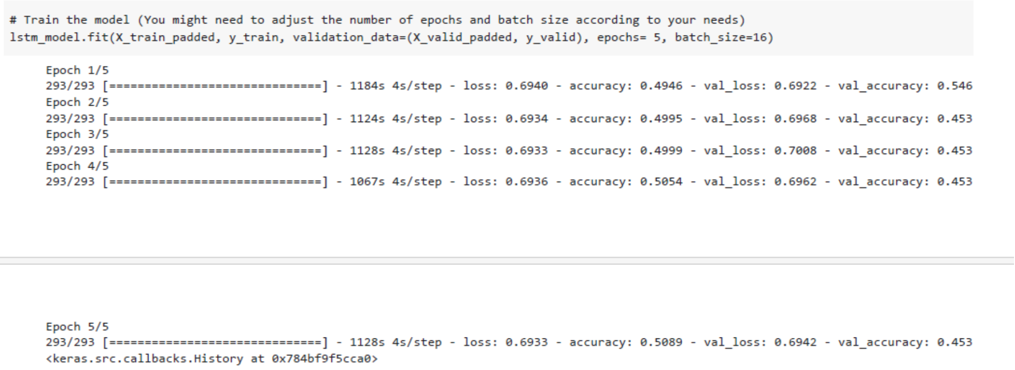
Evaluation of LSTM model

The results using just validation dataset for the first two epoch for the LSTM model are shown below in table 4.13.

Model	Dataset	Accuracy
LSTM	Validation dataset Epoch [1]	45.32
LSTM	Validation dataset Epoch [2]	54.68

Table 4.13 – Results of LSTM on validation datasets for epoch [1 & 2].

The results from the second run of the model using GPU with 5 epochs ran successfully but did not yield great results on validation and test dataset as shown in figure 4.12. The results of the same are shown below in table 4.14.



```
# Train the model (You might need to adjust the number of epochs and batch size according to your needs)
lstm_model.fit(X_train_padded, y_train, validation_data=(X_valid_padded, y_valid), epochs= 5, batch_size=16)

    Epoch 1/5
    293/293 [==============================] - 1184s 4s/step - loss: 0.6940 - accuracy: 0.4946 - val_loss: 0.6922 - val_accuracy: 0.546
    Epoch 2/5
    293/293 [==============================] - 1124s 4s/step - loss: 0.6934 - accuracy: 0.4995 - val_loss: 0.6968 - val_accuracy: 0.453
    Epoch 3/5
    293/293 [==============================] - 1128s 4s/step - loss: 0.6933 - accuracy: 0.4999 - val_loss: 0.7008 - val_accuracy: 0.453
    Epoch 4/5
    293/293 [==============================] - 1067s 4s/step - loss: 0.6936 - accuracy: 0.5054 - val_loss: 0.6962 - val_accuracy: 0.453

    Epoch 5/5
    293/293 [==============================] - 1128s 4s/step - loss: 0.6933 - accuracy: 0.5089 - val_loss: 0.6942 - val_accuracy: 0.453
    <keras.src.callbacks.History at 0x784bf9f5cca0>
```

Figure 4.12 Execution of 5 epoch using GPU.

Model	Dataset	Accuracy	Precision	Recall	F1 score
LSTM	Validation dataset	45.31%	20.53%	45.31%	28.26%
LSTM	Test dataset	50.54%	25.50%	50.54%	33.94%

Table 4.14 – Results of LSTM on validation and test datasets.

4.5 Deployment

The first stage is to develop a plan for deploying outcomes (Clinton et al., 2000). Depending on the project objectives, the deployment stage might be difficult, such as developing a repeatable data mining process, or simple, such as publishing a report (Schröer et al., 2021). If the results are integrated into the organization's everyday activities, it will be critical to monitor and maintain the algorithms. The many models utilised in the research, particularly Clinical BERT, may be fine-tuned by modifying the model's parameters. The findings can be further enhanced by increasing the processing capability of the resources and then utilising a larger training dataset. As a result, the algorithm requires significant changes before it can be deployed.

CHAPTER 5: DISCUSSION

This chapter focuses on discussing the results of different models executed to predict hospital readmission of patients. Further the results were analysed and linked with respect to the literature from the previous studies for better understanding of the conducted analysis. Before the results are analysed and compared the entire flow of the project implementation is shown in figure 5.1 which is a flow chart.

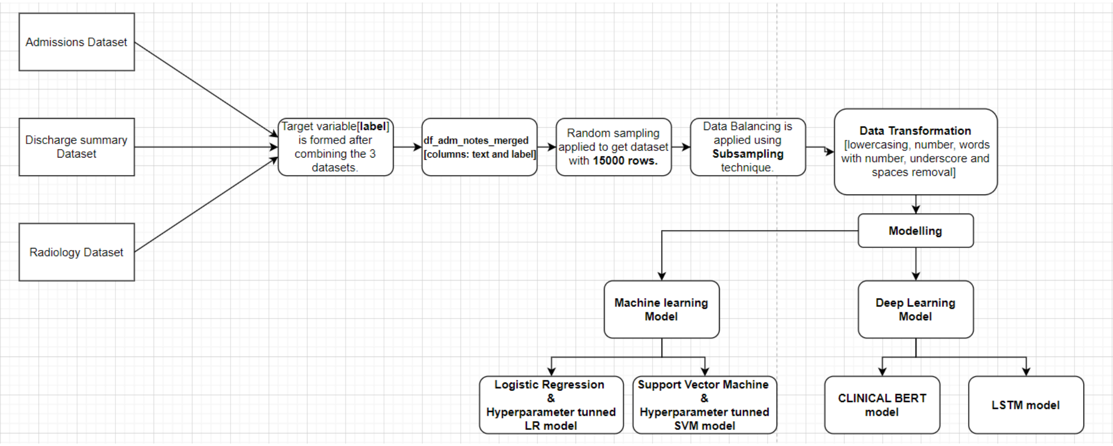

Admissions Dataset
Discharge summary Dataset
Radiology Dataset
Target variable[label] is formed after combining the 3 datasets.
df_adm_notes_merged [columns: text and label]
Random sampling applied to get dataset with 15000 rows.
Data Balancing is applied using Subsampling technique.
Data Transformation [lowercasing, number, words with number, underscore and spaces removal]
Modelling
Machine learning Model
Deep Learning Model
Logistic Regression & Hyperparameter tunned LR model
Support Vector Machine & Hyperparameter tunned SVM model
CLINICAL BERT model
LSTM model

Figure 5.1 – Flow chart of Project implementation.

5.1 Models Results

Table 5.1 presents the results of all the models executed in this study.

Machine Learning models SVM and logistic regression both were competitively efficient. In terms of accuracy the SVM model outperformed logistic regression on validation dataset by a slight edge of 0.20% while on test dataset the margin was of 0.40%. In terms of precision, the validation dataset has identical precision for class [0] for both models, but SVM is better by 1% for class [1]. While on test dataset the logistic regression model outperforms SVM by 1% for class [0] but SVM performs 2% better for class [1] while in terms of recall, the validation dataset has identical recall for class [1] for both models, but SVM is better by 1% for class [0]. While on test dataset the logistic regression model outperforms SVM by 1% for class [0] but SVM performs 1% better for class [1]. Also, with respect to the hyperparameter tunning the SVM model performs better than the logistic regression with an accuracy of 82.68%. To conclude, both models offer comparable performance on the given datasets, but SVM holds edge with respect to accuracy and performance for class [1].

Further the deep learning models LSTM and Clinical BERT had extremely opposite results. The accuracy of Clinical BERT model has drastically outperformed LSTM models. After 4 runs of Clinical BERT on the entire dataset for training, validation and test results were 64.94% and 67.16%, respectively. Meanwhile, after 5 runs of the LSTM model on the training dataset, validation and test results were 45.31% and 50.54%, respectively. The precision, recall and F1 score for Clinical BERT model is 61.07%, 62% and 61.73% respectively for validation dataset while for test its 68.54%, 65% and 66.59% respectively while for LSTM model is 20.53%, 45.31% and 28.26%

respectively for validation dataset while for test its 25.50%, 50.54% and 33.94% respectively.

The results are extremely low for the LSTM model as compared to Clinical BERT model. The low precision indicates the small chunk of instances that were predicted positive and were actually positive while recall highlights the percentage of actual positive instances that model identify correctly. Finally, F1 score suggest that there's a significant imbalance between precision and recall thus the ability to balance false positives and false negatives isn't very good. The extremely low results of LSTM model are due to the complexity of model, the model was trained on a training dataset with just 4685 rows for a maximum of 5 epoch on the GPU, the number of epochs could not be increased as even after using GPU the system was not responding and even crashed a few times. Thus, the complexity of data and model made it hard to run the model multiple times even on the GPU for high end results and were major constraints to yield good results specifically for deep learning model like LSTM in this project.

Model	Dataset	Accuracy	Precision	Recall	F1 score
Machine Learning Models					
Logistic Regression	Validation dataset	80.77%	85% [0] * 76% [1] *	78% [0] 84% [1]	82% [0] 80% [1]
Logistic Regression	Test dataset	82.58%	84% [0] 81% [1]	83% [0] 83% [1]	83% [0] 83% [1]
Hyperparameter tunned LR	Test dataset	81.79%	84% [0] 80% [1]	79% [0] 85% [1]	81% [0] 82% [1]
SVM	Validation dataset	80.97%	85% [0] 77% [1]	79% [0] 84% [1]	82% [0] 80% [1]
SVM	Test dataset	82.98%	83% [0] 83% [1]	82% [0] 84% [1]	83% [0] 83% [1]
Hyperparameter tunned SVM	Test dataset	82.68%	82% [0] 83% [1]	82% [0] 83% [1]	82% [0] 83% [1]

***Class [0] – Here 0 indicates no readmission**
***Class [1] – Here 1 indicates readmission occurred**

Model	Dataset	Accuracy	Precision	Recall	F1 score
Deep Learning Models					
Clinical BERT	Validation dataset	64.94%	61.07%	62%	61.73%
Clinical BERT	Test dataset	67.16%	68.54%	65%	66.59%
LSTM	Validation dataset	45.31%	20.53%	45.31%	28.26%
LSTM	Test dataset	50.54%	25.50%	50.54%	33.94%

Table 5.1 – Result Comparison of different models

5.2 Discussion

The main purpose of this study is to predict hospital readmission using clinical notes. The exploration of text classification using machine learning and deep learning models, the results showcase a synergy between the established methodologies and recent techniques. The traditional methods like random forest, logistic regression, SVM among many others have been widely used for hospital readmission prediction research according to various studies (Li et al., 2020; Michailidis et al., 2022). In the study conducted the accuracy of logistic regression model on validation dataset is 80.77% while on test its 82.58%. On the contrary the study by Li et al., (2020) highlighted use of logistic regression with penalization and noted an accuracy of 77.59%. These results shows that the model works efficiently for predicting readmissions on the new MIMIC-IV dataset with a slight improvement in the accuracy of model. Further, the SVM model in the current study achieved an accuracy of 80.97% on validation dataset and 82.98% on test dataset while the study by Michailidis et al., (2022) which compared different machine learning models found balance random forest to be better than SVM in their study while our results are still competitive and indicate the robustness of SVM in handling such tasks. Transitioning to deep learning models, the landscape of text classification has been significantly redefined. The recent surge in the use of transformer-based models like BERT for various NLP tasks has been profound (Durmus et al., 2019). The Clinical BERT model on the validation dataset achieved an accuracy of 64.94% and on the test dataset an accuracy of 67.16% while in the literature, (Huang et al., 2019) found Clinical BERT outperformed the standard BERT model with an AUROC score of 0.714 ± 0.018. Another study by (Bikram et al., 2022) found that Clinical BERT performed better than other deep contextual representation techniques. The accuracy achieved by this study is comparatively lower which is highly dependent on factors such as the size and quality of training data, hyperparameter tuning, and domain specificity. The LSTM model had a poor performance in this study but previous study by (Ashfaq et al., 2019) presents that LSTM model combined with other data sources performs better than other models. The accuracy in this study is at the lower end which may be due to the context of clinical measurements, face challenges due to potential delays and missing values in time series data (Sushrith et al., 2021). This study presented that traditional machine learning models achieved high accuracy as compared to deep learning models and it aligns with the discussion from (Futoma et al., 2015) who pointed that while deep learning models can capture complex nonlinearities and interactions in the data, they are sometimes more challenging to tune and interpret compared to traditional machine learning models.

5.3 Summary

The chapter started with an overview of the entire implementation process; further, the results of all the different models were analysed and compared, which highlighted that machine learning models still perform better than deep learning when the data provided is limited. The last section of this chapter compared the results of this study with the results of previous studies to understand if the model performance has been affected by the new dataset, as all the previous studies focused on the MIMIC-III dataset.

CHAPTER 6: CONCLUSION

This chapter summarises the project's primary objectives, presents the research contributions made and examines the potential additional study and development that could result from this research.

6.1 *Summary of the Research Project*

The study which focuses on predicting hospital readmission by creating models like Clinical BERT using MIMIC-IV dataset had certain objectives.

- The first objective is to understand the scope of the study and the research already done in this area. The studies presented that different techniques have been used for this prediction including traditional methods to more complex methods like neural networks and the comparative results of these methods are used as a base for understanding the issues that were left unnoticed in previous studies. The highlight from the reviews was that there was no major study done using MIMIC-IV dataset making our study unique from others. There is very little study done using clinical BERT, so it was worth exploring how the model works with the latest clinical notes dataset.
- The second objective of this study was to identify the best research approach for this study. There were different methodologies like SEMMA, KNN and CRISP-DM which were explored to understand their working. The comparison of different approaches identified CRISP-DM as the best choice for this study as it works efficiently on large datasets and helps in understanding the business and data for a better execution of project.
- The next objective was to identify the correct subset of the data for execution. Based on the previous studies it was identified that MIMIC-III had been used immensely for this prediction thus, to make the study unique we use MIMIC-IV dataset. The dataset is prepared by using 3 different datasets and even the target variable is calculated based on the features of the 3 datasets. Finally, the huge dataset is sampled with just 15000 rows which are further divided into train, valid and test datasets in the ratio 70%,15%, 15% respectively.
- The four objective is to execute different models on this dataset to predict hospital readmissions. The study executed 4 models, 2 of which were machine learning namely logistic regression and SVM while the other two are deep learning models LSTM and Clinical BERT.
- Finally, the last objective was to evaluate the results obtained from this study. The results clearly indicate that the machine learning model were quicker and more accurate on the small subsection of MIMIC-IV dataset whereas complex deep learning models performed poorly as due to lack of data models were not trained accurately.

The completion of these five objectives supported in the completion of the aim which focused on developing the clinical BERT model for predicting hospital readmission. The study also executed models like SVM, logistic regression and LSTM on the new MIMIC-IV dataset for a better comparative study as well as a comparison. As the aim was achieved the conclusion highlighted that deep learning models are extremely complex and time consuming therefore its necessary to run such models with high computational power and large dataset to get more accurate and generalized models.

6.2 *Research Contribution*

This study involves the use of MIMIC-IV dataset which makes it unique from the existing studies for hospital readmission. This study will act as a new contribution to the healthcare field with its results highlighting the hospital readmission prediction. Though the study utilizes only about 15,000 rows out of a total dataset with 167,258 rows from MIMI-IV the results obtained in this study are unique for every model while the previous studies presented results using different models but none of the studies used the MIMIC-IV dataset. The study contributed to the healthcare field by providing a model for predicting readmission, this can also be used for other predictions in healthcare field by simply fine tuning the model. It can also be used as a baseline for studies in field like finance, bio medical, bio medical literature by using just a different form of BERT model such as finBERT, BioBERT and SpanBERT, respectively. Thus, it’s interesting to see how this study can help in exploring different variants of BERT in the future.

6.2.1 Theoretical Implications

The study can be used as a theoretical guide for future researchers as it gives an insight into the different machine learning and deep learning models and their results on MIMIC-IV dataset. The researchers can use this as a baseline to either use MIMIC-IV dataset for any other area of research or to use the BERT model as there are different BERT models that can be used for implementing different studies. The limitations of this study will guide them to improve the results for their studies.

6.2.2 Practical Implications

The study has major practical implications in the field of healthcare, the model can be improved by training it with more data and computational power making it more generalized. Then it can be used by healthcare professionals and government officials to maintain the quality of the services provided by calculating the chances of readmission for the patients. It can also be modified and used by government to keep a track on the diseases that have a high impact on the population. If the study is modified using finBERT model it will have a great practical implementation in financial sector helping the employees of banks, multinational companies and even the finance sector of government.

6.3 Limitations

The limitation of this study is that only 15,000 rows of MIMIC-IV dataset are used, this resulted in the deep learning models performing less efficiently. The models could have performed better if more amount of data was used also the models could not train themselves for more than 5 epochs even after using GPU, this could be because of the quality and complexity of data.

6.4 Future Work

Based on the limitations of this study the future research needs to ensure that more data is used for training purpose and a more efficient GPU is used to ensure that the model could be trained multiple times to obtain better results and to form a more generalized model.

6.5 Personal Reflections

This study allowed me to gain insight into medical data analysis using a new dataset which was not explored much by previous research. It also helped me examine the efficiency of different machine learning and deep learning models especially Clinical BERT in predicting hospital readmission for patients. This experience not only enriched my technical skills but also raised ethical considerations regarding the use of patient data and model interpretability. Working on this project aided me in comprehending the various packages and libraries utilized in Python, as well as in formulating a suitable research objective. I have set a personal goal to explore additional Python libraries and to attain a deeper proficiency in the programming language. I would further want to explore the BERT model in different domains like finance and education.

REFERENCES

Aftan, S. & Shah, H. (2023) ‘A Survey on BERT and Its Applications’, in *20th International Learning and Technology Conference, L and T 2023*. [Online]. 2023 Institute of Electrical and Electronics Engineers Inc. pp. 161–166.

Alsentzer, E. et al. (2019) *Publicly Available Clinical BERT Embeddings*. [online]. Available from: http://arxiv.org/abs/1904.03323 (Accessed 30 August 2023).

Al-Rfou, R. et al. (2018) *Character-Level Language Modeling with Deeper Self-Attention*. [online]. Available from: http://arxiv.org/abs/1808.04444.

Ashfaq, A. et al. (2019) Readmission prediction using deep learning on electronic health records. *Journal of Biomedical Informatics*. [Online] 97.

Azevedo, A. & Santos, M. F. (2008) *KDD, semma and CRISP-DM: A parallel overview Business Intelligence-Implantation on Federal Institute of Triângulo Mineiro (IFTM) System View project KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW*. [online]. Available from: https://www.researchgate.net/publication/220969845.

Banerjee, D. et al. (2017) An informatics-based approach to reducing heart failure all-cause readmissions: The Stanford heart failure dashboard. *Journal of the American Medical Informatics Association*. [Online] 24 (3), 550–555.

Bikram, N. et al. (2022) *Hospital Readmission Prediction Using Clinical Admission Notes*. [Online] [online]. Available from: https://doi.org/10.1145/3511616.

Blinder, Y. (2017) *Predicting 30-day ICU readmissions from the MIMIC-III database Capstone Project Machine Learning Engineer Nanodegree*. [online]. Available from: https://mimic.physionet.org/.

Britz, D. et al. (2017) *Massive Exploration of Neural Machine Translation Architectures*. [online]. Available from: https://github.com/moses-.

Chandra, A. et al. (2023) Transformer-based deep learning for predicting protein properties in the life sciences. eLife 12.

Cheng, J. et al. (2016) *Long Short-Term Memory-Networks for Machine Reading*. [online]. Available from: http://arxiv.org/abs/1601.06733.

Clinton, J. et al. (2000) *CRISP-DM 1.0 Step-by-step data mining guide*.

Dai, Z. et al. (2019) *Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context*. [online]. Available from: http://arxiv.org/abs/1901.02860.

Devlin, J. et al. (2019) *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. [online]. Available from: https://github.com/tensorflow/tensor2tensor.

Durmus, E., Ladhak, F., & Cardie, C. (2019). *Determining Relative Argument Specificity and Stance for Complex Argumentative Structures*. https://doi.org/10.18653/v1/P19-1456

Fayyad, U. et al. (1996) The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM*. [Online] 39 (11), 27–34.

Futoma, J. et al. (2015) A comparison of models for predicting early hospital readmissions. *Journal of Biomedical Informatics*. [Online] 56229–238.

Ganesh, P. et al. (2021) Compressing Large-Scale Transformer-Based Models: A Case Study on BERT. *Transactions of the Association for Computational Linguistics.* [Online] [online]. Available from: http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00413/1964006/tacl_a_00413.pdf.

Hasan, O. et al. (2010) Hospital readmission in general medicine patients: A prediction model. *Journal of General Internal Medicine*. [Online] 25 (3), 211–219.

Hou, W. & Ji, Z. (2023) GeneTuring tests GPT models in genomics. *bioRxiv: the preprint server for biology*. [Online] [online]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/36993670.

Huang, K. et al. (2019) ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv:1904.05342.* [online]. Available from: http://arxiv.org/abs/1904.05342.

Huang, Y. et al. (2021) Application of machine learning in predicting hospital readmissions: a scoping review of the literature. *BMC Medical Research Methodology*. [Online] 21 (1),

Huang, Y. et al. (2022) Machine learning methods to predict 30-day hospital readmission outcome among US adults with pneumonia: analysis of the national readmission database. *BMC Medical Informatics and Decision Making*. [Online] 22 (1),

Hung, M. et al. (2020) Prediction of 30-day hospital readmissions for all-cause dental conditions using machine learning. *Risk Management and Healthcare Policy*. [Online] 132047–2056.

Jegan, C., Kumari, V. A., & Chitra, R. (2013). *Classification Of Diabetes Disease Using Support Vector Machine Identification and Rectification of Security Issues in IOT View project IDENTIFICATION OF MICRO-ANEURYSMS IN DIABETIC RETINOPATHY USING DEEP LEARNING View project Classification Of Diabetes Disease Using Support Vector Machine*. *3*, 1797–1801. https://www.researchgate.net/publication/320395340

Ji, S. et al. (2021) Does the magic of BERT apply to medical code assignment? A quantitative study. *Computers in Biology and Medicine*. [Online] 139.

Jiang, S. et al. (2018) An integrated machine learning framework for hospital readmission prediction. *Knowledge-Based Systems*. [Online] 14673–90.

Johnson, A. E. W. et al. (2016) MIMIC-III, a freely accessible critical care database. *Scientific Data*. [Online] 3.

Johnson, A. et al. (2022) *MIMIC-IV*. [online]. Available from: https://doi.org/10.13026/S6N6-XD98 (Accessed 26 August 2023).

Johnson, A. E. W. et al. (2023) MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*. [Online] 10 (1),

Koh, H. C. & Tan, G. (2011) Data Mining Applications in Healthcare. *Journal of Healthcare Information Management.* 19 (2),

Kovaleva, O. et al. (2019) *Revealing the Dark Secrets of BERT*. [online]. Available from: https://gluebenchmark.com/leaderboard.

Lamproudis, A. et al. (2022) *Evaluating Pretraining Strategies for Clinical BERT Models*. [online]. Available from: http://dsv.su.se/healthbank.

Lee, J. et al. (2020) BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. [Online] 36 (4), 1234–1240.

Lewis, P. et al. (2020) *Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art*. [online]. Available from: https://github.com/titipata/pubmed_.

Li, Q. et al. (2020) *How Good Is Machine Learning in Predicting All-Cause 30-Day Hospital Readmission? Evidence From Administrative Data*. [Online] [online]. Available from: https://doi.org/10.1016/j/jval.2020.06.009.

Lin, T. et al. (2021) *A Survey of Transformers*. [online]. Available from: http://arxiv.org/abs/2106.04554.

Liu, X. et al. (2021) *GPT Understands, Too*. [online]. Available from: http://arxiv.org/abs/2103.10385.

Lin, Y. W. et al. (2019) Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long shortterm memory. *PLoS ONE*. [Online] 14 (7),

Liu, Y. & Lapata, M. (2019) *Text Summarization with Pretrained Encoders*. [online]. Available from: http://arxiv.org/abs/1908.08345.

Lomet, D. B. et al. (2000) *Editorial Board Editor-in-Chief Associate Editors*. [online]. Available from: http://list.research.microsoft.com/scripts/lyris.pl?enter=debull.

Michailidis, P. et al. (2022) Forecasting Hospital Readmissions with Machine Learning. *Healthcare (Switzerland)*. [Online] 10 (6).

Palacios, H. J. G. et al. (2017) A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change. *Advances in Science, Technology and Engineering Systems*. [Online] 2 (3), 598–604.

Peters, M. E. et al. (2018) *Deep contextualized word representations*. [online]. Available from: http://allennlp.org/elmo.

Raffel, C. et al. (2020) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*. [online]. Available from: http://arxiv.org/abs/1910.10683.

Raffel, C. et al. (2019) *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. [online]. Available from: http://arxiv.org/abs/1910.10683.

Rahman, F. A. et al. (2014) Knowledge Discovery Database (KDD)-Data Mining Application in Transportation. *Proceeding of the Electrical Engineering Computer Science and Informatics*. [Online] 1 (1).

Rajkomar, A. et al. (2018) Scalable and accurate deep learning with electronic health records. npj Digital Medicine. [Online] 1 (1),.

Reddy, B. K. & Delen, D. (2018) Predicting hospital readmission for lupus patients: An RNN-LSTM-based deep-learning methodology. *Computers in Biology and Medicine*. [Online] 101199–209.

Ribeiro, M. T. et al. (2016) '"Why should i trust you?" Explaining the predictions of any classifier', in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [Online]. 13 August 2016 Association for Computing Machinery. pp. 1135–1144.

Romanov, A. & Shivade, C. (2018) *Lessons from Natural Language Inference in the Clinical Domain*. [online]. Available from: http://arxiv.org/abs/1808.06752.

Schröer, C. et al. (2021) 'A systematic literature review on applying CRISP-DM process model', in *Procedia Computer Science*. [Online]. 2021 Elsevier B.V. pp. 526–534.

Shafique, U. & Qaiser, H. (2014) A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA) Data Mining in Healthcare for Heart Diseases View project A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA) View project A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research 12 (1). [online]. Available from: http://www.ijisr.issr-journals.org/.

Sunitha, T., Chandravallika, M., Ranganayak, M., Suma Sri, G., Jagadeesh, T. V. S., & Tejaswi, A. (2020). *International Conference on IoT based Control Networks and Intelligent Systems (ICICNIS 2020) Predicting the Loan Status using Logistic Regression and Binary Tree*. https://ssrn.com/abstract=3769854

Sushil, M. et al. (2021) *Are we there yet? Exploring clinical domain knowledge of*

BERT models. [online]. Available from: https://play.google.com/store/apps/.

Sushrith, B. et al. (2021) A Case Study on Hospital Readmission Prediction Using Deep Learning Algorithms on EHRs. Turkish Journal of Computer and Mathematics Education 12 (3).

Vaswani, A. et al. (2017) *Attention Is All You Need*. [online]. Available from: http://arxiv.org/abs/1706.03762.

Vig, J. (2019) *Visualizing Attention in Transformer-Based Language Representation Models*. [online]. Available from: http://arxiv.org/abs/1904.02679.

Williams, R. J. & Zipser, D. (1989) A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. Appears in Neural Computation 1.

Wong, F. K. Y. et al. (2010) What accounts for hospital readmission? *Journal of Clinical Nursing*. [Online] 19 (23–24), 3334–3346.

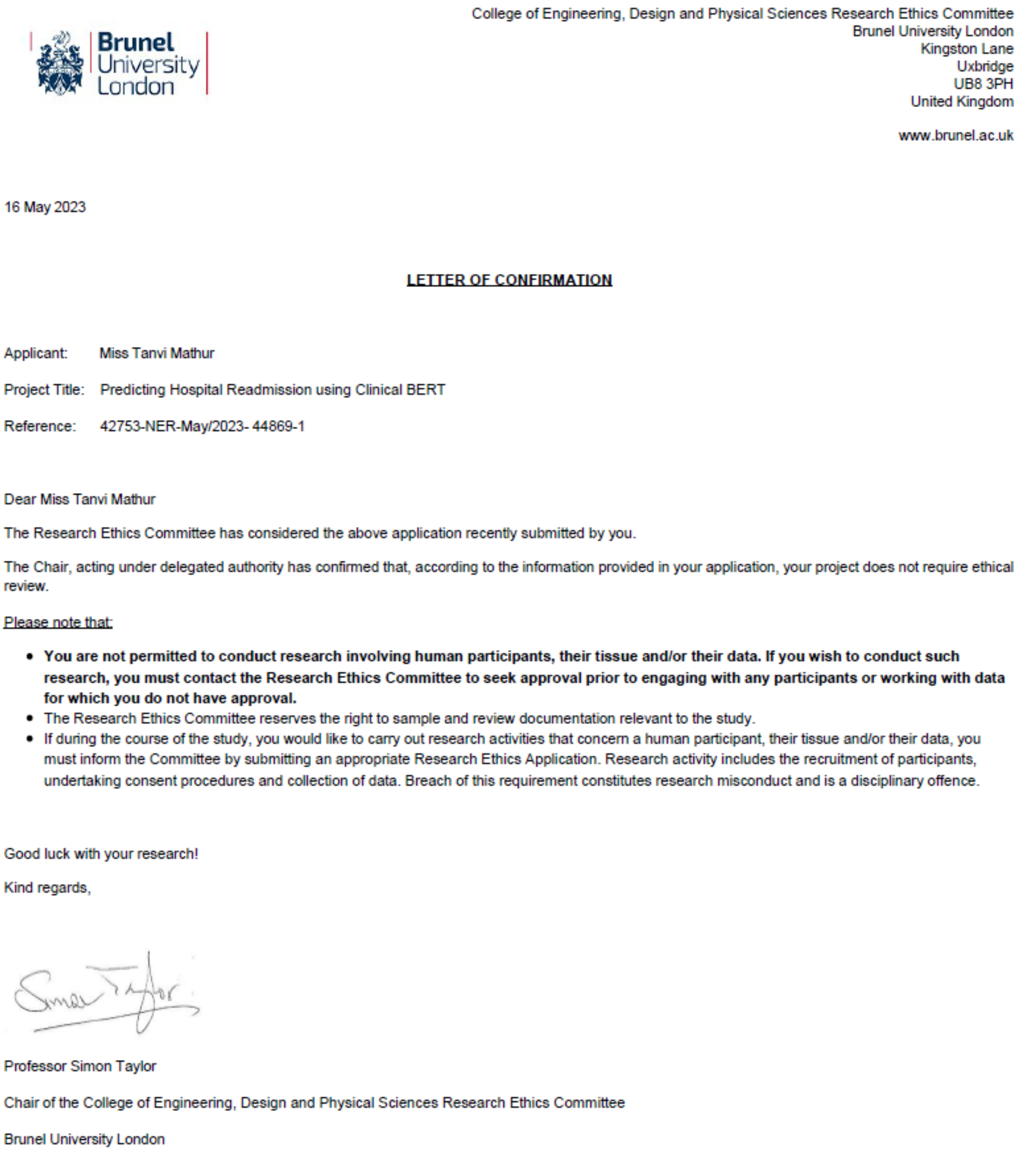
Wu, Y. et al. (2016) *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. [online]. Available from: http://arxiv.org/abs/1609.08144.

Yenduri, G. et al. (2023) *Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions*. [online]. Available from: http://arxiv.org/abs/2305.10435.

Youbi Idrissi, B. et al. (2022) Simple data balancing achieves competitive worst-group-accuracy. Proceedings of Machine Learning Research 140.

Yu Lin, F. & McClean, S. (2001) *A data mining approach to the prediction of corporate failure*. [online]. Available from: www.elsevier.com/locate/knosys.

APPENDIX A: ETHICAL *APPROVAL*



College of Engineering, Design and Physical Sciences Research Ethics Committee
Brunel University London
Kingston Lane
Uxbridge
UB8 3PH
United Kingdom

www.brunel.ac.uk

16 May 2023

LETTER OF CONFIRMATION

Applicant: Miss Tanvi Mathur

Project Title: Predicting Hospital Readmission using Clinical BERT

Reference: 42753-NER-May/2023- 44869-1

Dear Miss Tanvi Mathur

The Research Ethics Committee has considered the above application recently submitted by you.

The Chair, acting under delegated authority has confirmed that, according to the information provided in your application, your project does not require ethical review.

Please note that:

- **You are not permitted to conduct research involving human participants, their tissue and/or their data. If you wish to conduct such research, you must contact the Research Ethics Committee to seek approval prior to engaging with any participants or working with data for which you do not have approval.**
- The Research Ethics Committee reserves the right to sample and review documentation relevant to the study.
- If during the course of the study, you would like to carry out research activities that concern a human participant, their tissue and/or their data, you must inform the Committee by submitting an appropriate Research Ethics Application. Research activity includes the recruitment of participants, undertaking consent procedures and collection of data. Breach of this requirement constitutes research misconduct and is a disciplinary offence.

Good luck with your research!

Kind regards,

Simon Taylor

Professor Simon Taylor

Chair of the College of Engineering, Design and Physical Sciences Research Ethics Committee

Brunel University London

APPENDIX B: OTHER APPENDICES

The code implementation for this study is attached to the Appendix material section of wise flow. The attachment is a zip folder named **final_code.**
The zip folder has following files and folders:
Initial dataset folder: This contains 3 initial datasets namely admission, discharge and radiology which were used to form the single dataset and target variable for the modelling.

Dataset for models: This contains 6 datasets, 3 for machine learning models and LSTM models (train_final, validation_final, test_final) while the other 3 purely for Clinical BERT model (train_bert_final, validation_bert_final, test_bert_final).

Preprocessing_code.ipynb file: This is the python code file containing code for pre-processing of data.

Machine_learning_models.ipynb file: This is the python code file containing code for machine learning models (logistic regression and SVM).

clinicalBERTmodel.ipynb file: This is the python code file containing code for deep learning model Clinical BERT.

lstm_with4epoch8batchsize_withoutgpu.ipynb file: This is the python code file containing code for deep learning model LSTM for condition where GPU was not used just high RAM was used.

lstm_with5epoch16batchsize_withgpu.ipynb file: This is the python code file containing code for deep learning model LSTM for condition where GPU was used.