An Optimised Predictor For Patient Health Records While Ensing HIPAA Compliance

by

Stanley Matthew Das
Student ID: 21141014
Ashiqul Alam Student
ID: 21301336 Arif
Jawad Alvi Student
ID: 21301039

A project submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
January 2026.

Declaration

It is hereby declared that

1. The project submitted is my/our own original work while completing degree at Brac University.
2. The project does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The project does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Stanley Matthew Das
Student ID: 21141014

Ashiqul Alam
Student ID: 21301336

Arif Jawad Alvi
Student ID: 21301039

Approval

The thesis/project titled "An Optimised Predictor For Patient Health Records While Ensuring HIPAA Compliance" submitted by

1. Stanley Matthew Das (21141014)
2. Ashiqul Alam (21301336)
3. Arif Jawad Alvi (21301039)

of Summer, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in [Current Semester] Year.

Examining Committee:

Supervisor:
(Member)

Mr. Moin Mostakim
Senior Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Mr. Hamim Ibne Nasim
Adjunct Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Mr. Sifat Tanvir
Adjunct Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi
Associate Professor and Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

The increasing digitization of healthcare has led to predictive analytics becoming an essential tool for early risk detection and personalized patient care. This project introduces an optimized predictor for Patient Healthcare Records. This is a microservices-driven, AI-based system architected to analyze patient data while maintaining HIPAA (Health Insurance Portability and Accountability Act), ensuring scalability through Dockerized deployment. The system functions through three main phases: (1) Data processing through Optical Character Recognition (OCR), which extracts text and refines patient data from medical records; (2) Health Risk Prediction utilizing a Hidden Markov Model (HMM) for sequential health analysis and Neural Networks for predictive modeling; and finally, (3) Secure storage and Recommendations where the predictions are organized in a structured PostgreSQL database and accessed via a web/mobile platform built with HTML and CSS. This design guarantees effective, privacy-conscious, and AI-enabled healthcare analytics, delivering real-time insights for healthcare professionals and providing them with a streamlined, scalable, and secure method for health risk prediction, supporting proactive medical decision-making.

Keywords: Healthcare Digitization, Predictive Analytics, Patient Health Records, Microservices Architecture, AI-based Risk Prediction, HIPAA Compliance, Dockerized Deployment, Optical Character Recognition (OCR), Hidden Markov Model (HMM), Neural Networks, Secure Data Storage, PostgreSQL, HTML CSS, Real-time Healthcare Insights

Acknowledgement

We would like to express our appreciation and gratitude to our supervisors who have guided and supported us along the project stages. Their encouragement, feedback, support and insights were essential to this work and its successful accomplishment. We owe our success to the Almighty in providing us the strength, patience and perseverance to persevere and accomplish the completion of our work. Lastly, we would like to thank our families, friends and fellow group members who have been supportive, cooperative and encouraging along this journey.

Table of Contents

Chapter 1

Introduction

1.1 Background

The increasing adoption of Electronic Health Records (EHRs) has opened up a unique opportunity for predictive analytics within healthcare to identify risks at an early stage, tailor individual treatment plans, and enhance clinical decision-making. By digitizing patient histories, test results, prescriptions, and other clinical notes, EHRs have transformed these elements into dynamic repositories of up-to-date medical information. However, this shift also introduces both technical and ethical challenges concerning privacy and data security. The application of machine learning and artificial intelligence (AI) in these sensitive domains must consider a complicated landscape that includes information fragmentation, unstructured data, and compliance with rigorous privacy regulations, with a key area of focus being the Health Insurance Portability and Accountability Act (HIPAA). Examining a real-world example, chronic conditions such as IBS (Irritable Bowel Syndrome) can be incredibly challenging to manage. Individuals suffering from IBS often endure many years of symptoms and various medical treatments, resulting in a collection of handwritten food diaries, scanned test results, clinic visit documentation, and physician notes. These data sources are typically stored in formats that are not interoperable across healthcare providers and systems. Consequently, clinicians must manually review this unstructured information to identify patterns and assess treatment efficacy, as well as detect early risks. This process is not only inefficient and mentally exhausting but also prone to errors, leading to delays in effective interventions and causing frustration for both patients and healthcare professionals.

The proposed project introduces a centralized AI-driven prediction service, SAA, which aids healthcare professionals. It converts digitized or handwritten medical records into actionable health information. The project's framework incorporates Optical Character Recognition (OCR), Named Entity Recognition (NER), a Hidden Markov Model (HMM), and a neural network operating within a Docker environment. Every stage of the process is crafted with an emphasis on data security and user accessibility. SAA provides real-time risk assessments via a secure web interface, enabling clinicians to utilize personal, privacy-respecting decision support while fully complying with HIPAA regulations. By connecting unstructured medical data to health predictions, SAA seeks to empower healthcare practitioners to operate more effectively, delivering dependable, privacy-focused solutions that improve patient outcomes and ease the difficulties associated with data interpretation.

1.2 Problem Statement

The future of healthcare relies heavily on electronic health record (EHR) systems. However, these systems lack interoperability, making it challenging to implement advanced analytics. Non-structured data, including scanned documents and handwritten notes, can hinder the process of accessing critical clinical information. Current AI systems do not consistently adhere to the standards set by HIPAA privacy regulations, and their rigid architecture lacks the necessary flexibility to adapt to various clinical applications. We aim to address these limitations by developing a modular, HIPAA-compliant predictive system that:
1. Automatic extraction and anonymization of text information in the unstructured medical records by using OCR and NER
2. Forecasts health risks based upon explainable time-based models (e.g. HMM and NN).
3. Stores and displays findings in secured databases and API layers in accordance with HIPAA regulations
4. Allows real-time communication by means of scalable back-end front-end architecture.
In this way, the project proposes an effective and legally acceptable model of providing smart, centralized health risk forecasts within the real clinical environment.

1.3 Objective

The main goal of the current project is to design a HIPAA-compliant AI-based system that would be able to predict the health risks of the patients with references to their historical and real-time health data. Modularity, scalability and privacy are the drivers of the system. The particular aims are:
1. Preprocessing of data: Optical Character Recognition (OCR) of medical records and turn the unstructured health data to obtain and digitize the medical records into structured information.
2. To use predictive models like Hidden Markov Models (HMM) for analysing sequential data and Neural Networks to classify the risks.
3. To ensure that HIPAA compliance is maintained at every stage of the collection, processing, storage, and transmission of data by establishing the secure database infrastructures and access strategies.
4. The system must be scalable and portable in Dockerized microservices architecture.
5. To provide an interactive, secure frontend for healthcare professionals that is developed using HTML and CSS. to provide real-time expertise and healthcare advice.
Accomplishing these objectives, the proposed system is aimed at empowering early and proactive decision-making within the clinical context, where the priority is put on the flexibility of data privacy and deployment.

Chapter 2

Literature Review

2.1 Preliminaries

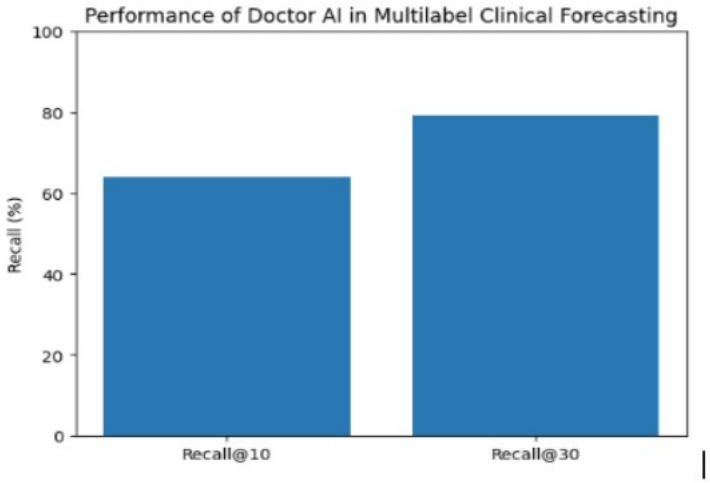
During the last several decades, the use of AI in healthcare has seen a significant uptake, especially in the use of such applications as EHRs, risk prediction, and health diagnostics. Some works of scientific papers are taken into account in this literature review in the OCR, HMM, neural networks, secure healthcare systems, and scalable architectures.

2.2 Review and Summary of Existing Research

AI in Healthcare Prediction

AI and machine learning are increasingly becoming significant in today's healthcare landscape, particularly in the area of predictive modeling of electronic health records (EHRs). One of the earliest recurrent neural network (RNN) models developed for predicting medical events based on time series data from patient records is known as Doctor AI, created by Choi et al.[2].This model achieved a recall rate of 64 percent for predictions made within 10 days and 79 percent for those within 30 days, utilising EHR data collected over eight years from 263,706 patients.
Recall counts the number of correctly identified real positive outcomes, including the correct identification of the future diagnosis, that the model was able to make. Simply put, it informs us of all conditions that truly happened, what percentage it is that the model is able to predict. Thus, a recall of 79 implies that almost 8 out of 10 future medical events were well-captured by the model when predicting further into the future. The paper notes that not all predictions came in the top, but the model tended to contain the correct diagnosis or medication in the list of its predictions. This is especially crucial within a clinical context since a doctor will usually think of several potential diagnoses before arriving at one. Hence, although the correct state may be ranked second or third on the list, the model still contributes real clinical value.
The increase in recall from 10 to 30 demonstrates that the model is adept at identifying pertinent clinical patterns and that there are numerous accurate predictions within the confidence range of 11-30. This indicates that RNNs are capable of making effective multilabel clinical predictions, aligning access to correct medical decisions within a limited selection of top predictions with the needs of clinical

decision-making.



Miotto et al.[4] developed Deep Patient, which is an unsupervised deep learning model trained on more than 700,000 electronic health records (EHR) and tested on 76,214 patients across 78 disease prediction tasks. The model demonstrated an average AUC-ROC of 0.773, which was much higher than the baseline RawFeat model (AUC= 0.659). The AUC-ROC is the abbreviation of the Area Under the Receiver Operating Characteristic curve. The ROC curve is a curve that shows the tradeoff between sensitivity (recall) and specificity (preventing false alarms). The AUC (area under the curve) is a number that provides one summative measure of performance: 0.5 represents random guessing, whereas 1.0 represents perfect classification. In that way, the AUC of 0.773 demonstrates that Deep Patient was significantly more precise than chance and outperformed traditional feature-based models in the separation of patients who would either develop a disease or not. F1-scores and precision were also reported in the study. Precision measures how many conditions of all the conditions that the model predicted were actually correct. F1-score is a balance between precision and recall and can be used in cases where both false positives and false negatives are important in healthcare. Deep Patient achieved 15% and an F1-score increase over older methods by 54% . This means that it not only identifies correctly higher number of patients at risk (high recall), but also undertakes fewer errors (high precision), which is a good balance between the two.

Time Interval = 1 year (76,214 patients)			
		Classification Threshold = 0.6	
Patient Representation	**AUC-ROC**	**Accuracy**	**F-Score**
RawFeat	0.659	0.805	0.084
PCA	0.696	0.879	0.104
GMM	0.632	0.891	0.072
K-Means	0.672	0.887	0.093
ICA	0.695	0.882	0.101
DeepPatient	**0.773***	**0.929***	**0.181***

Time Interval = 1 year (76,214 patients)			
	Area under the ROC curve		
Disease	**RawFeat**	**PCA**	**DeepPatient**
Diabetes mellitus with complications	0.794	0.861	**0.907**
Cancer of rectum and anus	0.863	0.821	**0.887**
Cancer of liver and intrahepatic bile duct	0.830	0.867	**0.886**
Regional enteritis and ulcerative colitis	0.814	0.843	**0.870**
Congestive heart failure (non-hypertensive)	0.808	0.808	**0.865**
Attention-deficit and disruptive behavior disorders	0.730	0.797	**0.863**
Cancer of prostate	0.692	0.820	**0.859**
Schizophrenia	0.791	0.788	**0.853**
Multiple myeloma	0.783	0.739	**0.849**
Acute myocardial infarction	0.771	0.775	**0.847**

Conventional recurrent neural networks (RNNs) like Doctor AI [2] showed that sequence modelling could be useful in healthcare prediction. Nevertheless, the basic RNNs exhibit vanishing gradient problem and therefore do not easily remember long patient histories. In a bid to address this weakness, Long Short-Term Memory (LSTM) networks have been invented, which have gating mechanisms that regulate information flow across time. Formally, at each time step t, given input $\mathbf{x}_t$, hidden state $\mathbf{h}_{t-1}$, and cell state $\mathbf{c}_{t-1}$, the LSTM is defined as follows:

$$f_t = \sigma(W_f[\mathbf{h}_{t-1}, \mathbf{x}_t] + b_f) \quad \text{(1) Forget gate} \tag{2.1}$$

$$i_t = \sigma(W_i[\mathbf{h}_{t-1}, \mathbf{x}_t] + b_i) \quad \text{(2) Input gate} \tag{2.2}$$

$$\tilde{c}_t = \tanh(W_c[\mathbf{h}_{t-1}, \mathbf{x}_t] + b_c) \quad \text{(3) Candidate memory} \tag{2.3}$$

$$\mathbf{c}_t = f_t \odot \mathbf{c}_{t-1} + i_t \odot \tilde{c}_t \quad \text{(4) Updated cell state} \tag{2.4}$$

$$o_t = \sigma(W_o[\mathbf{h}_{t-1}, \mathbf{x}_t] + b_o) \quad \text{(5) Output gate} \tag{2.5}$$

$$\mathbf{h}_t = o_t \odot \tanh(\mathbf{c}_t) \quad \text{(6) Hidden state output} \tag{2.6}$$

where σ is the sigmoid activation function, and $\odot$ denotes element-wise multiplication.These gates allow the network to decide which information to forget, what to store, and what to output at each time step.

Lipton et al. (2016) [3] used LSTMs on multivariate clinical time-series data to make multilabel diagnoses of 128 conditions. They demonstrated that LSTMs using target replication (deep supervision at each time step) reported better results than logistic regression and multilayer perceptrons (MLPs) using a dataset of 10,401 pediatric intensive care unit episodes. Their best model achieved:

1. Micro AUC = 0.8450 vs. logistic regression baseline (0.8074)
2. Macro AUC = 0.7842 vs. logistic regression baseline (0.7218)
3. Micro F1 = 0.2800 vs. baseline (0.2221)
4. Macro F1 = 0.1365 vs. baseline (0.1029)

These advancements indicate that LSTMs learn the short- and long-term clinical dependence more efficiently than the simple models. Expanding on this, Rajkomar et al. (2018) [6] have shown how deep learning can be applied on a large scale using LSTMs and CNN-LSTM hybrids on a collection of more than 216,000 records of adult patients across two large hospitals. They used a combination of convolutional layers in their architecture to find local temporal patterns in EHRs and LSTMs to model long-term dependencies. Convolutional component is provided as:
First, convolutional layers extract local temporal features:

$$h_t^{(CNN)} = f(W * x_{t:t+k} + b) \quad \text{(7) Convolutional layer output} \tag{2.7}$$

where W is the convolutional filter, $x_{t:t+k}$ represents a temporal input window, and f is a nonlinear activation function (e.g., ReLU).
The extracted features are then passed to the LSTM:

$$h_t^{(LSTM)} = \text{LSTM}(h_t^{(CNN)}, h_{t-1}^{(LSTM)}) \quad \text{(8) LSTM output} \tag{2.8}$$

Finally, the model outputs a classification or prediction:

$$\hat{y} = \sigma(W_y h_T^{(LSTM)} + b_y) \quad \text{(9) Final prediction} \tag{2.9}$$

This architecture allows CNN to handle local feature extraction and LSTM to handle long-term sequence dependencies, suitable for temporal patient data. Across

predictive tasks, Rajkomar et al. reported strong performance:
In-hospital mortality AUROC = 0.93–0.94

1. 30-day unplanned readmission AUROC = 0.75–0.76
2. Prolonged length of stay AUROC = 0.85–0.86
3. Discharge diagnosis AUROC = 0.90

Although CNN- LSTM structures are effective in local temporal trends and long-term associations, they are limited to unidirectional sequence modelling in which antecedent data is the only input which is used to make a current forecast. In order to eliminate this limitation, recent studies have analyzed bidirectional recurrent designs. In Ullah et al. (2021) [9], the ECG-based arrhythmia classification was done using CNN-BiLSTM, where convolutional layers were used to extract morphological signal features and BiLSTM to extract backward and forward temporal variations. Their findings showed better classification performance as compared to regular CNN-LSTM architectures, which highlight the significance of bidirectional context in the medical time-series analysis.
A Bidirectional Long Short-Term Memory (BiLSTM) network is formally a network which is made up of two parallel LSTM layers which process the input sequence forward and backward direction. Given an input sequence $\{x_1, x_2, \ldots, x_T\}$, the forward LSTM computes the hidden states as:
First, convolutional layers extract local temporal features:

$$h_t^{(CNN)} = f(W * x_{t:t+k} + b) \qquad (7)\ \text{Convolutional layer output} \tag{2.10}$$

where W is the convolutional filter, $x_{t:t+k}$ represents a temporal input window, and f is a nonlinear activation function (e.g., ReLU).
The forward LSTM processes the extracted features in chronological order:

$$\overrightarrow{h_t} = \text{LSTM}_f\left(h_t^{(CNN)}, \overrightarrow{h_{t-1}}\right) \tag{2.11}$$

Similarly, the backward LSTM processes the sequence in reverse order:

$$\overleftarrow{h_t} = \text{LSTM}_b\left(h_t^{(CNN)}, \overleftarrow{h_{t+1}}\right) \tag{2.12}$$

The final Bi-LSTM representation at time step t is obtained by concatenating the forward and backward hidden states:

$$h_t^{(BiLSTM)} = \left[\overrightarrow{h_t}; \overleftarrow{h_t}\right] \tag{2.13}$$

Finally, the model outputs a classification or prediction:

$$\hat{y} = \sigma\left(W_y h_T^{(BiLSTM)} + b_y\right) \qquad (9)\ \text{Final prediction} \tag{2.14}$$

This bi-directional formulation allows consideration of the antecedent and subsequent clinical setting in representation generation, and thus giving it significant benefits when it comes to healthcare time-series analysis, where the analysis of the current status of a patient may depend on subsequent observation.
The findings are all quite convincing to support the idea that BLSTM-based architectures are very effective in healthcare prediction, motivating their inclusion in our

system.

Hidden Markov Models (HMMs) have long been used in medical research to model disease progression and temporal dependencies in patient data. HMMs give an interpretable probabilistic structure unlike deep learning methods, which tend to be black boxes, and in this way, they model the health of a patient as a chain of the occurrence of hidden states. This recursive computation makes Hidden Markov Models (HMMs) suitable for real-time patient monitoring, where new data updates the probability of the patient being in a particular health state.

Recent research has demonstrated the power of HMMs and their extensions in healthcare:

1. AttDMM (Attentive Deep Markov Model, 2021) [10] introduced attention mechanisms into a deep Markov framework for ICU patient risk scoring. Using the MIMIC-III dataset, AttDMM achieved an AUROC of 0.876, outperforming standard LSTM baselines by effectively combining latent Markov states with neural attention. This work highlights the potential of augmenting interpretable probabilistic models with deep architectures for clinical risk prediction.

2. The comorbidity-HMM (2021) [12] suggested that the co-evolution of diabetes and chronic liver disease could be modeled with a coupled HMM. The coupled model introduced non-homogeneous transition probabilities as opposed to the traditional HMMs, which assumed independent trajectories, and modeled disease interactions. The experiments showed that the coupled HMM was much better fit to patient progression data compared to independent HMMs, with higher likelihood scores and more real-world disease progressions.

3. Continuous-Time HMMs (CT-HMMs) [8] extend traditional HMMs to handle irregularly sampled medical events, which are common in Electronic Health Record (EHR) data. Instead of discrete time steps, CT-HMMs define transition intensities λ_{ij} between states, with transition probabilities computed as:

$$P(s_{t+\Delta t} = j \mid s_t = i) = [\exp(\Lambda \Delta t)]_{ij} \quad (15) \text{ Continuous-time transition} \qquad (2.15)$$

where Λ is the matrix of transition rates. Applications to Alzheimer's disease, COPD, and glaucoma progression showed AUC values in the 0.75–0.85 range, demonstrating the ability of CT-HMMs to capture disease dynamics even with sparse and irregular data. Together, these recent studies show that HMMs continue to play a crucial role in healthcare analytics. While neural models like LSTMs provide higher raw predictive accuracy, HMMs offer transparent, state-based representations of disease trajectories. For this reason, our system integrates HMMs with deep neural networks, combining interpretability and predictive power for patient health risk forecasting.

Nguyen et al.[11] examined how convolutional neural networks (CNNs) could be used in cardiovascular risk stratification. The relevance of CNN architectures in analyzing large scale patient data in both structured and imaging-based settings is brought out in their study. Zhang et al.[16] introduced a secure deep-learning framework with the focus on privacy-preserving computation in clinical risk classification.

The research supports the feasibility of realizing excellent predictive performance coupled with real-time compliance validation and encrypted model inference. JD-Supra[19] also indicated that in addition to accuracy, AI systems in the current healthcare environment should be expected to be explainable, fair, and transparent to regulatory authorities. These findings and results justify the use of interpretable models, e.g. Hidden Markov Models (HMMs), in combination with neural networks (CNN-BiLSTM) in this project to enable time-sensitive and privacy-compliant patient risk prediction.

OCR and Data Extraction from EHRs

The extraction of structured information from scanned paper forms is one of the barriers in working with patient records. The optical character recognizer (OCR), especially the Tesseract program (Smith, 2007)[1] has allowed the conversion of clinical text and hand written notations into machine readable forms. However, OCR is not sufficient to work, but also requires Natural Language Processing (NLP) to label and organize valuable clinical data, particularly in privacy-sensitive analytics. Zhou et al.[7] went even further and applied Named Entity Recognition (NER) modeling of electronic medical records to identify changing Protected Health Information (PHI) including names of patients, their date of birth (DOBs), addresses, and contact information. The model had entity-level F1 scores of over 0.90 and was very consistent in detecting and filtering sensitive information in clinics. Yadav et al.[15] suggested a system for redaction PHI based on Optical Character Recognition (OCR) (in scanned images) and NLP (process language). Their policy is able to analyze and detect sensitive data, and therefore, it meets the HIPAA requirements. Although it is not used in this project, transformer-based models, i.e., RoBERTa and ClinicalBERT, have demonstrated good results in the context of clinical de-identification tasks. As an example, Meaney et al.[13] obtained more than 99% overall accuracy and 96.7% precision and recall on the i2b2 dataset de-identification benchmark with RoBERTa to exceed the state-of-the-art PHI anonymization. Such positive results provide a promising future direction for further efforts to enhance the preprocessing component of such systems as ours.
Nevertheless, the current project takes a more lightweight path of NER that trades off effectiveness against computational efficiency and is very well-suited to be deployed into resource-constrained (or real-time) healthcare settings.

HIPAA Compliance and Secure System Design

Security of the data and compliance with laws are very important since AI applications in the healthcare field manage highly sensitive patient data. HIPAA (1996) Health Insurance Portability and Accountability Act contains technical and administrative requirements on the ways Protected Health Information should be treated. Halapeti[18] has proposed a blueprint of HIPAA-compliant AI, which includes such fundamental aspects as access control, encryption, secure backups, and audit logs. Moreover, Foley Lardner[17] outlined optimal practices for implementing AI in a clinical setting, which include field-level encryption, JWT-based authentication, and role-based access control (RBAC). These guidelines directly influence the design of our system. Simultaneously, a Reuters report from 2025[21] highlighted the shifting

enforcement landscape surrounding the HIPAA security rule, noting an increased focus on risk assessment, secure audit logs, and compliance with encryption standards within AI workflows.
From an academic viewpoint, the JASON report titled "Artificial Intelligence for Health"[5] underscored the significance of traceability, modular system design, and secure data management as fundamental components of trustworthy AI in healthcare. Motivated by these suggestions, this project incorporates secure APIs, multilayered access control, and auditing processes to guarantee a privacy-conscious and legally compliant framework for clinical prediction activities.

Scalable Deployment and System Architecture

Healthcare systems in the modern setting must be not only scalable but also modular, especially when dealing with real-time, secure implementations of AI. Docker was created by Solomon Hykes in 2013, and later made open source in 2014, has now become the favored choice for developers to encapsulate services in a way that they operate as intended, both correctly and securely regardless of where they are deployed. When it comes to our initiative, we decided to containerize a number of modules, including OCR, prediction, and API management, so that the overall system became much more manageable. Key elements of data governance in healthcare AI, including the capacity for auditing, model version control, and monitoring data flows, were highlighted by Sweeney et al.[14] and influenced our design decisions. Our collaborative work has enabled us to create a system that is both practical and fully compliant with contemporary data security standards.

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

The architecture is deployed as a microservices system comprising primary components deployed through Docker containers. The technical details consist of a three-step process of processing the health records of patients and predicting their diseases.

Technical Requirements:
The first, the Preprocess Service is created with Fast API version 0.111.0 with uvicorn 0.30.1 as the ASGI server. The service uses PyMuPDF (1.24.7) to extract text via PDF, Pillow (10.4.0) to process images, pytesseract (0.3.10) to read printed characters, and spaCy (3.7.5) to identify names in the text. The medical documents uploaded to the service were processed to retrieve text and find protected health information. The second step is the Patient Store Service, which is based on FastAPI 0.111.0, SQLAlchemy 2.0.31 to operate with the database, psycopg2-binary 2.9.9 connectivity to PostgreSQL, and cryptography 43.0.1 to perform encryption operations. A PostgreSQL database is the persistence database that is used to store patient records, documents, extracted entity, and structured clinical events. The service uses audit logs of any read and write operations. The Inference Service is the third stage that uses FastAPI 0.111.0 with HTTPx 0.28.1 to make HTTP client requests. It has NumPy 1.26.4, joblib 1.4.2, scikit-learn 1.5.1, hmmlearn 0.3.2 (operating the hidden Markov Model), and TensorFlow 2.15.1 (inferring deep learning functions) in the machine learning stack. One of the loaded model bundles used in the service includes the trained CNN-BiLSTM deep model, HMM feature extractor, fusion MLP, tokenizer, multilabel binarizer, and label mappings. Auth Service offers a service that implements JWT-based authentication using FastAPI 0.111.0, pydantic 2.7.4 to support data validation, and python-jose 3.3.0 to perform operations with JSON Web Tokens and passlib 1.7.4 to hash passwords. The Gateway service uses Caddy as a reverse proxy to offer TLS termination using a local certificate authority and referrals to internal services. The Web Portal service is a server-rendered user interface to co-ordinate document uploading, processing, and predicting workflow.

Data Requirements:
The system handles the data from the synthetic electronic health record created on the Synthea platform. The sample size of the training set is 47, 013 patients with an average of 2.12 disease labels per patient in 25 different conditions represented by codes of the SNOMED-CT. The data is divided and made into training, validation, and test set in a 70:15: 15 proportions, making a total of 32,909 training patients, 7052 validation patients, and 7052 test patients. The sequences of patient events are tokenized with a maximal length of a token 512 and the HMM component works with sequences containing 25 or more tokens.

Deployment Requirements:
The system use Docker and Docker Compose to make deployments in containers. It is built to execute on regular CPU infrastructure, but is also capable of using GPU acceleration in situations where a GPU enabled Docker image and host GPU support exist. The secure deployment mode makes services available via the gateway on port 8443 and the development mode makes services available on port 8000 to 8004 individually. Environment variable is configured using environment variables found in a configuration file.

3.2 Societal Impact

The current system attempts to address the social issue of improving healthcare delivery through predictive analytics without compromising patient privacy. It predicts several simultaneous conditions based on the health records of the patients and allows early risks to be identified, which can inform clinical decisions and provide the opportunity to intervene earlier and treat patients more effectively. The multi-label prediction feature allows it to identify patients with co-occurring disorders, which is a significant strength in the management of complex chronic diseases that are often clustered together.

The HIPAA-compliant architecture shows how sophisticated AI-based health prediction can be implemented with strong privacy protections, thus dispelling the fears of society about the use of sensitive medical information. The design satisfies the societal requirements of human control over medical AI usage by being a clinical decision-support system, as opposed to an independent diagnostic system. As a result, healthcare specialists retain the influence regarding clinical decisions and at the same time benefit on AI-produced risk assessment and recommendations.

The use of artificial Synthea data as a means of development and testing can allow the system to be developed and tested without the need to have access to actual patient information, thus preventing privacy concerns in the research and development process. This plan will enable the development of AI technology in healthcare and ensure patient privacy in the development of the system. However, the use of synthetic data will potentially restrain the ability to apply the existing performance measures of the system to the real clinical environment, and the system must be

externally tested on genuine electronic health records of working healthcare organizations before it can be applied in any patient care facility.

3.3 Environmental Impact

The environmental footprint of the system depends mostly on the size of the computational resources needed to train the model and perform inference. The training process of the deep learning model was run in 12 epochs with a per-epoch training time of between 291 and 332 seconds during run time on the CPU infrastructure. The entire training schedule involving the CNN-BiLSTM architecture, HMM feature generating, and fusion layer training can be run on a standard server hardware that does not necessarily need any specialized graphics hardware, but any access to a GPU would make training times and energy usage shorter. The containerized microservices reuse allows effective deployment of resources via isolable services. Services are scalable separately depending on demand which means that computation resources can be deployed only on demand instead of having over-sized monolithic systems. The Docker containers have made it easier to deploy on a shared infrastructure where several applications can be co-located and hence the total hardware footprint is lower than when one application is dedicated by deploying a dedicated server. In the case of inference operations, the system works effectively with the predictions of the patients, and the model can produce the prediction of an individual patient within seconds using CPU hardware. This inference efficiency implies that the current operational energy use to provide prediction to clinicians is fairly low. The design of the system to run on local or institutional infrastructure instead of necessarily connecting to the cloud all the time makes the network transmission overhead and the related energy consumption low. The open-source software in the project such as FastAPI, PostgreSQL, TensorFlow, and Docker makes the software environment less expensive to the environment in terms of the infrastructure of software licensing. The use of synthetic Synthea data instead of the need to transfer large amounts of data to develop the product also reduces energy levels of data transmission to the research stage.

3.4 Ethical Issues

The first ethical issue relates to the processing of the information about the health of the people that is protected in accordance with the HIPAA regulation. The system provides several privacy controls, such as field-level encryption of sensitive information in the database, JWT-based authentication to identify the user, role-based access control to ensure that only authorised employees gain access to patient data, and extensive record keeping by audit logging to monitor all access to patient information. Such technical practices deal with the ethical responsibility to protect the privacy of patients and the medical records confidentiality.

The ethical principle of maintaining the human agency in medical decision-making is addressed in the framework of the system as a decision-support tool, but not as an autonomous decision-diagnostic tool. Results produced by the artificial-intelligence models are not supposed to replace, but complement clinical judgement. This will

also make sure that the healthcare professionals do not lose the ability to decide on patient care and be able to take into account such aspects of the patients as their preferences, clinical situations and individual factors that an organized health record might fail to capture.

The use of data of Synthea which is not real raises questions as to the validity of the performance claims which the evaluation will consider. Since the models were tested and trained on only artificial data created based on statistical disease models, the accuracy and performance metrics could not be used as accurate measures of performance in the real world. This should be evidently stated ethically and any purportations on clinical usefulness should be duly put in perspective pending external confirmation on actual patient data. Implementation of the system in a real clinical environment would necessitate approval of an institutional review board, informed consent procedures where necessary, and stringent validation of the system that it is safe and effective in the target clinical setting.

The imbalance in class between the performance of the model and the prevalent conditions with which the system was more accurate than the rare diseases leads to the ethical issues of such imbalance in the health disparities. When the system does not work well in the occasional, but severe cases, this may lead to patients with unusual illnesses being diagnosed or given the treatment late. Ethical implementation would require tracking per-condition performance with special focus given to uncommon conditions but of clinical importance and telling clinicians clearly about the limitations of the system in use with particular conditions. The system is currently lacking mechanisms to measure or counter bias among demographical groups. It is important to assess the means of performance stratified by the age, sex, race, ethnicity, and other demographic factors of the patients to reveal possible gaps that may cause unequal healthcare services provision. Such fairness analyses and a strategy to counter bias would be ethical use of the system in case there is disparity.

3.5 Standards - if applicable

This system is in compliance to the provisions of the Health Insurance Portability and Accountability Act of 1996 (HIPAA) on the security and privacy of electronic protected health information. Technical safeguards are used to achieve compliance, and comprise encrypting data delivered by the system, access controls that restrict access to data to authorized users, audit logs that document every data access and manipulation, and user authentication, which ensures that the user is authentic before gaining access to the system. The diagnostic codes that are used in the system are modeled based on SNOMED-CT, a standard medical terminology that is used across the globe. The codes constituting the twenty five disease labels include: 38341003 of Hypertension, 44054006 of Diabetes, 230690007 of Stroke, 53741008 of Coronary Heart Disease and many more codes that represent the list of acute and chronic diseases. Standardized terminology is used and this makes it easier to interoperate with other medical information systems which use SNOMED-CT coding.

The microservices architecture uses the RESTful API construction relying on the standard HTTP methods and status codes. Authentication is introduced through an Oauth 2.0 password grant flow, a universal protocol of secure user authentication over web services. The session management is implemented with the use of the JSON Web Tokens, which follow the JWT specification of the safe representation of claims between parties. FastAPI is used to build all the services and has auto-generated API documentation that conforms to the OpenAPI Specification and hence, makes it easier to integrate the service with other systems and tooling that support the standard. The storage of data is done in a PostgreSQL database that satisfies SQL standards in the operations of a relational database. The deployment of containers is handled by Docker that adheres to the Open Container Initiative guidelines of container formats and runtimes.

3.6 Project Management Plan

The project was structured into three main development phases, which met a three-stage pipeline architecture. The initial stage was based on the deployment of the preprocessing service of document ingestion, OCR text extraction, and named entity recognition. The second step included the implementation of the patient store service, which included the schema design of the database, data persistence operations, implementation of encryption and audit logging. The third step promoted the development of the inference service that combined model training and evaluation and deployment, and integration of CNN-BiLSTM and HMM models.

The model development was iterative, starting with baseline implementations with the use of TF-IDF features and Logistic Regression as well as Linear SVM to design a performance benchmark. The model development in the form of deep learning was achieved with a series of consecutive architecture choice, hyperparameter optimization (learning rate, threshold, etc.), and training a model over twelve epochs with the validation-based selection of checkpoints. At the same time the HMM component was created, which involved the setting of latent states and the requirements put on minimum sequence length. After training the single components, the fusion layer was added and it was optimized on validation set.

To assess the value of the results, the evaluation framework was determined using the fixed 70:15:15 train-validation-test split on the 47,013-patient dataset. Measurements of evaluation included subset accuracy, micro-averaged F1 score, macro-averaged F1 score and per-label precision and recall. The methodology used in testing consisted of systematic comparison with the baseline models and the detailed analysis of the per-label performance of the 25 disease categories.

The deployment architecture was done concurrently with model development, a microservices structure was used with Docker Compose settings both in secure and development modes. In secure mode, TLS termination is obtained through the Caddy gateway using local certificate authority, and in development mode services are directly exposed to allow debugging and repeated development.

3.7 Risk Management

Technical Risks: The current system is prone to technical threats that relate to the generality of its predictive model. The training and evaluation were done using artificial data that was obtained using the Synthea simulator, and this means that the effectiveness of the model on real electronic health records (EHRs) is not conclusive. Synthea builds patient trajectories using statistical disease models, which may not be a sufficient representation of the complexity, inherent noise, and variability of real clinical data. Mitigation therefore requires extrinsic validation over actual EHR data of a wide range of healthcare organizations before launching it into clinical practice.

The reliance of optical character recognition (OCR) to digitise the scanned medical records poses a threat of extraction errors that might further spread to the following predictive models. The quality of the documents used, the scanning resolution and the availability of handwritten annotations are some of the factors which determine the performance of OCR. The modern system was created using structured CSV output of Synthea and has not been subjected to thorough testing of structured CSV output of authentic scanned documents. Mitigation measures must thus be thoroughly tested with representative samples of real clinical documents and confidence-scoring mechanisms of OCR output should be incorporated.
The imbalance in classes in the training corpus, i.e. some conditions represented by much more examples than the others, presents a risk of the poor performance of the rare diseases. This model reached a macro-F1 of 0.4989, which is even lower than the 0.5099 of the baseline Linear SVM, and thus indicates that there are issues with the model in correctly classifying the infrequent conditions. The risk is reduced by overseeing per-label performance indicators and clearly communicating to the users the circumstances under which the model displays lower levels of reliability.

Privacy and Security Risks: The storage and use of the protected health information (PHI) pose a risk to privacy. Although the system has encryption, access controls and audit logging, any information technology platform managing PHI is prone to data breach or unauthorized access. The mitigation measures thus involve infrequent security tests, penetration-tests, the installation of intrusion detection tools, maintenance of extensive audit trails that enable law enforcement to investigate the occurrence in case of suspected violations and the application of stipulated incident-respondent measures.
The existing application of authentication using JSON Web Token (JWT) provides session management; however, it is specific to research and demonstration use, as opposed to production. The deployment of operations would require a combination with enterprise-level authentication systems, the creation of multi-factor authentication systems, and the use of secure key-management systems, e.g. the use of a hardware security module or a dedicated key-management service.

Clinical Risks: The subset accuracy that was found 0.2548 indicates that about three-quarters of the patients made at least one incorrect prediction among their profile of the disease. This increases the chances of misdiagnoses in case the clinicians

rely on the predictions which are not complete or are not accurate when they are used as a clinical tool. The mitigation measures applied to this risk include viewing the system as a decision-support tool and not as a self-governing diagnostic agent and offering user training that highlights the limitation of the system, persuading clinicians to use its output as a single part of a collection of inputs when making clinical decisions.
The model shows high accuracy and low accuracy in identification of certain conditions meaning that it may overlook patients who actually have those diseases. The fact that this phenomenon can create a danger of false reassurance when negative predictions are considered as the certain exclusions. Therefore, it is crucial that a targeted user training and clear communication on per-condition reliability are mitigation measures.

Operational Risks: The use of a microservices architecture creates an increased complexity in operation in comparison to monolithic designs. The implication of interdependency of services is that the breakdown of one component may affect the total system availability. This risk can be mitigated by undertaking regular health checks, offering graceful degradation pathways wherever possible, and implementing monitoring infrastructure that is able to detect failures in the service. Also, containerized deployment allows quick restarts and effective recovery processes.
The dependency of the system on an external model bundle means that coordination is expected to be extremely careful so that updates do not conflict with the inference service and the related model artifacts. Strong version management procedures as well as strict testing procedures are thus inalienable to prevent implementation of incompatible model components.

3.8 Economic Analysis

The economical aspects of the system include the cost of development and the possible costs of operations to deploy the system. It has utilized open-source software elements, such as Fast API, PostgreSQL, TensorFlow, scikit-learn, Docker, and the Synthea synthetic data generator, to avoid the cost of licenses on the underlying technology. Computation resources needed to train the model on CPU infrastructure is a one-time cost that would be incurred in the early development and later periodic retraining. The twelve-epoch training program, where each epoch lasts between 291 and 332 seconds, can be trained in around sixty minutes of computation in a typical server server training. The inference operations prove to be computationally efficient, which would facilitate the cost-effective implementation. The model can make predictions of a single patient in seconds on CPU infrastructure; therefore, a single server can support a significant number of prediction requests of a large group of patients in an average clinical day. The microservices architecture enables scaling of sing components with demand instead of scaling the entire system and may potentially lower infrastructure costs compared to monolithic architectures which have to be scaled as one unit.

Docker-based containerized deployment can be deployed on a wide range of infrastructures, whether on-premises (servers), in a specific cloud (such as a private or public cloud), or on a different cloud application. Such flexibility enables the in-

stitutions to choose deployment models that they can adopt without compromising the current infrastructure and cost frameworks. The system will be compatible with institutional servers without the need to pay constant subscription fees of cloud services even though organizations may choose to deploy cloud services.

The use of synthetic Synthea data in development addresses cost related to acquisition of real patient data, including data use contracts, clearing institutional review board requirements, and data preparation. However, this development cost-saving will have to be weighed against the need to have validation on actual electronic health record data prior to clinical implementation, which would involve the cost of data accessing and validation research, and may need further refinement of the model. The possible savings of costs generated by clinical implementation would come through the earlier detection of diseases to preventive treatment potentially saving the downstream treatment costs, more effective clinical workflows created by the risk-based prioritization of patients, and better resource allocation due to the screening of high-risk patients, who would need more rigorous monitoring. Nonetheless, these benefits could only be quantified through prospective studies which ascertains real clinical and economic impact after the implementation of the system, which has not been done in the past.

The system development investment is a one-off investment, but the operation cost would include the maintenance of the infrastructure, retraining of the models with the new data availed, continuous security monitoring and updates as well as maintenance, and user support and training. The modular design allows a system to be incrementally upgraded by making changes to single components without requiring a full system upgrade, which has the potential to lower the maintenance costs in the long term.

Chapter 4

Proposed Methodology

4.1 Design Process or Methodology Overview

This project shall take the modular and layer-based approach in designing a HIPAA-compliant health risk prediction system. Our approach to the methodology was built around five points: (1) data collection and processing, (2) predictive model, (3) secure (private) storage, (4) user interface, as well as deployment architecture. Each component will be constructed in a scalable manner, real world application keeping in mind with a high regard for data security. The layering architecture makes the system easier to manage, integrate and deploy in the real world healthcare setting.

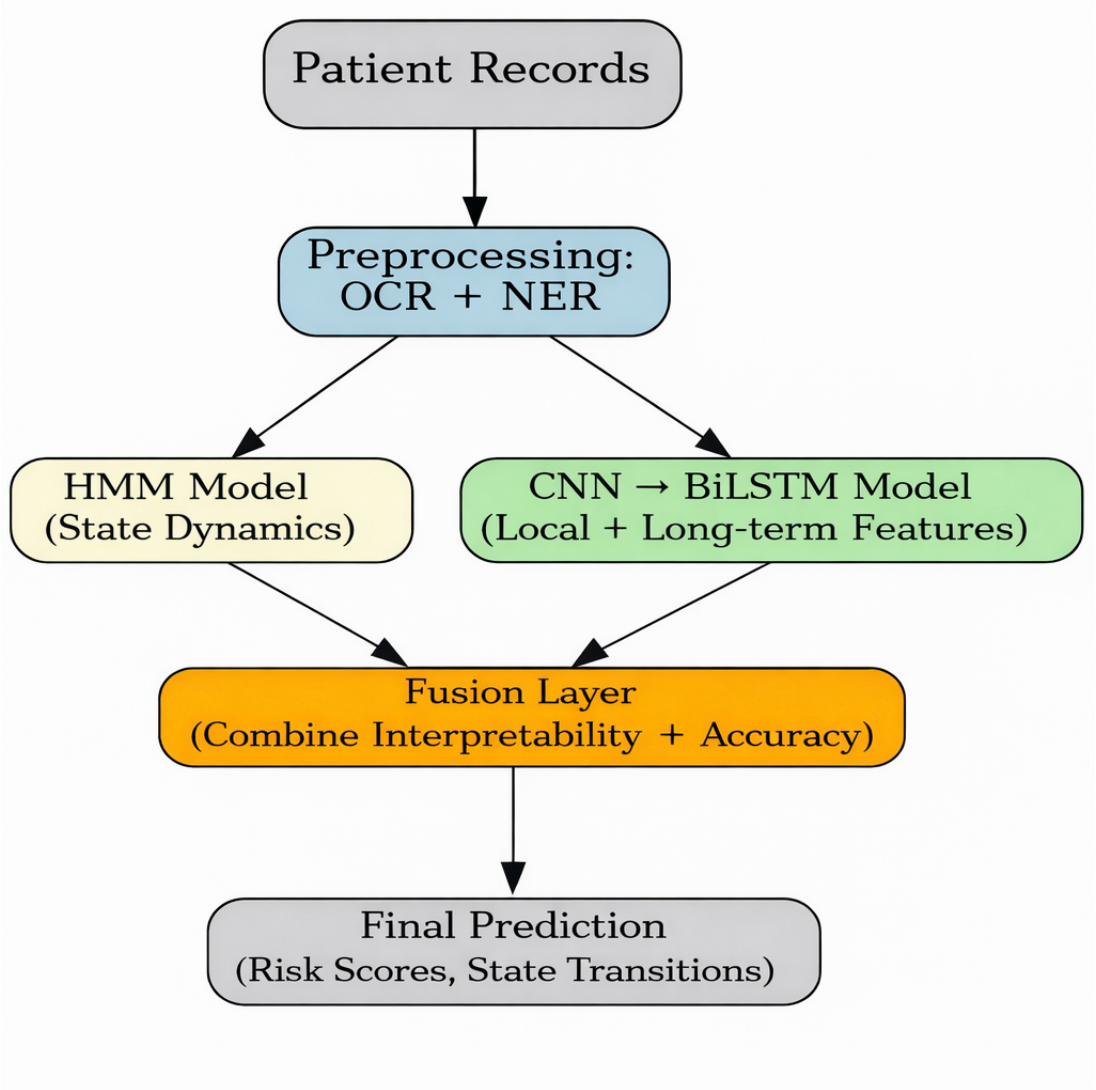
4.2 Data Collection

The first step of our strategy includes obtaining the patients' data using electronic health records, both de-identified and synthetically generated data. In the real-life scenario, this will be investigated by looking at the scanned health records and clinical notes recorded. Images of the document are further converted to plain text using Tesseract OCR to extract the plain text of scanned materials. The retrieved text will be preprocessed using normalization, lowercasing, removal of punctuation, and tokenization. It will use a Named Entity Recognition (NER) process in order to remove sensitive data, including patient names, dates, and addresses to comply with HIPAA regulations. The result will be a clean denormalized dataset which we will parse into the predictive modeling layer. This two-step process will make the information easily amenable to analysis as well as privacy-compliant, providing a responsible basis on which AI will be used in the healthcare sector.

4.3 Predictive Modeling

Two main algorithms will be used to make up the model layer of the system. In this study, the predictive modeling aspect is a combination of probabilistic and deep learning models to provide interpretability, robustness, and high predictive accuracy in clinical decision making. The proposed pipeline, as shown in below uses a dual-model pipe, which consists of a Hidden Markov Model (HMM) and a Convolutional Bidirectional Long Short-Term Memory (CNN–BiLSTM) network. The two models are trained concurrently using the same preprocessed health record information,

and then the output of the models is fused using a fusion layer to yield the final prediction on the level of health risk of patients and state transition.



The modeling process starts with the stage of the preprocessing of the data, in which the raw patient records are processed with Optical Character Recognition (OCR) and Named Entity Recognition (NER) to structured digital text. They then take the processed data as inputs to the two modelling branches. The HMM branch models time dynamics of states, which are the latent health states e.g. healthy, mild or severe. All state transitions are characterised by probabilities based on sequential clinical observations, which enables the model to acquire the changes of the state of a patient over time. This probabilistic element also increases the interpretability of the system by connecting observable events (e.g. the progression of the symptoms) with underlying health states using emission and transition probabilities. Simultaneously, the CNN-BiLSTM branch is aimed at identifying both local and long-term sequential features on the same input data. The convolutional layers identify both spatial and temporal patterns within sequences, i.e. repeated clinical terms or combinations of symptoms, whereas the LSTM layers identify long-term dependencies i.e. how the disease develops with time. The CNN-BiLSTM model produces a list of dense features that captures better contextual and time-dependent dependencies

in comparison with the traditional shallow models. This branch is therefore useful in attaining better predictive performance based on deep neural feature learning.
The results of the two branches are then interpreted into a fusion layer that integrates the interpretability of the HMM and the accuracy of the CNN-BiLSTM model. The fusion is executed using methods of weighted averaging and ensemble, which combine probabilistic and neural forecasting to produce an overall output decision that is a compromise between transparency and accuracy. The integrated model then offers clinicians interpretable results like risk scores and state transition probabilities, which can be utilised to intervene proactively in the medical case.
The models are trained and evaluated with a 70/10/20 ratio of training, validation and testing data sets. The system will be built to manage the multi-classification of health, with hospitals categorising patients based on the risk level of a diagnosis. The HMM parameters are optimised with the help of the Baum-Welch algorithm, and the CNN-BiLSTM model is trained with the help of categorical cross-entropy (loss) and adaptive (learned) learning rate scheduling. The two models are compared in terms of accuracy, weighted F1-score and Macro F1-score to determine the performance consistency among unbalanced classes. The hybridised system thereby makes use of the temporal probabilistic models and deep sequential learner to provide a scalable, hybrid, and HIPAA-compliant predictive system of patient health risk assessment.

4.4 Secure Storage and Data Handling

The computational results and predictions will be securely stored on a PostgreSQL database. Any other identifiers and health summaries will be encrypted fields in this data model. Moreover, the database is access-controlled so that the information about patients in the database can be accessed only by the clearly specified group of individuals (e.g., doctors and administrators). Standard audit logs are also kept to hold on to accountability and also to identify any suspected unauthorized entry. Such capabilities are required when the HIPAA compliance rates are to be achieved, more so when several healthcare providers have to operate on the same systems.

4.5 Web Interface and API Layer

The system interacts with the healthcare professionals via a minimal, basic web interface which is rendered by the server and uses HTML and CSS. This interface is user-friendly and fast, and it will allow clinicians with a clean and efficient method of uploading patient records, predicting health risks, and risk summaries. The web interface is designed in a manner that makes its user experience to be user-friendly and healthcare professionals will therefore be able to have an easy time navigating through the system to acquire the information that they require. The backend is driven by FastAPI a Python framework that is renowned because of its high-performance as well as its ability to be asynchronous. The backend consists of a microservices architecture with each service focused on a certain task, including OCR (Optical Character Recognition) to extract the text, NER (Named Entity Recognition) to anonymize the patient data, and health risk prediction. All the services are independent with each other and interact smoothly through the backend API.

The uploaded documents (e.g., scanned medical records or PDFs) are processed by the OCR service and a part of it is extracted to create pertinent information, and the NER service tries to make sure the HIPAA compliance and remove sensitive data, e.g., patient names and dates of birth. Once the data is handled, the health risk prediction service applies learning models such as CNN-BiLSTM and HMM to forecast the possible health risks of the patient on the basis of his past medical history. These predictions are then given by the system and presented on the frontend interface. FastAPI is used to construct the API layer, where routing, data processing, and model inference requests made by the frontend are processed. The API will communicate with the PostgreSQL database, where patient records, health risk forecasts, etc. will be stored. The system is designed in a way that sensitive information is stored in a secure way and can only be accessed by authorized users where role-based access control (RBAC) is enforced to control access to data.

This system has security as its primary concern and authentication is done using JSON Web Tokens (JWT). Upon the log-in of healthcare professionals to the system, a JWT token is issued and should be added to the further API requests. This makes sure that protected endpoints and sensitive data can be accessed by only authorized persons. Moreover, all the information between the frontend and the backend is encrypted (when transmitting and when at rest) and, therefore, in line with the HIPAA requirements.

4.6 Containerization and Deployment

Docker containers are used to roll out the entire system with every service operating in an isolated container. This containerized design allows modularity, where the development, testing, and maintenance of every service can be performed on an independent basis like OCR, NER, and health risk prediction. Docker will bring about the ability to deploy the system to various environments (both local healthcare servers and cloud-based infrastructure, such as AWS, Azure, or Google Cloud) without reconfiguring it. This design will ensure portability because all dependencies are contained in the containers and the system is environment-agnostic and compatibility problems are removed.

Docker Compose is utilized to control and coordinate the various services. This tool enables easily configuring, starting, and managing all containers so that services could be able to communicate and to collaborate as one system. Docker Compose also enables easy scaling of individual services to meet more demand like the introduction of more instances of the prediction service when there is a need. This horizontal scaling will give the ability to scale the system so that the system can cope with increased traffic or data loads effectively without compromising the performance of the system. This modular deployment system is very scalable, flexible and easy to maintain. Services may be updated or substituted without interfering with the entire system, and when new needs or technology arise, the system may be easily expanded or modified. Finally, the application of Docker makes sure that the system is not only efficient in the present moment but can expand and develop to address the future needs.

Chapter 5

Result Analysis

5.1 Performance Evaluation

Evaluation of the proposed HIPAA-compliant system of health risk prediction was performed based on various metrics that were used to measure the success of the model in multi-label disease prediction in the context of patient health records. The evaluation utilized an extensive testing procedure whereby the dataset was divided into training, validation, and test sets with each comprising 70: 15: 15 ratios. Out of 47,013 patients in the dataset, this amounted to 32,909 patients to be trained, 7,052 to be validated, and 7,052 to be tested. There was a total of 25 distinct disease labels included in the full dataset and each patient had an average of 2.12 disease labels, indicating that the prediction task was multi-label with each patient potentially having more than a single medical diagnosis.
The main evaluation criteria were subset accuracy, which is the percentage of samples in which all label predictions are accurate at the same time, micro-averaged F1 score, which is an equal weighted average of each label prediction and therefore focuses more on the performance on more frequent conditions, and macro-averaged F1 score, which gives equal weight to each label prediction and therefore puts more focus on the performance on more frequent conditions. Other metrics were per-label precision and recall values in order to comprehend model performance on individual disease conditions. The sequence input to the models was set to take at most 512 tokens per patient, as suggested by the sequence shapes of train at 32,909 by 512, validation at 7,052 by 512, and test at 7,052 by 512.

The test methodology was to train a deep CNN-BiLSTM model as the main sequence modeling model, an HMM feature extractor to learn the temporal patterns and then a fusion layer that fused outputs of both models. The deep model training was done in 12 epochs and model checkpoint was saved on validation micro-F1 performance. The validation measures improved gradually over the 12 training epochs.

SNOMED-CT Code	Disease Label
10509002	Acute bronchitis
15777000	Prediabetes
162573006	Suspected lung cancer
195662009	Acute viral pharyngitis
230690007	Stroke
239873007	Osteoarthritis of knee
36971009	Sinusitis
38341003	Hypertension
40055000	Chronic sinusitis
410429000	Cardiac Arrest
428251008	History of appendectomy
429007001	History of cardiac arrest
43878008	Streptococcal sore throat
44054006	Diabetes
44465007	Sprain of ankle
444814009	Viral sinusitis
53741008	Coronary Heart Disease
62106007	Concussion with no loss of consciousness
64859006	Osteoporosis
65363002	Otitis media
68496003	Polyp of colon
72892002	Normal pregnancy
74400008	Appendicitis
75498004	Acute bacterial sinusitis
87433001	Pulmonary emphysema

Table 5.1: Disease Labels and Corresponding SNOMED-CT Codes

Each epoch took different training times depending on computational load and the execution of call backs. Binary accuracy of single label prediction was also monitored as part of the training process, it rose to 0.9158 in epoch 1 to 0.9447 in epoch 12 and validation loss, which dropped to 0.2164 in epoch 1 to 0.1728 in epoch 12.
The top weights were reinstated on the basis of the validation micro-F1 score of 0.5946 at threshold 0.30 which was also equivalent to a macro-F1 of 0.4982 and subset accuracy of 0.2408 on the validation set. HMM component was trained on sequences, whose length (considered as a minimum requirement) met the requirement, and 32,616 sequences of 32,909 training sequences passed the requirement of having at least 25 events. HMM feature extractor generated 10 dimensional features of each patient, the shape of training features was 32,909 by 10, validation features 7,052 by 10 and test sets 7,052 by 10.
The last fused model that fused the deep CNN-BiLSTM predictions and HMM features via a learnable fusion layer had validation micro-F1 of 0.5932, macro F1 of 0.4928, and subset accuracy of 0.2558 at threshold 0.30. In the held out test set, the fused model had a subset accuracy of 0.2548, micro-F1 of 0.5971 and macro-F1 of 0.4989. It should be mentioned that the explicit test set evaluation results of the standalone deep model had not been reported in the experimental logs, and instead

Epoch	micro-F1	macro-F1	Threshold
1	0.4817	0.1622	0.30
2	0.5255	0.2786	0.25
3	0.5507	0.3677	0.25
4	0.5638	0.3782	0.30
5	0.5786	0.4466	0.25
6	0.5849	0.4486	0.30
7	0.5874	0.4593	0.30
8	0.5863	0.4753	0.30
9	0.5904	0.4789	0.30
10	0.5943	0.4826	0.30
11	0.5941	0.4907	0.30
12	0.5946	0.4982	0.30

Table 5.2: Validation micro-F1, macro-F1, and threshold for each epoch

Epoch	Training Time (seconds)	Binary Accuracy	Validation Loss
1	292	0.9158	0.2164
2	295	0.9202	0.2114
3	299	0.9254	0.2058
4	332	0.9301	0.1991
5	318	0.9336	0.1942
6	298	0.9381	0.1885
7	294	0.9411	0.1826
8	308	0.9424	0.1795
9	293	0.9431	0.1769
10	297	0.9439	0.1745
11	296	0.9444	0.1730
12	291	0.9447	0.1728

Table 5.3: Training Time, Binary Accuracy, and Validation Loss for Each Epoch

the only validation set results of the deep model were micro-F1 equal to 0.5946, macro-F1 equal to 0.4982, and subset accuracy equal to 0.2408 at a threshold of 0.30. The final evaluations were consistently done at the threshold of 0.30 which was the most optimal decision limit to use during the validation to apply probability predictions to label assignments in binarity.

5.2 Analysis of Design Solutions

The design solutions adopted in this paper had three major elements that worked synergistically to support a HIPAA-compliant multi-label predictive system of diseases based on structured patient events. The former consisted of the data preprocessing and model input preparation pipeline which converted synthetic Synthea patient data into event sequences in a structured format. The second component was the predictive modeling framework that involved deep CNN-BiLSTM model and

the HMM feature extractor. The third element was the fusion mechanism and the overall system architecture in the secure deployment. The nature of each of these design solutions was different when tested on the validation and test dataset.

The deep CNN -BiLSTM model served as the key sequence-modeling part. This model has a micro-F1 of 0.5946, macro-F1 of 0.4982 and a subset accuracy of 0.2408 at the best threshold of 0.30 or so on the validation set. The architecture took sequences of up to 512 tokens per patient, using convolutional layers to obtain local temporal features, and bidirectional LSTM layers to obtain long-range dependencies both forward and backwards in time. Training was done in 12 epochs and the best validation performance was achieved on epoch 12 in terms of restoration of optimal weights. In the training process, there was a stable learning, and the validation loss value dropped by 0.2164 to 0.1728 in epochs 1 to 12 whereas the binary accuracy of individual label predictions has been rising by 0.9325 to 0.9452 respectively. The first 12 epochs had the initial learning rate of 3.0×10^{-4}

The fact that the validation loss is reduced and the binary accuracy is simultaneously increasing as the epochs go shows that the model was able to learn sequential patterns within the patient data. That micro- F1 of 0.5946 is higher than macro- F1 of 0.4982, means that the model is more realistic on the frequent labels than on the rare labels, which is because macro- F1 is equal-weighted by all labels regardless of the amount of their support, whereas micro -F1 is mainly determined by the common conditions. The subset accuracy of 0.2408 on the validation data means that on average one patient of every four had its full disease profile predicted correctly, the multi-label prediction task being a difficult one with 25 possible labels.
The HMM feature extractor was another type of complementary modelling that added explicit time structure to the prediction pipeline. The HMM was trained to be able to learn latent health states based on patient event sequences, with sequences containing at least 25 events to be trained, which produced 32,616 training sequences that were valid, out of the total of 32,909 patients. The extraction process of the features produced ten-dimensional vectors of probabilistic temporal patterns.
The structure of the fusion architecture was to combine the output of the two modeling strategies by using a trainable multilayer perceptron model that combined the output of the 25 dimensional probability vectors of the deep model with the 10 dimensional HMM feature vectors. This combination method gave a validation micro-F1 of 0.5932, macro-F1 of 0.4928, and subset of 0.2558 at threshold 0.30. The fused model achieved a micro-F1 of 0.5971, macro-F1 of 0.4989 and subset accuracy of 0.2548 on the test set. In comparative analysis, the subset accuracy of the fused model 0.2558 was higher than the validation subset accuracy of deep model 0.2408, indicating a better representation of co-occurrence patterns of multiple labels. A slightly reduced validation micro-F1 (0.5932 vs. 0.5946) and macro-F1 (0.4928 vs. 0.4982) point to the fact that the fusion model took a slightly different trade-off of precision and the recall.

High-precision, low-recall labels, including label 10509002 (precision 0.97, recall 0.17) can be viewed as evidence of the model being conservative in its predictions, and that it was accurate on cases when predictions have been made but it has missed many

Label ID	Precision	Recall	F1-Score	Support
162573006 (Suspected lung cancer)	1.00	0.91	0.95	275
15777000 (Prediabetes)	0.80	0.94	0.86	2,338
38341003(Hypertension)	0.74	0.87	0.80	1,933
44054006 (Diabetes)	0.92	0.79	0.85	481
87433001 (Pulmonary emphysema)	0.98	0.71	0.83	272
10509002 (Acute Bronchitis)	0.97	0.17	0.30	885
239873007 (Osteoarthritis of knee)	0.90	0.14	0.24	264
428251008 (History of appendectomy)	1.00	0.04	0.08	342
74400008 (Appendicitis)	1.00	0.04	0.08	342
195662009 (Acute viral pharyngitis)	0.44	0.09	0.15	1,026
40055000 (Chronic sinusitis)	0.35	0.34	0.34	1,582
444814009 (Viral sinusitis)	0.43	0.73	0.54	1,754

Table 5.4: Per-Label Performance Based on Test Set Classification Report

true positives. The labels with low precision and low recall include label 195662009 (precision 0.44, recall 0.09), which implies that this label is especially difficult to predict because of an ambiguous or a rare occurrence of clinical presentation. On the other hand, labels with high precision and high recall such as label 162573006 (precision 1.00, recall 0.91) represent the conditions that have a unique temporal structure that the model was able to capture.

5.3 Final Design Adjustments

The process of development was carried out with multiple iterations and modifications carried out to streamline the performance of the models according to the results of the validation sets and the model behavior analysis. The main modifications were related to threshold choice, training period, architectural set-ups, and fusion policy. The most notable modification was the choice of the best threshold of decision to convert the probability outputs to binary label prediction in the multi-label classification task. The training procedure tested various threshold values at the various epochs. Different thresholds were explored by the early epochs: threshold 0.30 was used in epoch 1, threshold 0.25 in epochs 2, 3 and 5, and threshold 0.30 consistently used in epochs 4 through 12. Final model threshold was decided at threshold 0.30 by the performance of validation set optimization and this threshold was applied to all final assessments on both the validation and test set. Interpretation: 0.30 as opposed to the typical default of 0.50 can be viewed as a design choice to permit the class imbalance occurring in medical data whereby many conditions are characterized by relatively low base rates. The training was set to 12 epochs with checkpoint saving mechanism to save the best model weights according to validation micro-F1 performance. Validation micro-F1 increased steadily with the epoch starting with 0.4817 in epoch 1 up to 0.5946 in epoch 12 whereas the macro-F1 increased steadily

to 0.4982 in epoch 12. The optimal weights of epoch 12 with validation micro-F1 equal to 0.5946 were reinstated as the final model setup. The training was done at a 3.0e-04 learning rate across the 12 epochs. Interpretation: The steady increase of all 12 epochs with no obvious plateau may be viewed as an indication that the model utilized the entire training time without exhibiting overfitting throughout the training time that was examined.

The sequence length setup was configured at 512 tokens per patient to the extent that the distribution of patient occurrences of the event history was analyzed in the dataset. This structure was a tradeoff between having enough history of the patient to learn the temporal patterns and having computational tractability of LSTM-based sequence modeling. Interpretation: The selection of 512 tokens may be viewed as a design tradeoff that would fit most patient histories without much truncation that would cause the loss of valuable time information. The HMM component set-up had set the minimum selection length of training sequences at 25 events. This filtering parameter left 32,616 training sequences and only 6909 patients overall. The HMM identified 10-dimensional feature vectors that had the time dependent patterns such as the state transition probabilities as well as the state occupancy distributions. Interpretation: The 25 event minimum requirement could also be interpreted as a design choice to make sure that the HMM was trained on sequences that are long enough to model any state transition, but not very short ones which may not have enough information to reliably predict state-based patterns. The fusion layer was created as a weightable integration representation as opposed to a less complex fixed weighting system. The fusion layer was trained to fuse the 25 dimensional deep model probabilities and the 10 dimensional HMM feature vectors with the help of optimization in the validation set. This model design allowed the model to find the best integration strategies that may be different under different labels as opposed to using the same weighting in all the conditions. Interpretation: The trainable fusion method can be viewed as permitting the model to infer which element (deep learning or HMM) gives more trustworthy data to every particular circumstance, according to the patterns of the data of validation.

Component	Configuration/Decision	Rationale/Interpretation
Threshold Selection	Epoch 1: 0.30, Epochs 2, 3, 5: 0.25, Epochs 4-12: 0.30	Chosen threshold of 0.30 to accommodate class imbalance, common in medical datasets where many conditions have low base rates.
Training Duration	12 epochs with checkpoint saving	Ensured continuous improvement without overfitting, with final weights restored from epoch 12.
Sequence Length	512 tokens per patient	Balanced computational tractability with sufficient temporal context for LSTM-based sequence modeling.
HMM Configuration	Minimum sequence length of 25 events for training	Ensured meaningful state transitions and reliable patterns from sequences with sufficient temporal span.
Fusion Layer	Trainable integration mechanism combining deep model and HMM features	Allowed model to optimize which component (deep learning or HMM) contributes more reliable information for each label.

Table 5.5: Summary of Key Development Decisions

5.4 Statistical Analysis

The performance of the models was statistically evaluated based on classification measures based on the test and validation sets. The test set comprised 7,052 patients and 14, 989 label instances spread over 25 disease categories, therefore, used as a hold-out dataset to determine the performance. The number of per-label support ranged between 110 samples (label 75498004), and 2,338 samples (label 15777000). Micro-averaged measures provided an aggregate measure of performance weighted by frequency of labels. Precision, recall and F1-score of the test set were 0.65, 0.55 and 0.60, respectively, and micro-averaged. This F1-score of 0.60 is almost equal to the reported micro-F1 of 0.5971 separately, with the range of rounding error. The resulting micro-averaged values are a performance that is skewed to more common conditions, which have a larger contribution to the metric overall as they support it more.

Macro-averaged measures provided a different view point which considers all labels as equally significant irrespective of their frequency. Macro- averaged precision, recall and F1-score achieved on the test set were 0.80, 0.44 and 0.50 respectively. The F1- score of 0.50 matches the independently recorded macro-F1 of 0.4989, once again within round-off error. The significantly greater macro-averaged precision versus recall reflects an asymmetry in the labels, i.e. when the model makes a

Metric	Test Set Value
Micro-Averaged Precision	0.65
Micro-Averaged Recall	0.55
Micro-Averaged F1-Score	0.60
Macro-Averaged Precision	0.80
Macro-Averaged Recall	0.44
Macro-Averaged F1-Score	0.50
Weighted Average Precision	0.72
Weighted Average Recall	0.55
Weighted Average F1-Score	0.55
Samples Average Precision	0.64
Samples Average Recall	0.58
Samples Average F1-Score	0.58
Subset Accuracy	0.2548

Table 5.6: Summary of Key Evaluation Metrics

prediction of a particular label it is correct more than it is complete in capturing all the true positives.

Weighted-average measures give a third perspective that takes into consideration label-wisely support in computing the mean. The weighted-average precision, recall and F1-score were 0.72, 0.55 and 0.55 correspondingly on the test set. These values fall between micro and macro averages as would be the case under the weighting scheme.

Sample-based averages, which are calculated by averaging measures per sample and then averaging over the entire dataset, generated a precision of 0.64, recall of 0.58, and F1 -score of 0.58. These numbers could be explained as the performance of the model in a patient-centric angle with the accuracy of prediction per patient as an independent variable.

The test set subset accuracy, which involves the labels of all the patients being accurate, was 0.2548. In this regard, about a quarter of the test patients were able to spontaneously identify their entire disease profile perfectly. Since the set of all potential label configurations in 25 disease categories is exponential, this degree of subset accuracy can be considered a significant performance, and it also demonstrates that it is not an easy task to obtain ideal multi-label predictions.

The F1-score per-label distribution depicted a high degree of heterogeneity among disease conditions. The labels with F1 -scores above 0.85 were label 162573006 (0.95), label 15777000 (0.86), and label 44054006 (0.85). F1 -scores of 0.50 to 0.80 included label 44465007 (0.76), label 62106007 (0.78), label 53741008 (0.69), label

72892002 (0.68), and label 65363002 (0.66). Labels with an F1-score below 0.40 included label 10509002 (0.30), label 40055000 (0.34), label 36971009 (0.34), label 195662009 (0.15), label 239873007 (0.24), label 410429000 (0.34), label 428251008 (0.08) .

Label ID	**F1-Score**	**Category**	**Support**
162573006 (Suspected lung cancer)	0.95	High	275
15777000 (Prediabetes)	0.86	High	2,338
44054006 (Diabetes)	0.85	High	481
44465007 (Sprain of ankle)	0.76	Medium	185
62106007 (Concussion with no loss of consciousness)	0.78	Medium	115
53741008 (Coronary Heart Disease)	0.69	Medium	574
10509002 (Acute bronchitis)	0.30	Low	885
195662009 (Acute viral pharyngitis)	0.15	Low	1,026
428251008 (History of appendectomy)	0.08	Low	342

Table 5.7: Per-Label F1-Scores and Support

5.5 Comparisons and Relationships

The contrastive assessment of different modelling methods provides the idea of the inherent characteristics of various predictive paradigms used in the field of multi label disease prediction. Baseline models define attainable performance metrics through traditional machine-learning methods, but the deep-learning and hybrid approaches depict the benefits of the complex sequence modeling and temporal pattern recognition.

The TF-IDF with Logistic Regression baseline is a methodology that assumes sequences of patient events are bags of words and learns linear decision boundaries per label. This model achieved a subset accuracy of 0.1921, micro-precision of 0.8909, micro-recall of 0.4084 and micro-F1 of 0.5601 on the test set. The precision of 0.8150, a recall of 0.3843 and F1-score of 0.4684 were a precision, a recall, and the F1-score respectively. This high accuracy compared to the recall indicates that the logistic regression estimator was a conservative predictor, where the accuracy was attained when it made one of the labels but frequently missed a large number of the true positives.

The TF-IDF Linear SVM base model has different performance features as compared to the logistic regression. Linear SVM on the test set had a subset accuracy of 0.2094, micro-precision of 0.9062, micro-recall of 0.4265 and micro-f1 of 0.5800. Its macro-averaged measures were precision of 0.8431, recall 0.4295 and F1-score of 0.5099. Linear SVM has higher values when compared to logistic regression: subset accuracy increases to 0.2094 (up to 0.1921), micro-F1 increases to 0.5800 (up to 0.5601), and macro-F1 increases to 0.5099 (up to 0.4684). Such advantages can be explained by the fact that the maximum-margin principle in SVM seems to provide a superior generalization in comparison with the maximum-likelihood model used by logistic regression.

Linear SVM baseline validation and test partitions demonstrate high consistency, with an accuracy of 0.2028, micro-F1 of 0.5797 and macro- F1 of 0.5100 on validation, and 0.2096, 0.5801 as well as 0.5099 on the test, respectively. This consistency suggests that the validation data was not overfitted and the baseline models were properly regularized.

Deep CNN 2- Bi LSTM model is a vastly different strategy as compared to the classical baselines. It has a score of micro-F1 of 0.5946, macro-F1 of 0.4982 and subset accuracy of 0.2408 at a decision threshold of 0.30. Clear test-set outcomes were not mentioned. The deep model (micro-F1=0.5946) and Linear SVM (test micro-F1=0.5800) have better micro-F1 metrics than the deep model (macro-F1=0.4982) and Linear SVM (macro-F1=0.5099), respectively. In addition, the validation subset accuracy of deep model (0.2408) is more than the test subset accuracy of Linear SVM (0.2096). This trend of increased micro-F1 and decreased macro-F1 suggests better performance on common conditions at the cost of uncommon conditions but the two are compared using different data fractions (validation and test).
The hybrid fused model that incorporates CNN-BiLSTM prediction with hidden-markov-model (HMM) temporal characteristics through a learned fusion layer had a test subset accuracy of 0.2548, micro-F1 of 0.5971 and macro- F1 of 0.4989. The fused model has better test measures (0.2548 vs. 0.2096) and micro-F1 (0.5971 vs. 0.5800) than the Linear SVM baseline, but has a low macro-F1 (0.4989) than Linear SVM (0.5099).

A comparative analysis of the validation findings of the deep model and the fused model shows the influence of the integration of features of HMM. Validation of the deep model gives micro-F1= 0.5946, macro-F1= 0.4982 and subset-accuracy= 0.2408 whereas validation of the fused model gives micro-F1= 0.5932, macro-F1= 0.4928 and subset-accuracy= 0.2558. Therefore, the validation micro-F1 of the fused model is slightly less (0.5932 vs. 0.5946) and the macro-F1 is slightly less (0.4928 vs. 0.4982) in comparison to that of the deep model, but the accuracy of its subset is higher (0.2558 vs. 0.2408). This is a trade-off which demonstrates that fusion can boost the ability to perform joint multi-label predictions at the expense of single label performance.

The fused model per-label results depict certain outcomes of the test set of the fused model. Label 162573006 had an accuracy of 1.00, recall of 0.91 and F1-score of 0.95 on 275 samples. Label 15777000, which contains 2 338 samples, achieved a precision

of 0.80, recall of 0.94 and F1-score of 0.86. Label 44054006 (481 samples) has a precision of 0.92, a recall of 0.79, and an F1 -score of 0.85. The precision, the recall, and the F1-score of 87433001 (272 samples) were 0.98, 0.71, and 0.83, respectively. Other labels showed varying trends. Label 428251008 and 74400008 with 342 samples each had a precision of 1.00, recall of 0.04, and F1 -scores of 0.08. Label 195662009 was the most precise with 1026 samples with a recall of 0.09 and F1 score of 0.15. Label 40055000 with 1,582 samples achieved a precision of 0.35, recall 0.34, and F1 -score of 0.34. The recall of the labels is extremely low, together with the high accuracy of 428251008 and 74400008 means that the model is highly confident in predicting the conditions that it predicts but missing the majority of the real instances. The poor result of label 40055000, although its support is the second largest, indicates that the classification problem is not only due to the sample size but also to certain temporal pattern peculiarities inherent in the condition.

Model	**Subset Accuracy**	**Micro-F1**	**Macro-F1**
TF-IDF + Logistic Regression	0.1921	0.5601	0.4684
TF-IDF + Linear SVM	0.2094	0.5800	0.5099
Deep CNN-BiLSTM (Validation)	0.2408	0.5946	0.4982
Fused Hybrid Model	0.2548	0.5971	0.4989

Table 5.8: Comparison of Model Performance on Test Set

5.6 Discussions

The experimental findings show that the hybrid CNNBiLSTM and HMM fusion model had test micro-F1 of 0.5971, macro-F1 of 0.4989, and subset accuracy of 0.2548 when trained on synthetic patient data. These results highlight the potential as well as the drawbacks of using the deep learning structures in conjunction with probabilistic time models to predict multiple labels of diseases. Interestingly, the fused model boosted subset accuracy, which was 0.2096 at baseline with the Linear SVM, to 0.2548, which enhanced the prediction of entire patient disease profiles. Conversely, the macro-F1 score went down to 0.4989, which showed some challenges in the correct classification of rare conditions. This pattern indicates that although the hybrid method is able to effectively model co-occurrence associations between several illnesses in single patients, it requires a significant amount of training cases to represent with dependable outcomes on uncommon ailments.

The clinical implications revolve around the deployment strategy. Subset accuracy of 0.2548 suggests that between three to four patients in every five had at least one incorrect prediction in their profile of the disease, which in turn, points to the system being a more appropriate clinical decision support system as opposed to an autonomous diagnostic system. The model is capable of ranking patients based on their level of risk, identifying high-risk patients to undergo clinical re-evaluation, or making possible diagnosis suggestions to be reviewed by the physician. Such human-

AI partnering model is consistent with the current practice where AI systems are developed to enhance clinical judgment but not to replace it.

The fact that high precision and low recall are being observed in terms of specific conditions (e.g., labels 428251008 and74400008, which have a precision of 1.00 and a recall of 0.04) indicates that the model has taken a conservative set of prediction criteria. This kind of strategy can enable the prioritisation of cases that need to be attended to immediately without causing many false alarms.

However, these findings have several serious limitations in the generalisability and applicability of the findings. First amongst them is the use of artificial Synthea data, which is produced using statistical disease models. Electronic health records in the real world have a high degree of divergence, which is characterized by errors in coding, missing encounters, inconsistent documentation, and administrative confouders, which do not occur in synthetic data. Synthea-based temporal patterns rely on simplified disease-progression models that might not be sensitive to the complexity of real disease-progressions, which are affected by patient behaviour, treatment compliance and environmental influences.

The problem of the imbalance of classes is inherent. The lower macro of F1 of 0.4989 compared to the 0.5099 of Linear SVM baseline shows that the deep-learning method has difficulty with rare cases, although it is good at common ones. This fact indicates that deep-learning architecture needs to have large numbers of training examples per condition to utilize the full pattern-learning capacity. This in turn requires close observation of per-condition performance, especially in rare yet clinically important diseases, when used clinically. Future research needs to examine the concept of transfer-learning where large, general medical datasets are fine-tuned with models originally trained on large datasets, which may increase their performance on rare conditions.

The microservice-based system architecture with Docker containerisation and privacy-focused features of the HIPAA-based system creates the basis of deployment with privacy. Nonetheless, in order to make the shift between the existing research prototype and the production one, more protective measures must be taken, including business partner contracts, written security policy and procedures of incident-response, and ongoing security compliance.

Chapter 6

Probable Limitations and Future Work

6.1 Probable Limitations and Future Work

Despite the fact that the existing system facilitates centralized data processing and is HIPAA compliant due to encryption and access controls, all data needs to be centralized so that it can be aggregated to use in the training process. This minimizes the overall application of the system in hospitals and regions. Recently, it has been theoretically proven that federated learning allows training models in different institutions without exchanging raw data (Markaicode, 2025)[20]. Federated learning is not implemented in this work but it is deemed to be updated in future to enhance the privacy and flexibility of the system. Clinical de-identification tasks using Transformer-based NER models, including RoBERTa and ClinicalBERT, have demonstrated improved accuracy in clinical de-identification tasks. Future work would be to incorporate such advanced models and make the process of extracting the information of the patients more precise.

Chapter 7

Conclusion

The present study provides an end-to-end pipeline that is capable of converting unstructured health data to actionable clinical insights with an AI-driven and privacy-first architecture. The system ensures the solution to both technical and ethical issues hindering the development of digital healthcare through the integration of the most advanced machine learning methods into a modular and Dockerized deployment framework: Optical Character Recognition (OCR), Named Entity Recognition (NER), Hidden Markov Models (HMM), and Neural Networks. One of the most important contributions of this work is that it has paid attention to HIPAA compliance and has made sure that each step (data collection to prediction delivery) is as per the high privacy and security standards that are needed to process Protected Health Information (PHI). Based on the best-practice guidance provided by the top-tier research (e.g., Cyber Defense Advisors, Halapeti[18], Sweeney, et al.), this system features field-level encryption, access control, audit logs, and JWT-based security of the API, which makes it applicable in most clinical settings. The prediction pipeline is based on the use of time-aware models, e.g. HMMs to detect the sequence and deep learning models to classify the risk, which is mutually beneficial in terms of interpretability and performance. References to the studies (e.g., Doctor AI and Deep Patient) support the potential of such models to be more helpful than the traditional methods, especially in the conditions of EHR-based diagnosis. Systems architecture-wise, the microservices and containerization (using Docker) approach add scalability, portability, and ease of maintenance to the mix, which are key characteristics in real world deployment in hospitals, clinics, and mobile care systems. This modularity allows updates to be independent, cross platform, and distributed computation in the healthcare networks. To have an even greater impact, possible future integrations might incorporate federated learning, real-time alerting, and advanced NER models like ClinicalBERT. Finally, the system can fill the gap between unstructured records and structured insights and keep the data privacy high.

Bibliography

[1] R. Smith, "An overview of the tesseract ocr engine," in *Proc. 9th Int. Conf. Doc. Anal. Recognit.*, vol. 2, Curitiba, Brazil, 2007, pp. 629–633.

[2] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," in *Proc. Mach. Learn. Healthcare Conf.*, 2016, pp. 301–318.

[3] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell, "Learning to diagnose with lstm recurrent neural networks," *arXiv preprint arXiv:1511.03677*, 2016.

[4] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," *Sci. Rep.*, vol. 6, no. 1, p. 26 094, 2016.

[5] JASON, "Artificial intelligence for health," MITRE Corporation, U.S. Dept. of Health and Human Services, Tech. Rep., Dec. 2017.

[6] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, et al., "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, no. 18, pp. 1–10, 2018.

[7] L. Zhou, G. Hripcsak, and C. Friedman, "Named entity recognition in electronic medical records," *J. Amer. Med. Informat. Assoc.*, vol. 26, no. 9, pp. 738–746, 2019.

[8] S. D. Fox and L. Chen, "Continuous-time hidden markov models for chronic disease progression analysis," *Journal of Biomedical Informatics*, vol. 112, p. 103 621, 2020.

[9] "An automated system for arrhythmia detection using cnn–bilstm," *Comput. Biol. Med.*, vol. 136, p. 104 700, 2021.

[10] J. Cheng, Y. Liu, and H. Zhang, "Attentive deep markov model for interpretable risk prediction in intensive care units," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 11, pp. 4128–4139, 2021.

[11] P. Nguyen, T. Nguyen, and T. Do, "Cnn-based risk stratification in cardiology: A comprehensive approach," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 9, pp. 3349–3359, Sep. 2021.

[12] M. Zhao, R. Wang, and Q. Chen, "Comorbidity-hmm: Coupled hidden markov models for disease progression with comorbidities," *Artificial Intelligence in Medicine*, vol. 121, p. 102 183, 2021.

[13] C. Meaney, W. Hakimpour, S. Kalia, and R. Moineddin, *A comparative evaluation of transformer models for de-identification of clinical text data*, arXiv:cs.CL, 2022.

[14] L. Sweeney, I. S. Kohane, and J. W. Smoller, "Data governance and privacy in ai-driven health systems," *ACM Digit. Health*, vol. 6, no. 2, pp. 112–124, 2022.

[15] R. Yadav, M. Singh, and R. Patil, "Ocr and phi detection in clinical text using hybrid deep learning models," *J. Biomed. Inform.*, vol. 135, p. 104 274, 2023.

[16] Y. Zhang, C. Lin, and J. Wang, "Secure deep learning for health risk classification," *IEEE Trans. Med. Imag.*, vol. 43, no. 1, pp. 201–212, 2024.

[17] Foley & Lardner LLP, "Hipaa compliance for ai in digital health," White Paper, Tech. Rep., May 2025.

[18] A. Halapeti, "The basics of data privacy in healthcare ai: Making sense of hipaa-compliant machine learning," *AI Privacy J.*, vol. 12, no. 1, pp. 34–50, 2025.

[19] JDSupra, *Ai regulation and fairness in healthcare*, Online; accessed 2025, 2025.

[20] Markaicode, *Ai agents in healthcare: Solving hipaa-compliant diagnostics with federated learning*, Online; accessed 2025, 2025.

[21] Reuters, "New legal development herald big changes for hipaa compliance in 2025," *Reuters Health Technology Briefing*, Apr. 2025.