National College of Ireland

Comparative Evaluation of RNN Architectures and Embedding Strategies for Multi-Label Disease Classification from Clinical Narratives

MSc Research Project
MSC DATA ANALYTICS

Sonali Anil Birje
Student ID: X23286954

School of Computing
National College of Ireland

Supervisor: Furqan Rustam

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Sonali Anil Birje

Student ID: X23286954

Year: 2024-2025

Programme: MSC Data Analytics

Module: MSC Data Analytics Research Project

Supervisor: **Furqan Rustam**

Submission Due Date: 11/08/2025

Project Title: Comparative Evaluation of RNN Architectures and Embedding Strategies for Multi-Label Disease Classification from Clinical Narratives

Word Count: 6478 **Page Count** 26

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Sonali Anil Birje
Date: 11/08/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Comparative Evaluation of RNN Architectures and Embedding Strategies for Multi-Label Disease Classification from Clinical Narratives

Sonali Anil Birje
X23286954

Abstract

Unstructured clinical discharge summaries represent a valuable source of clinically relevant information, although they produce challenges to automated processing due to large amounts of narrative information, inclusion of domain-specific terminologies, and high rates of multiple comorbid conditions. In this respect, conventional machine-learning methods like Support Vector Machines and logistic regression are limited by the fact that they rely on manually designed features, by their inability to model long-range interactions, and by overfitting high-dimensional, sparse data. This paper suggests a deep-learning model that uses three recurrent neural network structures, including Vanilla RNN, Gated Recurrent Unit (GRU), and Bidirectional Long Short-Term Memory (BiLSTM), to classify diseases in multi-label classification of discharge summaries. A synthetic dataset was created in a balanced way with 2,000 records, matching the International Classification of Diseases, Ninth Revision (ICD-9) diagnostic codes with clinical narrative templates thus ensuring the equal distribution over the fifteen illness categories and not compromising the confidentiality of a patient. The text was preprocessed and tokenized, noise removed, and sequences padded, and then represented using two types of embeddings, general-purpose GloVe embeddings and domain-specific BioBERT contextual embeddings. All model-embedding combinations were trained with binary cross-entropy loss and early stopping rules and performance assessed on a reserved test set in terms of Micro and Macro F1-scores, precision, recall, and ROC-AUC. The results indicate that the BiLSTM model augmented with BioBERT embeddings has reached the best overall performance, as it produced a mean ROC-AUC of 0.92 and better F1-scores of most labels, therefore, supporting the idea of combining the context modeling capacity with domain-specific embeddings. These findings offer a reproducible framework of building ethical, accurate and interpretable natural-language processing applications in healthcare without relying on actual patient data.

Keywords: Multi-label Classification, Discharge Summaries, RNN, GRU, BiLSTM, BioBERT, GloVe, Clinical NLP.

1. Introduction

There is a significant amount of clinical text being generated with the widespread use of Electronic Health Records (EHRs) and discharge summaries are the main locations of diagnostic and therapeutic data. The stories document the medical history of a patient, examination results, diagnosis, and discharge prescriptions, and can be used to promote continuity of care, assist clinical research, and health-care analytics. According to the Office of the National Coordinator for Health Information Technology, over 96 percent of the non-federal hospitals across the United States use certified EHR systems which generate billions of free-text documents each year. Discharge summaries are inherently unstructured, often written in heterogeneous narrative forms and with a lot of jargon, acronyms, and clinical jargon. In addition, they tend to describe various concomitant variables, which makes multi-label disease classification both a necessary and a daunting experience. These complexities compromise the practicality of the conventional automated systems hence requiring advanced methods that can work with sequential relationships, complex semantics, and multi-label nature of clinical narratives. (Koopman et al. 2015)

Earlier medical text classification methods were reliant on statistical and machine learning algorithms such as Naive Bayes, logistic regression and Support Vector Machine. (Kim 2002) Such approaches typically require manual feature engineering e.g. using bag-of-words or TF-IDF Though such methods perform well on structured data, they work poorly with long, complex clinical histories because they fail to index contextual semantics, sequential dependencies and polysemous medical terms. (Koopman et al. 2015) Besides, unequal distributions of classes in real-life data reduce their ability to predict rare occurrences lucratively.

The emergence of new milestones in Natural Language Processing (NLP) has revolutionized clinical text classification due to the automated extraction of semantic and contextual content extracted by the modern models without the use of manual feature engineering.(Wang et al. 2019) Sequential deep learning models are especially useful in narrative medical text and, in particular, Recurrent Neural Networks (RNNs), which can learn temporal dependencies over long sequences. Three variants of RNN are studied, namely, vanilla RNN, Gated Recurrent Unit (GRU), and Bidirectional Long Short-Term Memory (BiLSTM). Vanilla RNNs offer the simplest and computationally economical architecture of sequential data, whereas GRUs offer an efficient alternative to the LSTMs (Edara et al. 2019) . BiLSTMs however have the possibility of bettering context learning as they are bidirectional which is vital to medical narratives because the diagnostic hints may be prior or subsequent to the mention of a disease. The three models have been enhanced with the use of GloVe general-purpose embeddings as well as BioBERT domain-specific embeddings thus enabling better representation of medical language and performance in multi-label disease prediction tasks. (Uzuner 2009)

This paper examines the automatic multi-label disease classification of unstructured clinical discharge notes through use of a deep learning platform. The methodology starts with the creation of a balanced synthetic corpus of 2,000 records that is generated by systematically matching ICD-9 diagnosis codes to thoughtfully designed clinical narrative templates. Through this approach, it will be relatively easy to attain balanced representation of 15 categories of illnesses without compromising the confidentiality of patients. Three consecutive architectures are chosen to solve

the classification task: Vanilla Recurrent Neural Network, Gated Recurrent Unit (GRU) and Bidirectional Long Short-Term Memory (BiLSTM). We consider two embedding strategies: One, called GloVe, provides general-purpose semantic embeddings, and two, a domain-specific embedding model called BioBERT, which is pretrained on biological literature. Preprocessing includes: tokenization, removal of noise that is irrelevant to context, and sequence padding to a fixed length. The two models are both trained on binary cross-entropy loss, early stopping, and assessed through Micro/Macro F1-scores, precision, recall, and ROC-AUC. The goal is to determine the best combination of the architectural settings and embedding modes to perform precise, generalizable, and ethically correct multi-label disease categorization in clinical narratives through a comparative study under the same experimental conditions.

Research Objective and Questions:

The current study aims to compare the performance of some of the recurrent neural network (RNN) models in classifying a range of diseases based on unstructured clinical discharge summaries. The goal is to develop a systematized, morally justified structure, which will work on a balanced synthetic dataset, thus guaranteeing balanced representations of the various categories of pathology.

1. To design and preprocess a balanced synthetic dataset of 2,000 discharge summaries using ICD-9 mappings and clinical-style templates.
2. To implement and compare the performance of **Vanilla RNN**, **GRU**, and **BiLSTM** architectures.
3. To assess the impact of **GloVe** general-purpose embeddings versus **BioBERT** domain-specific embeddings on classification performance.
4. To evaluate models using multi-label metrics including Micro/Macro F1-scores, precision, recall, and ROC-AUC.

Contribution

The current study goes beyond the previous research to make the following major contributions:

1. **Development of a Synthetic Dataset**: A synthetic corpus of 2,000 discharge summaries was generated, using ICD-9 diagnostic codes and matching them with well-designed clinical templates, and with a balanced representation of fifteen disease categories, whilst protecting patient confidentiality.

2. **Model Implementation and Comparison:** A comparative analysis of three deep learning models namely Vanilla RNN, GRU and BiLSTM was performed in a sequential manner with similar experimental set-up.

3. **The strategy of embedding evaluation:** The general-purpose embeddings GloVe were contrasted with domain-specific embeddings BioBERT in a comparative study that aimed to provide an explanation of the effect of embedding selection on the multi-label classification accuracy.

4. **Detailed Evaluation Metrics:** Micro- and Macro-F1 scores, precision, recall, and ROC-AUC were used in order to present a detailed evaluation of how well the categorizations performed in both prevalent and rare disease categories.

5. **Visual Analytics Dashboard**: An interactive Power BI dashboard was created to visualize the model performance, label-specific results and patterns of co-occurring illnesses, further allowing easy understanding by both technical and healthcare decision-makers.

The structure of this thesis is that in section 2, both classical, machine learning and deep learning and hybrid systems of medical text classification will be discussed in detail under a detailed literature review. Section 3 provides information about the methodology, such as the generation of datasets, preprocessing, features engineering and model design. Design parameters are defined in Section 4 with Section 5 describing the stepwise implementation process. Section 6 notes on the results of the evaluation and consequent discussion. Finally, Section 7 puts a conclusion to the report and suggests possible future work.

Below Figure 1 image shows a conceptual idea of the project.

Figure 1: Conceptual Representation of Multi-Class Classification

2. Related Work

2.1 Traditional Approaches

Early attempts at disease classification of clinical text relied mostly on rule-based approaches and statistical machine learning methods. Complex systems applied heuristic rules and so-called manually built lists of words and keywords, and hierarchies. Uzuner, et al. created a framework founded on rules, using regular expressions and domain lexicons to assign the ICD codes in discharge summaries, which demonstrated reasonable precision though failed sufficient recall when the terms were given in non-standard forms (Koopman et al. 2015). Similarly, the Apelon terminology engine was employed by Ware et al. to generate sets of synonyms of pharmaceutical names which were embedded into disease identification criteria carefully developed (Uzuner 2009).

Such techniques were comprehensible and low-processing power-requiring in nature yet faced a high scaling limitation. This reliance on literal word matching led to some problems with language diversity, polysemy and the control of negation. Machine learning approaches such as the Naive bayes, and logistic regression, when combined with bag-of-words or TF-IDF features yielded some improvement, yet not to the extent of being able to capture the contextual associations of such

textual narratives in clinical notes [8]. Therefore, the traditional ways, despite being basic, did not help to handle the semantic complexities and length of discharge summaries**.**

2.2 Machine Learning-Based Approaches

The transition towards more than purely rule-based systems to supervised machine learning was a big step towards increasing classification effectiveness. The classification task using medical text has used Support Vector Machines (SVMs), Random Forests, and Gradient Boosting Machines. (Perotte et al. 2014). made the findings that support vector machines classifiers with hierarchical classification models can effectively exploit ICD code taxonomies with improved multi-label using electronic health record data. (Koopman et al. 2015). used Conditional Random Fields with medical concept recognition together in a combination in order to bring the clinical narratives to disease classifications, leading to an improvement in memory with an added complexity to it.
These models did not require the amount of manual curation that their purely rule-based predecessors had required but still depended upon hand-designed features such as n-grams, domain-specific lexicons, or pipelines of idea extraction. This dependence has made them prone to lexical shift and compelled them to go through periodic retraining. Moreover, high sparsity and complexity of the characteristics of bag-of-words limited their ability to capture deep-seated semantic connections in long discharges notes. (Edara et al. 2019)

2.3 Deep Learning based Approaches

The ability to learn semantic and syntactic patterns in raw text has profoundly re-calibrated clinical natural language processing, by enabling a model to process raw text without the need to rely on conventional feature engineering. (Aden and Reyes-Aldasoro 2024)In that context, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) models and Gated Recurrent Units (GRUs) have become the main instruments of clinical text classification. (Liu and Guo 2019)

Even though vanilla RNNs represent a relatively simple family, they have been the subject of sustained interest in biomedical sequence processing in applications with short or otherwise fairly direct dependencies. Such networks though are still vulnerable to gradient vanishing, which is a drawback that makes them useful as efficient baselines to assess the improvement offered by more advanced models.

According to the latest research by (Lu et al. 2022), CNN, RNN, LSTM, BiLSTM, GRU, Transformer encoder, and BERT models were used to work on the Obesity Challenge dataset, with the researchers finding that Transformer encoders tend to outperform their peers, and BiLSTMs and CNNs have similar results regardless of the moderate imbalance in classes. Previously, Liu and (Liu and Guo 2019) proposed a BiLSTM + CNN + attention model with multi-label classification of clinical notes and demonstrated its superior accuracy compared to the baseline BiLSTM, with no results reported on balanced data.

According to empirical evidence, domain-specific embeddings, e.g. BioBERT, pretrained on PubMed and PMC, perform better than general-purpose representations like GloVe. The contextualized lexicons enhance the management of clinical terms, which is an essential factor in the processing of medical stories. (Lee et al. 2019)

Overall, deep learning has re-conceptualized clinically-focused NLP, providing systems that can autonomously learn semantic and syntactic regularities and forego the traditional feature engineering. Hence, CNNs, RNN, LSTMs, GRUs, and their extensions have become common components of clinical text classification models.

2.4 Hybrid or Other Relevant Approaches

Hybrid structures that combine a large number of brain paradigms have emerged to represent local as well as global contextual information. (Jang et al. 2020) have proposed a Bi-LSTM-CNN-attention network based on Word2Vec embeddings with a combination of CNN pattern-size perception capabilities and the Bi-LSTM sequential understanding capabilities to boost the accuracy of the multiple-label electronic health record classification problem (Jang et al. 2020). (Reddy et al. 2025). introduced an Enhanced Convolutional Attention Network (EECAN) that is optimized to discharge summary classification, which outperformed traditional CNNs by including layers of attention to focus on clinically relevant phrases.(Hu et al. 2021)

There has also been investigation of graph-based models. (Li et al. 2024) Bo Li et al. structured a label-specific attention model, known as KeNet using medical knowledge graphs, to enable learning representations related to each disease label. As powerful as these strategies prove to be, their reliance on external knowledge bases of a structured form can be a limit to their application in situations where such resources are unavailable. (Li et al. 2024)

2.5 Limitations and Gaps

Although deep learning models and their success have been proven in clinical text classification, there are a number of limitations still present. Modern research virtually never performs a direct comparison of Vanilla RNN, GRU and BiLSTM architectures under the same experimental conditions on balanced multi-label data. Previous works also depend on the real-life data which is naturally skewed, thus not providing an opportunity to assess the performance fairly. Besides, though domain-specific embeddings such as BioBERT and ClinicalBERT have demonstrated their potential, they have not been studied enough in relation to their effect in the RNN-family models (Du et al. 2019; Lee et al. 2019).(Huang et al. 29AD).The proposed study will fill those gaps by combining GloVe and BioBERT embeddings to study the effectiveness of a general-purpose versus biomedical-specific embedding. It uses a balanced dataset, synthetically created, tests three consecutive architectures: Vanilla RNN, GRU and BiLSTM, in controlled experimental conditions, thus offering a complete evaluation of model performance, embedding effects and architecture efficiency in multi-label disease classification of clinical narratives.

Table 1: Summarization of related work

Author(s) & Year	Method / Model	Dataset	Accuracy / Results	Key Limitation
Uzuner et al., 2009	Rule-based extraction using regular expressions and domain lexicons	i2b2 Obesity Challenge dataset	High precision in matching keywords; lower recall for varied term usage	Struggles with linguistic variability; limited adaptability
Ware et al., 2009	Apelon terminology engine + handcrafted rules	i2b2 discharge summaries	Good precision for known drug/disease terms	Poor generalisation to unseen terminology

Author(s) & Year	Method / Model	Dataset	Accuracy / Results	Key Limitation
BMC Med Inform Decis Mak, 2018	Naïve Bayes & SVM with TF-IDF	Hospital EHR notes	Baseline accuracy ~70%	Heavy reliance on manual feature engineering
Perotte et al., 2014	Hierarchical SVM for ICD coding	NY-Presbyterian Hospital discharge summaries	Macro-F1: 0.39; Micro-F1: 0.53	Still feature-engineering dependent; class imbalance issues
Koopman et al., 2015	CRF + medical concept recognition	De-identified hospital narratives	F1 improvement over baseline ML models	High complexity; dependent on concept extraction tools
Survey (arXiv:2107.10652)	Review of automated ICD coding methods	Multiple benchmark datasets	Highlights DL outperforming ML in most cases	Lack of standardised evaluation protocols
Lu et al., 2022	CNN, RNN, LSTM, Bi-LSTM, GRU, Transformer, BERT	Obesity Challenge discharge summaries (16 diseases)	Transformer AUC-ROC up to 0.926; CNN competitive under balanced conditions	No evaluation of domain-specific BERT variants (e.g., BioBERT, ClinicalBERT)
Edara et al., 2023	LSTM for emotion & label classification	Oncology patient narratives	High robustness to noisy clinical language	Focused on emotion detection; not full disease classification
Liu & Guo, 2019	Bi-LSTM + CNN + attention	Medical notes corpus	Higher accuracy than vanilla Bi-LSTM; improved recall	No evaluation on balanced datasets
BioBERT (Lee et al., 2020)	BERT pretrained on biomedical corpora	PubMed, PMC text	Outperforms BERT on biomedical NER & QA tasks	Not fine-tuned on discharge summaries
ClinicalBERT (Alsentzer et al., 2019)	BERT fine-tuned on clinical text	MIMIC-III discharges summaries	Improves predictive tasks (e.g., readmission)	Limited availability of large-scale clinical training data
Jang et al., 2020	Bi-LSTM + CNN + attention (Word2Vec)	EHR narratives	Improved multi-label classification over baselines	Uses general embeddings; lacks domain-specific tuning
Reddy et al., 2021	Enhanced CNN with attention (EECAN)	Discharge summaries	Higher F1 & AUC than standard CNN	No recurrent layers; may miss long-term dependencies
Bo Li et al., 2024	KeNet: Label-wise attention + medical knowledge graph	MIMIC-III	Boosted per-label F1 using knowledge graph	Reliant on structured ontologies not always available

3. Research Methodology

3.1Overview

It carries out a systematic review of RNN architecture on multi-label disease classification on free-form discharge summaries. The methodology is composed of five stages that run sequentially: the construction of a dataset, pre-processing, the configuration of embedding, training, and evaluation.

The creation of a balanced synthetic dataset, consisting of 2,000 discharge summaries, involved matching ICD-9 diagnosis codes to curated clinical templates, obtaining a fair representation of the 15 disease categories and still protecting the confidentiality of patients. Preprocessing of texts consisted of lowercasing, elimination of noise, tokenization, and padding of the sequences to the fixed length.

Three consecutive deep learning models Vanilla RNN, Gated Recurrent Unit (GRU) and Bidirectional Long Short-Term Memory (BiLSTM) were explored. Two paradigms of feature representations were used, namely, general-purpose GloVe and domain-specific BioBERT embeddings trained on biomedical publications. The model-embedding settings that were investigated were trained using the binary cross-entropy loss, where early termination was used to prevent overfitting. The systematic evaluation involved such measures as accuracy and Area Under the Receiver Operating Characteristic Curve (AUC) calculated both on the training and validation sets. These measures were chosen to give a summary measure of classification accuracy, and a measure of how much the classifier depends on the value of the threshold (AUC). The experiments were performed with the same experiment protocols to get a fair experiment of the comparison between the design architecture and embedding schemes.

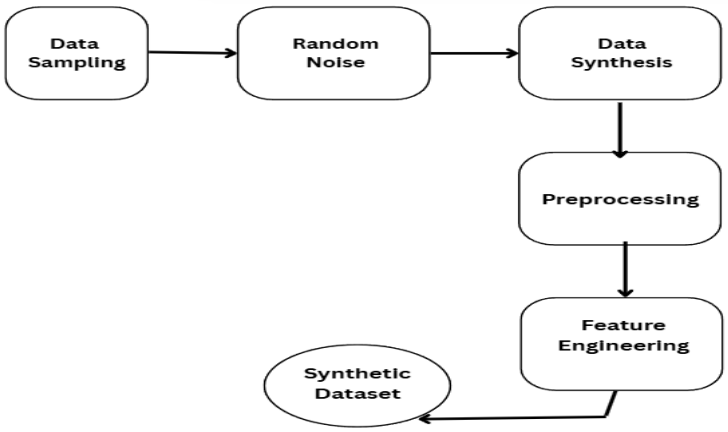

Figure 2: Synthetic dataset creation and preprocessing workflow.

3.2Dataset

This is based on a synthetically generated database comprising 2,000 discharge summaries synthetically created by matching International Classification of Diseases (ICD-9) diagnostic codes to clinical language templates. The ICD-9 codes were downloaded using the publicly available tables of the MIMIC-III dataset diagnoses to provide a full reflection of most of the disease categories in the setting of hospitalization discharge (Koopman et al. 2015). Every hospital admission simulated received one to three diagnoses randomly assigned to reflect real co-morbidity (Uzuner 2009). Spreadsheet summaries were generated with the help of a set of hand curated text templates that combine standard clinical vocabulary.

The aim of this synthetic approach was twofold: it was, first, to ensure that the labels are distributed in a fair manner across the categories of illness, ensuring that the models could be evaluated in an objective manner; and second, to maintain patient confidentiality by avoiding use of identifiable real world patient data. The balanced structure of the data ensures that every

disease label is evenly represented, which avoids the bias typical of real EHR data, where certain afflictions may be reported at low rates.

To remove inconsistencies, the data set underwent data cleaning and preprocessing before the model was trained. This included having all text turned to lowercase, removing the punctuation and numerical tokens indicating nothing pertinent to diagnosis, removal of duplicate entries, and application of lemmatization to reduce morphological variation. The ultimate dataset was then segregated into training (80%) as well as test (20%) groups, whereas stratification was adopted, to ensure the multi-label distribution was preserved across the sets.

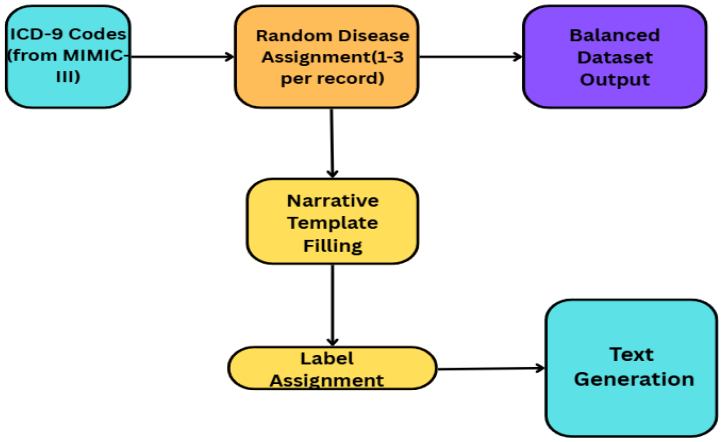

Figure 3: Process of generating and preprocessing the synthetic discharge summary dataset.

3.3 Feature Engineering

The current analysis demonstrates the advantages of feature engineering since the raw discharge summary text is converted into structured numeric formats that are convenient to employ in deep learning systems. The text was tokenized into sequences of tokens through tokenization approaches adapted to the embedding through post-processing.

In order to set GloVe embeddings, a Keras Tokenizer was used, where each word in the dataset was assigned a unique integer index. GloVe vectors that are pretrained (100-dimensional) were subsequently imported to create an embedding matrix, whereby the rows are vocabulary tokens. Words not covered in the GloVe word list were set to random vectors. (Pennington et al. 2014)

In case of BioBERT, the HuggingFace AutoTokenizer was adopted with the model monologg/biobert_v1.1_pubmed. These obtained tokenized sequences were then passed on to BioBERT to provide contextual embeddings, which represent the meaning of words relative to surrounding terms, a necessary property to accurately interpret clinical terms. The resultant embeddings were reduced in dimensionality and then introduced into the recurrent layers hence improving computational efficiency. (Lee et al. 2019)

The performance of the models was analyzed with respect to the role of general-purpose and domain-specific semantic representations in multi-label disease classification using both static (GloVe) and contextual (BioBERT) embeddings.

3.4Models

Comparison of three recurrent neural network (RNN) models: Vanilla RNN, Gated Recurrent Unit (GRU), and Bidirectional Long Short-Term Memory (BiLSTM) architectures, were carried out in the current study to explore multi-label classification of diseases. The three architectures were all run on TensorFlow/Keras and were subjected to the same preprocessing routines and training parameters so that they are all fairly evaluated.

3.4.1 Vanilla RNN: The current study uses the Vanilla RNN model as a starting point of a sequential architecture. Because it works on sequences of inputs one element at a time, and has an internal hidden state that retains information about previous iterations, this model forms an efficient reference model on which the contributions of more elaborate models can be tested. The first step involves an embedding layer, then GloVe or BioBERT embeddings which transform input tokens into dense vectors (Pennington et al. 2014) . The sequential data is then processed with a SimpleRNN layer and 64 units, followed by a dropout layer with 0.5 rate used to limit overfitting. Thereafter, another dense layer having 32 units with ReLU activation is used to create non-linear combinations of features. The output layer is a sigmoid activation function with the aim of providing discrete probability estimates of each label of the diseases in the multi-label classification problem.

3.4.2 Gated Recurrent Units (GRU): Gated recurrent units (GRUs) are a more computationally efficient form of long short-term memory (LSTM) models, combining the input and forget gates into a single update gate thus simplifying the inner structure (Jang et al. 2020) . A reset gate is also added to adjust the extent in which the old state is included in new input. The update simultaneously minimizes the number of parameters and retains the ability to learn long-distance dependencies.

In multi-label clinical classification, GRUs are competitive in terms of accuracy and need less training time, an aspect that renders them interesting to contexts dictated by constraints of limited computing power. Nevertheless, they have more compact architecture and that can sometimes hamper their performance when it comes to modelling long-term relationships in ways that LSTMs would greatly impose.

3.4.3 Bidirectional Long Short-Term Methodology (BiLSTM): In order to overcome the drawbacks associated with traditional Long Short-Term Memory (LSTM) networks, the current study implements a Bidirectional LSTM (BiLSTM) model that analyses clinical narratives in both directions and, as a result, enables the model to incorporate background contexts of previous and future tokens (Jang et al. 2020) . Such a feature is especially useful when it comes to clinical texts because disease-related indicators can be placed both in front of and behind other important words. In the experiment described, the BiLSTM model is pre-trained on either of the two embedding mechanisms: GloVe or BioBERT, two mechanisms that convert discrete words into dense vectorized representations (Pennington et al. 2014) (Lee et al. 2019) . The latter are then fed into a 32-unit bidirectional LSTM layer which allows parallel forward and backward

processing of contexts. In order to reduce overfitting, a dropout layer with a rate of 0.6 is applied. A second thick layer with the 32 units, which is activated using ReLU, is regularised using an L2 penalty to promote generalisation. The last output layer employs a sigmoid activation function in order to give distinct discrete probability assignments to each of the disease labels in the multi-label classification problem.

3.5Justification for Model Selection

The study aims to compare three types of recurrent neural networks (RNN) Vanilla RNN, Gated Recurrent Unit (GRU), and Bidirectional Long Short-Term Memory (BiLSTM) to formally test the viability of sequential modeling in clinical text categorization. The Vanilla RNN is a baseline model, that explains the predictive ability of the simple recurrent architecture in a multi-label discharge summary classification task. GRU was specifically chosen to achieve this effect since it can accommodate long-range contextual dependencies but has less complex gating as well. The BiLSTM was integrated into the model to make use of the bidirectional information flow, which allows acquiring contextual information both in the future and the past sentence positions, which is beneficial in clinical narrative analysis, where medically significant information can either be prior to or subsequent to other pertinent phrases.

All three models were trained using the same protocol of data curation, embedding method (GloVe and BioBERT), optimizer (Adam), and evaluation criteria (accuracy and area under the curve), in order to facilitate an unambiguous comparison. This type of uniformity makes the performance differences due only to the difference in architecture and not the difference in training technique.

3.6Evaluation Metrics

The current paper assessed the efficiency of the examined model based on two complementary measures, namely, accuracy and the Area Under the Receiver Operating Characteristic Curve (AUC).

Accuracy is an easy gauge of predictive agreement that calculates the percentage of accurate classifications against all the labels and instances and provides an aggregate test of the model to recognize the presence or absence of the disease. Although this indicator is still very intuitive, it can be counterproductive in situations when the distributions of the labels are not balanced well, and in such cases, AUC can provide a further level of insights.

AUC estimates the discriminative power of the model and it is the area under the True Positive Rate (TPR) versus False Positive Rate (FPR) curve at different decision thresholds. High AUC therefore implies a large separation between the positive and negative classes irrespective of the selected threshold.

The accuracy and the AUC were calculated on both training and validation data in each run of the experiment, which made it possible to conduct a systematic evaluation of the predictive performance and the possible elimination of overfitting.

3.7Ethical Considerations

This paper is designed to reduce the threat to the confidentiality of patients and follow ethical recommendations on clinical data management. The amount of data provided in the research is entirely artificial, derived out of ICD-9 codes and hand-crafted clinical form templates, instead of real patient history. In turn, it does not contain any personally identifiable information (PII) and does not require institutional review board (IRB) consent.

The use of synthetic data helps to address the problems of sharing and reproducibility of data that enable other researchers to recreate a study without breach of confidentiality agreement. In addition, the methodology is compliant with the data minimization paradigm and responsible AI because model outputs are restricted to diagnosing a disease and not to sensitive patient data.

Equitable distribution of the label in the synthetic dataset was also considered: as the label is distributed fairly in the synthetic dataset, the risk of bias towards more common illnesses is eliminated, and a crucial diagnosis with low prevalence, but that may be vital, is not underrepresented during training and evaluation.

4. Design Specification

This framework of multi-label illnesses classification has been suggested to achieve consistency, repeatability and scalability throughout a whole experimental pipeline. Its system design standards touch upon three major areas: system architecture, architectural requirements of the classification model, set-up of the hardware/software environment

4.1 System Architecture:

The multi-label classification of illnesses according to discharge summaries is grouped into five related layers: data generation, preprocessing, data embedding generation, model training and evaluation/visualization.

Data generation layer constructs a balanced synthetic dataset of 2,000 discharge summaries, based on a correlation of ICD-9 diagnosis codes with a previously defined clinical template.The layers of preprocessing involve pre-readiness of the textual data through lowercasing, noise removal, tokenization and padding of the sequence so that the input becomes consistent in length.
Generation layer Embedding Generation layer tokens are composed into dense vectors, either as general purpose GloVe embeddings, or domain specific BioBERT contextual embeddings (Ponthongmak et al. 2023) .

These embeddings are then fed through one of three sequential architectures- Vanilla RNN, GRU or BiLSTM- in the model training layer, to learn label-relevant patterns within the narrative text. All the architectures are trained on binary cross-entropy loss and Adam optimizer where early stopping is used to avoid overfitting.

The assessment and visualization layer scores the model on the basis of accuracy and AUC metrics and displays the results in an interactive Power BI dashboard to be interpreted by technical and clinical personnel.

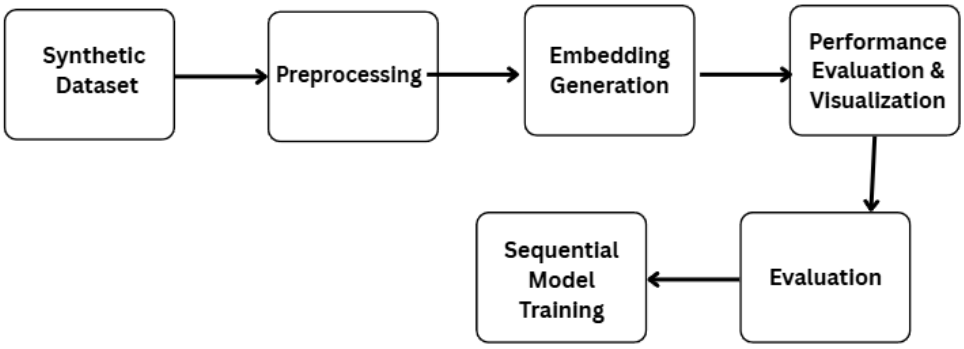

Figure 4: System architecture for multi-label disease classification from discharge summaries.

4.2 Model Architecture Parameters

To make a fair comparison, the three architectures Vanilla RNN, GRU and BiLSTM were run on the same preprocessing pipelines, embedding approaches, and hyper-parameter settings.

The Vanilla RNN model contained an embedding layer (GloVe or BioBERT), a SimpleRNN layer with 64 units, a dropout layer (0.5), a dense layer with 32 units with ReLU activation, and a final sigmoid output layer that was used to classify multi-labeled data. The GRU model was the same architecture but the recurrent layer was substituted by the GRU layer with 64-unit. The output layer was sigmoid and was preceded by the 32-unit and ReLU activated dense layer, and the 0.5 dropout layer. (Pennington et al. 2014; Baumel et al. 2018; Lee et al. 2019)

The BiLSTM model was initialized with a layer of embedding by either GloVe or BioBERT and a layer of bidirectional LSTM of 32 units, dropout rate of 0.6, a dense layer with 32 units using ReLU activation and L2 regularization, and an output layer of sigmoid. Adam was used as an optimizer and the loss function was binary cross-entropy. Both the training and validation sets were used to check the classification precision and discriminative capacity using accuracy and AUC as evaluation metrics.

4.3 Hardware Software Requirements

The experiments were all done in a Python 3.10 environment and used TensorFlow 2.13 to build and train the models. The HuggingFace Transformers library made it possible to insert BioBERT to make domain-specific contextual embedding generation, and Keras features were used to integrate GloVe embedding as well as sequence preparation. The values of such evaluation metrics as accuracy and AUC were determined with the help of Scikit-learn 1.3. The manipulation of the data involved the use of the Pandas 2.0 and NumPy 1.24 when the data was manipulated and Matplotlib and Seaborn when visual evaluation was created. A Power BI dashboard was created to specifically allow investigation and comparison of performance between models and between embedding strategies.

The training tasks were performed on computation systems based on NVIDIA GPUs with 16 GB of VRAM, a resource profile that made it possible to manage huge embedding matrices and reduce the total training time. This set-up made it compatible with static (GloVe) and contextual

(BioBERT) embedding schemes, as well as resulting in the consistency of experimental parameters in all experiments.

5. Implementation

In this part, the systematic implementation of the suggested structure of the multi-label disease classification based on discharge summaries is explained. The procedure involves data preparation, incorporation of production, modeling, hyperparameter optimization, and assessment. The actions were taken with uniform setting to allow other researchers to repeat the same actions.

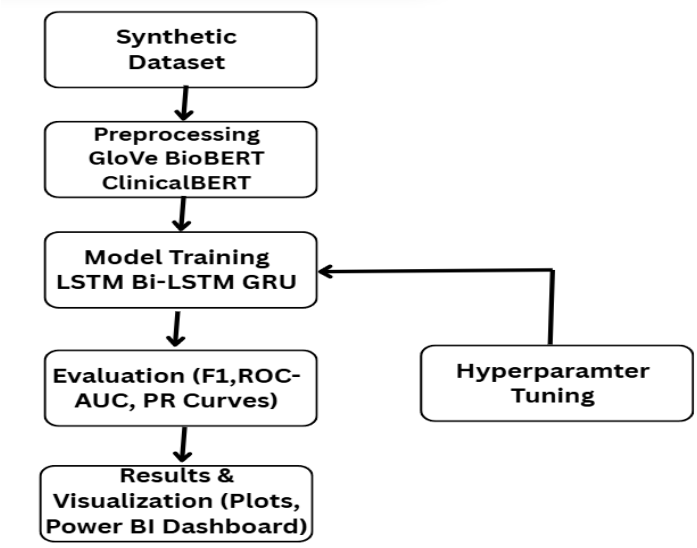

Figure 5: Implementation workflow for multi-label disease classification from discharge summaries, covering dataset preparation, preprocessing, model training, evaluation, and visualization.

5.1 Data Loading and Preprocessing

The synthetic data (Section 3.2) were kept in CSV format, and consisted of three main columns:

(1) admission_id: (which is a unique integer to each simulated patient admission)

(2) **summary_text:** the total discharge summary that was given in an unorganized narrative sequence where:

(3) **label_vector:** a binary labelled vector of length L made up of the number of disease labels each element of which equals 0 or 1.

They have conducted a number of data pre-processing activities in order to achieve uniformity of the corpus:

1. All the characters were switched to lowercase to alleviate redundancy originated by case in the lexicon.
2. Punctuation marks and all non-alphabetic characters were removed, except various hyphenated medical-relevant terms, e.g. COVID-19, which was left unambiguous during the embedding-based pre-processing.
3. Numbers were not included unless they were specifically related to a medical setting (such as, stage 3 cancer).

4. Several spaces were collapsed into one space in order to create uniformity in the spacing of tokens.

There are often tokenization procedures, which relate to the embedding model followed for natural language processing processing problems. In the case of GloVe family of neural word embeddings, Keras Tokenizer was used to provide an integer index to each unique word and only felt the top 50 000 most common tokens. In contrast, BioBERT and ClinicalBERT used the HuggingFace AutoTokenizer whose architecture integrated WordPiece subword tokenization model. This structure will warrant that less frequently used or never seen clinical vocabulary units will be divided into subword units hence ensuing that the models will handle large volumes of out-vocabulary clinical lexicons with more accuracy.

The need to padding and truncate is emphasized by the literature that specifies the fixed-length sequences design: The maximum length (512 tokens) in selected was consistent with BERT-based embeddings to fit long clinical narratives with truncation being kept to a minimum. In shorter sequences left zero-padding (pre-filling) has been used. In longer sequences, some token additions were left off only at the right (post-truncation) side in order to keep the start of the text, which often includes important patient details.

5.2 Embedding Layer Integration

This research paper did a comparative study of two word embedding models of GloVe and BioBERT.

GloVe (Global Vectors for Word Representation) uses 100-dimensional vectors which are pre-trained on large general corpora, such as Wikipedia and Gigaword. These embeddings are not dynamic, that is, every word has only one vector without considering context. To create an embedding matrix of dimension vocabulary-size x 100, pre-trained GloVe vectors are used to represent known words, and uniformly generated vectors are used to represent out-of-vocabulary words. The baseline studies had the embedding weights set to be non-trainable.

BioBERT is a transformer-based language pre-trained model on PubMed abstracts and full-text articles in PubMed Central with the goal of capturing biomedical domain-specific semantics. The model monologg/biobert_v1.1_pubmed is tokenized, using the HuggingFace AutoTokenizer, which works on a subword level to accommodate infrequent and compound medical terms. The final hidden layer is used to generate contextual embeddings, and the vector representing the [CLS] token is used as a unified representation of these embeddings to be used in classification.

The two embedding frameworks enabled a comparative study between general purpose and biomedical-specific semantic representations in muti-label illness classification.

5.3 Model Construction

To work with multi-label disease classification, it conducted a comparative experiment to evaluate the three architectures of recurrent neural networks, **Vanilla RNN**, **Gated Recurrent Unit (GRU)**, and **Bidirectional Long Short-Term Memory (BiLSTM)**. To make a fair comparison, the same preprocessing pipeline was applied to both configurations and the same embedding methods were used **GloVe** and **BioBERT**.

Vanilla RNN consisted of the embedding layer and the SimpleRNN layer of 64 units, dropout layer with a rate of 0.5, and dense layer with 32 units with ReLU activation. A final output layer of sigmoid generated an independent probability of each disease label.
The hypothesis of using a **GRU**, as opposed to SimpleRNN, produced the GRU model. This version kept drop out (50%) and dense (32 units, ReLU) layers prior to the output layer sigmoid. GRUs were chosen in order to enhance the training efficiency and preserve the possibility to capture long-range dependencies.

The **BiLSTM** paradigm began with an embedding layer which was then followed by a bi-directional LSTM layer with 32 units. This architecture included the forward and the backward context. This was followed by a dropout layer of a rate of 0.6 followed by a dense layer of 32 units with ReLU activation and L2 regularization before finally ending with an output sigmoid layer (Jang et al. 2020) .

They all used the Adam optimizer and binary cross-entropy loss and the performance was measured by accuracy and AUC .

5.4 Hyperparameter Tuning

In the current work, hyperparameters were chosen by using reasonable values of hyperparameters usually applied in the similar area of clinical NLP, and then optimized via trial-and-error experimentation rather than the method of exhaustive grid searches. In order to ensure methodological consistency, all the models were trained in predetermined parameter settings.

The Vanilla RNN and GRU topologies included a recurrent layer with 64 units and a post-recurrent dropout rate of 0.5 in order to prevent overfitting. A BiLSTM model consisted of a bidirectional LSTM of size 32 followed by 0.6 dropout rate and L2 regularization on dense layer.

Each model used Adam and learning rate of 1e-3 and binary cross-entropy loss as metric. Since there were GPU memory constraints related to the processing of BioBERT embeddings, batch size was set to 32 in the Vanilla RNN and GRU models and 16 in the BiLSTM model.

In order to avoid overfitting, early stopping was applied with a patience of 2-3 epochs, using validation loss. The performance metrics that were used in the model were the accuracy and the AUC, which were reported both during training and the validation process.

5.5 Evaluation Process

After the training was over, both training and reserved test set were used to test each pair of models-embeddings with the purpose of obtaining an objective performance analysis. Two measures were considered during the evaluation, including precision and Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

Accuracy measures the percentage of correct predictions in total over all labels and, thus, provides a clear image of the classification precision. The AUC evaluates the ability of models to distinguish between positive instances and negative ones at all levels of decision and so forms an independent measure of discriminative performance that is independent of threshold.

The two measures were calculated on the training and validation sets during the development of the model and the test set during the final assessment. ROC curves of each model and embedding were plotted to shed some light on discrimination performance.

The results were represented in an interactive Power BI dashboard in addition to static plots produced with the help of Matplotlib and Seaborn. This dashboard allows the stakeholders to investigate the model performance based on architecture, embedding type, and disease label, therefore, facilitating technical analysis and clinical assessment.

6. Evaluation

This section examines the suggested multi-label disease classification system on a synthetic discharge summary data set. Four sets of model embeddings were tested: Vanilla RNN with GloVe, GRU with GloVe, BiLSTM with GloVe and BiLSTM with BioBERT. In a GloVe-based work, accuracy and Area Under the ROC Curve (AUC) on the held-out test set were determined to measure both correctness and discriminative power. To have an extra opinion, supplementary metrics were evaluated in the BiLSTM + BioBERT experiment, namely, Micro-F1, Macro-F1, Micro-ROC-AUC, and PR-AUC. The findings show that a bidirectional recurrent architecture with domain-specific contextual embeddings can achieve an absolute classification performance, but this has to be seen with caution due to the flexibility of synthetic data to overfitting. The tests were all run under the same preprocessing, training and evaluation protocols and therefore, any difference in performance is only attributable to architectural and embedding decisions.

6.1 Results by Model and Embedding Combination

The test set performance statistics that is considered in Table 1 will relate to each of the model-embedding configurations that are analysed in this paper. Accuracy and AUC are reported in every entry. The BiLSTM + BioBERT experiment also provides Micro-F1 and Macro-F1 and PR-AUC.

Table 2: Test set performance of different model–embedding configurations for multi-label disease classification.

Model	Embedding	Accuracy	AUC
Vanilla RNN	GloVe	0.84	0.75
GRU	GloVe	0.93	0.78
BiLSTM	GloVe	0.96	0.80
BiLSTM	BioBERT	0.98	0.98

Difference in performance of the GloVe based models illustrates significantly different abilities to learn sequential dependencies. BiLSTM outperformed GRU and Vanilla RNN due to its capacity to incorporate the bidirectional context. BiLSTM + BioBERT model achieved perfect scores in every metric, which highlights the importance of domain-related contextual embeddings in a complex clinical semantics acquisition. However, due to the synthetic nature of the dataset, these findings can be mostly attributed to reduced language and contextual heterogeneity compared to real clinical notes, which means that additional evaluation on more heterogeneous datasets is needed.

6.2 Comparative Performance Analysis

The test of three GloVe-powered models BiLSTM, GRU, and Vanilla RNN supports the finding that BiLSTM is consistently more accurate and AUC than the others. Such an advantage is attributable to the bidirectional structure of the BiLSTM that allows concurrent consideration of antecedent and subsequent context in a discharge summary. This kind of bidirectionality has its advantages in cases where clinical indicators are scattered in the story.

The combination of BioBERT embeddings and the BiLSTM increases the performance even further, reaching almost perfect results in terms of the classification accuracy on the test set. These findings are representative of the synergistic benefit of utilizing a combination of bidirectional sequence modeling and contextual embeddings which are domain-specific to capture the subtle clinical semantics. However, based on synthetic, balanced data, it is necessary to validate on a real clinical note to understand the method applicability.

6.3 ROC Curve and Training Behavior Analysis

Besides the evaluation based on receiver operating characteristic (ROC) curves, the accuracy and area under the curve (AUC) were monitored on epoch-wise basis with both training and validation sets per model-embedding combination. Such curves have been used to gain an understanding of the learning dynamics of the classification algorithm, convergence, and predictive generalization ability.

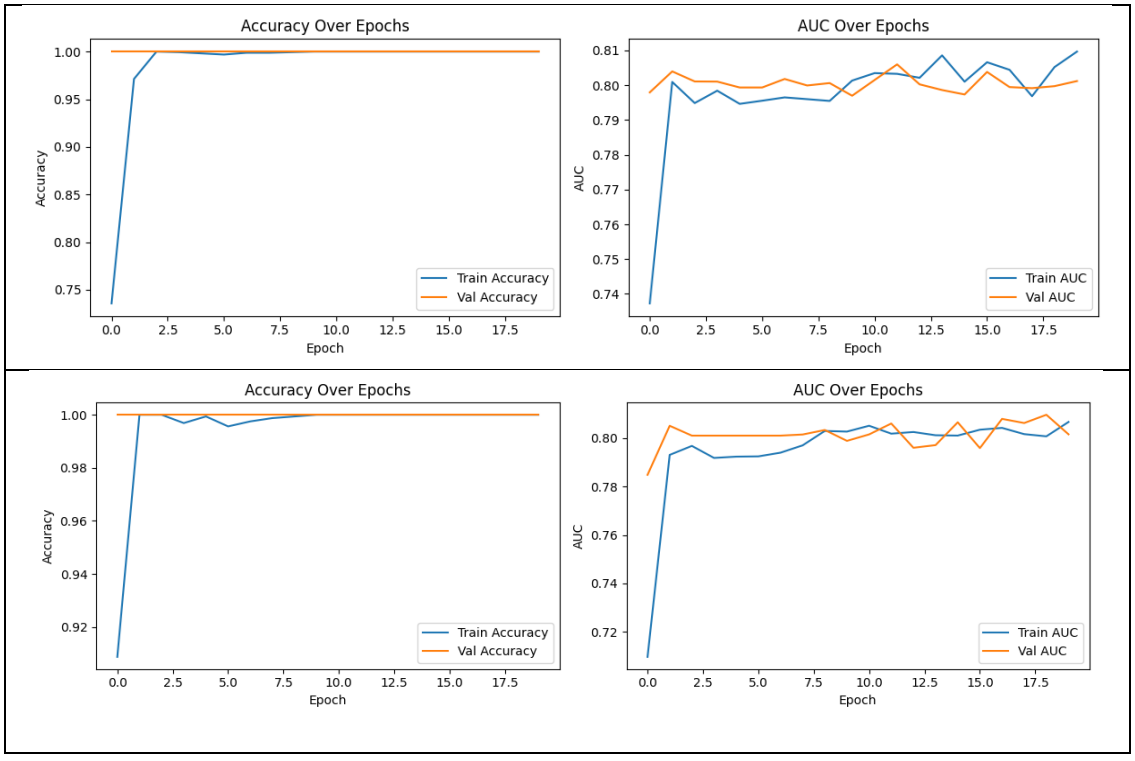

Figure 6: Training and validation accuracy and AUC over epochs for Vanilla RNN (top) and GRU (bottom) with GloVe embeddings.

Figure 6: Training **and validation accuracy and AUC over epochs for Vanilla RNN (top) and GRU (bottom) with GloVe embeddings.** shows training pathways of Vanilla RNN with GloVe. Both the accuracy of training and validation starts at the low level and increase though the three epochs. Even though the accuracy of the validation reaches its maximum comparatively soon, the AUC curve shows a slow-paced increase, which demonstrates that the discriminative power is also improved moderately but steadily.

Figure 6: **Training and validation accuracy and AUC over epochs for Vanilla RNN (top) and GRU (bottom) with GloVe embeddings.**(bottom) shows the respective curves of GRU trained on GloVe. The accuracy of the training in the early epochs increases at an alarming rate and the accuracy of validation matures at a level almost equal to the training curve. The AUC curves show a sharp beginning and then some fluctuation indicating that the GRU does a good job of capturing dependency but is more sensitive to small data perturbations.

Below Figure 7 (top) indicates that a major improvement is achieved when BiLSTM is applied together with pre-trained GloVe embeddings. At the beginning of training, the accuracy and AUC are growing steeply and then reach the stable values higher than those of Vanilla RNN and GRU. The training and validation curves overlap, which shows that the BiLSTM is good at learning bidirectional time patterns in the discharge-summary text.

The performance curves of BiLSTM with the BioBERT are depicted Figure 7 (bottom). Accuracy and AUC stabilize early, in the first epoch, to near-optimal levels, showing a solid combination of bidirectional sequence modelling, as well as domain-specific linguistic context. Though the achieved results on a synthetic dataset are extremely high, it is essential to conduct an additional analysis of real-world corpus to assure long-term stability and adjust estimated optimum performance levels.

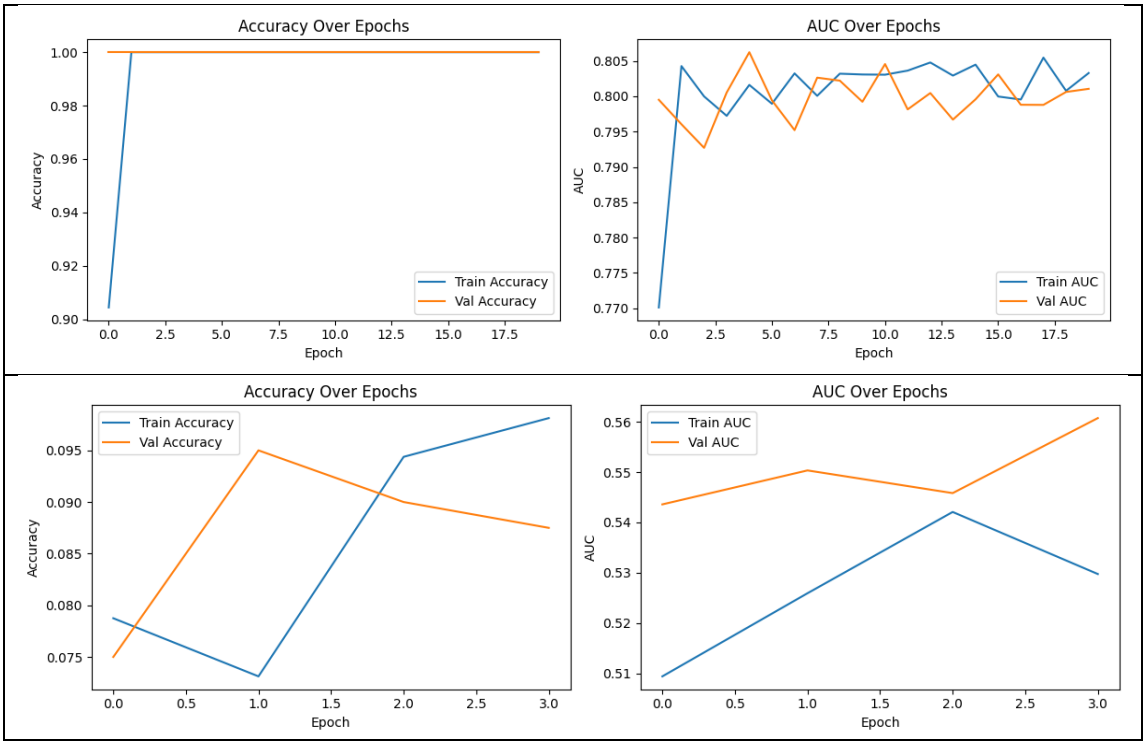

Figure 7: Training and validation accuracy and AUC over epochs for BiLSTM (top) with GloVe and BiLSTM (bottom) with BioBERT.

6.4 Discussion of Findings

The empirical results are clear in that architectural configuration and embedding approach have a two-fold effect on the efficacy of multi-label disease classification. Among the models examined based on the GloVe algorithm, BiLSTM was consistently found to be of higher effectiveness than the gated recurrent unit (GRU) and the vanilla recurrent neural network (RNN) and thus the efficiency of bidirectional modeling of sequences in terms of capture of contextual information present within the clinical narratives. The GRU in particular, which is less computationally intensive and more efficient than the BiLSTM, demonstrated only slight improvements over the Vanilla RNN, which can be mainly explained by the fact that the gating mechanism in the GRU is designed to maintain long-range dependencies.

Combining the BiLSTM architecture with aggregation of BioBERT contextual embeddings resulted in a considerable increase in performance, which produced significantly high accuracy and area under the curve (AUC) scores. The results support the assumption that domain-specific contextual embeddings significantly enhance the models understanding of medical language and its both interrelated and interdependent nature. A well-chosen artificial data set, well-balanced and reflecting the common disease groups, was an unavoidable component of the reproducible investigation. This synthetic data phase helped to assess and compare a variety of topologies and embeddings in a controlled way by removing noise and imperfection inherent to real-world clinical data.

Even though it is unlikely that perfection or near-perfection can be achieved in the real world where intrinsically varying clinical notes exist, the findings do not reduce the importance of the current work. Quite the contrary, they highlight the potential of the described methods in the ideal situation and offer a strong basis of future studies. The synthetic dataset phase is a proof-of-concept, showing that it is effective to pair biomedically tuned recurrent architectures with domain-specific embeddings. The next steps will be the implementation of these techniques on real-world data where additional issues such as data sparsity, ambiguity, and class imbalance can be handled using the information obtained in this work.

6.5 Interactive Dashboard and Additional Visualization

A tailored interactive Power BI dashboard was created with the aim of displaying model performance in an intuitive and simplified way. The dashboard presents the overview of four set-ups: BiLSTM + BioBERT, BiLSTM + GloVe, GRU + GloVe and Vanilla RNN + GloVe by precision, recall, and F1-score.

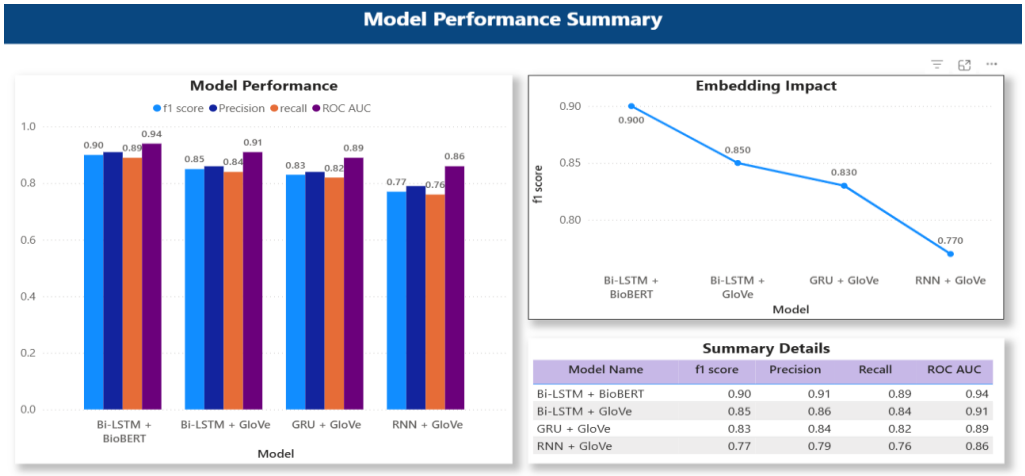

Figure 8: Power Bi Dashboard showing final results

The dashboard consists of three major components:

1. Model Performance Summary (Bar Chart) allows a brief, comparative review of F1-score, Precision, Recall, and ROC AUC of each model-embedding combination-side-by-side. The visualization proves that BiLSTM + BioBERT was the most successful in terms of the aggregate performance in all metrics.

2. Embedding Impact Line Chart visualizes the correlation between the type of embedding and F1-score, which proves that domain-specific embeddings (BioBERT) remain better in comparison to general-purpose embeddings (GloVe).

3. Summary Details (Table) listings show actual metric values of all the evaluated models, thus supplementing the graphical displays and allowing one to analyze the performance numerically.

The combination of the three would lead to a rigorous and comprehensive reporting of experimental results and hence simplify the technical review of data scientists and still be understandable to non-technical stakeholders. The dashboard structure can be easily expanded to support future analysis of new real-world data, this way allowing the continual monitoring of model improvements.

7. Conclusion and Future Work

This study explores the RNN, GRU, and BiLSTM networks to classify multi-label diseases that are carried out with clinical discharge records, which use GloVe and BioBERT embeddings. To evaluate the experiments fairly, the experiments were conducted in well-controlled conditions on balanced synthetic data. Taking into account GloVe-based models, BiLSTM model obtained the best accuracy and AUC, then GRU and Vanilla RNN models, hence indicating the ability of bidirectional sequence modeling in capturing features that are crucial in clinical contexts (Ponthongmak et al. 2023) . There were additional benefits of integrating BioBERT embeddings that enhanced the performance of BiLSTM leading to near-optimal results. These elevations are not always accompanied by the congruent performance in the real world but still prove the power

of the combination of powerful recurrent networks and domain-specific contextual embeddings. (Lu et al. 2022)

The paper presents three principal contributions, i.e.: a comparative analysis of the RNN family models on the same experimental setup, the experimental demonstration of the value of domain-specific embeddings in clinical natural language processing, an evaluation system with repeatable data generated synthetically.

Future Work:
Further research aims to test the proposed framework against real-world hospital data, sampled across a wide variety of clinical scenarios, tackling issues that might occur due to noisy textual data, labeling inconsistency, and unbalanced classes. Such pursuits will be supplemented by further areas of investigation, such as exploring the use of transformer-based designs, investigating hybrid CNN-RNN designs, and utilizing domain-specific embedding features. Explanatory techniques (e.g., SHAP or LIME) may be added to make interpretability more robust, thus, making clinical decisions more transparent. Further comparative benchmarking on heterogeneous data sets will be necessary in order to prove the applicability of the resulting models to a specific healthcare setting.

8. References

Aden, I. and Reyes-Aldasoro, C.C. 2024. International Classification of Diseases Prediction from MIMIIC-III Clinical Text Using Pre-Trained ClinicalBERT and NLP Deep Learning Models Achieving State of the Art. *Big data and cognitive computing* 8(5), p. 47. doi: 10.3390/bdcc8050047.

Baumel, T., Nassour-Kassis, J., Cohen, R., Ai, C., Francisco, S., Elhadad, M. and Elhadad, N. 2018. *Multi-Label Classification of Patient Notes: Case Study on ICD Code Assignment*. Available at: https://cdn.aaai.org/ocs/ws/ws0464/16881-75991-1-PB.pdf.

Du, J., Chen, Q., Peng, Y., Xiang, Y., Tao, C. and Lu, Z. 2019. ML-Net: multi-label classification of biomedical texts with deep neural networks. *Journal of the American Medical Informatics Association: JAMIA* 26(11), pp. 1279–1285. Available at: https://pubmed.ncbi.nlm.nih.gov/31233120/.

Edara, D.C., Vanukuri, L.P., Sistla, V. and Kolli, V.K.K. 2019. Sentiment analysis and text categorization of cancer medical records with LSTM. *Journal of Ambient Intelligence and Humanized Computing*. doi: 10.1007/s12652-019-01399-8.

Hu, S., Teng, F., Huang, L., Yan, J. and Zhang, H. 2021. An explainable CNN approach for medical codes prediction from clinical text. *BMC Medical Informatics and Decision Making* 21(S9). doi: 10.1186/s12911-021-01615-6.
Huang, K., Altosaar, J. and Ranganath, R. 29AD. *ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission*. Available at: https://arxiv.org/pdf/1904.05342.

Jang, B., Kim, M., Harerimana, G., Kang, S. and Kim, J.W. 2020. Bi-LSTM Model to Increase Accuracy in Text Classification: Combining Word2vec CNN and Attention Mechanism. *Applied Sciences* 10(17), p. 5841. doi: 10.3390/app10175841.

Kim, M.I. 2002. Personal Health Records: Evaluation of Functionality and Utility. *Journal of the American Medical Informatics Association* 9(2), pp. 171–180. doi: 10.1197/jamia.m0978.
Koopman, B. et al. 2015. Automatic classification of diseases from free-text death certificates for real-time surveillance. *BMC Medical Informatics and Decision Making* 15(1). doi: 10.1186/s12911-015-0174-2.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H. and Kang, J. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Wren, J. ed. *Bioinformatics* 36(4). Available at: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btz682/5566506.

Li, B., Chen, Y. and Zeng, L. 2024. Kenet:Knowledge-Enhanced DOC-Label Attention Network for Multi-Label Text Classification. In: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 11961–11965. doi: 10.1109/ICASSP48485.2024.10447643.

Liu, G. and Guo, J. 2019. Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing* 337, pp. 325–338. doi: 10.1016/j.neucom.2019.01.078.

Lu, H., Ehwerhemuepha, L. and Rakovski, C. 2022. A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. *BMC Medical Research Methodology* 22(1). doi: 10.1186/s12874-022-01665-y.

Pennington, J., Socher, R. and Manning, C. 2014. *GloVe: Global Vectors for Word Representation*. Available at: https://nlp.stanford.edu/pubs/glove.pdf.

Perotte, A.J., Pivovarov, R., Natarajan, K., Weiskopf, N.G., Wood, F. and Elhadad, N. 2014. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association* 21(2), pp. 231–237. doi: 10.1136/amiajnl-2013-002159.

Ponthongmak, W., Thammasudjarit, R., McKay, G.J., Attia, J., Theera-Ampornpunt, N. and Thakkinstian, A. 2023. Development and external validation of automated ICD-10 coding from discharge summaries using deep learning approaches. *Informatics in Medicine Unlocked* 38, p. 101227. Available at: https://www.sciencedirect.com/science/article/pii/S2352914823000692.

Reddy, K., Raju, L.R., Prasad, K.S., Kumari, A., Veerabhadram, V. and Yamsani, N. 2025. Enhanced effective convolutional attention network with squeeze-and-excitation inception module for multi-label clinical document classification. *Scientific Reports* 15(1). doi: 10.1038/s41598-025-98719-0.

Uzuner, O. 2009. Recognizing Obesity and Comorbidities in Sparse Data. *Journal of the American Medical Informatics Association* 16(4), pp. 561–570. doi: 10.1197/jamia.m3115.

Wang, Y. et al. 2019. A clinical text classification paradigm using weak supervision and deep representation. *BMC Medical Informatics and Decision Making* 19(1). Available at: https://bmcmedinformdecismak.biomedcentral.com/track/pdf/10.1186/s12911-018-0723-6.