AI-BASED CLINICAL DECISION SUPPORT SYSTEMS USING MULTIMODAL HEALTHCARE DATA

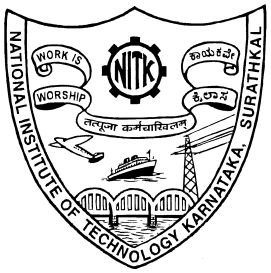
THESIS

Submitted in partial fulfillment of the requirements
for the award of the degree of

DOCTOR OF PHILOSOPHY

by

VEENA MAYYA
(Reg. No.: 187054IT004)



DEPARTMENT OF INFORMATION TECHNOLOGY
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA
SURATHKAL, MANGALORE - 575 025

JUNE 2022

DECLARATION

I hereby declare that the Research Thesis entitled "**AI-BASED CLINICAL DECISION SUPPORT SYSTEMS USING MULTIMODAL HEALTHCARE DATA**", which is being submitted to **National Institute of Technology Karnataka, Surathkal** in partial fulfilment of the requirements for the award of the degree of **Doctor of Philosophy** in **Information Technology** is a *bonafide report of the research work carried out by me.* The material contained in this Research Thesis has not been submitted to any University or Institution for the award of any degree.

Place : NITK Surathkal
Date : 14/06/2022

VEENA MAYYA
Reg.No.: 187054IT004
Department of IT,
NITK Surathkal.

CERTIFICATE

This is to certify that the Research Thesis entitled, "**AI-BASED CLINICAL DECISION SUPPORT SYSTEMS USING MULTIMODAL HEALTHCARE DATA**", submitted by **VEENA MAYYA (Reg. No. 187054IT004)**, as the record of research work carried out by her, *is accepted as the Research Thesis* submission in partial fulfilment of the requirements for the award of the degree of **Doctor of Philosophy**.

Place : NITK Surathkal
Date : 14/6/2022

Dr. Sowmya Kamath S
Research Guide,
Assistant Professor Grade I,
Department of IT,
NITK Surathkal.

Dr. Jaidhar C. D.
Chairman - DRPC,
Department of IT,
NITK Surathkal.

14/06/2022

CHAIRMAN - DRPC
Department of Information Technology
NITK Surathkal, Srinivasnagar P.O.
Mangaluru 575 025, INDIA

Dedicated to

My Mother, Husband and Kids

Who are always there for me to make my

journey special...

Acknowledgements

First and foremost, I would like to thank, from the bottom of my heart, my beloved guide, Dr. Sowmya Kamath S, for her eminent guidance, continuous inspiration, and unconditional support that I received throughout the course of my research. I consider it an honour to have worked under her supervision.

I would also like to extend my gratitude to my RPAC panel—Dr. Rekha S., Dept. of E&C and Dr. Kiran M, Dept. of IT, for all the constructive feedback they have given me, which has definitely enhanced my research. I thank all the co-authors for their valuable contributions in terms of time and expertise.

I thank all of the Department of Information Technology, NITK—all faculties and staff for numerous opportunities to learn and the support they provide at times of need. I thank NITK Surathkal as a whole for providing me with the necessary platform for my research and for the opportunity to attain my PhD I wholeheartedly thank MIT, MAHE, Manipal, for sanctioning the study leave and providing an opportunity to join the PhD course.

It is needless to say that this research venture would not have been possible without the blessings and cooperation of my family members. I thank my wonderful mother— Arundhati Mayya, mother in law— Laxmi Hebbar, husband— Shripada Hebbar, son— Pradyumna and daughter— Pranavi, for the eternal love, patience, and moral support they have showered on me throughout my life, even more importantly, during the period of my research. I also want to thank the rest of my family members, who have always prayed for me and supported me.

I extend my gratitude to my friends, who have inspired me to pursue my goals and supported me through difficult times. I also thank my labmates and fellow research scholars for their support, valuable input, and constant encouragement.

Last but not least, I thank all those who have helped or supported me in one way or the other in accomplishing the completion of my research and thesis.

Veena Mayya

Abstract

Healthcare analytics is a branch of data science that examines underlying patterns in healthcare data in order to identify ways in which clinical care can be improved – in terms of patient care, cost optimization, and hospital management. Towards this end, Clinical Decision Support Systems (CDSS) have received extensive research attention over the years. CDSS are intended to influence clinical decision making during patient care. CDSS can be defined as *"a link between health observations and health-related knowledge that influences treatment choices by clinicians for improved healthcare delivery"*.A CDSS is intended to aid physicians and other health care professionals with clinical decision-making tasks based on automated analysis of patient data and other sources of information. CDSS is an evolving system with the potential for wide applicability to improve patient outcomes and healthcare resource utilization. Recent breakthroughs in healthcare analytics have seen an emerging trend in the application of artificial intelligence approaches to assist essential applications such as disease prediction, disease code assignment, disease phenotyping, and disease-related lesion segmentation. Despite the significant benefits offered by CDSSs, there are several issues that need to be overcome to achieve their full potential. There is substantial scope for improvement in terms of patient data modelling methodologies and prediction models, particularly for unstructured clinical data.

This thesis discusses several approaches for developing decision support systems towards patient-centric predictive analytics on large multimodal healthcare data. Clinical data in the form of unstructured text, which is rich in patient-specific information sources, has largely remained unexplored and could be potentially used to facilitate effective CDSS development. Effective code assignment for patient clinical records in a hospital plays a significant role in the process of standardizing medical records, mainly for streamlining clinical care delivery, billing, and managing insurance claims. The current practice employed is manual coding, usually carried out by trained medical coders, making the process subjective, error-prone, inexact, and time-consuming. To alleviate this cost-intensive pro-

cess, intelligent coding systems built on patients' unstructured electronic medical records (EMR) are critical. Towards this, various deep learning models have been proposed for improving the diagnostic coding system performance that makes use of patient clinical reports and discharge summaries. The approach involved multi channel convolution networks and label attention transformer architectures for automatic assignment of diagnostic codes. The label attention mechanism enabled the direct extraction of textual evidence in medical documents that mapped to the diagnostic codes.

Medical imaging data like ultrasound, magnetic resonance imaging, computed tomography, positron emission tomography, X-ray, retinal photography, slit lamp microscopy, etc., play an important role in the early detection, diagnosis, and treatment of diseases. Presently, most imaging modalities are manually interpreted by expert clinicians for disease diagnosis. With the exponential increase in the volume of chronic patients, this process of manual inspection and interpretation increases the cognitive and diagnostic burden on healthcare professionals. Recently, machine learning and deep learning techniques have been utilized for designing computer based analysis systems for medical images. Ophthalmology, pathology, radiology, and oncology are a few fields where deep learning techniques have been successfully leveraged for interpreting imaging data. Ophthalmology was the first field to be revolutionized and most explored in health care. Towards this, various deep learning models have been proposed for improving the performance of ocular disease detection systems that make use of fundoscopy and slit-lamp microscopy imaging data.

Patient data is recorded in multiple formats, including unstructured clinical notes, structured EHRs, and diagnostic images, resulting in multimodal data that together accounts for patients' demographic information, past histories of illness and medical procedures performed, diseases diagnosed, etc. Most existing works limit their models to a single modality of data, like structured textual, unstructured textual, or imaging medical data, and very few works have utilized multimodal medical data. To address this, various deep learning models were designed that can learn disease representations from multimodal patient data for early disease prediction. Scalability is ensured by incorporating content based learning models for automatically generating diagnosis reports of identified lung diseases, reducing radiologists' cognitive burden.

KEYWORDS: Clinical Decision Support Systems, Natural Language Processing, Computer Vision,Machine Learning, Healthcare Informatics.

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
AUPRC	Area Under Precision Recall Curve
AUROC	Area Under Receiver Operating Characteristic Curve
CBOW	Continuous Bag-Of-Words
BLEU	BiLingual Evaluation Understudy
CDC	Centre for Disease Control and Prevention
CDS	Clinical Decision Support
CDSS	Clinical Decision Support System
CLAHE	Contrast Limited Adaptive Histogram Equalization
CNN	Convolutional Neural Network
CodiESP	Clinical Case Coding in Spanish
DL	Deep Learning
DNN	Deep Neural Network
DS	Dataset
EHR	Electronic Health Record
EMR	Electronic Medical Record
FFNN	Feed Forward Neural Network
FK	Fungal Keratitis
FN	False Negatives
FP	False Positives
FPR	False Positive Rate
FS	Feature Selection
GloVe	Global Vectors
GRU	Gated Recurrent Unit

continued ...

... continued

Abbreviations

HADM_ID	Hospital Admission Identifier
ICD-9	International Classification of Diseases, 9th Version
ICD-10	International Classification of Diseases, 10th Version
ICU	Intensive Care Unit
IR	Information Retrieval
MIMIC	Medical Information Mart for Intensive Care
ML	Machine Learning
MMDL	Multimodal Deep Learning
MRD	Medical Records Department
MSR	Multi-Scale Retinex
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NN	Neural Network
NPV	Negative Predictive Values
ODIR	Ocular Disease Intelligent Recognition
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SGD	Stochastic Gradient Descent
SoB	Shortness of Breath
SVM	Support Vector Machine
TN	True Negatives
TP	True Positives
TPR	True Positive Rate
WHO	World Health Organization

Nomenclature

Notations common to all Chapters

β_1, β_2	exponential decay rates
β	weight for precision or recall
F_1	harmonic mean of the precision and recall with $\beta = 1$
C	number of output classes/labels

Notations used in Chapter 4 and 5

$1\{\bullet\}$	indicator function
e	embedding size
k	convolution kernel size
f	number of feature maps

Notations used in Chapter 6

L	maximum pixel intensity in the current image channel.
I_{eq}	CLAHE enhanced image
I_{gs}	Gaussian filter enhanced image
$I_{ms-retinex}$	MSR enhanced image
loss($\hat{y}$, y)	loss computed between actual target labels (y) and predicted labels ($\hat{y}$)
$Kappa_{score}$	average of Cohen's kappa (measures inter-annotator agreement) for each label
P_o	empirical probability of agreement on the label assigned to any sample (the observed agreement ratio)
P_e	expected agreement (with labels randomly assigned), it is estimated using a per-annotator empirical prior over the class labels

Notations used in Chapter 8

$C(i)$	is the number of i-gram tuples in the candidate document
(t_i)	i-gram tuple in candidate C
$H_c(t_i)$	number of times (t_i) occurs in C
$H_{cj}(t_i)$	number of times (t_i) occurs in reference j of this candidate

PART I

Introduction and Background

Chapter 1

Introduction

Decision support is a crucial requirement in modern healthcare delivery systems. Through the analysis of a huge amount of patient data, the process enables the generation of solutions or usable insights that can be used to improve patient outcomes and healthcare process management. This process empowers clinicians, patients, staff, and other individuals with knowledge and individual-specific information at the right time to improve patient outcomes and healthcare processes. A clinical decision support system (CDSS) is a computer system designed to assist physicians in making clinical decisions during patient care (Berner, 2010). CDSS can be defined as *"a link between health observations and health-related knowledge that influences treatment choices by clinicians, for improved healthcare delivery"*. CDSS is an automated system that assists doctors in making clinical decisions and managing data, for a variety of purposes, like care quality improvement, diagnostic error prevention, adverse event prediction, similarity based outcome prediction, and so on.

Four key functions of electronic CDSS as defined by Perreault & Metzger (Perreault and Metzger, 1999) are listed below:

1. *Administrative:* CDSS can assist in tasks such as documentation, authorization of procedures, procedure and test ordering, clinical and diagnostic coding, and patient triage. CDSS may recommend a more refined selection of diagnostic codes to assist clinicians in picking the most appropriate ones (Mullenbach *et al.*, 2018). A CDSS could be developed to solve inaccuracies in ICD coding[1]. ICD is a medical coding taxonomy that is widely employed to describe patients' procedures and diagnostic conditions. CDSS could assist stakeholders in quickly locating diagnostic codes, which is an essential

[1]International Statistical Classification of Diseases and Related Health Problems, online `https://www.who.int/standards/classifications/classification-of-diseases`

requirement for epidemiology, billing, and managing insurance claims.

2. *Managing clinical complexity and details:* CDSS could perform tasks such as keeping patients on prescribed protocols, tracking orders (Agarwal *et al.*, 2018), referrals, follow-up, and preventive care (Njie *et al.*, 2015). CDSS has been found to improve adherence to clinical guidelines (Kwok *et al.*, 2009). This is significant since execution of standard clinical guidelines and treatment pathways in practice is difficult due to poor clinician adherence (Davis and Taylor-Vaisey, 1997; Cabana *et al.*, 1999). It is quite difficult for practitioners to study, comprehend, and implement new regulations in a timely manner. However, the rules may be encoded literally in the CDSS. Such CDSS may take several forms, including standardised order sets for a targeted case, alerts to a specific procedure for the required patients, and testing reminders. Additionally, CDSS may aid with patient management for research/treatment protocols, monitoring and order placing, referral follow-up, and assuring preventive care. CDSS may also alert doctors to contact patients who have not adhered to care regimens or are due for follow-up, as well as assist in identifying patients who meet certain criteria for research study.

3. *Cost control:* CDSS is cost-effective for health systems by monitoring drug prescriptions (Marcilly *et al.*, 2011), clinical interventions (Calloway *et al.*, 2013), and preventing duplicate or unnecessary testing (Procop *et al.*, 2015). CDSS can also lower the cost of healthcare delivery by optimizing in-patient length-of-stay (Dimagno *et al.*, 2014), avoiding unplanned re-admissions (Cox *et al.*, 2016), suggesting low-cost medication alternatives (McMullin *et al.*, 2004; Kuperman *et al.*, 2007). Additionally, CDSS may shorten the time required to complete diagnostic procedures while also lowering such procedures' cost (Algaze *et al.*, 2016). The CDSS could switch drug consultations automatically and without errors, hence boosting providers' safety, decreasing effort, and lowering costs (Pruszydlo *et al.*, 2012).

4. *Decision support:* CDSSs facilitate clinical diagnosis and treatment planning processes. They promote best practices, disease-specific guidelines, and population health management. This is especially advantageous in locations with a scarcity of established clinical professionals and a requirement for systems that augment specialised diagnosis. Given the known prevalence of diagnostic mistakes, especially in primary care (Singh *et al.*, 2014), there is a

lot of scope for CDSS to improve diagnostic decisions. By assisting in the selection of the most suitable tests, an interventional CDSS for image ordering can significantly reduce medical imaging utilization (Georgiou *et al.*, 2011; Blackmore *et al.*, 2011). Several CDSS have shown diagnostic performance comparable to that of human experts. CDSS can enable doctors to quickly determine whether previously authorised treatment options are feasible for a new patient, assisting in therapy selection (Valdes *et al.*, 2017; McNutt *et al.*, 2018). CDSS can assist in population health monitoring (Amirfar *et al.*, 2011) by combining technologies for extracting latent patterns and providing potentially actionable insights that can be used to help regulate a population's health and also influence public health policy.

CDSS have been classified and subdivided into a variety of categories, based on factors such as intervention time and delivery mode. Regardless, CDSS can be commonly classified as knowledge based systems or non-knowledge based systems (Berner, 2010). Knowledge based systems try to emulate an expert clinician/human's thinking using a set of *if-then-else* rules. Rules may be developed based on literature, practice, or patient-directed evidence. For example, a clinician might have suggested drug Y when a particular drug X is prescribed, based on a set of rules that were previously provided. Non-knowledge based systems are built using statistical pattern recognition, artificial intelligence based approaches (machine learning, deep learning) that extract the knowledge (recognizes the pattern) from electronic health records (EHR) during the training phase and automatically make decisions during the inference phase, thus eliminating the need for and dependency on predefined and complicated rules.

As the availability of EHR data abounds, CDSSs have been increasingly integrated into the modern healthcare system's workflow, allowing healthcare stakeholders to efficiently receive and act on system-generated recommendations. EHRs are seen as a significant step in streamlining the storage, management, and transmission of patient data in hospitals. The EHR comprises vital multimodal patient information such as medical history, diagnosis, prescriptions, treatment plans, vaccination dates, allergies, imaging, and laboratory test results. EHRs can be structured or unstructured with respect to the nature of the data. A structured EHR contains patient data in a predefined, consistent format, typically in the form of rows and columns (e.g., relational data or csv files), with keywords identifying and evaluating data values. The structured EHR includes numerical factors such as age, gender, height, weight, and the results of laboratory tests. In contrast, un-

structured patient data lacks discrete organization; free-form files, textual clinical notes, discharge summaries, scanned documents, and medical images are the most common types of unstructured data accessible in hospital settings. Around 80% of healthcare data remains unstructured and untapped after it is created, mainly due to the challenges of modelling the abundant latent knowledge available. Given the ever-increasing patient population every day, manual storage, search, retrieval, and analysis are highly cost-intensive, often inexact, time-consuming, and error-prone. It is crucial to develop intelligent CDSSs that could leverage the valuable information contained in unstructured healthcare data. This thesis, in general, focuses on the design and development of non-knowledge based CDSSs using multimodal unstructured healthcare data with an emphasis on administrative and decision support tasks.

1.1 Unstructured Multimodal Medical Data

Most healthcare data, whether generated by machines or humans, is unstructured in nature. Machine-generated data includes biosignal data from patient monitors in operating rooms and critical care units as well as information gathered by numerous medical imaging equipment. Additionally, wearable health monitoring gadgets create a wealth of data. Human-generated data comprises recordings of interactions between patients and healthcare practitioners, either in text, audio, or video formats. Free text clinical notes recorded by doctors and nurses while monitoring the condition of patients, as well as discharge summaries, are also examples of human-generated unstructured data. Unstructured data may be represented in a variety of ways and stored/recorded in myriad formats necessitating human analysis and processing by healthcare experts. The enormous volume of generated textual data every day in hospitals are manually analyzed for making clinical decisions and plan the treatment, epidemiology, billing, and managing hospital resources. Given the breadth and volume of available healthcare unstructured data, current data management approaches will no longer sufficient and will pose a slew of challenges. Some of them are discussed in detail next.

1.1.1 Text Interpretation

Clinical text data include significant medical jargon, abbreviations and misspelt words. Currently, most such textual observations are manually reviewed by trained experts and healthcare professionals to make clinical decisions. Few sample medi-

cal notes from MIMIC-III (Medical Information Mart for Intensive Care) database (Johnson *et al.*, 2016), comprising over $40,000$ patients' data, are listed in Table 1.1. As can be seen, each note is long (dots are used to clip many sentences from the original text), often contains spelling mistakes, and contains lot of medical jargon to indicate procedures and terminology that medical professionals use. Consider the diagnostic or procedural code assignment task that is mainly used for billing, and managing insurance claims. Based on clinicians' free-text notes and other patient records such as discharge summaries, doctors' notes, nursing notes, and other relevant sources, trained professional medical coders employed by the Medical Records Department(MRD) in hospitals, transcribe patient records into a set of appropriate medical diagnostic codes (from a potentially large number of choices of over $15,000$ codes). These medical coders utilize their medical domain expertise along with a plethora of coding rules and terminologies to facilitate mapping of a patient record to many diagnostic codes (one-to-many). Manual assignment of such codes is often subjective, error-prone, inexact, and time-consuming.

Table 1.1: Types of unstructured medical notes - samples.

Note category	**Text**
Radiology (CHEST PORTABLE AP)	CLINICAL INDICATION: Central venous catheter placement. Comparison is made to previous study of one day earlier. There has been interval placement of a left subclavian central venous catheter which terminates at the junction of the left brachiocephalic vein and superior vena cava. No pneumothorax is identified. The cardiac silhouette remains enlarged. There are bilateral pleural effusions and adjacent areas of increased lung opacification as well as lower lobes. . . . IMPRESSION: Central venous catheter terminates at confluence of left brachiocephalic vein and superior vena cava. No pneumothorax . . . .
Nursing progress note	Sinus tachycardia Short PR interval Possible anterior infarct - age undetermined Left atrial abnormality Inferior T wave changes are borderline Repolarization changes may be partly due to rate Low QRS voltages in limb leads Since previous tracing of [**2103-7-27**], no significant change

continued . . .

. . . continued

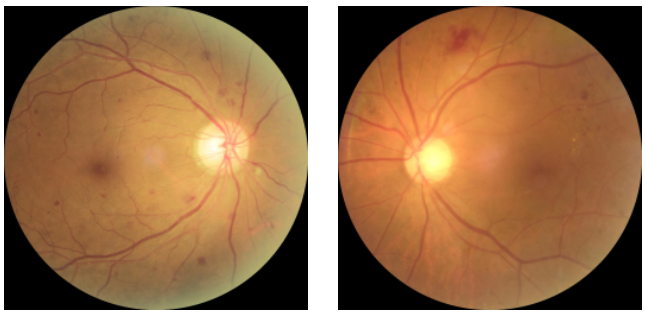
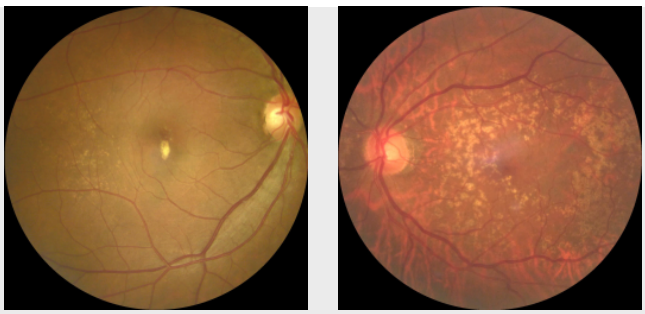
Note category	Text
Radiology (CHEST PORTABLE AP)	Cancer (Malignant Neoplasm), Hepatic (Liver) Assessment: Patient is more lethargic yesterday & today than he was on .... Action: He was made DNR/CMO tonight, per agreement of family. Response: Plan: Acute coronary syndrome (ACS, unstable angina, coronary ischemia) Assessment: Patient had acute SOB, midsternal chest pain, feeling that he was going to die @ [**2016**] when he rolled in bed onto bedpan & had BM. . . . Action: Given 100% high flow neb, 0.5 NTP & 0.25mg IV morphine. EKG done during SOB. Response: Pain & SOB relieved. No changes on EKG . . .
Pharmacy care note	Pharmacy Note Plasmapheresis effect on cyclophosphamide and mesna Patient scheduled to receive Total Plasma Pheresis on . . . Pre-Mesna and ondansetron should not be started until after the pheresis is complete. Final Recommendation: If patient receives TPP on . . . (off day of cyclophosphamide) could consider giving mesna 500 mg q3 hrs times 3 doses post pheresis. Any remaining mesna would be removed with pheresis. Continuous bladder irrigation with sodium chloride 0.9% could also be used to treat hemmoragic cystitis.
Discharge summary	Admission Date: . . . Discharge Date: . . . Date of Birth:. . . Sex: MService: Cardiac SurgeryCHIEF COMPLAINT: Chest pain, 3-vessel disease oncatheterization.HISTORY OF PRESENT ILLNESS: The patient is a 66-year-oldmale transferred from [**Hospital6 33**] to the [**Hospital1 346**] status post catheterization,revealing 3-vessel cardiac disease.The patient presented to . . . tothe point that he had chest pain with minimal exertion.PAST MEDICAL HISTORY:1. Known coronary artery disease, status postcatheterization 10 years ago at [**Hospital1 **].2. Heavy smoker.3. Hypertension.4. Gastroesophageal reflux disease/peptic ulcer disease.5. Wegener granulomatosis with complete resolution.6. Glaucoma. PAST SURGICAL HISTORY: . . . HOSPITAL COURSE: . . .

1.1.2 Image Interpretation

The images obtained in a clinical setting are often captured by a technician and may include a variety of undesirable artefacts resulting in poor diagnostic quality, necessitating re-scans and repeated patient visits. Consider the disease prediction

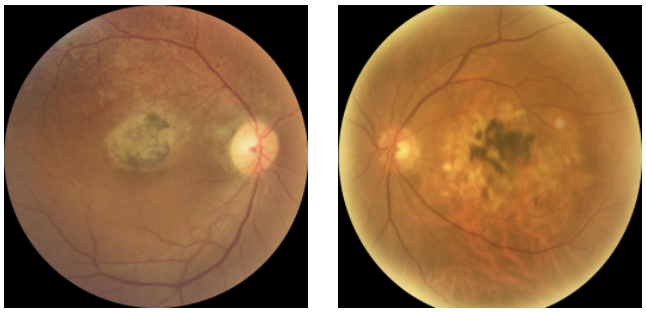
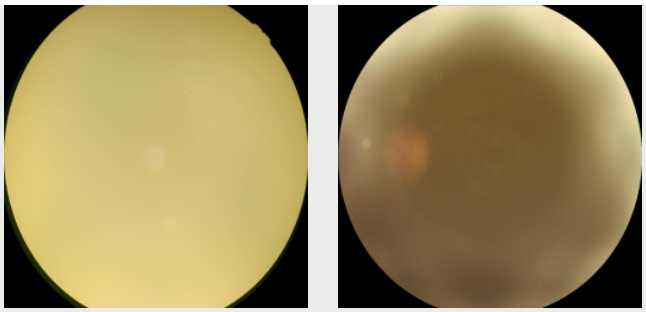
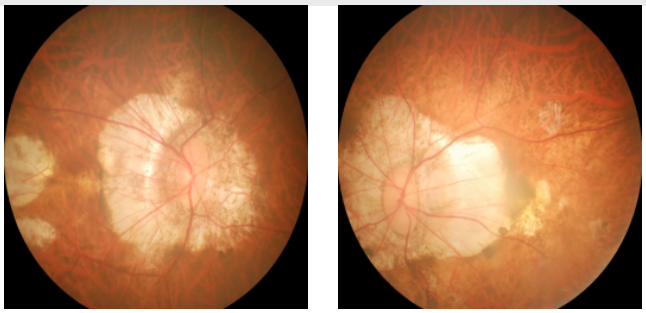
task utilizing the imaging data, imaging modalities present significant challenges due to the diverse textures, overlapping lesions, and diversity within the screening population. Currently, the majority of imaging data is examined and analyzed manually by experienced physicians and licensed professional for disease diagnosis. The decision making may also involve consideration of previous medical history, past procedures and other characteristics. Often decision-making is subjective, resulting in a significant degree of inter-and intra-physician variability in diagnosis. For instance, when a physician analyses an image several days later, they may diagnose it differently than they did before, or there may be discrepancies in opinion when the same image is evaluated by numerous physicians with varying domain expertise. Physicians' preferences for the appearance and quality of the image can change over time and can also differ. Scaling manual interpretation to an ever-growing patient population is likewise a challenge. There are several imaging modalities for every medical domain, and each modality has many more variants for each disease. Additionally, a single image may be related with several diseases. Table 1.2 lists a few examples of fundoscopy images used for detection of various chronic ocular diseases (COD) in the Ophthalmology domain, taken from the ODIR-5K dataset (ODIR, 2019).

Table 1.2: Sample fundoscopy images used for ophthalmological examination.

Image category	**Images**
Diabetic retinopathy	
Dry age-related macular degeneration	

continued ...

. . . continued

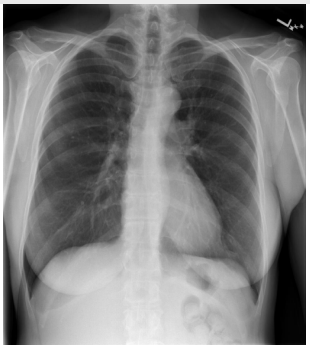
Image category	Images
Wet age-related macular degeneration	
Cataract	
Pathological myopia	

1.1.3 Multimodal Interpretation

For in-depth diagnosis, physicians often rely on a patient's data from multiple sources, such as lab test reports, clinical history notes, and imaging data. When one considers the amount of skill necessary to comprehend a single modality of data in detail, it is almost impossible for a single physician to master all modalities. Table 1.3 lists a few examples of chest X-ray images and the corresponding findings in the form of textual reports. For e.g., a radiologist is trained specifically to interpret radiological images but does not have as much knowledge of surgery or specialized fields like Ophthalmology. While a pathologist is well versed in pathology slide interpretation, he or she is often unaware of how to analyse cardiology ECGs. This demonstrates the need of involving diverse group of specialists in the analysis of patient multimodal data in order to develop an appropriate treat-

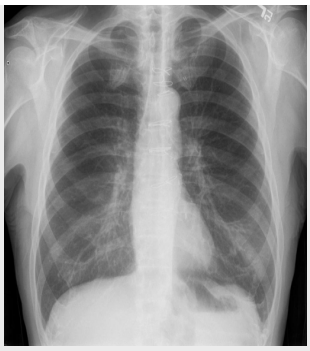
ment strategy. The problem is exacerbated in rural areas where people find it hard to have access to domain specialists in addition to shortage of primary care physicians. Consider the case of radiologists providing the findings of the X-ray images in the form of textual reports. For almost all diseases, radiology is the principal disease management tool. This increases the cognitive burden of radiologists due to the manual effort required to assess and report a large number of patient population. Table 1.3 lists a few examples of radiology X-ray images and the corresponding findings in the form of textual reports (extracted from Indiana University dataset (Demner-Fushman *et al.*, 2016)).

Table 1.3: Sample radiology images and the corresponding report.

X-ray report	**Image**
"The heart, pulmonary XXXX and mediastinum are within normal limits. There is no pleural effusion or pneumothorax. There is no focal air space opacity to suggest a pneumonia. There are degenerative changes of the thoracic spine. There is a calcified granuloma identified in the right suprahilar region. The aorta is mildly tortuous and ectatic. There is asymmetric right apical smooth pleural thickening. There are severe degenerative changes of the XXXX."	
"The lungs are clear. The heart and pulmonary XXXX are normal. The pleural spaces are clear. Mediastinal contours are normal."	
"The lungs are clear. There is no focal airspace consolidation. No pleural effusion or pneumothorax. Heart size and mediastinal contour are within normal limits. There are degenerative changes of the spine."	

continued . . .

. . . continued

X-ray report	Image
"The lungs are hyperaerated suggestive of chronic obstructive pulmonary disease. No focal lung consolidation. No pleural effusion. No definite pneumothorax. Heart is not enlarged. Postsurgical changes with mediastinal clips and XXXX XXXX."	

1.1.4 Data Retrieval and Storage

Given the large volumes of unstructured data generated during healthcare delivery, effective processing and storage requires a very robust and efficient infrastructure. In order to provide unstructured data in usable formats for advanced analytics applications, high-quality metadata must be generated and stored for enabling a well-organized and accessible repository. For instance, codes or keywords that describe an image or textual data need to be entered. At present, these processes are predominantly manual, and thus highly cost and effort intensive. Without such well-organized consumable metadata, it is difficult to support intelligent decision support application powered by the latent knowledge in historical patient data.

Moreover, unstructured data is continuously generated over time and at each point of patient engagement, demonstrating Big data characteristics. As a result, it is essential to design intelligent healthcare data management systems that address these challenges, for enabling automated CDSS based on multimodal healthcare data. AI-based CDSS may be used to analyse enormous amounts of patient-specific data in order to improve healthcare systems and manage effective care delivery. Predictive modelling, preventative modelling, intelligent retrieval, automated treatment/concept extraction and recommendations for physicians and patients are just a few of the applications that have the potential to revolutionize and dramatically improve healthcare practices.

1.2 Motivation

Presently, most clinical decisions, including the initial disease screening, are mostly manually performed by clinical experts based on various sources of patient data.

Manual screening of common cohort diseases like ageing-related diseases, hypertension, cardiovascular diseases, osteoporosis, diabetes mellitus etc, are handled by experienced specialists, due to which dealing with the growing population of patients who are at risk is already an uphill task. It has been reported that more than 80% of patients visit first-tier clinics for trivial cases (Wang *et al.*, 2021). In developing countries where the rural population often lacks access to sophisticated care facilities, high-end medical devices, highly experienced clinicians etc., automating clinical decision support can aid clinicians in early diagnosis of diseases and reduce the burden to some extent by reducing the preliminary manual routine tasks. The amount of unstructured multimodal healthcare data being routinely collected as a part of the EHR provides an unprecedented opportunity to develop an automated CDSS. CDSS is intended to aid healthcare workers in the normal course of their activities by assisting them with tasks that require data and knowledge manipulation (Coiera, 2003).

Recent breakthroughs in computer vision (CV), natural language processing (NLP) and AI technologies have resulted in a dramatic evolution of CDSS in many ways. Though AI-based CDSS perform well for some clinical tasks, very few existing works in the literature provide evidential support to visualize the inference results. For CDSS to be adaptable in real-world circumstances, providing a transparent, explainable decision (even if it is wrong) is considerably more acceptable than putting forth a highly accurate, non-transparent decision, primarily due to the trust barrier between doctors and automated systems. Though AI-based CDSS are capable of recognizing clearly obvious lesions in input medical imaging data that often occur in later stages, the objective is to enable automated systems to learn minute lesions for reliable early detection. While several researchers have independently addressed various clinical tasks, a user-friendly comprehensive framework usable in clinical settings is a critical requirement to validate the systems' efficacy.

1.3 Prevalent Challenges

EHR based CDSS is an evolving technology with the potential for wide applicability to improve patient outcomes and healthcare resource utilization. However, the development of CDSSs using multimodal healthcare data poses significant challenges, which vary depending on how closely the CDSS is tied to what the clinician already intends to do. For example, suppose the clinical task is the identification of lesions in a medical image that indicate the presence of a disease. In that case,

the availability or creation of large, annotated data corpora is an essential burden, requiring significant manual effort from domain experts. Suppose the task is disease code prediction from diagnostic reports. In that case, processing and analyzing the unstructured nature of the medical text and then dealing with the extensive medical jargon is a challenging NLP task. The unstructured nature of textual medical data and images presents many challenges to exploiting their full potential in supporting intelligent clinical decision making. It is necessary to overcome these challenges in order for CDSS to be used in real-time clinical applications. Listed below are a few challenges that need to be addressed while developing an AI-based CDSS built on multimodal healthcare data.

- **Data quality:** Healthcare data is often acquired from several sources and annotated by physicians with varied degrees of experience; AI-based CDSS should be robust enough to manage any discrepancies or outliers caused by these differences in data quality and expert opinion. Along with system quality, data quality is the crucial component of an AI-based CDSS.
- **Data imbalance:** Healthcare data are often extremely skewed in terms of class distribution, with thousands of normal instances but just a few tens of abnormal cases. AI-based CDSS must address such class imbalance problems.
- **Domain knowledge requirement:** The systems need to be trained to understand physiology and learn the domain knowledge latent in the multimodal data, to enable useful recommendations.
- **Margin for prediction errors:** The negative consequences associated with AI-based CDSS can be severe, specifically for disease detection and treatment decision making systems. The acceptable error rate for an AI-based CDSS is task-dependent. For example, it is tolerable to miss a few pixels while segmenting an organ in an input image, but even a single false negative instance might be devastating when dealing with life-threatening disease detection tasks.
- **Natural language complexity:** The CDSSs that make use of textual input data must be capable of recognizing characters/words, comprehending intent, and drawing inferences. Additionally, the system must be trained to understand abbreviations, acronyms, shorthand terms, and misspelt words that are pervasive in healthcare practice.

- **Dealing with medical images:** Most medical image modalities have a monotonous, monochromatic background, and CDSS that are built on such input images must thus be capable of identifying even the tiniest abnormalities, particularly for early detection systems.
- **Varied processing requirements:** Multimodal healthcare data is often collected from a variety of sources (e.g., reports, notes, and imaging data from multiple devices). Each data modality may need its own preprocessing and AI approaches.
- **Deep neural model design:** Choosing suitable hyperparameters (for example, loss function, activation function, and layer count) from the wide range of possibilities that are best suited for the intended medical task and the input data.
- **Interpretability:** Inability to provide a comprehensive explanation of the projected outcomes may cause skepticism among healthcare professionals. Machine learning and deep learning based CDSSs are considered black boxes, and in domains like healthcare, it is often required to provide information on how and why the system recommended a certain outcome. Thus, interpretability is already a crucial aspect in predictive analytics based CDSSs.

1.4 Prevalent Challenges specific to Indian Scenario

India, the world's second most populous country, has made significant progress towards the establishment of nation-wide healthcare systems for its citizens. Infectious diseases like Smallpox and Polio have been eradicated by successful large-scale public awareness and education efforts, boosting the life expectancy of inhabitants (Reddy *et al.*, 2011). However, India's health outcomes remain inadequate; the burden of preventable disease remains a big concern. The Indian healthcare system is diversified, consisting of a huge number of hospitals of varying sizes and units administered by the state and central governments (Gopal, 2019). In this system, patient care is mainly delivered through primary/community healthcare centre (PHC/CHC), secondary healthcare centre (district hospital), and tertiary healthcare centre (Bagchi, 2008). The private sector provides more than three-quarters of healthcare services, with roughly 80% of them being modest with fewer than 50 beds. The government's supervision is restricted since a large portion of

healthcare providers are in the informal sector. As a result, it is unlikely that any legislative attempts to mandate EHR adoption will succeed outside the parliament (Ramaswamy *et al.*, 2022).

Over the past few years, several major initiatives have been introduced to address these lacunae. The Government of India has undertaken the ambitious Ayushmaan Bharat scheme under which unilateral health covergae for over 50 crore people will be rolled out. The adoption of technology to encourage hospitals to adopt digital EHRs also will help in streamlining patient data management processes. It also offers insurance coverage for patients, but the approach is severely limited in that it only covers In-Patient episodes of specific patients in hospitals that have been approved. It disregards the remainder of the patient's journey, including wellness, outpatient consultations, rehabilitation, follow-up, and recovery (Ramaswamy *et al.*, 2022). Through the digital India initiative under the AI task force (India, 2020), the government has initiated to invest in the development of AI based healthcare systems. The primary challenges identified by this initiative are the development of an EHR repository and training of healthcare stakeholders with newer technologies. Currently, the majority of a patient's imaging data is stored digitally. There is an abundant opportunity for multidisciplinary academic and industry collaborations in order to build a conducive digital healthcare ecosystem. The digitization of a country's whole healthcare system and the adoption of EHR can contribute significantly to its economics, growth, and even industrialization.

1.5 Summary

Intelligent systems capable of modelling, analyzing, and learning inherent patterns from multimodal patient data has evolved to be a crucial requirement in modern healthcare scenarios. This chapter explored the need for automated CDSS, and the challenges and issues associated with the process of developing CDSS that utilise multimodal healthcare data. Addressing unstructured data management issues in healthcare for effective management and utilization of rich patient data is the need of the day. By enumerating the prevalent challenges affecting the development of intelligent, non-knowledge based CDSS that utilize multimodal data, a significant opportunity for designing and developing a comprehensive framework for overcoming the challenges resulting due to the high volume and variety of healthcare data were identified. These aspects can then be utilized to provide actionable insights for enhancing the process of healthcare delivery.

1.6 Thesis Organization

The rest of this thesis is organized as follows.

- In Chapter 2, an extensive literature review on CDSS in healthcare and the observed research gaps are elucidated.
- In Chapter 3, based on outcomes and gaps learned from the existing literature, the research problem addressed in this thesis is formally defined. The scope of this research and a brief description of the proposed methodologies are also provided in Chapter 3.
- Chapters 4 and 5 cover the details of the proposed approaches for patient-specific predictive analytics using textual healthcare data.
- In Chapters 6 and 7, multi-task CDSS built on diagnostic imaging based healthcare data are discussed in detail.
- Chapter 8 discusses the proposed approaches for AI-based CDSS using multimodal healthcare data.
- Chapter 9 presents concluding remarks on the extensive research work carried out and prospects of future research in the area.

Chapter 2

Literature Review

EHRs have emerged as a fundamental requirement for enabling modern healthcare management systems and improving the quality and efficiency of healthcare delivery. Over the last decade, adoption of EHR systems in hospitals has increased dramatically in developed countries, owing to legislation such as the US Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009, which provided hospitals and providers with $30 billion in incentives to adopt EHR systems (Kruse *et al.*, 2018). EHRs are useful data sources for comparative effectiveness research, efficiency improvement, and cost reduction (Cowie *et al.*, 2017). Advances in new imaging techniques and the availability of large volumes of diagnostic medical multimodal data have opened a new era of automated clinical decision-making systems. This thesis focuses primarily on the development of CDSS using unstructured multimodal EHR data. To further comprehend the research gaps, a detailed study of AI-based CDSSs was conducted, encompassing the various data modalities that can be used to capture patient information – textual, imaging, and multimodal healthcare data.

2.1 Clinical Textual Data-driven CDSSs

EHR systems store data associated with each patient, including structured textual data like demographic information, diagnoses, and laboratory tests, as well as unstructured textual data such as prescriptions, radiological notes, clinical notes, and discharge summaries (Birkhead *et al.*, 2015). AI-based CDSS are non-knowledge based intelligent systems that have garnered significant research interest for better understanding disease dynamics and enabling personalised care. Many works have incorporated machine learning and deep learning models (refer Fig. 2.2) for various clinical analytics tasks such as inpatient mortality/sepsis prediction, hospital

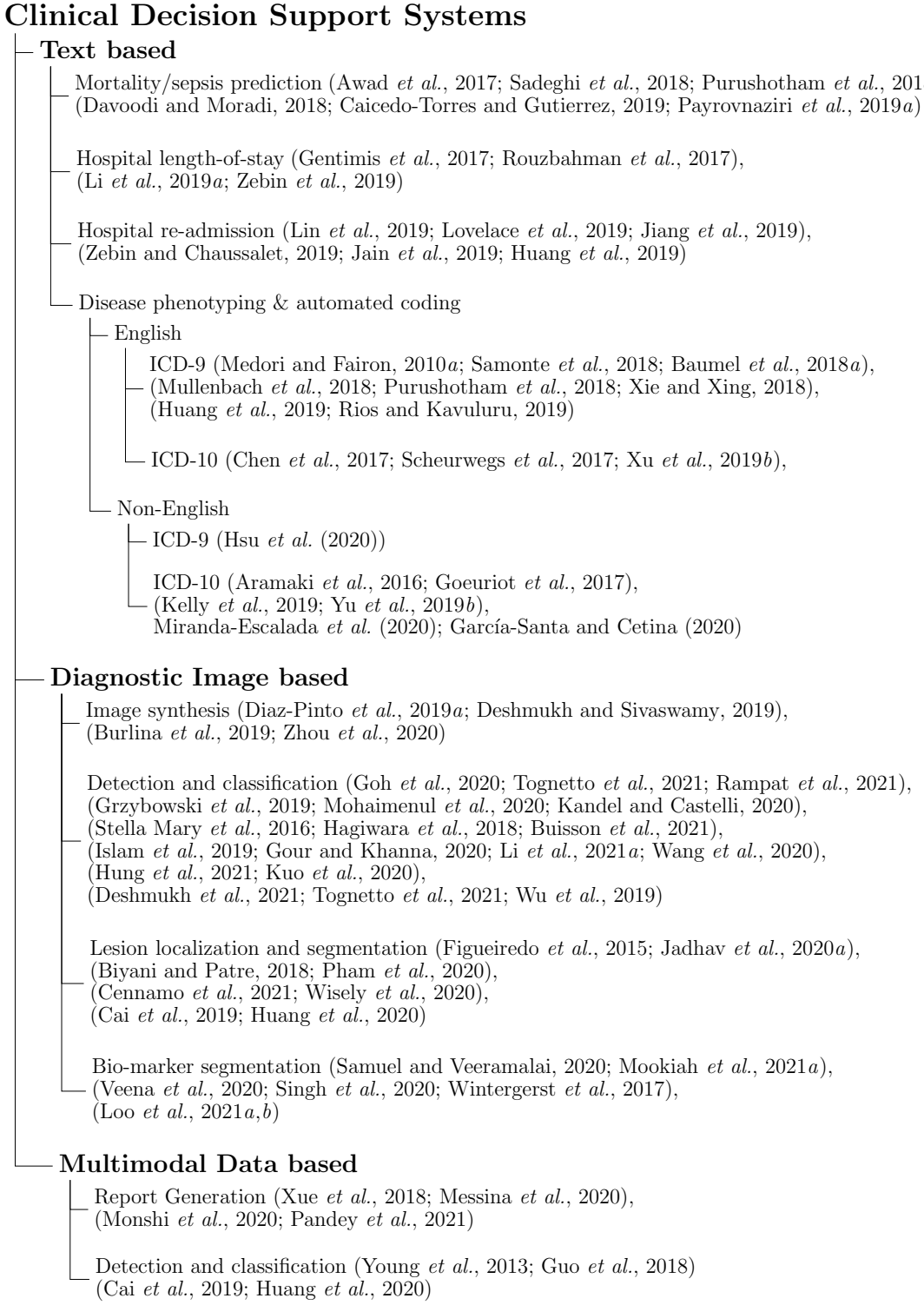


Figure 2.1: Categorization of clinical decision support systems

length-of-stay prediction (Gentimis *et al.*, 2017), disease phenotyping (Rashidian *et al.*, 2018), automated coding (Zeng *et al.*, 2019*a*). Fig. 2.1 presents state of the art in these fields, each of which will be discussed in more detail in subsequent sections.

2.1.1 Inpatient Mortality/Sepsis Prediction

Multiple factors lead to mortality/sepsis in patients admitted to the Intensive Care Units (ICU) of hospitals, making early diagnosis challenging for physicians. Advanced assessment of patients' mortality risks in ICUs is thus of great importance. Several severity scores are already in use as mortality risk estimation tools, which were developed based on a large sample of medical and surgical patients. APACHE (Acute Physiology and Chronic Health Evaluation) (Knaus *et al.*, 1985; Zimmerman *et al.*, 2006), SAPS (Simplified Acute Physiology Score) (Le Gall *et al.*, 1984; Moreno *et al.*, 2005), and SOFA (Sequential Organ Failure Assessment) (Vincent *et al.*, 1996) are some parametric scoring systems used for mortality risk estimation. Recently, machine learning and deep learning based models have been proposed to predict mortality. Most of these existing methods use either single structured data (recorded during admission) or time series structured data (recorded at various stages after admission).

Davoodi and Moradi (2018) proposed a deep fuzzy rule-based system that employed a modified supervised fuzzy k-prototype clustering algorithm for fuzzy rule generation. They initially generated fuzzy pre-clusters from the training data. Then, deep learning models with fully connected layers were used to learn the fuzzy rules in every layer. However, a major drawback of this method was that the fuzzy matrix needed to be generated during the testing phase too, which is a time-consuming task. Darabi *et al.* (2018) experimented with CNN and gradient boosted trees on demographic factors along with the procedure and diagnostic codes for disease code prediction based on structured data. However, a significant limitation of this work is its dependency on the availability of expert-assigned diagnostic codes for all ICU patients. Sadeghi *et al.* (2018) made use of heart rate signals obtained during the first hour after ICU admission and experimented with decision tree algorithms, their focus being only patients admitted to CCUs (Coronary Care Units). Suresh *et al.* (2018) proposed a method that used the 24-hour time series data of vital signs. The authors experimented with the LSTM autoencoder to learn embeddings from time series data and a supervised learning method that performs multiple tasks, including mortality prediction. However,

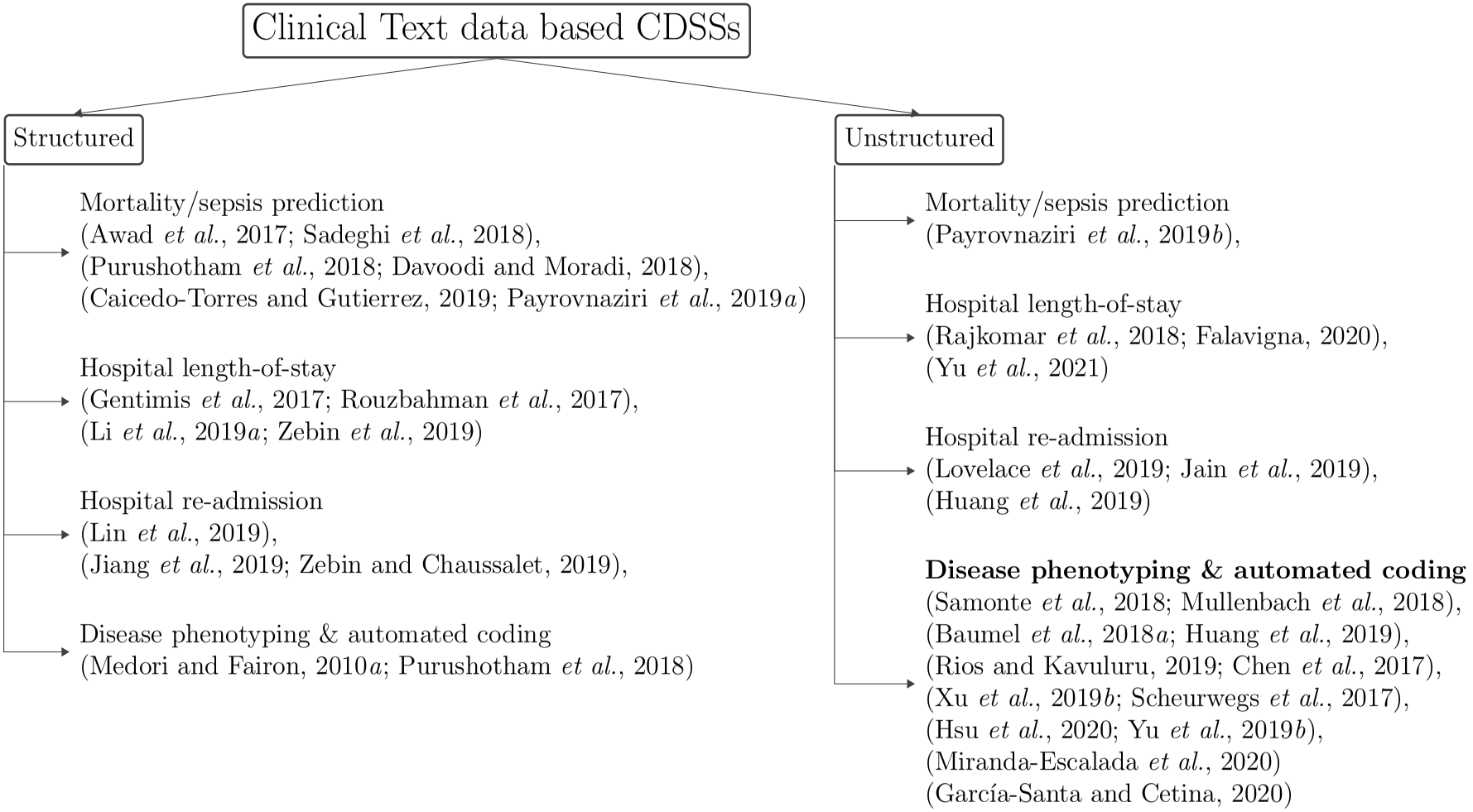


Figure 2.2: Summary of CDS systems built on clinical textual data

interpretation of results was difficult due to the ensembling of multiple stages.

Caicedo-Torres and Gutierrez (2019) used ConvNets to experiment with 22 different concepts (including vital signs and demographic features) and provided evidence support through data visualization. Their model uses only structured data, and there is scope for further experimentation with unstructured data. Payrovnaziri *et al.* (2019*a*) used both unstructured (discharge summaries) and structured patient data for performing myocardial infarction based mortality prediction. However, they used a very small corpus (around 5K) and non-standard patient data, making it significantly difficult to adapt or benchmark their work in real-time clinical setups. A study by Awad *et al.* (2017) revealed ample scope for the development of robust systems that can make use of large critical care environment databases. Table 2.1 summarizes the aforementioned state-of-the-art mortality risk prediction techniques discussed.

Table 2.1: Summary of mortality/sepsis prediction systems

Author	**Methodology**	**Remarks**
Davoodi and Moradi (2018)	Proposed a Deep Fuzzy Rule based System that uses hierarchical fuzzy classifiers.	Training requires pre-clustering, which increases the execution time during testing. Also, a partition matrix needs to be computed, resulting in longer prediction time.
Darabi *et al.* (2018)	Proposed CNN and gradient boosted trees on demographic factors along with procedure and diagnostic codes.	The codes for ICU patients must be assigned by expert clinicians.
Sadeghi *et al.* (2018)	More than 12 statistical and signal-based features extracted from the patient's heart signal within the first hour after ICU admission were given for eight ML classifiers.	Focused only on coronary care unit (CCU) patients. The combination of numerous vital signs in addition to the heart rate signal can help obtain a better understanding of the cause of mortality.

continued ...

... continued

Author	Methodology	Remarks
Scherpf *et al.* (2019)	Used demographic and laboratory measurements at different intervals and RNN for classification.	Maximum achieved is 47.0%, so there is scope for improvement.
Suresh *et al.* (2018)	Used 24-hour time series vital signs data and two model approaches.	Result interpretation and adaption to a real-time system is difficult due to multiple stages.
Caicedo-Torres and Gutierrez (2019)	Used ConvNets on 22 different concepts (vital sign and demographic features) and includes visualization of data.	Prediction is done for 48 hour vital sign structured data details. There is scope for improvization (AUC 0.870396).
Payrovnaziri *et al.* (2019*a*)	Used both unstructured (discharge summary) and structured data.	Only Myocardial Infarction mortality prediction was undertaken. The process of structured data generation is unclear, hence benchmarking is not possible.

2.1.2 Hospital Length-of-stay Prediction

The term "*Hospital length-of-stay (LoS)*" typically refers to the number of days that an inpatient may have to stay in a healthcare facility during a single admission. An accurate prediction of the length of stay in advance can assist hospital administrators and clinicians to manage available medical (staff, devices, etc.) and non-medical resources (beds, rooms, etc.) efficiently and provide cost-effective healthcare services. It has been reported that US hospital stays cost the health system at least $377.5 billion per year (Torio and Moore, 2016). Ineffective resource management has significant repercussions for the hospital, as patient billing is based on procedures undergone rather than the number of days stayed, as per recent legislation in the US. Hence, quantifying the LoS risk of patients so that those with high LoS can be recommended an optimized treatment plan to lower their chances of hospital-induced infections is a critical requirement.

Towards this objective, Gentimis *et al.* (2017) used patient demographic factors

(age, ethnicity) and procedure and diagnosis codes to predict hospital stays (short, medium, and long), achieving around 80% accuracy. Though they reported good performance, assigning diagnostic and procedural codes requires medical expertise and is thus more challenging. Rouzbahman *et al.* (2017) used linear and logistic regression on K-Means clustered patient record features such as medication, labs, charts, demographics, ICD codes, and ICU stay tables. They achieved 74.1% accuracy, but their method was time-consuming as each test data was matched with all 99 clustering scenarios with the corresponding classifier and then averaged to find the LoS. Based on patient demographics, vital signs (blood pressure, temperature, heart rate) during admission, and diagnosis codes (17 groups of ICD 9 codes), Zebin *et al.* (2019) predicted short (<7 days) and long (>7 days) hospital stays. They used an auto-encoder with dense layers for prediction and achieved 77% accuracy in assigning diagnosis codes. Their methods use a feature selection technique along with deep learning, making it difficult to interpret the results.

Li *et al.* (2019*a*) used 49 of 72 patient features, including demographic information, ICU information, surgical information, drug information, and laboratory parameters, from over 1200 patients. These features were chosen by experienced doctors and trained nurses. The authors used Box-Cox and anti Box-Cox transformations to remove the LoS skews and a LASSO regression model with 10 fold cross validation to predict the LoS. However, the size of the data considered by them was too small, and the results obtained were insufficient for adaption in a real hospital scenario. Large-scale experiments with standard benchmark datasets can present important insights into how healthcare big data in both structured and unstructured formats can help support active decision making. Table 2.2 details the existing techniques for the prediction of the hospital length of stay of patients.

Table 2.2: Summary of hospital length of stay estimation CDSS.

Author	**Methodology**	**Remarks**
Gentimis *et al.* (2017)	Used patient demographic factors like age, ethnicity, and procedure and diagnosis codes.	Achieves around 80% accuracy. Assigning procedure and diagnosis codes itself requires medical expertise.

continued ...

... continued

Author	Methodology	Remarks
Rouzbahman *et al.* (2017)	Used medication, lab, chart, demographics, ICD codes, and ICU stay tables and linear and logistic regression on k-Means clustered features.	Achieves 74.1% accuracy, but it is a time consuming process as each test data set was matched with all 99 clustering scenarios with the corresponding classifier and then averaged to find the LoS.
Zebin *et al.* (2019)	Used patient demographic factors, vital signs (blood pressure, temperature, heart rate) during admission and diagnosis codes (17 groups of ICD-9 codes) to predict short (< 7 days) and long hospital stays (> 7 days).	Achieves around 77% accuracy in assigning diagnosis codes. But the assignment itself requires medical expertise.
Li *et al.* (2019*a*)	About 9 out of 72 features that include demographic, ICU data, surgical, drug and laboratory parameters were selected by experienced doctors and trained nurses for 1,200 patients.	The data considered is too small to be adapted for a real hospital scenario.

2.1.3 Hospital Re-admission Prediction

Unplanned re-admission of a hospitalized patient is often a reason for patients' exposure to hospital-induced infections. It can result in difficulty in planning optimal consumption of available medical resources. Conversely, premature hospital discharge may potentially expose patients to relapse, leading to avoidable extreme health deterioration. Identifying patients with higher risk rates plays a key role in reducing hospital re-admissions. An efficient CDSS can have a significant impact by assisting hospitals and physicians by identifying patients with high re-admission probabilities and thereby preventing inadvertent, early, and potentially life-threatening discharges from hospitals. Most of the existing literature is

built on structured data for predicting hospital re-admissions.

For assessing hospital re-admission scenarios, Lin *et al.* (2019) used chart events, ICD-9 embeddings, demographic information of the patients, and a combination of LSTM and CNN model features to achieve an AUROC score of 0.791. Lovelace *et al.* (2019) used a CNN trained on structured patient data represented as a StarSpace embedding to predict 30-day re-admission. They achieved an AUROC of 0.71. Zebin and Chaussalet (2019) employed the same dataset and methodology proposed by Lin *et al.* (2019) and improved the performance to an AUROC score of 0.821. They did not provide any in-depth details or analysis as to what modifications were made and what resulted in the improvement over Lin *et al.* (2019)'s model.

Huang *et al.* (2019) used Bidirectional Encoder Representations from Transformers (BERT) to create embeddings of clinical notes and discharge summaries, then summed the last few layers of probabilities to predict hospital re-admissions. They achieved an AUROC of 0.760, but summing the probabilities may not be suitable for capturing long sequence word relations, and hence there is significant scope for improvement. Jain *et al.* (2019) employed an LSTM model with attention and reported that the usage of attention improves the prediction performance. The authors highlight the difficulty in building an interpretive model with the observed attention weights. Table 2.3 lists existing techniques for designing hospital re-admission prediction systems.

Table 2.3: Summary of hospital re-admission prediction CDSS.

Author	**Methodology / Pros**	**Limitation**
Lin *et al.* (2019)	Used chart events, ICD-9 embeddings, and demographic information of the patients, and a combination of LSTM and CNN model features.	Achieves AUROC of 0.791, there is scope for improvement.
Lovelace *et al.* (2019)	Proposed CNN with StartSpace embedding for clinical notes to predict 30-day re-admission.	Achieves AUROC of 0.71, but there is still scope for improvement.

continued ...

... continued

Author	Methodology	Remarks
Jiang *et al.* (2019)	Used ICD-9 codes for readmission prediction and proposed an interpretable regularizer for logistic regression.	Achieves 0.71 AUROC with a proposed regularizer. But assigning ICD9 codes needs expertise and domain knowledge.
Zebin and Chaussalet (2019)	Used same dataset and methodology as in Lin *et al.* (2019).	Claims to achieve 0.821 AUROC, but it is not clear about the modifications done as compared to Lin *et al.* (2019) to achieve the improvement in accuracy.
Huang *et al.* (2019)	Employed Bidirectional Encoder Representations from Transformers (BERT) to train on clinical notes and discharge summaries.	Achieved AUROC of 0.760, but summing the probabilities may not be suitable for capturing long sequence word relations.
Jain *et al.* (2019)	Proposed LSTM with attention network.	Achieved AUROC of 0.71, but there is still scope for improvement.

2.1.4 Disease Phenotyping and Automated Coding

A significant part of the daily workflow tasks of clinicians is spent on EHR reviews to predict whether the patient has a medical condition. Phenotyping involves finding patients with specific conditions or outcomes. It is primarily concerned with forecasting whether a patient has a medical condition or identifying those at risk of developing one. ICD (International Classification of Diseases) is a widely used nomenclature for disease phenotyping in clinical care and research. The majority of the available literature employs the ninth and tenth revisions of the ICD, i.e., ICD-9 and ICD-10 codes for phenotyping. Automated phenotyping is mainly a multi-label classification task that classifies an EHR record based on the output ICD codes. Automated phenotyping can reduce the time spent on EHR reviews by clinicians.

Automated diagnostic coding of patient records has been a field of active and substantial study since the 1990s. Several significant contributions to solving

the automated ICD code assignment task have emerged ever since. These works can be broadly classified into – *(a)* rule-based systems (Mykowiecka *et al.*, 2009; Farkas and Szarvas, 2008), *(b)* primitive learning-based systems involving Bayesian classifiers, nearest neighbors, and relevance feedback (Pakhomov *et al.*, 2006*a*; Kavuluru *et al.*, 2015), and *(c)* explainable intelligent systems (Domingues *et al.*, 2019; Mullenbach *et al.*, 2018). The works could be categorized on the basis of ICD-9 or ICD-10 codes, depending on the predicted revision codes. Due to the ICD-9 coding system's wide acceptance status among current clinical datasets and hospitals alike, most existing works (Purushotham *et al.*, 2018; Huang *et al.*, 2019; Mullenbach *et al.*, 2018) have reported their performance on ICD-9 code assignment. However, with the recent shift towards ICD-10 coding, certain works (Xu *et al.*, 2019*b*; Wang *et al.*, 2020*b*) employed the much convoluted ICD-10 coding taxonomy.

With the latest advancements and success in deep neural modelling, ConvNets have been widely utilized to facilitate the classification of various free-text documents (Liu *et al.*, 2018), including voluminous unstructured healthcare records Si and Roberts (2019). Researchers have recently studied the significance of ConvNet-based methods for automated diagnostic code assignment based on free-text critical care discharge summaries (Huang *et al.*, 2019; Baumel *et al.*, 2018*b*; Mullenbach *et al.*, 2018; Teng *et al.*, 2020; Ji *et al.*, 2020; Vu *et al.*, 2020). In critical healthcare applications such as CDSSs, trust is rooted in more than just their performance; such systems also need to justify and explain their actions based on the principles that present the dynamics of the concerned domain. In an attempt to develop explainable intelligent systems, researchers aim to combine neural models such as ConvNets and recurrent networks with an attention mechanism (Baumel *et al.*, 2018*b*; Mullenbach *et al.*, 2018). Baumel *et al.* (2018*b*) proposed a hierarchical neural attention model to discern relevant portions of a given free-text document that corresponded to a specific ICD-9 code label, based on which a deep neural gated recurrent unit (GRU) was trained to enable the clinical task of automated coding.

Mullenbach *et al.* (2018) proposed a convolutional attention network to facilitate multi-label classification of ICD-9 codes, advancing the field of explainable predictive systems. The authors benchmarked their prediction performance using $8,921$ unique ICD-9 codes, including $6,918$ diagnostic codes and $2,003$ procedural codes. To encode the hierarchy of ICD-9 codes and facilitate diagnostic coding, Xie and Xing (2018) utilized LSTM networks with attention on the *diagnosis description* portion of the discharge summaries. Huang *et al.* (2019) evaluated and

benchmarked the performance of several existing deep neural models, including feed-forward neural networks, ConvNets, LSTMs, and GRUs, on patient discharge summaries, for the clinical prediction task of ICD-9 coding. Additionally, the authors also benchmarked their performance using traditional machine learning classifiers, including logistic regression and random forest. Table 2.4 summarizes the recent advancements in phenotyping/automated coding models based on EHRs.

Table 2.4: Summary of CDSS for phenotype/ICD code assignment

Author	**Methodology**	**Remarks**
Baumel *et al.* (2018*b*)	Used hierarchical GRU (HA-GRU) to predict all ICD-9 codes.	Achieves accuracy of 0.4072 for all codes.
Catling *et al.* (2018)	Extracted only history of presenting illness section from discharge summary text.	Achieves 0.686 F1-score, there is scope for improvement.
Rashidian *et al.* (2018)	Used demographics, lab results, and medications using machine learning and feed forward neural network.	Achieved 0.80 F1-score for two label coding system.
Huang *et al.* (2019)	Used whole discharge summaries and benchmarks LSTM, CNN, and GRU models with top10 and top50 codes.	Achieved maximum of 0.7090 for top 10 and 0.33 for top 50 codes.
Zeng *et al.* (2019*a*)	Used learning done for an automatic MeSH indexing task for code prediction.	Achieved maximum of 0.42 micro F1-score

While most existing works have concentrated on English based clinical input, very few works employ non-English textual inputs (Refer Fig. 2.1). Miranda-Escalada *et al.* (2020) promoted the development and evaluation of medical coding systems for medical documents in Spanish. The authors discuss a wide variety of methods used for preprocessing and modelling clinical text for enabling ICD-10 assignment, specifically, pre-trained language models and word embeddings (BERT, BETO, FastText, etc.). Supervised classification models like SVM, random forests, and logistic regression were benchmarked against deep neural models

such as LSTM, CNN, and BERT by Polignano *et al.* (2020). The authors combined the BERT pre-trained embeddings with a BiLSTM, CNN, and self-attention-based classifier and reported a F1-score of 0.308. Moons and Moens (2020) experimented with variants of CAML, which achieved a 0.76 F1-score. A dictionary-based approach was proposed by Cossin and Jouhet (2020). However, their algorithm failed to detect a term if its words were not in the right order or if they were nonadjacent.

Schäfer and Friedrich (2020) compared state-of-the-art transformer-based models, such as BioBERT and ClinicalBERT, with the XLNet BERT model. They augmented the data with documents from the MIMIC-III database (Medical Information Mart for Intensive Care) and achieved a 0.432 F1-score for the top-100 most frequent ICD codes. Rishivardhan *et al.* (2020) experimented with BERT, RoBERTa, Electra, and XLNet transformer models and reported the highest 0.33 F1-score with the Electra BERT variant. Costa *et al.* (2020) presented two approaches for code extraction based on conditional random fields (CRFs) and a pre-trained BETO model, fine-tuned on a Named Entity Recognition(NER) task for ICD code prediction. The authors evaluated the models by considering the classification task as a NER task, i.e., all presented metrics are on a token-by-token basis. Eslami *et al.* (2020) proposed a semi-hierarchical multi-label classification approach using pre-trained multilingual BERT and achieved a 0.21 F1-score. Blanco *et al.* (2020) proposed similarity match based algorithms that make use of pre-trained BERT embedding, and reported a 0.09 F1-score for the prediction of ICD diagnostic codes.

2.2 Diagnostic Imaging Data-driven CDSSs

Early diagnosis is crucial to improving treatment strategies and reducing the risks associated with the disease prognosis. Medical imaging data like ultrasound, magnetic resonance imaging, mammography, computed tomography, positron emission tomography, retinal photography, slit-lamp biomicroscopy, histopathology slides, X-ray images, dermoscopy images, etc., play a crucial role in the early detection, diagnosis, and treatment of diseases. Presently, most imaging modalities are manually interpreted by expert clinicians and trained experts for clinical diagnosis. With the exponential increase in the volume of chronic patients, this process of manual inspection and interpretation increases the cognitive and diagnostic burden on these healthcare professionals. As per a report of the Institute of Medicine at the National Academies of Science, Engineering, and Medicine, diagnostic errors contribute to approximately 10% of patient deaths (Institute of Medicine,

2015). It may not be solely due to physician performance. A lack of healthcare workflow management systems with adequate support for automated management of diagnostic images also plays a significant role.

Recently, machine learning and DL techniques have been utilized for designing computer based analysis systems for medical images. Automatic classification, localization of normal anatomy, detection, segmentation, and registration are some of the common tasks explored using AI techniques. Ophthalmology, pathology, radiology, and oncology are a few fields where AI-based techniques have been successfully leveraged for interpreting imaging data. Ophthalmology is the first field to be revolutionized and most explored in the healthcare domain. The imaging data captured during routine visits play a crucial role in disease diagnosis and treatment approaches. Additionally, the patient data collection procedure used in ophthalmology is not associated with significant adverse effects or high risks. Thus, a focused study was carried out to understand the scope of existing work in the ophthalmology field.

2.2.1 Imaging Modalities in Ophthalmology

Numerous AI-based CDDS have been proposed for monitoring, diagnosing, and treatment planning with respect to ocular diseases using imaging modalities such as ultra-wide-field fluorescein angiography (UWFA), optical coherence tomography (OCT), and optical coherence tomography angiography (OCTA), retinal fundoscopy, slit-lamp biomicroscopy, confocal microscopy, and fluorescein angiography (FA) among others. Each of these modalities is extensively used in hospitals for pre-screening, prognosis mapping, etc.

1. *Retinal Fundoscopy:* The fundus image is a frequently used ophthalmic imaging method in which optical cameras create magnified images of the retinal tissues; these retinal images are useful for monitoring, diagnosing, and planning therapy for various ocular problems.

2. *Slit-lamp Biomicroscopy:* The slit-lamp images are acquired using a high-intensity light source apparatus, which enables examination of both the anterior and posterior segments of the eye. It is primarily used to illuminate a large portion of the eye during routine ophthalmic exams, mainly when the thickness of the corneal layers is a concern for diagnosing or considering surgery.

3. *Confocal Microscopy:* Confocal microscopy is a non-invasive technique that

enables the in vivo assessment of structural alterations at the cellular level in a variety of ocular surface diseases. In vivo confocal microscopy has been used to examine the cornea, Meibomian gland, bulbar and palpebral conjunctiva, and lacrimal gland in the dry eye field.

4. *Fluorescein Angiography:* FA is an invasive diagnostic procedure that examines the blood flow in the retina and choroid using a special dye and camera. The leakage of dye in the later frames of the angiogram is used to diagnose and manage vitreoretinal diseases.

5. *Ultra-wide-field Fluorescein Angiography:* UWFA imaging is a relatively recent technique that includes color and red-free photography, fluorescein angiography, and fundus autofluorescence. In order to get a wider field of view, wide-field and ultra-wide-field imaging have been developed.

6. *Scanning Laser Ophthalmoscopy:* SLO imaging employs confocal laser scanning microscopy for diagnostic imaging of the retina and cornea. It aids in the diagnosis of glaucoma, AMD, and other retinal degenerative conditions.

7. *Optical Coherence Tomography:* OCT is a non-contact, non-invasive optical image-based diagnostic method that offers extensive information on the morphology of the retina and aids in the diagnosis of a variety of macular diseases.

8. *Optical Coherence Tomography Angiography:* OCTA is a novel non-invasive imaging technology that uses motion contrast imaging to get high-resolution volumetric blood flow information. It has a broad range of potential applications in the diagnosis and treatment of retinal vascular disorders.

Among these imaging modalities, more than 57% constitutes that of fundoscopy, 19% of OCT/OCTA, 5% of confocal microscopy, 5% of SLO, 4% of FA, 1% of slit-lamp datasets, and 8% of other modalities (external eye, movies, etc.) are available in the public domain (Khan *et al.*, 2021). In this thesis, fundoscopy and slit-lamp biomicroscopy images were considered due to the simplicity of acquisition, low cost, minimal risk, and widespread availability. The slit-lamp images capture the front surface of the eye while retinal fundoscopy captures the back wall of the eye. Thus, diagnostic imaging of these modalities aids in the early diagnosis of most ocular diseases. The sample fundoscopy and slit-lamp biomicroscopy images are depicted in Fig. 2.3.

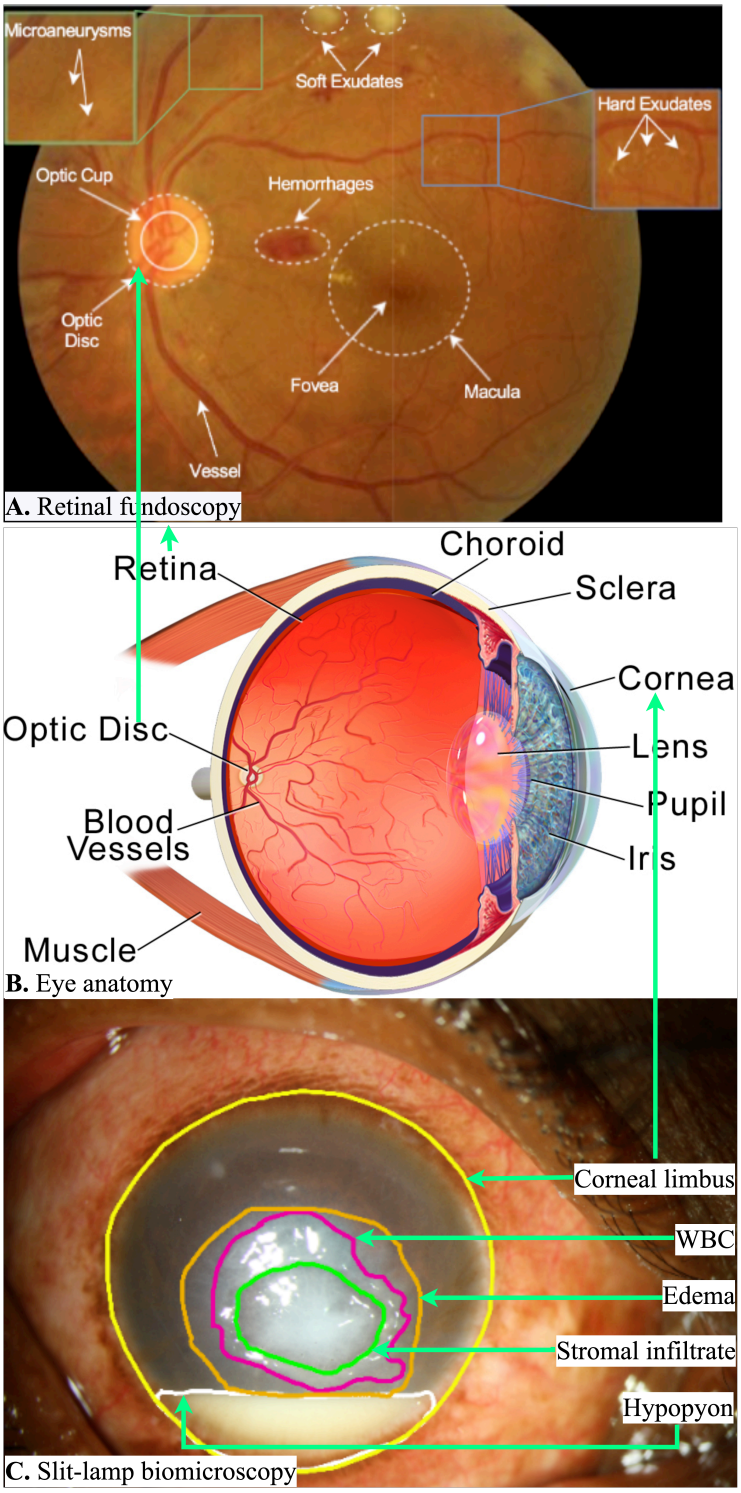


Figure 2.3: Sample retinal fundoscopy (Li *et al.*, 2021*b*) and slit-lamp biomicroscopy image (Loo *et al.*, 2021*a*).

2.2.2 CDSSs using Fundoscopy Images

Intelligent systems for retinal disease diagnosis and management have received significant attention over the last decade. Typically, such decision-making systems have been developed for fundoscopy image synthesis, ocular disease detection, and lesion & biomarker segmentation. Few of these works are highlighted in Fig. 2.4 and are discussed in detail in the subsequent subsections.

2.2.2.1 Image Synthesis

For effective learning performance, generalizable deep neural models require a large number of labelled images. Image synthesis may be used to increase the number of fundus images and also enhance the quality of the images. A variety of Generative Adversarial Networks (GAN) based image synthesis techniques are widely used to deal with the limited number of training images. Guibas *et al.* (2017) proposed a two stage hierarchical synthesis process. The first stage utilized deep convolutional GAN (DCGAN) to generate the vessel structure from the input noise vector. These generated vessel structure masks were then given to conditional GAN (CGAN) to generate corresponding fundus images. In their study, authors were able to generate fundus images with simple features like general colour, shape, and lighting. But generated images lacked complicated structures such as the boundaries of the optic disc, minute lesions, and realistic vessel structure. Costa *et al.* (2018) proposed a similar approach, but the networks were jointly trained by combining the loss functions associated with each stage. Encouraging results were observed for the synthesis of initial stage DR (0-2). But, incorrect vessel segmentation masks were generated for the images in the later stages (3-4) of DR.

Zhao *et al.* (2018) proposed Tub-GAN for generating fundus images with a binary mask of a tubular structure. The authors also proposed a variant, Tub-sGAN, that considers the style of the input while generating the output fundus images. Though the authors were able to maintain the same tubular structures and the position of the optic disc in synthesized images, the boundaries were often not as clear as those of the real images. DCGAN was proposed by Diaz-Pinto *et al.* (2019*a*) to synthesize the optic disc (OD) region in order to improve the performance of glaucoma detection systems. Synthetic images were obtained with a well-defined optic disc shape but with a focus on only glaucoma classification. Deshmukh and Sivaswamy (2019) synthesized the Optic Nerve Head (ONH) region in fundus images using a deep learning technique. The authors employed a combination of B-spline registration and GAN to generate high quality images

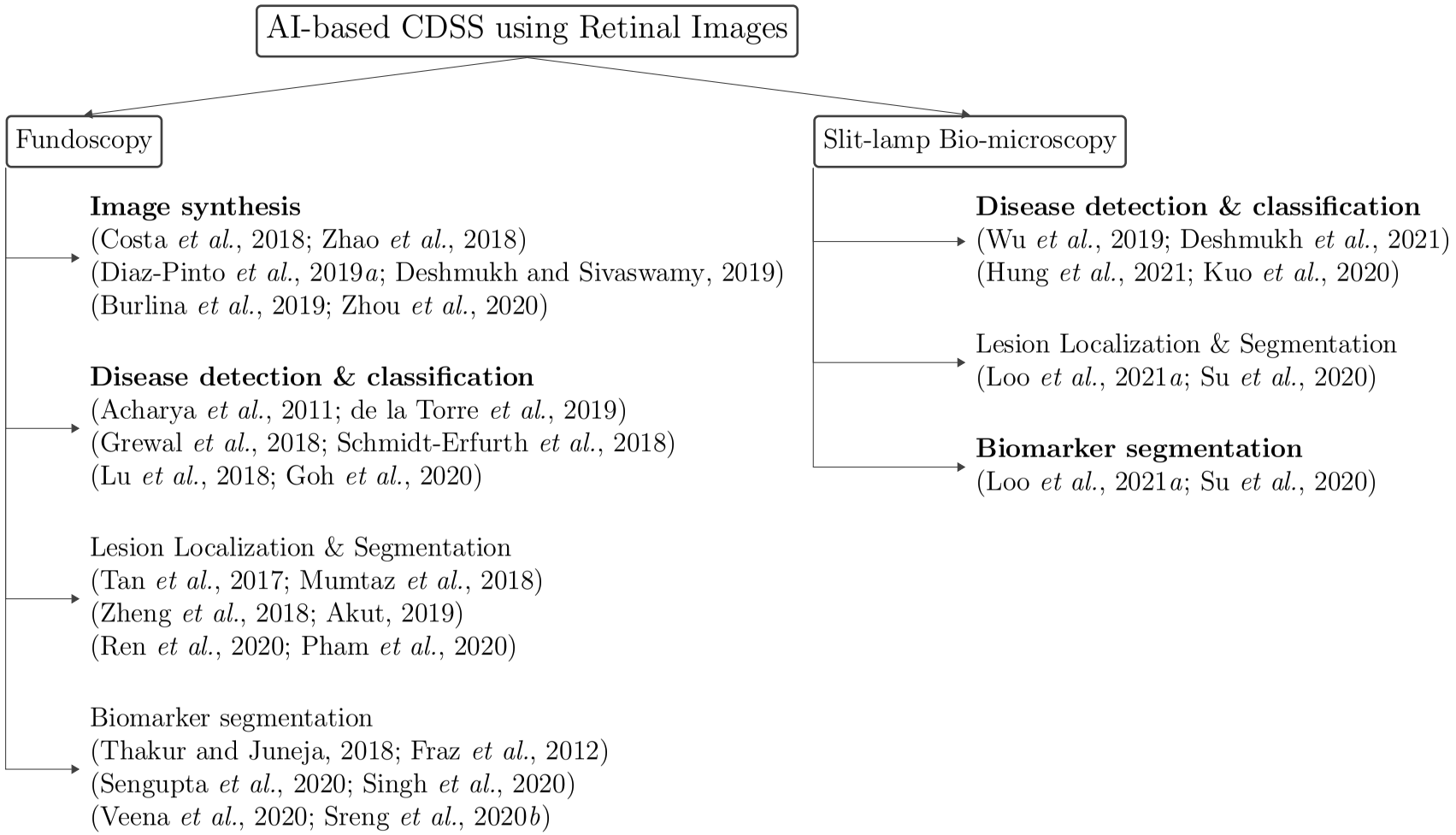


Figure 2.4: Diagnostic imaging based CDSSs

utilizing optic disc (OD), optic cup (OC), and vessels from arbitrarily different fundus images. Zhou *et al.* (2020) proposed a generative adversarial network for diabetic retinopathy (DR-GAN). It is capable of generating high-resolution images from DR grading and vascular or lesion information. The authors improved the DR grading system by using these generated synthetic images. A progressive generative adversarial network (ProGAN) is utilized by Burlina *et al.* (2019) for synthesizing referable and non-referable AMD images. It can be observed that most existing methods generate images for a single ocular condition, although there is potential for generating images for several diseases. Additionally, synthesizing should be generic, requiring just the ocular condition and the number of images to be generated as input. Table 2.5 summarizes the recent advancements in retinal image synthesis.

Table 2.5: Summary of CDSS for retinal image synthesis

Author	**Methodology**	**Remarks**
Guibas *et al.* (2017)	Proposed a two stage GAN network to synthesize retinal fundus images from the input noise vector.	Fundus images with simple features like general colour, shape, and lighting were obtained, but failed to obtain complicated structures such as the boundaries of the optic disc, minute lesions, and realistic vessel structure.
Costa *et al.* (2018)	The retinal vessel mask generation module and the vessel mask to retinal image module were jointly trained by combining the loss functions associated with each task.	Even though the generated images and their associated vessel masks were visually consistent, they lacked realism. Vessel masks often revealed aberrant discontinuities, and a clear distinction between veins and arteries did not seem to exist.

continued ...

. . . continued

Author	Methodology	Remarks
Zhao *et al.* (2018)	Proposed Tub-GAN for generating fundus images with a binary mask of a tubular structure.	Though the same tubular structures and the position of the optic disc in synthesized images were maintained in the synthesized images, the boundaries were often not as clear as those of the real images.
Diaz-Pinto *et al.* (2019*a*)	Proposed DCGAN to synthesize the optic disc (OD) region in order to improve the performance of glaucoma detection systems.	Synthetic images were obtained with a well-defined optic disc shape but with a focus on only glaucoma classification.
Deshmukh and Sivaswamy (2019)	Employed a combination of B-spline registration and GAN to synthesise high quality images utilizing optic disc (OD), optic cup (OC), and vessels from arbitrarily different fundus images.	Achieved 3.6% higher DICE score by using 200 synthetic images supplementing the training set for optic cup segmentation task. Focused only on improving the performance of the glaucoma classifier.
Zhou *et al.* (2020)	Proposed a generative adversarial network for generating high-resolution images from DR grading and vascular, or lesion information.	DR grading system performance improved by using these generated synthetic images.
Burlina *et al.* (2019)	Proposed a progressive generative adversarial network (ProGAN) for synthesizing referable and non-referable AMD images.	Over 90% of the synthetic images were deemed to be of sufficient quality. Focused only on improving the performance of the AMD classifier.

2.2.2.2 Ocular Disease Detection

Chronic ocular diseases (COD) such as myopia, diabetic retinopathy (DR), age-related macular degeneration (AMD), glaucoma, and cataract can affect the ability to see and can often lead to severe vision impairment or blindness if detected late or left untreated. According to WHO estimates, the global population suffering from myopia may exceed 3.3 billion by 2030, while those suffering from AMD, glaucoma, and DR will reach 243.3 million, 95.4 million, and 180.6 million, respectively (WHO, 2019). Thus, regular screening is a significant step towards early diagnosis and better disease prognosis and risk management, especially in the case of vulnerable individuals like the elderly or diabetic patients. Given the ever-increasing patient population every day, manual screening is highly time-consuming, and the treatment capacity is very limited. Thus, it is crucial to develop intelligent computational systems that accommodate these needs by facilitating automated early detection of COD at the patient level.

Most early computer-aided screening methods used digital image-processing based techniques (IPT) (Joshi *et al.*, 2011). Later, supervised machine learning techniques (ML) were developed, which extract features using predefined rules or statistical and structural metrics (Acharya *et al.*, 2011). Over the years, research directions have shifted towards end-to-end, intelligent predictive systems that use the predictive power of deep neural networks, owing to their data-driven-feature learning capabilities. Convolutional neural networks (CNN) have shown promising performance in detecting COD like glaucoma, DR, and AMD, using color fundoscopy images. de la Torre *et al.* (2019) proposed a pixel-wise score propagation for visualization of the prediction system. Though heatmaps can be used to locate lesions in significantly differential images, their method fails to provide an ophthalmologist-level interpretability mechanism for early disease prediction systems. Wang and Yang (2018) proposed a regression activation map for visualization of the key features of retinal fundus images. Pratap and Kokil (2019) extracted pre-trained AlexNet features and classified fundus images to detect cataracts using an SVM classifier. Similarly, residual network features were extracted by Gargeya and Leng (2017) and DR was detected using a gradient boosting classifier. Due to the use of multiple stages, such systems are difficult to interpret. Ting *et al.* (2017) trained three distinct ensemble VGGNet networks to detect DR, AMD, and glaucoma. Usage of multiple networks increases both training and inference time.

Grewal *et al.* (2018) reviewed DL techniques with a focus on the need for

interpretable prediction systems. Schmidt-Erfurth *et al.* (2018) highlighted the need for multi-step algorithms that first detect certain clinically known features and then predict or classify based on those features.

Table 2.6: Summary of diagnostic imaging data based CDSS for disease prediction

Author	**Methodology**	**Remarks**
Wang and Yang (2018)	Proposed a regression activation map for visualization of the key features.	With visualization using heat maps, it is difficult to interpret the prediction in a real-time screening application.
de la Torre *et al.* (2019)	Proposed pixel-wise score propagation for visualization of the prediction system.	Though the heat map could locate lesions in significantly differential images, it fails to provide ophthalmologist level interpretability for early disease prediction systems.
Pratap and Kokil (2019)	Extracted pre-trained AlexNet features and classified fundus images to detect cataracts using SVM	Due to the use of multiple stages, such systems are difficult to interpret.
Gargeya and Leng (2017)	Extracted features using residual network and classified fundus images to detect DR using a gradient boosting	Due to the use of multiple stages, such systems are difficult to interpret.
Ting *et al.* (2017)	Trained three distinct ensemble VGGNets to detect DR, AMD, and glaucoma.	Usage of multiple networks increases both training and inference time.

2.2.2.3 Lesion Localization and Segmentation

In fundoscopy images, microaneurysms (MA), exudates (EX), haemorrhages (HE) and drusen are the commonly occurring lesions in the case of DR patients. These artefacts are marked in a sample fundus image (shown in Fig. 2.3). Accurate early detection and marking of such lesions in the fundus image can help reduce further complications and vision loss. There are several impediments to the segmentation of MA, including the presence of additional lesions with similar color, clarity, background texture, very low contrast, and changes in image illumination. Recently,

CNNs with rectified linear activation function (ReLU) and max pooling have been widely used for MA detection (Tan *et al.*, 2017; Khojasteh *et al.*, 2018; Eftekhari *et al.*, 2019; Akut, 2019). Also, U-Net based CNN networks have also been proposed for MA detection (Chudzik *et al.*, 2018; Kou *et al.*, 2019). Here, the MA mask image/patch is generated using convolution and de-convolution layers for the given preprocessed fundus image/patch. Though several AI-based approaches have been proposed for MA detection and segmentation, a systematic review of existing literature focusing on the diagnostic use of automated MA detection and segmentation for early DR diagnosis has not been explored so far.

Haemorrhages (HE) is one of the most apparent indications of DR. Thus, accurate detection or segmentation of HE is critical for DR diagnosis. HE (along with other lesions) is often rather modest in size, with their pixels accounting for just a small fraction of the whole image, leading to a class imbalance problem. Typically, soft and hard EX are used to diagnose DR. The segmentation barriers include a lack of contrast, a range of sizes, and a resemblance to other lesions. Accurate EX detection is therefore critical for prompt treatment. Several image processing based approaches are proposed for accurately detecting red lesions and segmenting them (Kar and Maity, 2018; Murugan, 2019; Jadhav *et al.*, 2020*b*). Orlando *et al.* (2018) combined features obtained from CNN and domain knowledge contributed by experts for the detection of red lesions. Maqsood *et al.* (2021) proposed a 3D CNN to detect hemorrhages, using the features with a pretrained VGG. Robust features were selected using a multi-logistic regression entropy distribution function, and the selected features were fused and classified using an extreme learning machine. The feature fusion is time-consuming as it involves comparing all possible pairs, and also, the direct interpretation is challenging due to the use of multiple stages. Zheng *et al.* (2018) augmented the minority class images with a conditional generative adversarial network (cGAN) and proposed an ensemble convolutional neural network (MU-net) based on a U-net structure for EX detection. Sudha and Ganeshbabu (2021) applied saliency detection in combination with active contour approximation to segment exudates, microaneurysms, and hemorrhages from digital fundus images.

Lesions like drusen are primarily used to aid in the diagnosis of AMD. The primary challenges associated with drusen segmentation are that their yellowish-white hue is comparable to that of the fundus image and OD; they often exhibit uneven brightness and interference from other biomarkers such as blood vessels. Also, drusen frequently have irregular forms, and their borders may be obscured. Pham *et al.* (2020) devised a multi-scale deep learning model for segmenting drusen

that incorporates both global and local information. Ren *et al.* (2020) proposed a unified deep framework that merged adaptive, collaborative similarity learning with a deep learning model capable of learning discriminative features for drusen segmentation. Though several AI-based systems exist for lesion identification and segmentation, there is still a need for the development of robust approaches capable of detecting/segmenting multiple lesions using a single neural model. Table 2.7 summarizes recent techniques that were adapted lesion localization CDSS using retinal fundoscopy images.

Table 2.7: Summary of imaging data based CDSS for lesion localization

Author	**Methodology**	**Remarks**
Tan *et al.* (2017)	Resized the images to 51 × 51 × 3 and proposed a four layer CNN model to extract MA, EX, and haemorrhages.	Validated primarily on a small single database with 149 images.
Khojasteh *et al.* (2018)	Proposed a patch based probability map generation using four CNN layers with sixteen feature maps.	Involves two phases and also, post-processing was required to detect the DR lesions.
Akut (2019)	Utilized image processing techniques to obtain the MA candidates mask and then applied YOLO to detect MA.	Focused only on MA detection and localization.
Orlando *et al.* (2018)	Lesion candidates were retrieved using a four layer CNN was trained to characterize each red lesion candidate.	63 hand-crafted features were also utilized for the screening system, which makes it difficult to adapt to a real hospital scenario.
Zheng *et al.* (2018)	Proposed a modified path based U-net (MU-net) for exudate detection.	Validated primarily on a small single database with 22 images.

continued ...

. . . continued

Author	Methodology	Remarks
Sudha and Ganeshbabu (2021)	Applied saliency detection in combination with active contours approximation to segment exudates, microaneurysms, and hemorrhages from digital fundus images.	Focused only on improving the performance of the DR classifier.
Pham *et al.* (2020)	Devised a multi-scale deep model for segmenting drusen, incorporates both global and local information.	A Dice score of 0.542 was achieved when validated primarily on a small database with 78 test images.
Ren *et al.* (2020)	Proposed a unified deep framework that merged adaptive collaborative similarity learning for drusen segmentation.	Validated primarily on a small database with 9 test images.

2.2.2.4 Biomarker Segmentation

Several significant biomarkers are seen in the fundus image, including the OD, OC, blood vessels, the macula, and the fovea (please refer to Fig. 2.3). Segmentation of retinal blood vessels is critical for diagnosing a variety of ocular conditions, including diabetic retinopathy and glaucoma. Several AI-based approaches have been proposed for automatic vessel segmentation for the detection of these diseases. Most of these works are covered in recent survey studies Fraz *et al.* (2012); Sengupta *et al.* (2020); Singh *et al.* (2020); Mookiah *et al.* (2021*b*); Chen *et al.* (2021). The authors emphasize that although the techniques have been validated primarily on a single small database, more robust approaches that can be validated on several datasets, including a significant number of images, are the need of the day. Additionally, a comprehensive analysis of the effect of vascular stricture on the predictive performance for ocular disease detection has not been conducted. The cup-to-disc ratio (CDR) is a well established and commonly utilized criterion for diagnosing glaucoma. The analysis of current approaches Thakur and Juneja (2018); Sreng *et al.* (2020*b*); Veena *et al.* (2020) demonstrates that deep learning based techniques for OD & OC segmentation and classification are more accu-

rate than traditional image processing techniques. A recent study by Rudnicka *et al.* (2020) suggests that various regions like the arterioles, venules, etc., are also associated with high-tension open-angle glaucoma. The performance of CNN using only optic disc cropped regions has been experimented with, a systematic experimental evaluation of other preprocessing methods in CNN has yet to be undertaken. Table 2.8 summarizes recent techniques that were adapted biomarker segmentation CDSS using retinal fundoscopy images.

Table 2.8: Summary of imaging data based CDSS for biomarker segmentation

Author	**Methodology**	**Remarks**
Liskowski and Krawiec (2016)	Proposed a patch based approach with 3 CNN layers for vessel segmentation.	Validated primarily on a small single database with $<$ 40 images.
Maji *et al.* (2016)	Proposed and ensemble of 12 CNN networks for vessel segmentation.	Validated primarily on a small single database with $<$ 40 images.
Sengür *et al.* (2017)	Proposed two layer CNN along with a a novel reinforcement learning approach for vessel segmentation. The reinforcement learning enabled a reduction in the training time (fewer epochs).	Validated primarily on a small single database with $<$ 40 images.
Al-Bander *et al.* (2018)	Utilized DenseNet with fully convolutional network for optic disc and cup segmentation.	Used the OD centre provided in ground truth data for RoI segmentation, which makes it difficult to adapt to a real hospital scenario.
Yu *et al.* (2019*a*)	Utilized residual U-Net for optic disc and cup segmentation.	Validated on multiple datasets. The performance dropped for images with severe disc atrophy, especially seen in pathological myopia retinal images.

2.2.3 CDSSs using Slit-lamp Bio-microscopy Images

The cornea is a transparent layer of tissue covering the front surface of the eye that acts as a window, allowing light to enter the eye. Slit-lamp examination of the ocular surface, particularly the cornea, conjunctiva, and anterior chamber, is widely used in the diagnosis of cataract (Wu *et al.*, 2019; Deshmukh *et al.*, 2021; Tognetto *et al.*, 2021) and corneal infections and tumors commonly termed as Keratitis (Hung *et al.*, 2021; Kuo *et al.*, 2020). For cataract grading, Li *et al.* (2020) utilized Faster R-CNN (Ren *et al.*, 2015) for RoI segmentation and then used transfer learning with ResNet (He *et al.*, 2016). Xu *et al.* (2019*a*) localized the nuclear region using Faster R-CNN (Ren *et al.*, 2015) and then fed these regions to ResNet-101 (He *et al.*, 2016) to assess the cataract severity level based on the nuclear region's photometric appearance. Liu *et al.* (2017) cropped the RoI using Candy edge detection and the Hough transform. Then used transfer learning with AlexNet (Krizhevsky *et al.*, 2012) CNN to classify into normal and paediatric cataracts.

Automated diagnosis of corneal infections has received little research attention from the research community. As per statistics published as part of the 2019 World Vision Report (WHO, 2019), infection is the most common cause of corneal ulcers (called keratitis), and at least 4.2 million people worldwide are reported to suffer from corneal opacities. Corneal opacity is caused by a variety of conditions that cause the cornea to scar or become opaque. Microbial keratitis (MK) or infectious keratitis (IK) is the primary cause of corneal opacification and the fifth leading cause of visual impairment in the developing world (Ung *et al.*, 2019). If such infections are not detected and treated early, they can cause irreversible corneal blindness due to perforation, endophthalmitis, and panophthalmitis (Anutarapongpan and Brien, 2014; Schein, 2016; Maharana *et al.*, 2016; Tananuvat *et al.*, 2021). FK, in particular, is challenging to treat at later stages and may necessitate surgery. The gold standard method for diagnosing FK is corneal scraping with microbiological culture-sensitivity testing. However, this is a time-consuming laboratory procedure (Ferrer and Alió, 2011).

Fungal organisms are slow growing and may not be florid in the early stages of FK. Fungal cultures may also have limited sensitivity due to the scant quantity of material accessible from corneal scrapings, which may in turn lead to false-negative results (Ferrer and Alió, 2011). Slit-lamp examination of the ocular surface, particularly the cornea, conjunctiva, and anterior chamber, is widely used in the diagnosis of MK. However, the findings of corneal staining combined with

slit-lamp biomicroscopy are heavily reliant on the grader's clinical knowledge. It has been reported that correctly differentiating between bacterial keratitis and FK is a challenging process, even for trained corneal experts, and is often misdiagnosed in more than 30% of the cases (Dalmon *et al.*, 2012). Furthermore, certain clinical signs typically attributed to FK may also be of bacterial or protozoal origin, thereby complicating the diagnostic process. Automated grading of FK images could overcome these limitations by lowering physician burden and improving patient prognosis through early diagnosis. However, very few studies have used AI to enable the early diagnosis of FK using digital slit-lamp images. Loo *et al.* (2021*a*) proposed a modified version of mask R-CNN (region-based CNN) called SLIT-Net for the segmentation of ocular structures and biomarkers using 133 MK digital slit-lamp images. However, the authors did not address the problem of the classification of keratitis based on microbial etiology.

Xu *et al.* (2020*b*) developed a patch-level deep model to classify IK by manually segmenting the infectious regions of slit lamp microscopy images. Manual segmentation of corneal areas is a time-consuming task and needs expertise. Kuo *et al.* (2020) collected a total of 288 microbial laboratory-confirmed images, which included 114 FK and 174 non-FK slit-lamp biomicroscopy images. The authors used DenseNet (Huang *et al.*, 2017) for their experiments and performed five fold cross-validation to classify between FK and non-FK. They reported an average accuracy of about 70% and also highlighted that incorporation of transfer learning and region of interest (RoI) cropping processes could contribute to further improvements in the performance. Furthermore, RoI segmentation results in uniform clarity in images, which also helps enhance performance while also enabling accurate evidence for trustworthy prediction. Recently, Hung *et al.* (2021) used U^2 Net (Qin *et al.*, 2020) to segment the cornea in slit-lamp biomicroscope images and then incorporated transfer learning to classify these images into bacterial and FK. The authors directly utilized the U^2 Net segmented corneal regions and achieved an average diagnostic accuracy of 80%. Table 2.9 summarizes recent techniques that were adapted for CDSSs built on slit-lamp images.

Table 2.9: Summary of CDSS built on slit-lamp images

Author	**Methodology**	**Remarks**
Acharya *et al.* (2010)	Proposed a feed forward neural network using the centroids of K-means clustering to classify between normal, post-cataract, and cataract using slit lamp images.	Validated primarily on a small single database with $<$ 50 images. It is challenging to determine the cluster size for a large dataset, which makes it difficult to adapt to a real hospital scenario.
Li *et al.* (2020)	Utilized Faster R-CNN for RoI segmentation of 14 regions and then used transfer learning with ResNet to classify the pathological features of the segmented ocular lesions into 22 categories.	Involves segmentation of 14 regions and then classified these regions on the basis of pathological features. Validated primarily on a small validation data and there is still scope for improvement of performance.
Xu *et al.* (2019*a*)	Localized the nuclear region using Faster R-CNN and then fed these regions to ResNet-101 to assess the cataract severity level based on the nuclear region's photometric appearance.	Achieved 81.5% accuracy, there is still scope for improvement of performance.
Kuo *et al.* (2020)	Used DenseNet to classify between FK and non-FK with five fold cross-validation using 288 images.	Achieved an average accuracy of about 70% and highlighted that incorporation of transfer learning and region of interest (RoI) cropping processes could contribute to further improvements in the performance.

continued ...

... continued

Author	Methodology	Remarks
Xu *et al.* (2020*b*)	Developed a patch-level deep model to classify IK by manually segmenting the infectious regions of slit lamp microscopy images.	The manual segmentation of corneal areas is a time-consuming task and needs expertise.
Hung *et al.* (2021)	Used U^2 Net to segment the cornea in slit-lamp biomicroscope images and then incorporated transfer learning to classify these images into bacterial and FK.	Achieved an average diagnostic accuracy of 80%, and there is still scope for improvement of performance.

2.3 CDSSs using Multimodal Healthcare Data

EHRs are being adopted widely in the healthcare field and are composed of multimodal data, including textual and imaging data. Most existing literature utilizes a single modality of data, such as structured textual, unstructured textual, or imaging medical data. Very few works make use of multimodal medical data. Medical multimodal disease classification, medical report generation, multimodal fusion for single/multitask predictions are a few tasks that made use of multimodal data from EHRs. Multimodal data is often used in report generation and clinical diagnostic tasks as well.

Generating medical reports from multimodal EHR data is time-consuming and requires extensive expertise in practice. DL models have shown promising results for the generation of short reports that use textual and imaging data. Zhang *et al.* (2017) proposed a medical image diagnosis network (MDNet) to establish a direct multimodal mapping between medical images and diagnostic reports that generates diagnostic reports given an input bladder image, retrieves images by symptom descriptions, and visualizes attention, to justify the network diagnosis process. Jing *et al.* (2018) proposed a co-attention model based on tags and imaging features to localize regions containing abnormalities and generate a report for them using hierarchical LSTM. They achieved a BLEU score of 0.517 for a single word (BLEU1) on the CheXpert dataset. For the same dataset, Yuan *et al.* (2019) proposed synthesizing multi-view (frontal and lateral views) information

by applying a sentence level attention model and enforcing the encoder to extract consistent features with a cross-view consistency (CVC) loss. With the combination of late fusion with medical concepts, they could improve the BLEU score to 0.529.

Xue *et al.* (2019) proposed a recurrent technique for utilizing CNN with the LSTM model. Additionally, the proposed multimodal approach combined the encoding of frontal and lateral X-ray images with the first generated sentence to provide an attention input that guides the generation of the following sentence, ensuring that created sentences remain coherent. Liu *et al.* (2019*a*) proposed the encoder-decoder DL model that first predicts the topics to be represented in a diagnostic report, then conditionally generates the sentences corresponding to these topics. They utilized an image encoder along with both sentence and word decoders to generate the report for a chest X-ray image. They achieved a 0.313 BLEU1 score for the MIMIC-CXR dataset. Messina *et al.* (2020) reviewed deep learning algorithms for the automatic report generation task and discovered that the most existing works used the Indiana University dataset (Demner-Fushman *et al.*, 2016). Monshi *et al.* (2020) surveyed AI-based radiology report generation and highlighted the need for reasonable explanations for DL model outcomes. Pandey *et al.* (2021) studied the challenges associated with AI based systems that combine medical imaging & NLP and found that converting the unstructured information into structured information is very tedious and costly.

Recent research has proved that multimodal data contributes to the development of comprehensive diagnostic systems that aid in effective decision making. Jin *et al.* (2018) jointly trained time-series signals and unstructured clinical text representations to predict the in-hospital mortality risk for ICU patients and were able to improve the performance by 2% AUC. They proposed a multimodal architecture comprising LSTM for time-series data and Doc2Vec embedding for clinical notes taken from the MIMIC-III dataset. The output from these layers is then fused with the dense layer to predict mortality. Guo *et al.* (2018) used multi-domain imaging data to predict Alzheimer's disease progression. They used demographic features, MRI imaging data, longitudinal cerebrospinal fluid (CSF), and cognitive performance bio-markers. Furthermore, they fused the features at the last but one layer and achieved a 2% improvement in performance when compared to using only a single modality of the data separately. Young *et al.* (2013) used a combination of MRI, FDG-PET, CSF and APOE data to classify mild cognitive impairment (MCI-s) (stable) and MCI-c (convert to AD) patients. They used an SVM classifier for fused features and achieved a 3% AUC improvement

when compared with single imaging data.

Cennamo *et al.* (2021) utilized multimodal imaging data that included multicolor imaging, EDI-OCT, and ultrasonography to detect the vascular and structural features of choroidal metastasis. Wisely *et al.* (2020) utilized multimodal retinal imaging data and could achieve better performance for symptomatic Alzheimer's disease detection. However, the authors used a very small dataset of 159 patients, which makes it difficult to adapt to a real hospital scenario. Cai *et al.* (2019) conducted a survey on multimodal data driven CDSS and discussed the methods for multimodal data fusion used in the development of intelligent healthcare systems. The authors highlighted the need for knowledge-based reasoning for DL model outcomes. Huang *et al.* (2020) reviewed AI-based imaging and EHR diagnostic CDS systems. The authors observed that multimodal fusion significantly enhances performance over single modality models for CDS utilizing medical imaging tasks. Table 2.10 presents a summary of recent works that have utilized multimodal healthcare data for CDSS development.

Table 2.10: Summary of multimodal data based CDSSs

Author	**Methodology**	**Remarks**
Jing *et al.* (2018)	Used co-attention on tags and imaging features to localize regions containing abnormalities and generate a report using hierarchical LSTM.	Achieved BLEU score of 0.517 for a single word (BLEU1) on the CheXpert dataset, but there is still scope for improvement.
Xue *et al.* (2019)	Used a recurrent technique of combining CNN with LSTM model.	Combined the encoding of the X-ray frontal and lateral images with the first generated sentence to provide an attention input that guides the generation of the following sentence, ensuring that created sentences remain coherent. A smaller dataset (with high class imbalance) was used for training the network, which caused missing abnormal descriptions.

continued ...

. . . continued

Author	Methodology	Remarks
Liu *et al.* (2019*a*)	Used the encoder-decoder DL model that first predicts what topics will be discussed in the report, then conditionally generates the sentences corresponding to these topics. Also utilized an image encoder along with both sentence and word decoders to generate the report for a chest X-ray image.	Achieved a 0.313 BLEU1 score for the MIMIC-CXR dataset. Still, there is scope for improvement.
Yuan *et al.* (2019)	Synthesized multi-view information by applying a sentence level attention model and using cross-view consistency (CVC) loss.	With the combination of late fusion with medical concepts, they could improve the BLEU score to 0.529.
Guo *et al.* (2018)	Fused demographic, MRI imaging data, longitudinal cerebrospinal fluid (CSF), and cognitive performance bio-markers features to predict Alzheimer's disease progression.	Achieved a 2% improvement in performance when compared to using only a single modality of the data separately.
Young *et al.* (2013)	Fused a combination of MRI, FDG-PET, CSF and APOE features using SVM classifiers to classify mild cognitive impairment (MCI-s) (stable) and MCI-c (convert to AD) patients.	Achieved a 3% AUC improvement when compared with single imaging data.

2.4 Outcome of Literature Review

After an extensive survey of existing literature, several research gaps were identified, specifically in the area of CDSSs using textual, imaging, and multimodal

healthcare data. The previously discussed text-based CDSS works have a significant limitation in that they focus on structured patient data and processed EHRs, which are mostly standardized and widely used in Western countries. However, currently in India (and other developing countries), the structured and processed EHR adoption rate is very low. On the other hand, most hospitals and healthcare centres are equipped with computer systems to store patient data in an unstructured/semi-structured form for purposes like billing, pharmacy, lab reports, etc. Designing techniques to consume this clinical data for enabling intelligent CDSS doctors and hospital personnel could be a huge contribution, given the prevailing conditions in the Indian healthcare ecosystem. Very few existing works provide evidence to support the inference results, as was discovered. For the CDSS to be adaptable in real-world circumstances, providing a transparent, explainable decision is considerably more acceptable than putting forth a highly accurate, non-transparent decision. While most existing works have concentrated on English based clinical input, very few works employ non-English textual inputs. Based on these observations, a focus on this problem was made, with the intent of exploring patient-specific interpretive and predictive analytics using textual healthcare data.

Over the last two decades, CDSSs have been a focus of active study, and the approaches utilized have evolved over time. Machine learning and DL models have already been shown to be the most accurate and effective. CDSSs that make patient-specific predictions are critical in the area of clinical healthcare. These prediction systems not only assist patients with diagnostic predictions and reminders, but also ensure that healthcare personnel, particularly physicians, have a comprehensive view of the patients' medical history at a glance. In addition, healthcare personnel can use the trained system's recommendations for diagnostics and medical tests that should be performed to make more informed treatment decisions. Recent works by (Purushotham *et al.*, 2018; Harutyunyan *et al.*, 2019*a*; Sheikhalishahi *et al.*, 2020; Zhang *et al.*, 2020) have proposed deep learning-based approaches for providing CDS by predicting a patient's illnesses and risks. While some of these approaches (Zhang *et al.*, 2020; Sheikhalishahi *et al.*, 2020) showed promising results, there is undoubtedly scope for improvement in terms of patient data representations, neural network designs, and interpretability. Thus, developing more effective systems that deliver effective patient-specific predictive analytics based on unstructured textual healthcare data is an avenue that is explored in depth in this thesis.

The study of retinal imaging based CDSSs revealed a major need for improv-

ing the performance of existing DL models in multi-label disease classification problems that require binary classification of multiple diagnostic labels, each label indicating a specific disease. Another finding from the detailed analysis was the strategies used to preprocess and feature model of the clinical data prior to its use in prediction model-based CDSSs. It was observed that experimenting with various preprocessing strategies could enable effective feature learning for identifying the minute lesions. Typically, medical imaging data is resized and cropped before being input to a training algorithm, without an appropriate RoI segmentation strategy to allow accurate evidence for trustworthy prediction. As can be seen, the majority of present approaches provide medical images for a single ocular condition, while there is potential for several diseases to be generated. Additionally, the synthesis process should be general, requiring just the ocular condition and the number of synthesized images.

The EHR comprises vital multimodal patient information such as medical history, diagnosis, prescriptions, treatment plans, immunization dates, allergy information, imaging, and laboratory test results. After an extensive survey of existing multimodal based CDSS literature, it was noted that radiology is the principal disease management tool. This increases the cognitive burden of radiologists due to the manual effort required to assess and report on a large number of patient populations. As a result, it is critical to design intelligent CDSSs capable of leveraging the important information included in multimodal healthcare data from radiology. The performance of the existing approaches for report generation tasks that use multimodal radiology data for report generation may be further increased by developing superior textual and imaging feature modelling methodologies and prediction model architectures. A critical necessity is to build and construct a complete framework that allows quick and cost-effective illness screening using multimodal healthcare data in a user-friendly and unobtrusive way. Hence, this is also one of the issues proposed to be addressed in this work.

2.5 Summary

In this chapter, various existing approaches and models that have been proposed to develop AI-based CDSS systems were discussed. Existing approaches that focus on building CDSS using several types of healthcare data were grouped into three categories: those that make use of textual data, those built on diagnostic imaging data, and systems that utilize multimodal healthcare data. The extensive review of the existing literature revealed a definite need for varied CDSS deployments

capable of extracting latent knowledge embedded in a wide variety of healthcare data for tasks ranging from disease prediction to hospital mortality prediction to medical record management. The availability of large amounts of imaging data in the field of ophthalmology, specifically fundoscopy and slit-lamp imaging data used for the diagnosis of widely prevalent common retinal and corneal diseases, revealed a significant opportunity for the development of novel preprocessing and image modelling strategies for CDSS development.The availability of huge unstructured text-based healthcare data reveals a substantial scope for the design of novel preprocessing and language modelling strategies for CDSS development. Based on these observations, the scope of the problem and the problem statement addressed in this thesis were defined (discussed in Chapter 3). The proposed methodologies designed to address these identified research gaps are discussed in brief in Chapter 3 and in detail in later chapters of this thesis.

Chapter 3

Problem Description

3.1 Background

In the preceding chapter, a detailed review of existing literature with an emphasis on AI-based CDSS using multimodal healthcare data was presented. Additionally, prevalent issues and criteria for enhancing CDSS were outlined. This chapter discusses the identified research gaps, specifically in the domain of disease detection and prediction, utilizing textual, diagnostic imaging, and multimodal healthcare data. A problem statement is formulated to address the identified research gaps. The scope of the proposed research methods and a brief overview of the methodologies used to tackle the formally defined problems are also discussed.

3.2 Research Gaps

Based on a review of existing approaches for developing CDSSs, it is clear that the effectiveness of developed CDSSs is heavily dependent on how clinical data is modelled and represented, as this is a critical requirement for effective prediction models. While several existing works discuss computer-aided CDSS based on healthcare imaging data, a comprehensive examination of early detection CDSS approaches and the associated challenges has yet to be undertaken.

- CDSS models that make use of textual healthcare data fail to achieve acceptable performance for automated coding assignment tasks. There is ample scope to enhance the existing DL models to improve their performance both in terms of predictability and interpretability.
- Automated coding assignment systems mostly make use of clinical notes, mainly written in English. There is a need to develop a CDSS that comprises

multilingual clinical notes.

- Although DL models perform well for some clinical tasks, very few provide evidential support to visualize the inference results. For CDSS to be adaptable in real-world circumstances, providing a transparent, explainable decision (even if it is wrong) is considerably more acceptable than putting forth a highly accurate, non-transparent decision.
- Due to the difference in camera settings, the captured medical images may have different image dimensions, contrast, illumination, light incident angle, etc. Most existing methods perform training and testing using one specific dataset. There is a need to develop cross-data verifiable and robust models to deal with data from different distributions.
- In developing countries, sophisticated devices are often not available in rural hospitals. Thus, there is a need for developing automated image quality enhancement algorithms that enable the use of images captured using portable, low-cost devices but still provide clinical diagnostic quality similar to that offered by high-end appliances.
- To improve model performance for early disease detection tasks, there is significant scope for experimenting with various preprocessing strategies that enable effective feature learning for detecting minute lesions.
- Supervised DL with CNN models and deeper architectures requires huge volumes of annotated/labelled images. Such acquisition of images is expensive and requires the extensive annotation services of expert physicians. Semi-supervised Generative Adversarial Networks (GAN) (Goodfellow *et al.*, 2014) that can learn from limited data have seen limited exploration.
- There is scope for improving the performance of existing DL models in multi-label disease classification problems requiring binary classification of multiple diagnostic labels, each label indicating a specific disease.
- There is a significant requirement to design and develop a comprehensive framework that enables rapid and cost-effective disease screening using multimodal healthcare data in a user-friendly and unobtrusive manner.

3.3 Scope of the Work

With these observations and with the aim of bridging the observed gaps, the research work presented in this thesis has contributions in six major aspects, as listed below:

1. To design and develop CDSSs built on unstructured EHRs, for improved prediction accuracy in assigning diagnostic codes to medical records.
2. Designing approaches for medical imaging based CDSSs, for improved interpretation for disease detection and grading.
3. To design and develop multi-task CDSSs using diagnostic imaging data for ocular disease diagnosis.
4. Adaption of GAN-based augmentation techniques for the generalization of the developed DL models.
5. Design of novel ensemble of preprocessing approaches with CNNs for improved prediction accuracy in detecting ocular diseases.
6. Designing a comprehensive CDSS using multimodal healthcare data to detect and manage onset and grading of lung disease in patients.

3.4 Problem Statement

Based on an understanding of the gaps observed during the comprehensive review of current literature in the area of AI-based CDSS using multimodal healthcare data, the research problem addressed by this thesis is defined as below:

> *"To design and develop AI-based clinical decision support systems using multimodal healthcare data, with support for interpretability and evidence-based diagnosis."*.

3.5 Research Objectives

Based on identified gaps and the defined problem statement, three research objectives have been defined and addressed in this thesis:

1. To develop techniques for patient-specific predictive analytics using textual healthcare data.
2. To design and develop multi-task clinical decision support system using diagnostic imaging based healthcare data.
3. To design and develop AI-based clinical decision support system using multimodal healthcare data.

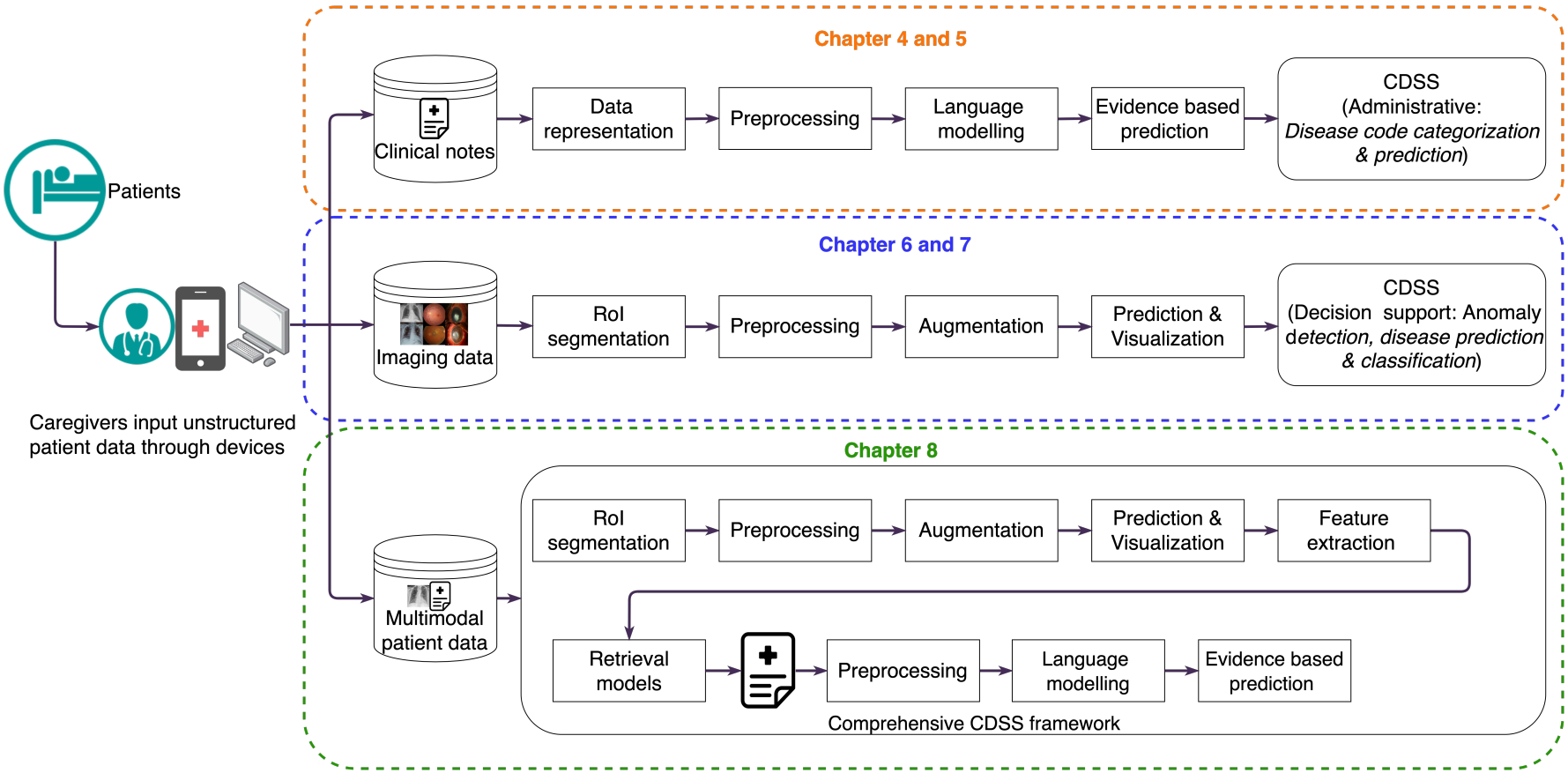


Figure 3.1: Scope of the proposed research work

3.6 Brief Overview of Proposed Methodology

Fig. 3.1 depicts the overall system architecture of the proposed AI-based CDSS for multimodal healthcare data. The contributions made towards each research objective are indicated by the thesis chapter(s) in which they are discussed in further depth. This section discusses a high-level overview of the proposed approaches.

3.7 Patient-specific Predictive Analytics with Unstructured Text based Healthcare Data

The utility of the rich, latent patient information embedded in unstructured clinical notes has been mostly overlooked, and its importance as a source of valuable disease-specific information has been under-exploited. There is a tremendous opportunity to build AI-based CDSS that are capable of directly absorbing unstructured clinical notes for predictive analytics. To address this, the effective textual feature modelling methodologies are combined for deriving effective patient data representations that can be utilized for effective deep learning training and inference. Fig. 3.2 depicts the high-level workflow of the design of a patient-centric CDSS built on unstructured clinical notes.

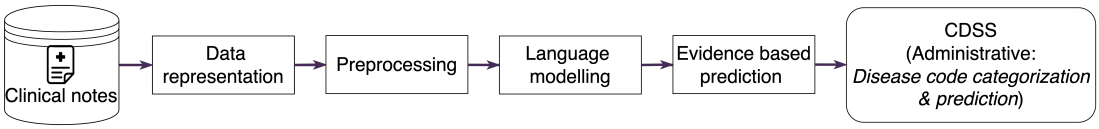


Figure 3.2: Predictive analytics with unstructured text data.

Chapter 4 discusses approaches for generating patient data representations using textual feature modelling techniques, as well as how they were used in ICD coding systems. *EnCAML*, a multi-channel, variable-sized convolutional attention model, was proposed to enable the clinical task of diagnostic code assignment as a multi-label classification problem. The proposed model enhances the predictability of the ICD codes by extracting multi-granular text snippets. The attention mechanism used in the model enables the selection of those segments that most contribute to the corresponding diagnostic code. Additionally, a qualitative analysis was performed to demonstrate the model's ability to capture crucial input tokens contributing to particular ICD-10 diagnostic codes from non-English clinical notes through the label attention transformer architecture ($\mathcal{LATA}$), which is described in detail in Chapter 5.

3.8 Multi-task CDSS using Diagnostic Imaging Data

As stated earlier, this thesis focuses on fundus and slit-lamp imaging data utilized in the ophthalmology domain to develop the multi-task CDSS. Initially, a systematic review of existing literature is carried out to examine the diagnostic use of automated MA detection and segmentation for early DR diagnosis. Initially, a focused study on existing early DR diagnosis techniques was carried out to understand their strengths and weaknesses. To address a few identified research gaps, a multi-task CDSS using fundus and slit-lamp imaging healthcare data was designed. The overall workflow is depicted in Fig. 3.3.

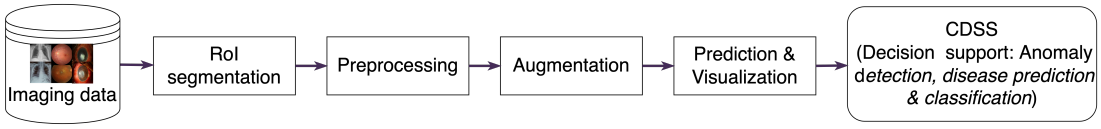


Figure 3.3: Multi-task CDSS using diagnostic imaging data.

Several preprocessing approaches were combined with CNN to detect chronic ocular diseases (COD) accurately, including myopia, diabetic retinopathy, age-related macular degeneration, glaucoma, and cataract. Chapter 7 provides a detailed discussion of various preprocessing methods applied to fundus images. For effective learning performance, generalizable deep neural models require a large number of labelled images. A variety of data augmentation techniques are widely used to deal with the limited number of training images. Batch-level and condition-level data augmentation techniques are incorporated for increasing the number of images used for training the neural models and are detailed in Chapter 7. In addition, a multi-scale convolutional neural network (KeratNet) was proposed for accurate segmentation of the corneal region to enable early fungal keratitis (FK) diagnosis (refer Chapter 6). A deep neural pipeline for corneal region segmentation followed by a CNN was proposed to differentiate between FK and non-FK classes.

3.9 CDSS using Multimodal Healthcare Data

A CDSS for predicting the presence or absence of lung diseases using diagnostic scans can be beneficial to healthcare professionals as well as patients. An attempt

was made to build such a model that leverages the wealth of information contained in expert reports along with the actual diagnostic scan data for learning disease representations of lung diseases. Figure 3.4 depicts the over workflow of the framework. A complete web-based framework was deployed on the cloud, to provide a highly usable platform for expert verified screening results. A comprehensive clinical decision support framework was deployed for quick and cost-effective early screening of lung diseases in a user-friendly and unobtrusive manner. The complete framework design and the methodologies are discussed in Chapter 8.

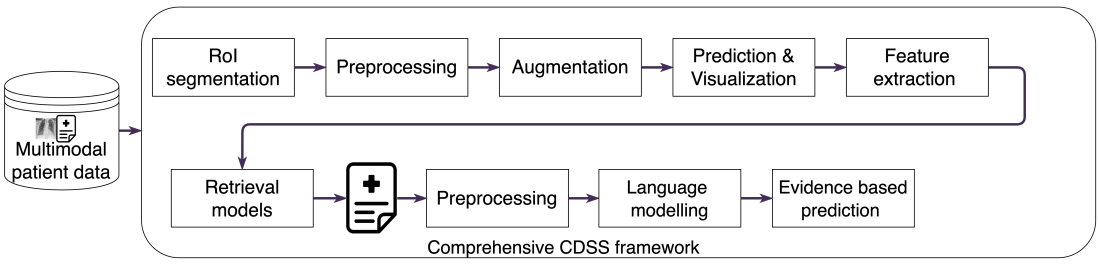


Figure 3.4: AI-based CDSS using multimodal healthcare data.

3.10 Research Contributions

In this research thesis, a framework for design and development of AI-based clinical decision support systems built using multimodal healthcare data, with support for interpretability and evidence-based diagnosis is presented. The objectives are to develop predictive analytics utilizing textual, diagnostic imaging, and multimodal healthcare data, thus aiding physicians to make clinical decisions. The major contributions of the research work are listed below.

- Development of a multi-channel, convolutional attention based neural model for automated diagnostic coding of unstructured patient discharge summaries.
- Design of $\mathcal{LATA}$ – Label Attention Transformer Architectures for automated ICD-10 coding of unstructured clinical notes.
- An empirical study of preprocessing techniques with convolutional neural networks for accurate detection of chronic ocular diseases using retinal fundoscopy images.
- Development of multi-scale convolutional neural network for accurate corneal

segmentation in the early detection of fungal keratitis using slit-lamp biomicroscopy imaging data.

- Design of multi-task deep neural networks for learning COVID-19 disease representations from multimodal disease data.
- Deployment of an AI-based clinical decision support system for learning COVID-19 disease representations from multimodal patient data.

3.11 Summary

In this chapter, the scope of the research work and the identified research gaps that are addressed in this thesis are presented, based on which the research problem for this research thesis was formally defined. The proposed approaches towards solving the defined problem were also discussed briefly, and are explained in detail in subsequent chapters.

PART II

Patient-specific Predictive Analytics using Textual Healthcare Data

Chapter 4

Automated Diagnostic Coding of Unstructured Discharge Summaries

4.1 Introduction

In hospitals, the International Statistical Classification of Diseases and Related Health Problems medical coding taxonomy (ICD-9[1] and ICD-10[2] are widely employed to describe patients' clinical conditions and associated diagnoses. These classification systems are maintained by the World Health Organization. The ICD is essentially a hierarchical classification that defines unique codes for patient conditions, diseases, infections, symptoms, causes of injury, and others. These unique diagnostic codes are assigned to patient records to facilitate clinical and financial decisions made by the hospital management for various tasks, including billing, insurance claims, and reimbursements (Jensen *et al.*, 2012; Li *et al.*, 2019*b*). Based on clinicians' free-text notes and other patient records such as discharge summaries, doctors' notes, nursing notes, and other relevant sources, trained professional medical coders employed by the MRD in hospitals transcribe patient records into a set of appropriate medical diagnostic codes (from a potentially large number of over $15,000$ codes). These medical coders utilize their medical domain expertise and a plethora of coding rules/terminologies to facilitate the mapping of a patient record to applicable diagnostic codes (one-to-many). Several publicly available large-scale healthcare datasets provide instances of patient data mapped to ICD-9 and ICD-10 clinical procedures/diagnostic codes.

Given the enormous volume of patient records generated every day in urban and rural hospitals alike, such manual coding processes are highly cost-intensive

[1] `https://www.cdc.gov/nchs/icd/icd9cm.htm`.
[2] `https://icd.who.int/browse10/2019/en`.

and often inexact, time-consuming, and error-prone (Chen *et al.*, 2017; Zeng *et al.*, 2019*b*). The additional costs incurred due to inaccurate coding and the financial investment towards improving diagnostic coding efficacy is estimated to be more than $25 billion per year (in the United States alone) (Lang, 2007; Farkas and Szarvas, 2008). Furthermore, automated systems reliant on structured electronic medical records (S-EMRs) find limited applicability in developing nations with a relatively low digitization rate. It is crucial to develop intelligent computational systems that accommodate these needs by facilitating automated diagnostic coding of *unstructured* patient records. Such a code assignment can be regarded as a multi-label classification problem involving binary classification of multiple diagnostic labels, with each code label pertaining to a specific diagnostic condition (recorded as a binary indicator).

Over the years, there has been a significant interest in developing and utilizing machine learning models to facilitate automated ICD coding as a multi-label classification task. Strategies and models utilizing Support Vector Machines (SVMs) (Ferrao *et al.*, 2013; Perotte *et al.*, 2013; Wang *et al.*, 2017), naïve Bayes (Pakhomov *et al.*, 2006*b*; Medori and Fairon, 2010*b*), nearest neighbors (Ruch *et al.*, 2008; Erraguntla *et al.*, 2012), unsupervised topic modelling (Perotte *et al.*, 2011; Dermouche *et al.*, 2016), and several others have been employed for the clinical prediction task. Recent surveys on the applications of deep learning approaches for the analysis of S-EMRs (Shickel *et al.*, 2018; Bizopoulos and Koutsouris, 2019; Domingues *et al.*, 2019) highlight the need for interpretability of predictions made and explainability of automated prediction systems. By understanding the input features that contribute to the output decisions, trust can be built in the predictions and recommendations enabled by such learned models, which is crucial in healthcare applications.

More recently, research on automated code assignment has been attempted by modelling the unstructured clinical text (Baumel *et al.*, 2018*a*; Huang *et al.*, 2019; Mullenbach *et al.*, 2018; Li and Yu, 2020; Vu *et al.*, 2020), thus, exploring the richness of patient-specific information in such free-text. While supervised learning approaches are applicable in cases of accessible large-scale annotated datasets, it is common for researchers to explore modelling approaches that are beneficial in targeted studies with minimal data resources. In this regard, deep neural models and modelling strategies, including the *DeepLabeler* (Li *et al.*, 2019*b*), Convolutional Networks (ConvNets) (Mullenbach *et al.*, 2018; Li and Yu, 2020; Teng *et al.*, 2020; Ji *et al.*, 2020), Long Short-Term Memory (LSTM) models (Xie and Xing, 2018), and transfer learning (Zeng *et al.*, 2019*b*; Rios and Kavuluru, 2019), have

been quite successful. However, the availability of healthcare clinical datasets are relatively abundant (e.g., PCORnet[3], Open NHS[4], eICU-Philips[5], MIMIC[6], VistA[7], ACS-NSQIP[8], and others (Marshall *et al.*, 2016)), owing to the volume of medical patient data generated day-to-day, thus promoting active healthcare research in modelling such data. Despite the abundance of data, only a limited number of these data sources include unstructured text-based patient diagnosis data, such as discharge summaries and nursing notes. Most state-of-the-art studies have utilized the standard, openly-available MIMIC-III (Medical Information Mart for Intensive Care) database (Johnson *et al.*, 2016), comprising over $40,000$ patients' data. Several researchers (Baumel *et al.*, 2018*a*; Mullenbach *et al.*, 2018) have attempted to utilize the predictive power of the machine and deep learning based models to enhance the diagnostic coding performance on the patient data available in the MIMIC-III database, making the database one of the most widely employed sources for performance benchmarking.

4.2 Problem Definition

Existing studies facilitating automated ICD-based clinical coding corroborate the critical nature of the task at hand. Moreover, the applicability, deployability, and adaptability of the proposed intelligent systems in real-world scenarios demand high performance (exceeding that of the manual clinical coders), both in code prediction and system explainability. However, the nature of the underlying data poses several modelling challenges, including the variety and sparseness of diagnostic codes, complex structural and temporal nature of unstructured data, and prolific use of medical jargon, limiting the reported performance in the existing works. Thus, the problem of accurate ICD code assignment remains a long-standing open research challenge in the fields of healthcare informatics and machine learning. To cope with the modelling complexities, specifically the vast imbalance in the code distribution across patient data, prior studies discarded medical records corresponding to less frequent diagnosis codes, thus reporting the performance of modelling the top$-k$ diagnostic procedures. Furthermore, several researchers and recent surveys on the use of deep neural approaches for patients' risk stratification

[3] `https://pcornet.org/data-driven-common-model/`.
[4] `https://digital.nhs.uk/data-and-information/data-collections-and-data-sets`.
[5] `https://eicu-crd.mit.edu/`.
[6] `https://mimic.physionet.org/`.
[7] `https://www.data.va.gov/widgets/4d7k-fkpu`.
[8] `https://www.facs.org/Quality-Programs/ACS-NSQIP/joinnow/data`.

(Shickel *et al.*, 2018; Bizopoulos and Koutsouris, 2019; Domingues *et al.*, 2019) highlight the urgent need for the interpretability and explainability of the proposed automated prediction systems. The problem to be addressed here is defined as follows:

> *Given the known issues arising due to the lengthier clinical notes, medical jargon, and large number ICD codes considered by traditional code assignment, design and develop approaches for effective automated ICD-9 code assignment, based on unstructured discharge summaries.*

In this study, the significance of interpretable intelligent healthcare solutions in ensuring the trustworthiness of the underlying computational clinical decision support systems is emphasized. This work proposes the Enhanced Convolutional Attention network for Multi-Label classification (*EnCAML*). The *EnCAML* model employs multi-channel, variable-sized convolution filters and multiple attention layers that reveal the associations of medical text with the predicted diagnostic code as a result of the interactions between the neurons. An attempt to dissect the black-box decisions facilitated by the proposed deep neural model by visualizing the associated clinical terms that contribute to the prediction of the respective disease code is made. Such analyses and interpretation of the obtained predictions can enhance the explainability of the proposed automated system. The key contributions of this work in advancing the efforts of the state-of-the-art can be summarized as follows:

- Design of *EnCAML*, a multi-channel, variable-sized convolution attention neural model that facilitates the reliable assignment of diagnostic codes using unstructured text-based patient discharge summaries, focusing on the interpretability and explainability of the neural system.
- Detailed analysis of the impact of the initial embedding layer on the overall performance of the proposed *EnCAML* model, using several state-of-the-art embedding approaches on voluminous discharge summaries.
- Extensive benchmarking results underscore the superior performance of the proposed *EnCAML* model compared to the current work on ICD-9 code prediction using MIMIC-III unstructured discharge summaries.

4.3 Motivating Example

To describe the prevailing conditions that emphasise the need for disease prediction CDSS based on unstructured clinical notes, let us consider scenarios where

a hospital has a full-fledged EHR system, and discharge summaries are recorded by physicians as notes. The MRD staff scans through the entire document (often longer than 3000 words) and assigns the ICD diagnostic and procedural codes for a particular patient based on historical observations and probability. Manually scanning the entire document is a time consuming task, and coding efficiency is often dependent on the expertise of trained medical coders. Effective coding of patient records in hospitals is an essential requirement for epidemiology, billing, and managing insurance claims. The additional costs incurred as a result of incorrect coding, as well as the financial investment in improving diagnostic coding efficacy, are estimated to be more than $25 billion per year (in the United States alone) (Lang, 2007; Farkas and Szarvas, 2008). This delay in processing and code assignment could be avoided if the ICD codes could be automatically generated using the clinical notes. The automated disease prediction CDSS could directly process the unstructured clinical notes recorded by the physician and provide them with a list of ICD codes along with highlighting of the texts that the system found relevant for predicting each of the codes. In this way, the MRD staff need not perform conversion to any predefined structure and, hence, has the advantage of significant savings in person-hours and reduced costs due to inaccurate code assignment.

In the subsequent section of this chapter, various approaches towards developing effective patient-specific ICD-9 code assignment models built on unstructured discharge summaries are presented. The contributions towards the defined problem are in the context of designing methods that can automatically process a variety of unstructured discharge summaries, with their differences in notation, usage of extensive medical jargon, acronyms, etc., and still be able to extract relevant disease-specific features, which can be leveraged for the purpose of automatic ICD-9 code prediction. The performance of the proposed models was compared to that of state-of-the-art ICD-9 code assignment built on unstructured patient discharge summaries using standard evaluation metrics.

4.4 EnCAML - Multi-label Convolutional Attention Model for ICD-9 Code Prediction

For the task of automated ICD-based code assignment, which is a multi-label problem, the MIMIC-III (v1.4) database was employed. In line with the existing works, this study benchmarks the performance using (a) *Top−k diagnostic codes*, covering over 76.93% ($k = 10$) and 93.60% ($k = 50$) of the database, (b) *Top−k*

diagnostic code categories, covering over 84.24% ($k = 10$) and 96.79% ($k = 50$) of the database, and (c) *All* $6,918$ *disease diagnosis codes*, corresponding to the discharge summaries of the patient cohort. Extensive benchmarking across several variations in the cohort data selection (presented as (a), (b), and (c) above) facilitated a detailed analysis of the obtained prediction performance with reference to existing models.

The MIMIC-III database is a comprehensive collection of diverse, clinical, and physiological healthcare data of critical care patients admitted to the Beth Israel Deaconess Medical Center, Boston, between June 2001 and October 2012. For this work, the discharge summaries corresponding to $46,520$ intensive unit patients are considered. In the given data, the occurrence of ICD-9 diagnostic codes associated with the extracted discharge summaries is highly imbalanced, indicating that the amount of data available to learn more infrequent codes is highly selective. Therefore, it is essential to understand the relevant portions of the clinical free-text that contribute towards the assignment of a particular diagnostic code. The subsequent sections describe the steps involved in extracting and preprocessing the unstructured text from the discharge summaries to facilitate ICD-9 code and category prediction.

The MIMIC-III database comprises 26 relational tables, and the required cohort data utilized in this study is extracted from two specific tables. A total of $52,726$ discharge summaries corresponding to various hospital admissions were extracted from the ***noteevents*** table, and the ICD-9 codes corresponding to these summaries were extracted from the ***diagnoses_icd*** table. Specific structural and linguistic details concerning the extracted discharge summary corpus are tabulated in Table 4.1. A cohort selection criteria in line with several state-of-the-art works (Mullenbach *et al.*, 2018; Huang *et al.*, 2019; Harutyunyan *et al.*, 2019*b*) is employed to enable comparative evaluation of the obtained performance. Accordingly, only the discharge summaries that correspond to the first hospital admission of a patient are considered, and the data from the subsequent admissions are discarded, as such conditions ensure risk assessment using the earliest detected symptoms.

To predict ICD-9 code categories, the corresponding diagnostic codes are grouped into categories based on the hierarchical nature of the ICD-9 coding taxonomy. This results in 942 code categories. The multi-label classification of discharge summaries is facilitated through pairwise comparison of the binary predictions with true code categories. The proposed *EnCAML* model is evaluated on various constructed datasets, hereby be referred to as: (a) *top-10-code*, for top-10

ICD-9 codes, (b) *top-10-cat*, for top−10 ICD-9 code categories, (c) *top-50-code*, for top−50 ICD-9 codes, (d) *top-50-cat*, for top−50 ICD-9 code categories, and (e) *all-codes*, for all 6,918 ICD-9 codes. Additionally, the approach is benchmarked using $k = 50$ most-frequent diagnostic (6,918) and procedural (2,003) codes, referred to as *top-50-dp-code*.

Table 4.1: Statistics of the discharge summaries corpus extracted from the MIMIC-III database for the clinical task of ICD-9 diagnostic code (and code category) prediction.

Parameter	**Total**	**Average**
Unstructured discharge summaries	52,726	–
Patients in the chosen cohort	46,520	–
Unique ICD-9 codes (chosen cohort)	6,918*	11.73
Unique words in the discharge summaries	150,854	606.465
Words in the discharge summaries	79,731,657	1,513.51
Words in the longest discharge summary	10,500	–
Words in the shortest discharge summary	51	–

**A total of 6,984 diagnostic codes were extracted from the MIMIC-III discharge summary corpus. However, post cohort selection and preprocessing 66 of these codes were removed.*

The discharge summaries obtained from the MIMIC-III database included duplicate entries, which were identified and deduplicated. The resulting data corresponds to 6,918 unique ICD-9 codes in total. Additionally, stemming from the manifold nature of the disease symptoms (e.g., *nephrolithiasis* (formation of kidney stones) caused by *hyponatremia* (low natrium presence in the blood), the dataset included multiple records per patient, mapped to different ICD-9 codes. To account for this, the content and diagnostic codes across multiple records of a patient are aggregated, thus enabling multi-label classification. Binary predictions as the target scores are used in this work, with a pairwise comparison of actual and predicted values.

4.4.1 Text Preprocessing

The next task is to transform the raw clinical text into a canonical form to account for the complex linguistic structure, medical jargon, and voluminosity of the clinical corpus. The discharge summaries obtained from MIMIC-III form a sizeable vocabulary of 150,854 words ($= |\mathbb{V}|$), and each summary consists of a variable number of tokens (see Table 4.1). In addition to this, multiple discharge

summaries maintained per patient add to the computational complexity and cost of training the underlying neural language models. Hence, it is vital to transform the corpus into a machine-processable format with a manageable vocabulary size. To enhance the manageability of the data, certain medications (e.g., discharge and transfer medications) and patient history sections (e.g., family and social history) are removed from the data. The procedure followed to facilitate such removal is described using Algorithm 1. Next, the punctuation marks and numeric tokens are eliminated, and character case folding is enabled. Additionally, all those tokens occurring in fewer than three summaries are tagged as *out-of-vocabulary* words.

Algorithm 1 Procedure employed for the removal of non-relevant medical jargon

1: Find tags ending with "*<string-1> <string-2>:*" using regular expressions
2: Filter-out all the tags ending with medications
3: Retain tags containing to-be-excluded keywords (e.g., discharge).
4: Store the extracted medication tags in a tags-specific database.
5: Repeat steps 1 through 4 to extract all the patient history tags.
6: **for each** *text* ∈ *discharge summary* **do**
7: **if** *text* contains a medication or history tag **then**
8: Extract the subsequent tag within the *text*
9: Remove content in the *text* between tags using regular expressions
10: **end if**
11: **end for**

To further normalize the content in the summaries, typographical error correction is applied to those tokens that are not present in the biomedical word embedding vocabulary (McDonald *et al.*, 2018). The biomedical word embeddings were trained with approximately $28,000,000$ articles comprising titles and abstracts obtained from the PubMed baseline 2018 collection[9], which accounts for a medical vocabulary of over $2,540,000$ terms. Utilizing the large PubMed vocabulary, the typographical errors of those tokens (ηs) whose Levenshtein distance (Wagner and Fischer, 1974) with the terms in the PubMed vocabulary (ρs) is less than three ($\approx 25,000$ tokens), are corrected. Levenshtein distance is computed as:

$$Lev_{\eta,\rho}(n,p) = \begin{cases} \max(n,p), & \text{if } \min(n,p) = 0, \\ \min \begin{cases} Lev_{\eta,\rho}(n-1,p) + 1 \\ Lev_{\eta,\rho}(n,p-1) + 1 \\ Lev_{\eta,\rho}(n-1,p-1) + \mathbf{1}\{\eta_n \neq \rho_p\} \end{cases} & \text{otherwise} \end{cases} \tag{4.1}$$

[9] `https://www.nlm.nih.gov/databases/download/pubmed_medline.html`.

where, $Lev_{\eta,\rho}(n,p)$ indicates the distance between the first n characters of η and first p characters of ρ (n and p are 1−based indices), and $\mathbf{1}\{\cdot\}$ denotes an indicator function. A few examples illustrating the use of Levenshtein distance for correcting the misspelled tokens are shown in Table 4.2.

Table 4.2: A few examples of misspelled tokens from the MIMIC-III discharge summary corpus, corrected using the biomedical word embedding vocabulary from (McDonald *et al.*, 2018).

Observed token	**Corrected token**	**Observed token**	**Corrected token**
abcsess	abscess	abdominall	abdominal
anixety	anxiety	arrhythmnia	arrhythmia
calcificed	calcified	calcicum	calcium
calcifed	calcified	cardiogolist	cardiologist
cardiololgy	cardiology	coronoray	coronary

4.4.2 Embeddings for Clinical Text

Word embeddings allows individual words to be represented as real-valued vectors in a predetermined vector space. The clinical texts were converted to vector space by employing a Continuous Bag-of-Words (CBoW) Word2vec embedding model (Mikolov *et al.*, 2013), trained on the underlying corpus. Table 4.3 lists the parameters utilized in generating the word embeddings. The learning rate was fixed to a default value of 0.025 (same as that of the base Word2Vec model presented by Mikolov *et al.* (2013)), and the number of iterations was set to 10. The optimal embedding size was empirically determined by experimenting with varying embedding sizes of 50, 100, and 200. The implementations of the Word2Vec model available in the Python Gensim library (Řehůřek and Sojka, 2010) were utilized in generating the embeddings. Additional details, including the rationale behind choosing the CBoW Word2Vec model over other recent neural word embedding approaches, like BERT, are discussed in Section 4.4.4.

4.4.3 Clinical Text Modelling

The *EnCAML* convolutional attention network was designed to enhance the predictability of diagnostic codes corresponding to a given discharge summary while enhancing the ease of model interpretability and performance explainability. A

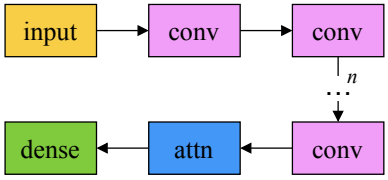
Table 4.3: Parameters of the Word2Vec models employed to effectively represent the extracted and cleaned discharge summaries.

Parameter	**Value(s)**
Number of iterations	10
Vocabulary size of the summaries without medical jargon removal or typographical error correction[b]	51,917
Vocabulary size of the summaries post processing using Algorithm 1	45,268
Vocabulary size of the summaries post processing using Algorithm 1, followed by typographical error correction	42,170
Employed word embedding sizes	$\{50; 100; 200\}$
CBoW context window size	5
Learning rate of the neural model	0.025

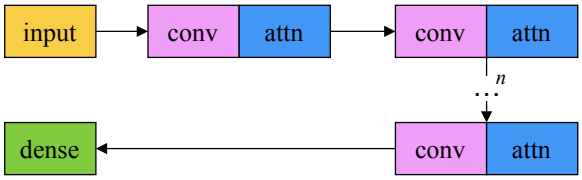
[b]At this stage, all the numeric tokens are removed, infrequent tokens marked as out-of-vocabulary words, and summaries are truncated to a maximum of 2,500 tokens (as done in Mullenbach *et al.* (2018)).

linear combination of the features (rather, feature weights) weighted by the convolutional filter, convolves the input representation into a more informative feature. Smaller kernel sizes were preferred over the larger sizes, as they captured the desired amount of context without over or undershooting. However, choosing larger kernel sizes could be beneficial when handling highly context-dependent data, as is the case in most healthcare applications. The proposed *EnCAML* neural model utilizes variable-sized multi-channel (parallel) convolution filters to ensure the choice of an appropriate kernel size. Attention weighting was employed after the convolution layer to highlight the text snippets that were responsible for mapping the respective summary to a diagnostic code. This mimics the actual diagnostic procedure followed at hospitals. Based on the observations, the proposed model facilitates enhanced predictability and interpretability over the alternate variants as depicted in Figure 4.1. The overall architecture of *EnCAML* model is presented in Figure 4.2.

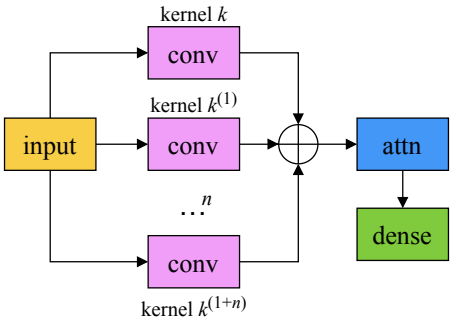
Let $\mathcal{D}^{(d)} = \{t_1^{(d)}; t_2^{(d)}; \ldots; t_L^{(d)}\}$ be the d-th ($d \in \{1; 2; \ldots; D\}$) discharge summary of length $L = |\mathcal{D}^{(d)}|$ ($\leq 2,500$) comprising tokens $t_i^{(d)}$s, each represented as an $e-$dimensional embedding. The token embeddings adjacent to the token of interest (i.e., the context) are combined using the convolution operation with a filter



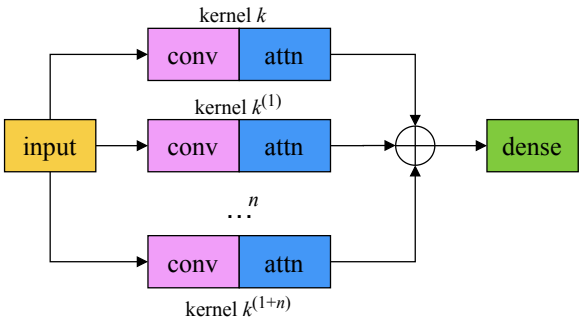
(a) Neural model with $(n+1)$ sequential convolution units.



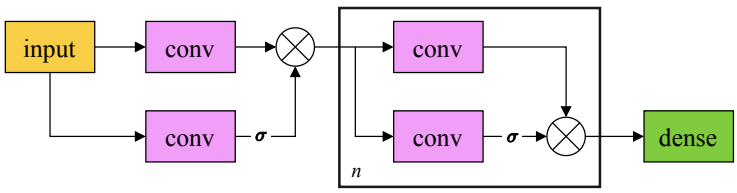
(b) Neural model with $(n+1)$ sequential convolutional attention units.



(c) Neural model with $(n+1)$ parallel convolution units with same or different kernels ($\oplus$ denotes concatenation).



(d) Neural model with $(n+1)$ parallel convolutional attention units with same or different kernels ($\oplus$ denotes concatenation).



(e) Neural model with $(n+1)$ gated convolution units ($\otimes$ denotes element-wise multiplication).

Figure 4.1: Convolutional attention neural model variants for the task of diagnostic code prediction as multi-label classification. The architecture in (d) with different kernels across parallel convolutional attention units forms the basis for the proposed *EnCAML*.

$F_k \in \mathbb{R}^{f \times e \times k}$, where, f is the number of feature maps ($\mathcal{F}_j$s) and $k \in \{3; 5; 7; 9\}$ is the kernel size. Each feature map $\mathcal{F}_j \in \mathbb{R}^L$ and the entire convolution operation over the discharge summary $\mathcal{D}^{(d)}$ results in (four) matrices H_ks of dimension $\mathbb{R}^{f \times L}$ for each kernel size k. It can be noted that pooling across the length of the summary is not performed to ensure no loss in information, i.e., different portions of the summary could be relevant to different diagnostic codes. Next, the process of diagnosis at hospitals (and manually annotating the patient records) is mimicked by narrowing down the entire discharge summary to a specific textual portion that contributes the most towards the respective diagnostic code. Towards this,

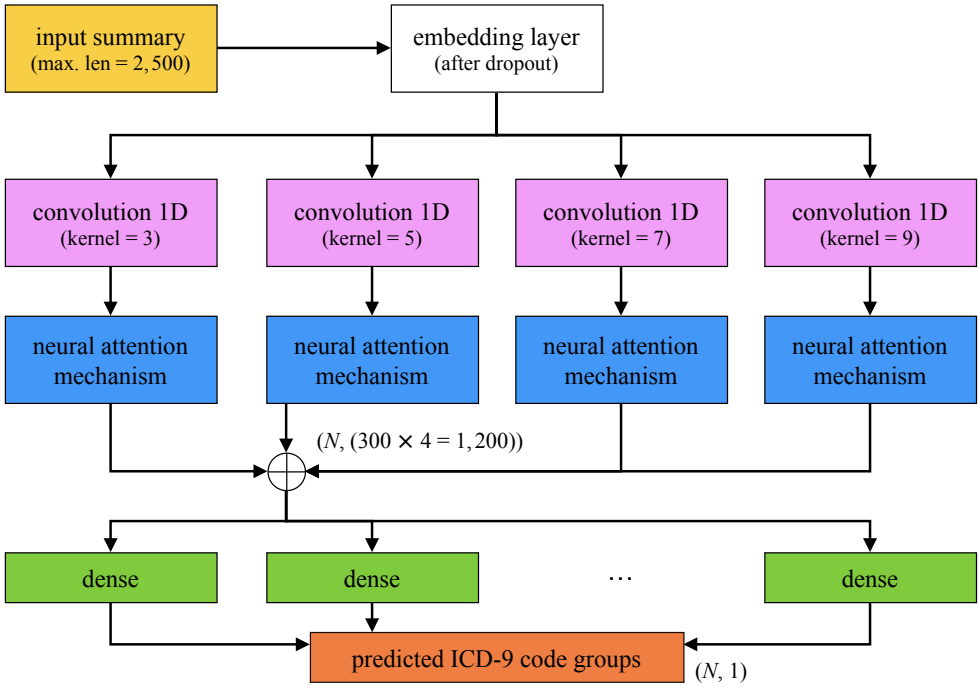


Figure 4.2: Proposed multi-channel, convolutional attention neural architecture

an attention mechanism is applied per code to highlight the text snippets in the convolution output matrices. The attention weights a_c for a code c are computed using the trainable vector parameter $u_c \in \mathbb{R}^f$ as $a_c = \text{softmax}(H_k^\mathsf{T} \cdot u_c)$. The attention weights a_c can help visualize which tokens contribute to code c. The final output representations obtained using the attention vector result in (four) matrices $A_k \in \mathbb{R}^{f \times N}$, one per kernel size, where N is the number of output codes (here, $N \in \{10; 50; 6,918\}$).

To facilitate the classification task of diagnostic code prediction, individual classifiers are built atop the $\oplus\{A_k\}\ \forall k$ vector representations ($\oplus$ denotes concatenation). modelling diagnostic codes independently instead of employing a single prediction layer is beneficial, as the model parameters are fine-tuned independently at the penultimate layer, thus enhancing the predictability of the *EnCAML* model. This way, the neural model can effectively learn and generalize what features best contribute to a particular diagnostic code. Therefore, a fully-connected layer with a sigmoid activation function is employed to facilitate binary code prediction, i.e., $\hat{y}_c = \text{sigm}(W^\mathsf{T}(H \cdot a_c) + b)$, where W and b are the corresponding weight matrix and bias vector, respectively. The neural model is trained to minimize binary cross-entropy loss using the Adam optimizer (Kingma and Ba, 2015). Additionally, an early stopping criterion to mitigate any overfitting of the model is also employed. While modelling for the prediction of diagnostic codes among all 6,918

codes, a single linear layer is employed as opposed to individual code-specific classifiers, to lower the computational overhead incurred in training a large number of independent classifiers.

The choice of the threshold (θ) on the sigmoid activation layer regulates the predictive performance of the proposed automated diagnostic coding system. Most of the existing studies (Huang *et al.*, 2019; Mullenbach *et al.*, 2018) round-up the obtained output values to the closer of 0.0 and 1.0 (i.e., an implicit threshold of 0.5), while others, including Li *et al.* (2019*b*), empirically determine the optimal threshold through experimentation with $\theta \in [0.1, 0.95]$. In this study, the Fisher-Jenks Natural Breaks algorithm (Jenks, 1967) is employed to find an optimal threshold that maximizes the predictability of $\hat{y}$. The algorithm aims at determining the most-suitable arrangement of values into different classes, i.e., the natural breaks in the data, by minimizing the intra-class variance while maximizing the inter-class variance. These natural breaks can be precomputed from the training data to be employed while testing. In this study, both *code-level* and *data-level* threshold values are computed. For instance, computing the code-level threshold for the diagnostic code `414.01` (*coronary atherosclerosis of native coronary artery*) would involve detecting the natural breaks in $\hat{y}_{\texttt{414.01}}$, i.e., the most optimal threshold that can cluster the training data into $+$ and $-$ classes. Alternately, computing the data-level threshold involves the use of $\hat{y}_c \ \forall c \in y$ to best group the input data according to the distribution of the output classes. The implementations of the algorithm available in the Python Jenkspy[10] library is employed to find the optimal classification threshold values. The generated data-level thresholds for the *top-10-code* prediction task is depicted in Fig. 4.3.

4.4.4 Word Embeddings and Predictability of *EnCAML*

The employed word embedding neural network determines the representation of the underlying clinical text, thereby effectively capturing the document's semantics. By extension, it seems intuitive to establish that vector representations capturing a higher level of document semantics (e.g., intra-word associations mined using self-attention) would outperform more simplistic approaches. However, it must be noted that a flexible and robust classification model must be capable of generalizing over minimalistic representations, as well as learning adequately from highly semantics-specific representations without over-representing the underlying patterns. To analyze the impact of the choice of initial word embedding on

[10] `https://github.com/mthh/jenkspy`.

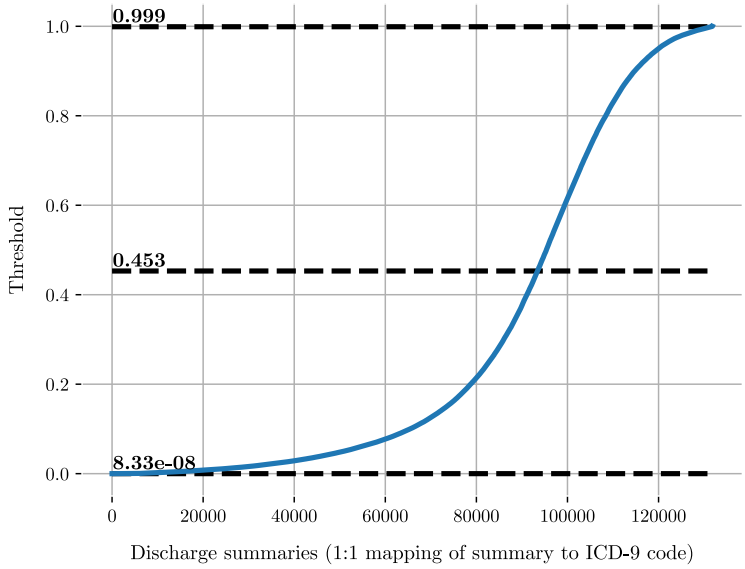


Figure 4.3: Data-level threshold values obtained using the Fisher-Jenks Natural Breaks algorithm for *top-10-code* data category.

the proposed *EnCAML* neural classification model, several state-of-the-art word embedding approaches were experimented with, including Word2Vec (skip-gram and CBoW variants), fastText (skip-gram and CBoW variants), and BERT (pre-trained on clinical corpora and fine-tuned).

The Word2Vec (or a close variant) neural model for generating word embeddings (Mikolov *et al.*, 2013) has been widely employed in modelling clinical text across several state-of-the-art studies (Li *et al.*, 2019*b*), owing to its ability to capture the text semantics in a simple yet efficient manner. However, models reliant on Word2Vec approaches often cluster all the out-of-vocabulary words into a single vector representation, defaulted for all unknown tokens. In this regard, the more flexible fastText neural model (Bojanowski *et al.*, 2017) aims at representing the unknown tokens as some combination of known sub-tokens, thus overcoming the limitations of the Word2Vec model. Finally, a more advanced self-attention-based BERT model (Devlin *et al.*, 2019*a*) captures the context of the given token from both left-to-right and right-to-left, aiming to extract the exact intended semantics of the underlying text, which would otherwise go unnoticed. The BERT embeddings are obtained for the entire MIMIC-III discharge summaries using service framework that hosts BERT as a sentence encoder service. The pre-trained checkpoints obtained while training on clinical texts, released by Alsentzer *et al.* (2019), and those obtained while training on PubMed abstracts, released by Peng *et al.* (2019), are used to start the service and thereby avoid computationally in-

tensive re-training. Word2Vec and fastText models are employed through the implementations available in the Python Gensim library (Řehůřek and Sojka, 2010), while BERT embeddings are generated using the openly available BERT-as-service framework[11].

Since the BERT (base model) outputs a vector embedding of 768 dimensions, the same embedding size is employed while modelling Word2Vec and fastText models for comparison. Furthermore, the Word2Vec and fastText embeddings are deployed with a window size of five, trained for 30 iterations over the corpus. The obtained performance (measured as micro F_1 score) of proposed *EnCAML* model for various neural embeddings is presented in Table 4.4. Additionally, the micro F_1 score obtained using a random Xavier uniform initialization of $768-$dimensional vector per token as the baseline, is presented also in Table 4.4. It can be observed that the CBoW variants of Word2Vec and fastText models always outperform their skip-gram counterparts. One possible interpretation for this behaviour could be that predicting a target word, given the neighbouring noisy context, is far simpler than predicting the exact noisy context for a given target token. Despite the fastText CBoW variant achieving the highest performance, the speedup obtained for Word2Vec models is nearly ten-fold ($10\times$) at a similar (i.e., insignificantly lower) performance. There was no drastic improvement in performance when the FasText model is used compared to Word2Vec. This could be attributed to the fact that medical terms cannot be split into many subwords. Considering these findings, vector representations output by the CBoW Word2Vec were chosen for the embedding layer to model the input discharge summaries.

Xavier uniform initialization at random, also provides comparable performance with respect to other more sophisticated models, as observed from Table 4.4. This corroborates that the values of initial embedding vector components play little to no role in enhancing the predictability of the *EnCAML* model. The proposed *EnCAML* model employs multiple attention layers, thus enabling the learning of per-code attention weights over training samples. Therefore, initialization of input vectors with pre-trained embedding weights is quite redundant and cost-intensive (requiring additional storage space of up to 1.5 GiB). The robustness of the proposed *EnCAML* model over other state-of-the-art models lies in its ability to learn from and generalize over the input discharge summaries in a rather end-to-end fashion. Hence, it is arguable that such a system could enable rapid prototyping and deployability in real-world scenarios, especially in modelling noisy clinical data obtained from the hospitals of developing nations, which are far less

[11] `https://github.com/hanxiao/bert-as-service`.

Table 4.4: Effect of initial word embedding choice on the overall predictive performance of the proposed *EnCAML* model, recorded using discharge summaries of the *top-10-cat* data category.

Embedding model	F_1 **micro**
Skip-gram Word2Vec	0.7784
CBoW Word2Vec	**0.7811**
Skip-gram fastText	0.7811
CBoW fastText	**0.7821**
Fine-tuned BERT (clinical texts + PubMed abstracts)	0.7760
Pre-trained BERT (Alsentzer *et al.* (2019))	0.7729
Xavier uniform initialization	0.7668

ideal than the standard datasets utilized in academic research.

4.5 Experimental Results and Discussion

Extensive performance evaluation was undertaken, both in terms of predictability and interpretability, on extracted MIMIC-III datasets. The proposed *EnCAML* deep neural model was implemented using Python PyTorch library (Paszke *et al.*, 2019*a*). All the experiments, training, and validation were performed using a server running Ubuntu OS with 56 cores of Intel Xeon processors, 128 GiB RAM, 3 TB hard drive, and two NVIDIA Tesla M40 GPUs, running CUDA v10.1. The model hyperparameters were tuned using optimal values obtained from prior studies and verified through experimental validation. The results of the hyperparameter tuning are summarized in Table 4.5. The predictive and interpretive superiority of the proposed *EnCAML* approach was validated over several state-of-the-art benchmarks with the chosen optimal hyperparameters.

The datasets were grouped into train, validation, and test sets exactly as reported by the respective state-of-the-art studies to enable accurate benchmarking of the obtained performance. For datasets with diagnostic codes and rolled-up categories (*top-10-code*, *top-50-code*, *top-10-cat*, and *top-50-cat*), the 50-25-25 split from the train-validation-test-HADM_IDs set reported in (Huang *et al.*, 2019) was employed. Also, while modelling the *top-50-dp-code* dataset, the hospital admission identifiers from the train_50-HADM_IDs set reported in (Mullenbach *et al.*, 2018) were employed. For the code prediction task employing all $6,918$ codes (*all-*

Table 4.5: The hyperparameter ranges and the experimentally-determined optimal values for the proposed *EnCAML* neural model (‖ denotes parallel operation).

Hyperparameter	**Experimental value(s)**	**Optimal value(s)**
Embedding sizes (e)	$\{50; 100; 200\}$	100
Kernel sizes (k)	$\{1 \parallel 3 \parallel 5 \parallel 10; 3 \parallel 5 \parallel 7 \parallel 9\}$	$3 \parallel 5 \parallel 7 \parallel 9$
No.of feature maps (f)	$\{100; 200; 300; 400\}$	300
Dropout probabilities	$\{0.2; 0.3; 0.5; 0.8\}$	0.2
Learning rates	$\{1e{-}4; 3e{-}4; 1e{-}3; 3e{-}3\}$	$1e-4$
Exponential decay rates	$\beta_1 = 0.9$; $\beta_2 = 0.999$	$\beta_1 = 0.9$; $\beta_2 = 0.999$

codes dataset), a 90-to-10 train-to-test split is used, enabling maximum training instances to ensure model generalizability across a large number of target classes. As stated earlier, the early stopping criterion (tolerance of five epochs) is incorporated while training, to overcome possible overfitting of the deep neural model.

For validating the performance of the proposed model, the standard metrics of micro-averaged and macro-averaged F_1 scores (Tsoumakas *et al.*, 2010) were employed. The F_1 (more generally, $F_{\beta=1}$) aims to seek a balance between precision and recall and is interpreted as a weighted harmonic mean of the two. Therefore, models with relatively higher F_1 scores are expected to enhance the predictability of the system. Since the F_1 measure accounts for the true and false positives (TP and FP) as well as true and false negatives (TN and FN), it is often regarded to be more indicative than the standard accuracy score.

$$F_{\beta=1} = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}} \tag{4.2}$$

where, the values of precision and recall are micro-averaged or macro-averaged over the target output classes (here, codes and code categories).

Finally, the performance of *EnCAML* is also benchmarked using micro-averaged and macro-averaged Area Under the Receiver Operating Characteristic curve (AUROC). Since the ROC curve is a probability curve (plotted as sensitivity against the fall-out), the area under the curve represents the measure of class separability, i.e., a quantitative measure of the capability of the model in distinguishing between target codes or code categories. By analogy, the higher the value of the AUROC value, the better the model is at distinguishing between patients with

and without corresponding diseases.

The performance of *EnCAML* model is benchmarked against several state-of-the-art studies. As stated earlier, six data categories from the obtained MIMIC-III corpus are curated to facilitate exhaustive comparison. For the *top-k-code* ($k = 10, 50$) data categories, the discharge summaries mapped to the top$-k$ ICD-9 diagnostic codes are employed in benchmarking. On the other hand, for the *top-k-cat* ($k = 10, 50$) data categories, the ICD-9 diagnostic codes are rolled-up to three digits (e.g., `225.2` (*benign neoplasm of cerebral meninges*) and other codes within the `225.x` class are rolled-up into the `225` category (*benign neoplasm of brain and other parts of nervous system*)) and extracted the discharge summaries corresponding to top$-k$ categories. The performance benchmark is presented using the combined set of most-frequent diagnostic and procedural ICD-9 codes just like most of the existing works, represented as *top-50-dp-code* data category. Finally, the *EnCAML* model is evaluated, which is trained on all the 6,918 ICD-9 diagnostic codes observed in the obtained MIMIC-III cohort, under the *all-codes* data category.

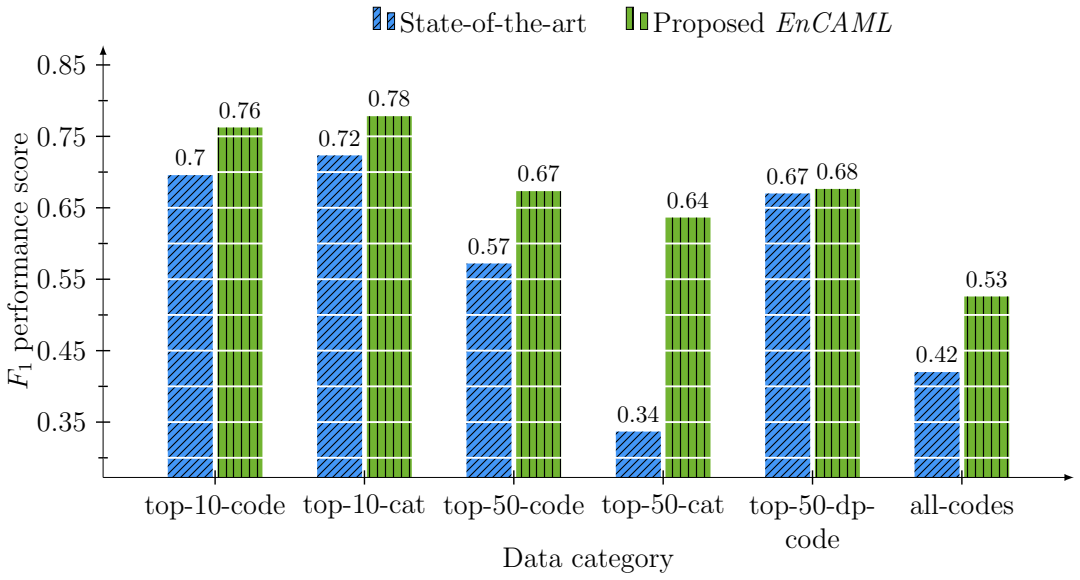


Figure 4.4: Benchmarking the proposed *EnCAML* approach against best-performing state-of-the-art models for the corresponding curated data category (measured as macro or micro F_1 score)

From the neural model training perspective, the *EnCAML* model starts to converge (performance saturation) between 34 to 36 epochs, while the single-channel convolutional attention model proposed by Mullenbach *et al.* (2018) takes twice as long (i.e., 63 to 65 epochs). This fast convergence is attributed to the use of

Table 4.6: Performance benchmarking of the proposed *EnCAML* model against state-of-the-art works across all data categories.

Data category	**Study (model)**	F_1 **score**	
		macro	**micro**
top-10-code	**Proposed *EnCAML***	**0.7624**	**0.7772**
	Huang *et al.* (2019) *(GRU)*	0.6957	–[§]
	Samonte *et al.* (2018) *(hierarchical attention + topic modelling)*	0.6870	–[§]
	Rios and Kavuluru (2019) *(transfer learning)*	0.6200	–[§]
top-10-cat	**Proposed *EnCAML***	**0.7782**	**0.7840**
	Huang *et al.* (2019) *(GRU)*	0.7233	–[§]
top-50-code	**Proposed *EnCAML***	**0.6028**	**0.6733**
	Huang *et al.* (2019) *(GRU)*	0.3263	–[§]
	Guo *et al.* (2019) *(bidirectional LSTM)*	–[§]	0.5720
top-50-cat	**Proposed *EnCAML***	**0.6363**	**0.6908**
	Huang *et al.* (2019) *(LSTM)*	0.3367	–[§]
top-50-dp-code	**Proposed *EnCAML***	**0.6109**	**0.6764**
	Mullenbach *et al.* (2018) *(description-regularised CAML)*	0.5760	0.6330
	Mullenbach *et al.* (2018) *(single-channel CAML)*	0.532	0.6140
	Li and Yu (2020) *(multi-filter residual ConvNets)*	0.6060	0.6700
all-codes	**Proposed *EnCAML***	**0.0859**	**0.5258**
	Zeng *et al.* (2019*b*) *(transfer learning)*	–[§]	0.4200
	Li *et al.* (2019*b*) *(Doc2Vec + ConvNet + $\theta = 0.2$)*	–[§]	0.4080
	Baumel *et al.* (2018*a*) *(ConvNet)*	–[§]	0.4070

[§] *These metrics were not reported by the authors.*

multi-channel convolutions and per-label classifiers as opposed to a single linear layer. Furthermore, the task of multi-label classification of N diagnostic codes or code categories, facilitated through N binary classifiers, enables the neural model to generalize over relevant features that correspond to the underlying code more effectively. On the aspect of extracting features from the given data, the multi-channel variable-sized convolution filters extract crucial information from the underlying discharge summary at varying contexts, which are then searched attentively (through neural attention) for vital portions that are responsible for the corresponding output diagnostic code. The use of multi-channel convolution instead of a fixed-length filter enhances the model's flexibility in choosing the context of representation and relies entirely on the attention layer to segregate between the convolved outputs. Employing a pooled convolution output (as opposed to an attention-based aggregation) often results in a loss of information (relevant features corresponding to specific code labels), especially when classifying data with a large number of sparse and diverse target labels (e.g., *all-codes* data category), as observed with the use of traditional ConvNet models in (Li *et al.*, 2019*b*) and (Baumel *et al.*, 2018*a*). Additionally, the *EnCAML* model facilitates an unrestricted use of variable-sized filters resulting in variable-sized contexts that are weighed by attention, enhances the interpretability of the obtained neural predictions to a large extent.

An argument towards the effectiveness of recurrent neural models such as LSTM or GRU in modelling text corpora could be presented, as they have been shown to very well capture the dependencies within natural language text. However, in this case, most of the discharge summaries range between 500 to $2,500$ tokens in length (after truncating), thus, sequence models could experience severe vanishing gradient problems. The *EnCAML* model with multiple convolutional layers was able to adequately cope with such issues, as is evident from the observed high performance of *EnCAML* when compared to GRU (Huang *et al.*, 2019), LSTM (Huang *et al.*, 2019), and bidirectional LSTM (Guo *et al.*, 2019) models. Additionally, employing more sophisticated neural models such as BERT to handle the limitations with recurrent networks is also challenging, especially due to the high computational cost of training, exacerbated by its fixed input sequence length of 512 tokens (lower end of the discharge summaries length range), warranting additional runs to accommodate longer texts. It could be observed that the performance of *all-codes* was decreased owing to a large number of sparse and diversified target labels. Additionally, 96.79 percent of discharge summaries are covered by the top 50 codes, indicating that there are insufficient training

Table 4.7: Sample discharge summaries from the MIMIC-III corpus with vague and unusable information with respect to the mapped ICD-9 diagnostic codes, illustrating the intrinsic complexities in modelling unstructured clinical data.

ICD-9 code(s)	**Discharge summary**
`584.9`: Acute renal failure, unspecified	... see outside medical records for history of present illness, physical examination, pertinent laboratories, x-ray electrocardiogram, and other tests ...
`428.0`: Congestive heart failure, unspecified	... her discharge was delayed one day due to bed unavailability at rehab ...
`427.31`: Atrial fibrillation	... please see discharge summary record from outside medical record notes ...
`998.32`: Disruption of external operation (surgical) wound	... this addendum will serve to confirm that in addition to the previous discharge summary the admission diagnosis should be included ...
`401.9`: Unspecified essential hypertension	... this is an addendum to the initial discharge summary which was dictated when the patient remained in the hospital awaiting appropriate rehabilitation facility ...
`V45.81`: Aortocoronary bypass status	... please refer to the discharge summary dictated by myself with discharge date for content ...

summaries for the majority of uncommon codes.

The discharge summaries of the MIMIC-III database corresponding to the misclassifications from the *EnCAML* model were analyzed, in an attempt to explain the predictions output by the model. For the more severe false-negative scenarios (existing disease goes unidentified), it was observed that several discharge summaries under this category include minimal disease-specific reference text and several links to alternate sources of patient-specific information such as nursing notes or outside medical records. With little to no diagnostic-code-specific text in the underlying summary, the *EnCAML* model is unable to provide conclusive predictions. Several such sample discharge summaries and their associated ICD-9 diagnostic codes are documented in Table 4.7. In the more tolerant false-positive cases (nonexistent disease gets marked-up), the discharge summaries included pro-

longed patient histories that signaled the *EnCAML* model to mark-up the content within the history as evidence to predict the corresponding nonexistent ICD-9 diagnostic code as existent. Specific examples of discharge summaries falling into the false-positive category are highlighted in Table 4.9.

4.5.1 Evaluation of Interpretability

Table 4.8: Examples of patient discharge summaries extracted from the MIMIC-III database establishing the interpretability and explainability of *EnCAML*. The text snippets indicating the possibility of the respective ICD-9 diagnostic code in the discharge summary are highlighted in blue.

Parameter	Value
Extracted $n-$grams using attention weights	... mass he *received units of packed red blood* cells ... discharge diagnosis *upper gastrointestinal bleed* discharge ...
Extracted $n-$grams using Grad-CAM	... presented with **hematocrit drop** and had guaiac ... mass he **received units of packed red blood** cells ...
Top-3 tokens	bleed, drop, and hematocrit
Associated ICD-9 code	`285.1`: Acute posthemorrhagic anemia
Extracted $n-$grams using attention weights	... a history of *hypothyroidism morbid obesity* polycystic ovarian ... in the *evening levothyroxine* mcg oral ...
Extracted $n-$grams using Grad-CAM	... on **exertion paroxysmal nocturnal dyspnea orthopnea** ankle ... in the **evening** **levothyroxine** mcg oral ...
Top-3 tokens	levothyroxine, hypothyroidism, and levoxyl
Associated ICD-9 code	`244.9`: Unspecified hypothyroidism

This subsection presents details on the interpretability of the diagnostic code predictions facilitated by the *EnCAML* model, specifically through the attention layers of the neural model trained at the individual diagnostic code level. Table 4.8 presents sample patient discharge summaries extracted from the MIMIC-III database whose content is highlighted using the learned attention weights (a_cs) corresponding to the respective diagnostic code c. These highlighted tokens are considered most contributing towards the corresponding ICD-9 code by the *EnCAML*

model, and Table 4.8 also presents the top−3 tokens that are highly weighted by the neural system. The visualization of the text snippets demonstrates the effectiveness of the *EnCAML* model in learning the most relevant and vital keywords adequately to facilitate enhanced predictability of the corresponding ICD-9 codes. In cases of summaries containing extended patient histories with minimal disease-specific indicators, the attention mechanism seems to classify the patient history as if it were the current illness. Examples of such discharge summaries extracted from the MIMIC-III database, resulting in false-positive predictions, are tabulated in Table 4.9.

To benchmark the interpretability and explainability of the proposed *EnCAML* approach, the resultant attention output for a discharge summary was compared to that obtained using the Gradient-weighted CAM (Grad-CAM) (Selvaraju *et al.*, 2017) approach. Grad-CAM employs the gradients of a target class, flowing into the final convolution layer (before the attention layers in *EnCAML*), to produce a localization map highlighting the important candidate $n-$grams in the underlying summary for predicting the corresponding code. Since Grad-CAM allows for the visualization of all possible contributing $n-$grams, it spans a much broader aspect than the attention outputs of the *EnCAML* model. However, on the flip side, because attention outputs are quite narrowed down, they are more precise and depict accurate understanding of what the underlying deep neural model looks at. Upon experimentation, it was observed that the Grad-CAM and attention outputs were quite similar for most of the discharge summaries (see Tables 4.8 and 4.9 to compare attention and Grad-CAM outputs). The model interpretability is compared between *EnCAML* and the single-channel convolutional attention network proposed by Mullenbach *et al.* (2018) employing a kernel size $k = 10$. More recent studies (Teng *et al.*, 2020; Vu *et al.*, 2020) facilitated enhanced learning from external data sources such as Wikipedia knowledge, in addition to training on the discharge summaries, and showed some improvements in the predictability of the system. However, such external-data-based boosting approaches often trade-off model interpretability for higher prediction accuracy. The mappings between the underlying clinical text and the corresponding diagnostic codes are often blurred in such models. For clinical decision support systems to be adaptable in real-world scenarios, providing an explainable decision (even when incorrect) is far more acceptable than just producing a highly accurate black-box decision.

Table 4.9: Predictability and interpretability of *EnCAML* for sample patient discharge summaries extracted from MIMIC-III. Here, the predicted false-positive ICD-9 codes (shown as ~~strikethrough~~ text) are identified from the text snippets (marked in red; in the first summary), `401.9` corresponds to the term *hypertension*; in the second summary, `414.01` corresponds to the terms *coronary artery disease* and *cardiac catheterization*.

Parameter	Value
Extracted $n-$grams using attention weights	... complaint *giant paraesophageal hernia* major ... past medical history of pulmonary *hypertension* *depression lyme disease osteopenia* ...
Extracted $n-$grams using Grad-CAM	... diagnosis giant paraesophageal **hernia gerd** **hypertension** **osteopenia depression** ...
Predicted ICD-9 code(s)	`311`: Depressive disorder, not elsewhere classified `530.81`: Esophageal reflux `401.9`: ~~Unspecified essential hypertension~~
Actual ICD-9 code(s)	`518.81`: Acute respiratory failure
Extracted $n-$grams using attention weights	... and family history of *coronary artery disease* ... who presents for *cardiac catheterization* to evaluate ...
Extracted $n-$grams using Grad-CAM	... past medical history of prostate brachytherapy years ago ... and underwent **aortic valve** replacement ...
Predicted ICD-9 code(s)	`39.61`: Extracorporeal circulation auxiliary to open heart surgery `401.9`: Unspecified essential hypertension `414.01`: ~~Coronary atherosclerosis of native coronary artery~~
Actual ICD-9 code(s)	`39.61`: Extracorporeal circulation auxiliary to open heart surgery `401.9`: Unspecified essential hypertension

4.6 Summary

Effective coding of patient records in hospitals is an essential requirement for epidemiology, billing, and managing insurance claims. The prevalent practice of manual coding, carried out by trained medical coders, is often error-prone and time-consuming. Mitigating this labor-intensive process by developing diagnostic coding systems built on patients' EHRs is vital. However, developing nations

with low digitization rates have limited availability of structured EMRs, thereby necessitating a need for systems that leverage unstructured data sources. Despite the rich clinical information available in such unstructured data, modelling them is complex, owing to the variety and sparseness of diagnostic codes, the complex structural and temporal nature of summaries, and the prolific use of medical jargon. In this work, a context-attentive network was proposed to facilitate automatic diagnostic code assignment as a multi-label classification problem. The proposed model facilitates information aggregation across a patient's discharge summary via multi-channel, variable-sized convolutional filters to extract multi-granular snippets. The attention mechanism enables selecting vital segments in those snippets that map to the clinical codes. The model's superior performance underscores its effectiveness compared to the state-of-the-art on the MIMIC-III database. Finally, the enhanced interpretability of the prediction output of the EnCAML model was demonstrated using the learned per-code attention weights, thereby establishing the impact of the proposed model on instigating trust in intelligent healthcare systems.

Publications

(based on works presented in this chapter)

1. Veena Mayya, Sowmya Kamath S., Gokul S. Krishnan, Tushaar Gangavarapu, "*Multi-channel,convolutional attention based neural model for automated diagnostic coding of unstructured patient discharge summaries.*", Future Generation Computer Systems, Elsevier, Volume 118, 2021, Pages 374-391, ISSN 0167-739X, (SCI & Scopus, IF: 7.187) *(Online)*

Chapter 5

Automated Diagnostic Coding of Non-English Clinical Notes

5.1 Introduction

In the previous chapter, a detailed discussion of the methodologies designed for the ICD-9 coding task was presented. In 2019, the ICD-10[1] medical coding taxonomy (World Health Organization, 2004) was introduced, which is already widely adopted in hospitals to describe patients' procedural and diagnostic conditions. ICD-10 codes can represent the usage of many new and contemporary technologies, while ICD-9 codes often cannot to mapped to these or are completely missing. ICD-10 provides additional specificity, such as episode of treatment, body areas, etc. ICD-10 has a greater number of diagnostic codes (69,823) than ICD-9 (14,025) (Mainor *et al.*, 2019). Additionally, the ICD-10 update included combination codes (e.g., E10.21, diabetes mellitus with nephropathy) for the use of new technology and possibilities for specifying unique care contexts. The amount of time coding trainees can dedicate to studying ICD-10 is an issue, as well as the expertise of the trainers. It is also challenging to maintain multiple coding systems, depending on hospital strategies.

In view of the complications and wide spectrum of codes available, automated ICD-10 code and group assignment based on patients' case reports have recently received significant research interest. While English is the most common language in which patient notes, symptoms, etc., may be recorded by caregivers, a large volume of patient data may be in native languages. Modelling native language characteristics in use and the diversity of languages spoken globally poses a sub-

[1]International Statistical Classification of Diseases and Related Health Problems, tenth revision; accessible at `https://icd.who.int/browse10/2019/en`.

stantial challenge. Another significant challenge is scaling manual interpretation to an ever-expanding clinical texts. The use of automated CDSS to aid coding specialists is becoming more important in order to address these challenges. A recent effort to promote the development of such systems, CodiEsp (Miranda-Escalada *et al.*, 2020), comprises 1,000 clinical cases in English and Spanish languages that experts manually annotated with 2557 diagnostic and 870 procedural ICD-10 codes. The Spanish clinical cases provide minimal text evidence for supporting the clinical code mappings. In recent years, non-English clinical texts have been made publicly available for Japanese (Aramaki *et al.*, 2016), French (Goeuriot *et al.*, 2017) and German (Kelly *et al.*, 2019). Thus, there is ample scope for developing well-defined interpretable methodologies for modelling unstructured clinical texts written in languages other than English.

5.2 Problem Definition

Clinical coding consists of mapping medical documents into a structured or coded format utilizing globally accepted taxonomy and class codes (numeric or alphanumeric format). These codes often convey information about a patient's condition, symptoms, diagnosis, treatment, or reason for seeking medical assistance. This conversion of natural language clinical notes into structured data is crucial for eventual use in clinical treatment and other applications such as statistical analysis and decision-making, billing, or reimbursement. Clinical notes may be written in a native language (other than English) and may cover a range of medical topics. Medical coders convert patient information into a set of suitable medical diagnostic codes in hospitals. Due to the diversity of languages, such human coding techniques are very expensive, often inexact, time-consuming, and prone to errors. It is critical to build intelligent computational systems that address these issues by automating the coding of unstructured patient clinical notes written in non-English languages. Additionally, for these systems to be accepted and used in clinical setup, their output must be interpretable and explainable. The problem to be addressed here is defined as follows:

> *Given the challenges of lengthy clinical notes written in the native language, medical jargon, large number ICD codes, & manual effort, design and develop approaches for effective automated ICD-10 code assignment using unstructured non-English clinical notes.*

5.3 Motivating Example

To describe the prevailing conditions that emphasize the need for disease prediction CDSS based on unstructured non-English clinical notes, consider scenarios where a hospital has a full-fledged EHR system, and the clinical notes are recorded by physicians as notes in their native languages. The Medical Records Department (MRD) staff scans through the entire document and assigns, with probability, the ICD diagnostic and procedural codes for a particular patient. Scanning over the entire document is a time consuming task, and coding efficiency is often dependent on the expertise of trained medical coders. The MRD team should have a working knowledge of the native language in order to comprehend clinical notes written in that language. This delay in processing and code assignment could be avoided if the ICD codes could be automatically generated using the clinical notes written in the native language. The automated disease prediction CDSS could directly process the unstructured clinical notes recorded by the physician and provide them with a list of ICD codes along with highlighting of the texts that the system found relevant for predicting each of the codes. In this way, the MRD staff need not perform conversion to any predefined structure and, hence, has the advantage of significant savings in person-hours and reduced costs due to inaccurate code assignment.

In this chapter, various approaches for developing effective patient-specific ICD-10 code assignment models built on unstructured clinical notes are presented. The contributions towards the defined problem are in the context of designing methods that can automatically process a variety of multilingual unstructured clinical notes, effectively capturing the differences in notation, usage of extensive medical jargon, acronyms, etc., and still be able to extract relevant disease-specific features, which can be leveraged for automatic ICD-10 code prediction. The proposed model was benchmarked against state-of-the-art ICD-10 code assignment built on unstructured patient clinical notes to evaluate their adaptability and prediction performance.

5.4 $\mathcal{LATA}$ Model for ICD-10 Coding of Unstructured Clinical Notes

A wide variety of methods have been used for preprocessing and modelling clinical text for enabling ICD-10 assignment. Bidirectional Encoder Representations from

Transformers (BERT) (Devlin *et al.*, 2019*b*) has been adapted for several NLP tasks and has been shown to achieve state-of-the-art results in tasks such as text classification, question answering, named-entity recognition, machine translation. BERT is built on an architecture consisting of a stack of transformers, forming an encoder-decoder network. The encoder employs self-attention on the input text. The decoder could be designed as per the task at hand and the application to be supported.

Several approaches aiming to improve the basic BERT model with respect to its computational speed or prediction metrics have been proposed. XLNet (Yang *et al.*, 2019) enables learning bidirectional contexts by maximizing the expected likelihood of the factorization order to improve NLP task performance. DistilBERT (Sanh *et al.*, 2019) used a combination of language modelling, distillation and cosine-distance losses and a lighter BERT model that retains 97% of the BERT's language understanding capabilities. RoBERTa (Liu *et al.*, 2019*b*) fine-tunes the training hyperparameters of basic BERT for improved learning capabilities. DeBERTa (He *et al.*, 2020) includes a disentanglement of the attention mechanism to further enhance the mask decoder and achieve improved performance. MobileBERT (Sun *et al.*, 2020) included a knowledge transfer mechanism for further enhancement of the basic BERT model. Parameter reduction techniques are incorporated in ALBERT (Lan *et al.*, 2020) to lower memory consumption of BERT. Most BERT variants use the hidden state representation of the first token ([CLS]) for sentence classification tasks. This output is usually not a good summary of the input's semantic content due to the unavailability and also limitations of pre-trained BERT variants that use large medical text corpus in non-English languages.

In contrast to Miranda-Escalada *et al.* (2020)'s work, the use of label attention in BERT and its variants is seen as a potential way to improve input context learning for the given output classes. This is well-suited for the NLP classification task when the input training samples are limited, but the number of output classes is large, as is the case with the ICD-10 code assignment. Another key contribution is the use of a single tokenizer and similar configuration hyper-parameters for all BERT variants. This allows an in-depth understanding of the variations in the number of training parameters for each BERT variant. The proposed $\mathcal{LATA}$ (Label Attentive Transformer Architectures) model performs automated ICD-10 code assignment using the patients' case reports. The changes in the predictive performance of code assignment with reference to each BERT variant are analyzed in detail.

5.4.1 Data Preprocessing and Feature Extraction

The ICD-10 code prediction model was benchmarked on the patient cases available in the CodiEsp database. As CodiEsp consists of only 500 case reports for training, 250 for validation and 250 for testing, only diagnostic ICD-10 codes were considered. Along with train 500 cases, 195 validation cases that include the unique codes that are not in the train data split. With 695 training cases, the individual sentences (until the previous '.' and next '.') that include the annotated text are extracted.

Next, multiple labels are assigned for a sentence if the sentence includes the corresponding annotated texts. For example: *"de 74 años que ingresó en el hospital por obnubilación anuria tras presentar durante días dolor abdominal vómitos"* is assigned `r34:anuriar r52:dolor, r11.10:vómitos, r10.9:dolor abdominal` and `r40.1:obnubilación` ICD-10 codes. That included 4,745 full and short case reports, along with stemmed case reports for each case report to improve the proposed model's generalization. The 4,745 case reports are augmented using Google translation API by translating the document initially to English, French, Portuguese, Italian & German languages and then convert it back to Spanish. In total, that gives us 33,215 (4,745 x 7) case reports from the initial 695 training cases. Surprisingly, with this text augmentation approach, less than 0.05% of the total 33,215 case reports are repeated, mainly for concise case reports. Other basic preprocessing techniques were applied to deal with the limited vocabulary size. These include adding of space after each punctuation, removal of most ($>$ 13000 times) and least ($<$ 15 times) frequently occurring terms, single character terms, digits only terms etc. In total, this retains about 10,329 unique words in the final vocabulary, which is utilized for further analysis.

5.4.2 Clinical Text Modelling

The $\mathcal{LATA}$ model is built on existing variants of the BERT model. In addition, an attempt is also made to address the need for explainable CDSSs, through the visualization of attention weights learnt by $\mathcal{LATA}$ that reveal the associations of clinical note text with the predicted diagnostic code, further enhancing the trustworthiness of the proposed CDSS. The proposed $\mathcal{LATA}$ ICD-10 diagnostic code prediction model with label attention is an extension of the basic BERT model (see Fig. 5.1). Label attentions are applied for each BERT layer outputs. The concatenated result is then fed into the fully connected layer with the sigmoid activation function to predict the output ICD-10 codes. Multi-label soft margin

loss with Adam optimizer (learning rate of 0.001) is used to train $\mathcal{LATA}$. Early stopping is incorporated using precision on validation split to overcome overfitting issues, thereby saving the best model for predicting ICD-10 codes effectively for unseen case reports.

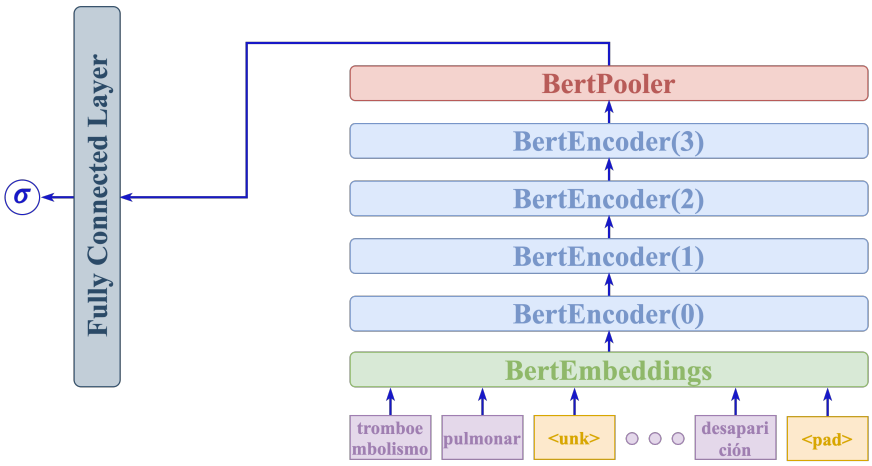


Figure 5.1: Architecture of basic BERT model.

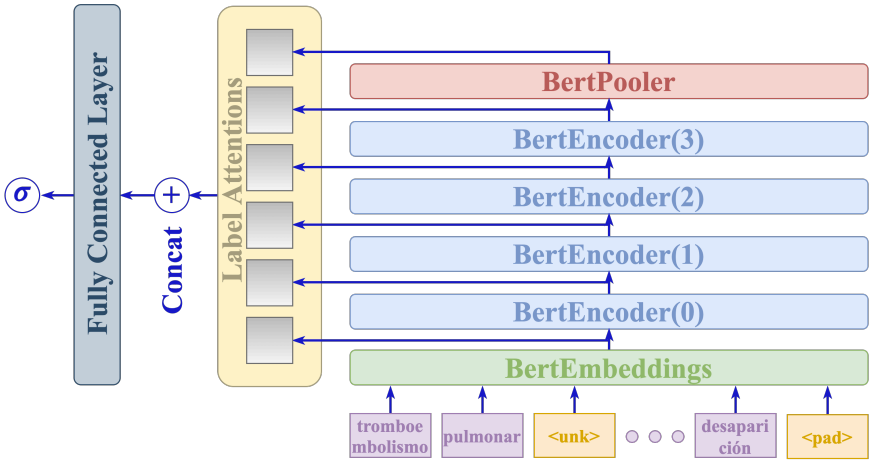


Figure 5.2: Proposed $\mathcal{LATA}$ model for ICD-10 code prediction.

The detailed high level architecture of $\mathcal{LATA}$ for basic BERT model is shown in Fig. 5.2. Let $\boldsymbol{BO} = [b_1, b_2, ..b_L]$ represent BERT layer outputs obtained with L hidden layers ($L = 4$ was used in this study). Each $\boldsymbol{b_i}$ represents the BERT layer output (including the embedding layer output) and $\boldsymbol{b_i} \in \mathbb{R}^{N \times h}$ where, N is length of the case reports and h is hidden size (embedding size) given for configuring BERT model. In ICD-10 coding task, multiple codes are assigned for each case report, and different parts of the report may be relevant for different codes. To

handle this, per-label attention was applied for each BERT layer output. i.e. $\alpha_i = softmax(\boldsymbol{A_{iweight}} \boldsymbol{b_i^T})$ and $O_i = \alpha_i \boldsymbol{b_i}$, where $\boldsymbol{A_{iweight}}$ is the label attention weight and $\boldsymbol{A_{iweight}} \in \mathbb{R}^{C \times h}$, for C codes (in this study, C=2194). Here $\alpha_i \in \mathbb{R}^{C \times N}$, so the label attention output dimension is given by $O_i \in \mathbb{R}^{C \times h}$. With this approach, for any length textual inputs, the model learns the label-wise contexts. All O_i's are then concatenated and given for final classification dense layer i.e. $\hat{y} = \text{sigmoid}(W_f^{\mathsf{T}}(\oplus O_i + b_f)$, where W_f and b_f are the final dense layer's weight matrix and bias vector, respectively.

To understand the implementation details of $\mathcal{LATA}$, consider a basic BERT with 4 hidden layers and an embedding size of 100. Let the clinical note text length for a batch be 2000 mapped to 50 ICD codes. In LATA, for every layer hidden output, the final output and encoding layer output (6 layers of dimension [2000, 100] each) label-wise attention is applied. Let U[i] be the label attention (of dimension [100, 50]). We are performing $softmax(U[i].weight.matmul(layerOutput .transpose(1, 2)))$ which results in alpha (of dimension [50, 2000]). Then this attention alpha is multiplied with layerOutput to get the label-wise weighted output i.e $m = alpha.matmul(layerOutput)$ and will be of dimension ([50, 100]). All the six layer's m are then concatenated along the last axis to generate M (of dimension [50, 600]). The final predicted y (of dimension 50) would then be computed as $y = final.weight.mul(M).sum(dim = 2).add(bias)$.

5.5 Experimental Results and Discussion

The experimental validation of the proposed model is presented in this section. Python PyTorch library (Paszke *et al.*, 2019*a*) was used to implement $\mathcal{LATA}$. All the experiments, training, and validation were performed using a server running Ubuntu OS with 56 cores of Intel Xeon processors, 128 GiB RAM, 3 TB hard drive, and two NVIDIA Tesla M40 GPUs, running CUDA v10.1. The hyperparameters of $\mathcal{LATA}$ are listed in Table 5.1. The optimal values were determined after performing extensive experimentation on validation data. The differences in the total number of trainable parameters (in Millions) with corresponding basic BERT sentence classification models are tabulated in Table 5.2.

For the CodiEsp challenge (Miranda-Escalada *et al.*, 2020), metrics like Mean Average Precision (MAP), micro-average precision, recall and F-score were used to evaluate the challenge solutions; thus, the same metrics were adapted to assess $\mathcal{LATA}$s' performance. The evaluation script released by the CodiEsp challenge

organizers[2] was utilized to obtain the results, considering only the codes available in the training and development/validation sets. Table 5.3 lists the results obtained with basic BERT variants. The basic BERT sequence classification models use the pooled output by taking the hidden state corresponding to the first token ([CLS]). Since the models were trained using available case reports (without pre-trained weights), very less MAP (0.01) was obtained. So, instead of considering a single token hidden state, max-pooling over all the tokens' hidden state (encoder output) was used before the final Dense layer.

Table 5.1: Experimentally-determined optimal hyperparameter values for $\mathcal{LATA}$.

Hyperparameter	**Optimal value(s)**
Embedding size/ hidden size/ dim (h)	128
Number of hidden layers (L)	4
Number of attention heads	4
Intermediate layer size	64
Hidden layer size	64
Max position embeddings	1200
Dropout probabilities	0.3
Learning rates	$1e-3$
Exponential decay rates	$\beta_1 = 0.9$; $\beta_2 = 0.999$

Table 5.2: Trainable parameters (in million)

Model	**$\mathcal{LATA}$ model**	**Base model**
BERT	6.532734	3.430418
ALBERT	**6.299838**	**3.197522**
DeBERTa	6.515454	3.413138
DistilBERT	6.515966	3.413650
MobileBERT	6.985342	3.883026
RoBERTa	6.532862	3.430546
XLNet	6.428286	3.325970

In addition to ICD-10 code assignment, the extent to which the proposed $\mathcal{LATA}$ models are interpretable and explainable were also explored. For this, the

[2]https://github.com/TeMU-BSC/codiesp-evaluation-script

Table 5.3: ICD-10 diagnostic code prediction results for variants of basic BERT.

Model	Precision	Recall	F1	MAP
BERT	0.405	0.354	0.376	0.219
ALBERT	0.439	0.263	0.329	0.184
DeBERTa	0.355	0.376	0.365	0.203
DistilBERT	**0.466**	0.316	**0.377**	**0.224**
MobileBERT	0.398	0.148	0.216	0.137
RoBERTa	0.409	0.342	0.373	0.22
XLNet	0.257	**0.387**	0.309	0.187

Table 5.4: ICD-10 diagnostic code prediction results for variants of $\mathcal{LATA}$.

Model	Precision	Recall	F1	MAP
BERT	**0.728**	0.556	0.63	0.434
ALBERT	0.689	0.555	0.614	0.424
DeBERTa	0.702	0.546	0.614	0.422
DistilBERT	0.688	0.563	0.619	0.424
MobileBERT	0.483	0.186	0.269	0.163
RoBERTa	0.719	**0.582**	**0.643**	**0.441**
XLNet	0.715	0.559	0.628	0.441

correctly recognized ICD-10 diagnostic codes and the *Layer 0* attention weights (α_0) for the predicted code were extracted using the codes' index. As a total of 2,194 diagnostic codes were used, where α_0 was of dimension (2194, N). For example, if the diagnostic code 5 (out of 2194) is predicted as 1, then $\alpha_0[5]$ represents the weights assigned for a case report tokens by $\mathcal{LATA}$. The influence of each input token on the predicted diagnostic code could be directly visualized using these weights. The input tokens with top-scoring attention weights (threshold=0.2) and their locations within the input case report text were prepared as per the CodiEsp evaluation script format. As can be seen from Table 5.4, direct interpretation is possible with $\mathcal{LATA}$, without requiring further processing. The selection of high-influence input tokens from a given patient report for each predicted ICD-10 code was directly derived using code-wise attentions. Results obtained with the evaluation script (explainability) for correctly predicted ICD-10 diagnostic codes (considering only train and validation) is tabulated in Table 5.6.

Table 5.5: Case study on clinical notes from the CodiEsp corpus demonstrating the predictability and interpretability of the proposed $\mathcal{LATA}$ model.

Parameter	**Value**
Interpretation using attention weights	... Consultó por hematuria macroscópica ... se apreció una masa de 5 cm en ... producía una ureterohidronefrosis en el *riñón izquierdo*. Realizamos una RTU-biopsia, objetivando un *tumor vesical* infiltrante de alto grado. ... Se realizó una cistoprostatectomía radical laparoscópica con derivación urinaria tipo Bricker en marzo 2005 ... En la revisión postoperatoria al mes de la cirugía, se redujo la ureterohidronefrosis izquierda manteniendo cierto grado de atrofia cortical, ... logrando una buena disección pélvica e identificación del muñón uretral. ... El tiempo quirúrgico fue de 180 minutos, con un sangrado menor de 80 cc. ... así como de las potenciales secuelas en términos de trastornos metabólicos, diarrea persistente e incontinencia y/o necesidad de autocateterismos por alto residuo post-miccional. ...
Actual expert-provided text evidence	hematuria, hematuria macroscópica, masa riñón, ureterohidronefrosis, sangrado, diarrea, incontinencia, *riñón izquierdo tumor*, *tumor vesical*
Predicted expert-provided text evidence	hematuria, hematuria macroscópica, masa, ureterohidronefrosis, sangrado, diarrea, incontinencia, RTU pélvica, cistoprostatectomía, atrofia
Predicted ICD-10 code(s)	`r31.9:` hematuria, unspecified, `r31.0:` gross hematuria, `n28.89:` other specified disorders of kidney and ureter, `n13.30:` unspecified hydronephrosis, `r58:` hemorrhage, not elsewhere classified, `r19.7:` diarrhea, unspecified, `r32:` unspecified urinary incontinence, `d49.4:` neoplasm of unspecified behavior of bladder, `c67.9:` malignant neoplasm of bladder, `n26.1:` atrophy of kidney (terminal)
Actual ICD-10 code(s)	`r31.9:` hematuria, unspecified, `r31.0:` gross hematuria, `n28.89:` other specified disorders of kidney and ureter, `n13.30:` unspecified hydronephrosis, `r58:` hemorrhage, not elsewhere classified, `r19.7:` diarrhea, unspecified, `r32:` unspecified urinary incontinence, `d41.02:` *neoplasm of uncertain behavior of left kidney*, `d41.4:` *neoplasm of uncertain behavior of bladder*

Note: *The colour legend used corresponds to attention weights for the predicted ICD-10 codes. The text in shades of green indicate true-positive predictions, while false-positive text predictions are marked using shades of red. The italic text represents the expert annotated texts which are not identified by* $\mathcal{LATA}$.

Though top-k tokens influencing the output ICD-codes (all predicted labels together) could be visualized with basic BERT variants, code-specific interpretation is quite challenging. Some sample test case reports along with predicted ICD-10 diagnostic codes and the extracted textual evidence are tabulated in Table 5.5, to illustrate the process.

Based on the results, it was observed that the ICD-10 coding system could handle both long and short case reports due to the usage of the proposed augmentation technique. Also, augmentation aids in reducing the overfitting of the

Table 5.6: ICD-10 diagnostic code explainability results for variants of $\mathcal{LATA}$.

Model	**Precision**	**Recall**	**F1-score**
BERT	0.800	0.442	0.570
ALBERT	0.583	0.098	0.168
DeBERTa	0.785	0.425	0.551
DistilBERT	0.802	0.453	0.579
MobileBERT	0.324	0.010	0.019
RoBERTa	**0.812**	**0.467**	**0.590**
XLNet	0.047	0.026	0.034

Table 5.7: Performance of state-of-art models on the CodiESP testset.

Model	Requires Pretraining?	Explainable?	**Diagnostic Performance**			**Explainability Performance**		
			Precisio	**Recall**	**F1-score**	**Precisio**	**Recall**	**F1-score**
N-gram CNN Encoder Tagawa *et al.* (2020)	✓	✗	0.123	0.522	0.199	-	-	-
BETO + LSTM + CNN + SelfAtt Polignano *et al.* (2020)	✓	✗	0.295	0.376	0.331	-	-	-
CAML + Hier Moons and Moens (2020)	✗	✗	0.124	0.064	0.084	-	-	-
Knowledge Graph + Multilingual BERT García-Santa and Cetina (2020)	✓	✓	0.767	0.699	0.731	0.704	0.634	0.667
Dictionary based Cossin and Jouhet (2020)	✗	✓	0.843	0.672	0.748	0.77	0.594	0.67
XLNet + BERT Schäfer and Friedrich (2020)	✓	✗	0.375	0.333	0.352	-	-	-
BiLSTM + CRF Ortega *et al.* (2020)	✓	✓	0.513	0.615	0.559	0.428	0.517	0.469
Tf-idf weighting Nunzio (2020)	✗	✗	0.373	0.76	0.5	-	-	-
Electra BERT Rishivardhan *et al.* (2020)	✓	✗	0.009	0.016	0.012	-	-	-
BiLSTM + CRF de la Iglesia *et al.* (2020)	✓	✗	0.75	0.624	0.682	-	-	-
Dictionary based Queipo-Álvarez and González-Carrasco (2020)	✓	✗	0.935	0.071	0.132	-	-	-
LM-BETO CM + PCS Costa *et al.* (2020)	✓	✓	0.551	0.743	0.633	0.534	0.562	0.548
Semantic similarity + Gradient BoostingAlmagro *et al.* (2020)	✓	✓	0.38	0.734	0.501	0.537	0.457	0.494
GRU Eslami *et al.* (2020)	✓	✗	0.014	0.045	0.021	-	-	-
BERTMatch + XGBoost classifier Blanco *et al.* (2020)	✓	✓	0.004	1	0.009	0.288	0.327	0.306
$\mathcal{LATA}$ (RoBERTa)	✗	✓	0.719	0.582	0.643	0.812	0.460	0.590

proposed model and thus improves the model's generalizability. From the neural model training perspective, though the number of training parameters for $\mathcal{LATA}$ was more than corresponding basic BERT variants, $\mathcal{LATA}$ variants converged earlier (10-11 epochs) than basic BERT variants (20-24 epochs). This fast convergence was due to the use of a label attention mechanism instead of a single pool layer for the encoder output. Thus, the training time almost remained the same for basic BERT and $\mathcal{LATA}$ variants. The inference time difference between $\mathcal{LATA}$ and basic BERT variants was negligible (5-6 milliseconds per test case report).

Furthermore, $\mathcal{LATA}$ variants always outperformed their basic BERT counterparts by a significant margin (Refer Table 5.3 and 5.4), owing to the enhanced predictability attributed by the label attention mechanism. The best results were obtained with the $\mathcal{LATA}$ model built on RoBERTa, which outperformed other basic BERT models by a significant margin of 33-49%. The methods and findings obtained by CodiEsp challenge participants are described in Table 5.7. The use of the label attention mechanism enhanced the model's flexibility in choosing the context of representation. Additionally, the $\mathcal{LATA}$ facilitated direct extraction of contexts from the input case report that were weighed by the label attention mechanism, thereby enhancing the interpret-ability of the obtained neural predictions to a large extent (as highlighted in Table 5.5). When most of the $\mathcal{LATA}$ generated false-negative cases (existent disease/context were not identified correctly) were analyzed, it was observed that the main reason was the lack of the training case reports for corresponding ICD-10 codes; mainly when less than 10 reports were available for the ICD-10 code (Refer row 2 of Table 5.5). The more tolerant false-positive cases (non-existent disease gets marked-up) analysis revealed that the patient case reports included long histories that signalled the $\mathcal{LATA}$ to mark-up the content within the history as evidence to predict ICD-10 diagnostic code (Refer row 1 of Table 5.5).

A detailed comparison was made with other comparable approaches to measure the potential of the EnCAML model for its suitability for deployment in real-world hospital scenarios. García-Santa and Cetina (2020) used a knowledge graph built on large-scale datasets (MIMIC-III, medical abstracts etc.) and then used pre-trained multilingual BERT to initialize the neural network. Also, they incorporated post-processing of data that involved multiple computationally intensive steps. The deployability and adaptability of such complex methodology is challenging in real-world scenarios. Cossin and Jouhet (2020) used a dictionary based approach. The normalized labels of the ICD-10 terminology and annotations

of the medical expert from the training set were used to construct the dictionary. A major limitation of their work is that, as the corpus size becomes larger, the dictionary construction takes time and is difficult to scale. Also, it is challenging to get accurate expert annotations. The authors also reported that their algorithm fails to detect a phrase if its tokens are not in the correct order or are nonadjacent. In contrast to these approaches, $\mathcal{LATA}$, adopted data augmentation and a simple post-processing pipeline to overcome these challenges, which increased its suitability for real-world applications to a significant extent. $\mathcal{LATA}$ directly uses augmented data without the need for pretraining on other large datasets. The real-time clinical EHR records could be directly fed into $\mathcal{LATA}$, and the model provides the possible diagnostic codes along with important terms that influenced the decision by directly using the label-wise attention weights.

5.5.1 $\mathcal{LATA}$ and BERT - Differences

The basic BERT sequence classification models use the pooled output by taking the hidden state corresponding to the first token ([CLS]). This output is usually not a good summary of the input's semantic content due to the unavailability and also limitations of pre-trained BERT variants that use large medical text corpus in non-English languages. Though top-k tokens influencing the output ICD-codes (all predicted labels together) could be visualized with basic BERT variants, code-specific interpretation is quite challenging.

$\mathcal{LATA}$ uses label attentions for every intermediate layer output in BERT to improve input context learning for the given output classes. The use of the label attention mechanism enhanced the model's flexibility in choosing the context of representation. Additionally, the $\mathcal{LATA}$ facilitated direct extraction of contexts from the input case report (using *alpha* as mentioned in Section 5.4.2) that were weighed by the label attention mechanism, thereby enhancing the interpret-ability of the obtained neural predictions to a large extent

5.6 Summary

Effective code assignment for patient clinical records in a hospital plays a significant role in standardizing medical records, mainly for streamlining clinical care delivery, billing, and managing insurance claims. The current practice employed is manual coding, usually carried out by trained medical coders, making the process subjective, error-prone, inexact, and time-consuming. To alleviate this cost-

intensive process, intelligent coding systems built on patients' structured electronic medical records are critical. Classification of medical diagnostic codes, like ICD-10, is widely employed to categorize patients' clinical conditions and associated diagnoses. In this work, a neural model $\mathcal{LATA}$, built on Label Attention Transformer Architectures is proposed for automatic assignment of ICD-10 codes. This work was benchmarked on the CodiEsp, a dataset for automatic clinical coding systems for multilingual medical documents, used in the eHealth CLEF 2020-Multilingual Information Extraction Shared Task. The experimental results revealed that the proposed $\mathcal{LATA}$ variants outperform their basic BERT counterparts by 33-49% in terms of standard metrics like precision, recall, F1-score and mean average precision. The label attention mechanism also enabled the direct extraction of textual evidence in medical documents that map to the clinical ICD-10 diagnostic codes.

Publications

(based on works presented in this chapter)

1. Veena Mayya, Sowmya Kamath S. and Vijayan Sugumaran, "$\mathcal{LATA}$– Label Attention Transformer Architectures for ICD-10 Coding of Unstructured Clinical Notes", In the proceedings of the 18th International Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE, Australia (Virtual), October 2021. (CORE Ranked) *(Online)*.

PART III

Multi-task CDSS using Imaging Healthcare Data

Chapter 6

Empirical Study of Preprocessing with CNN for Accurate Retinal Disease Diagnosis

6.1 Introduction

Chronic Ocular Diseases (COD) represent a class of noncommunicable chronic eye conditions that affect visual acuity or even cause vision loss if left untreated. As per to recent statistics (WHO, 2019), the global population suffering from myopia would reach 3.36 billion by 2030, while those suffering from age-related macular degeneration (AMD), glaucoma, and diabetic retinopathy (DR) will reach 243.3 million, 95.4 million, and 180.6 million, respectively. Early diagnosis is critical for reducing the risk of vision impairment, and regular screening is a significant step towards early diagnosis. However, the process adopted for screening is primarily manual investigation (Clarke *et al.*, 2018; Metsing *et al.*, 2018). This makes it impractical to scale, given the wide range of diseases and ever-growing patient population.

Color fundus images captured from digital fundus cameras are used for the early diagnosis of CODs. The transformations on these input images, before feeding them to the deep neural models, play a crucial role in improving the diagnostic performance of the system. Preprocessing reduces the possible noise in color fundus images such as irregular illumination, low contrast, unimportant features, etc., thus improving the performance of DL-based COD diagnosis. A few digital image preprocessing techniques reported in the literature are adaptable to all COD (hereby referred to as generic techniques). On the other hand, some preprocessing methods are only suitable for a particular ocular disease like DR, glaucoma, cataract, or AMD (referred to as specific preprocessing techniques).

6.2 Problem Statement

Comprehensive qualitative analysis of adapted preprocessing strategies has not been adequately explored in the existing literature. Preprocessing of input fundus images necessitates the use of computational resources. A few preprocessing methods may improve the predictive performance, while others may have the opposite impact. Some preprocessing techniques may be best suited for specific ocular diseases. Still, they may not meet the clinical needs in real-time, particularly in scenarios where there is a lot of variation in eye diseases. Thus, there is significant scope for conducting a comprehensive, systematic assessment of such techniques' relative strengths and weaknesses for quantifying their usefulness in automated COD diagnosis. Thus, the problem to be addressed here is defined as follows:

> *Given the wide range of diseases and the availability of abundant retinal imaging data, design and develop effective preprocessing and detection modelling approaches for COD detection based on retinal images.*

6.3 Motivating Example

To describe the prevailing conditions that emphasize the need for automated COD detection CDSS based on retinal images, consider scenarios where the retinal images are captured by physicians during routine patient visits or during mass screening campaigns in rural areas. Often decision-making is subjective, resulting in a significant degree of inter-and intra-physician variability in diagnosis. For instance, when a physician analyses an image several days later, they may diagnose it differently than they did before, or there may be discrepancies in opinion when the same image is evaluated by numerous physicians with varying domain expertise. Physicians' preferences for the appearance and quality of the image can change over time and can also differ. Scaling manual interpretation to an ever-growing patient population is likewise a challenge. These challenges in the manual screening process could be avoided if the ocular disease could be automatically identified using the retinal imaging data. The automated disease detection CDSS could directly preprocess the retinal images captured by the technician and provide them with a list of prediction scores for the CODs along with highlighting the regions that the system found to be relevant for detecting each disease. Thus, the technician may direct the patient to the appropriate physician according to the disease identified, thus decreasing diagnostic time, and could also immediately identify any photos with decreased quality. Automated systems can aid in the

early detection of COD via tele-ophthalmology in rural areas where there is a shortage of retina specialists.

In this chapter, approaches towards developing effective COD detection models built on retinal imaging data are presented. The contributions towards the defined problem are in the context of designing methods that can automatically preprocess retinal images, with their differences in lighting, contrast, lesion shapes, etc., and still be able to extract relevant disease-specific features, which can be leveraged for the purpose of automatic COD detection. The performance of the proposed models is compared to that of state-of-the-art COD detection systems built on retinal fundus images.

6.4 Empirical Study

The most common clinical signs based on which ophthalmologists identify the progression of DR in the fundus images are microaneurysms (MA), haemorrhages, exudates, thickening within one disc diameters from the foveal centre, and retinal neovascularization (Al-Diri *et al.*, 2019; Tan *et al.*, 2017). Thus, the lesion regions are segmented for building DL-based automated DR diagnostic systems. Damage to the optic nerve is the primary cause of loss of vision in glaucoma. Glaucoma can be detected by examining the abnormalities of the optic disk. Chalakkal *et al.* (2021) investigated the effect of fovea segmentation on macular oedema screening using DL-based transfer learning approaches. The authors critically analyzed the effects of limiting the RoI to the fovea and reported improved performance due to RoI segmentation than considering the entire fundus image. The progressive alterations in the retinal vessels are also crucial for identifying DR (Dubielzig *et al.*, 2010). Several works (Fu *et al.*, 2018; Diaz-Pinto *et al.*, 2019*b*; Zhao *et al.*, 2020; Sreng *et al.*, 2020*a*) employed cropping/segmentation of optic disk regions and then used CNN to diagnose glaucoma using color fundus images. Juneja *et al.* (2020) proposed a modified version of U-net (G-net) to segment the optic disk and cup region, after which they used the ratio of these areas to predict glaucoma. Zhao *et al.* (2020) adapted a template matching method for locating and cropping the bounding region around the optic disc and proposed MFPPNet to screen glaucoma. Modified U-Net (Ronneberger *et al.*, 2015*b*) is predominantly used for optic disk segmentation (Fu *et al.*, 2018; Diaz-Pinto *et al.*, 2019*b*; Civit-Masot *et al.*, 2020) and then transfer learning is applied on the cropped region to screen glaucoma. Pathan *et al.* (2021) used inpainting to eliminate the vascular structure before segmenting the optic disc/cup regions. The increase in protein aggregation in the

lens does not allow light to pass through the lens and may lead to cataract. Xu *et al.* (2020*a*) divided the input fundus images into eight local patches based on ophthalmologists' recommendations for automatic cataract grading using CNN. Using the unified rectangular fundus images, Zhang *et al.* (2019) extracted the high-level texture features using a CNN model and supervised SVM to grade the severity of the cataract. Imran *et al.* (2020) extracted the image green channel, resized the images, and used CNNs for feature extraction and SVM for cataract severity identification.

So far, the performance of CNNs trained on optic disc cropped regions has been experimented with; however, a systematic experimental evaluation of other preprocessing methods when used along with CNNs has yet to be undertaken. Thus, there is a wide scope for the experimental evaluation of other preprocessing methods using CNN for automatic cataract diagnosis. Though the choice of preprocessing techniques is dependent on the type and requirements of a particular ocular disease diagnosis, a thorough analysis of the effect of preprocessing on the efficiency of DL models has not been undertaken, to the best of our knowledge. The role of preprocessing in improving the performance of the DL-based diagnostic system is investigated in this study by experimenting with two specific preprocessing techniques (vessel segmentation and inpainting) and nine generic preprocessing techniques. Fig. 6.1 provides an overview of the training methodology, and further details regarding the preprocessing and augmentation methods used in the course of experiments have been elaborated below.

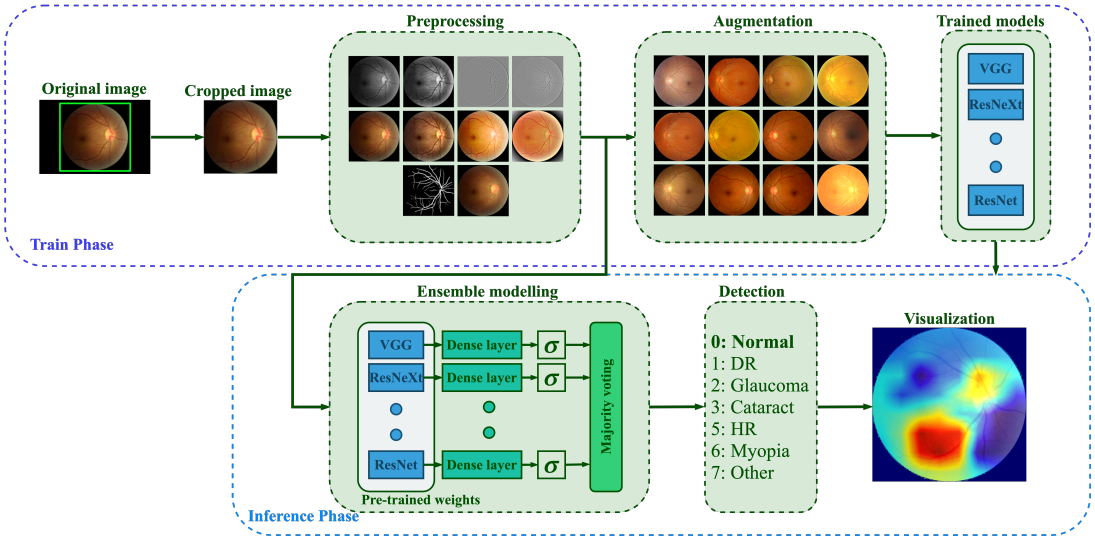


Figure 6.1: Methodology adopted for the empirical study

6.4.1 Automatic Region of Interest Segmentation

The gradient Hough transform was used to segment the foreground circular fundus region. The parameters were chosen based on the validation of 500 fundus images. The inverse of the ratio of accumulator resolution to image resolution is set to 1, and the minimum distance between the detected circles' centre coordinates is set to 20. The accumulator threshold value is set to 30, and the gradient value for edge detection is set to 50. The minimum radius (in pixels) is set to 1/4 of the input dimensions' minimum, while the maximum size is set to the input dimensions' maximum size. The bounding box is formed for the detected circular region, ensuring that the cropped region is within the image dimension. The appropriate circular area (from the detected circles) is chosen, ensuring that non-zero pixels outside the cropped region are within five rows and columns.

The automatic foreground cropping mechanism is described in detail in Algorithm 2. A few sample input fundus images and the corresponding foreground segments obtained after processing are shown in Fig. 6.2(c). Otsu thresholding (Otsu, 1979) using the largest contour crop method was also experimented with, but the foreground regions obtained were clipped in the darker images. With the proposed automatic segmentation approach, only the background regions are cropped without losing useful details (refer Fig. 6.2(a-c)).

6.4.2 Image Enhancement

To improve the visual quality of input fundus images, experiments were carried out using CLAHE (Zuiderveld, 1994) (on both green and RGB channels), Gaussian filter convolution, Multiscale Retinex (MSR) (Petro *et al.*, 2014), and multiscale residual block network (MIRNET) (Zamir *et al.*, 2020) approaches. The CLAHE enhanced image was obtained by applying the normalized intensity histograms P_n (see eq. (6.1)) to the image patches (5×5), which were defined as a matrix M ($r \times c$) with values (pixel intensity) ranging from 0 to $L-1$. The resulting image patch I_{eq} was defined by eq. (6.2). The contrast factor (or clip limit) to limit the slope associated with the gray-level assignment scheme in CLAHE was set to 2.

$$P_n = \frac{\text{Number of pixels with intensity n}}{\text{Total number of pixels}} \quad n = 0, 1 \ldots L-1 \tag{6.1}$$

$$I_{eq} = \lfloor (L-1) \sum_{n=0}^{M_{i,j}} P_n \rfloor \tag{6.2}$$

A Gaussian filter convolved (*blurred*) image was blended with the original

Algorithm 2 Automated fundus foreground region segmentation

Input: A sequence of fundus images.
Output: Foreground segmented images.

for each input fundus image **do**
  if zero-valued pixels < five rows/columns *(as per image size)* **then**
    Continue with next image
  end if
  Convert to grayscale, find Hough circles
  for each detected circle **do**
    Find bounding region using radius & centre ▷ *Ensured that bounding region is within image dimension*
    if non-zero pixels outside the crop region < five rows/columns *(as per input size)* **then**
      if current radius > previously detected radius **then**
        Update the centre and radius
      end if
    else
      Save the RoI segmented image
      Continue with next image
    end if
  end for
end for

image to improve image contrast. The resulting enhanced image I_{gs} is defined by eq. (6.3). $G(h, w)$ represents a Gaussian filter with a scale σ and $*$ the convolution operator. The parameter values were determined based on experimentation and are set as $\alpha = 4$, $\beta = -4$, $\sigma = 10$, and $\gamma = 128$. The MSR algorithm (Jobson *et al.*, 1997; Schivre, 2021) was adapted to improve the local contrast enhancement of the fundus images. MSR was implemented by Petro *et al.* (2014), based on the Retinex (Land, 1977) theory, which attempts to model human visual color perception. MSR improves the local contrast of the fundus image as per the desired scales ($sigma_n$) of the Gaussian kernels (refer eq. 6.4). N was set to three scales ($[5, 35, 150]$), and the weight factor associated with the Gaussian function (W_n) was chosen to be 1/3.

$$I_{gs}(h, w) = \alpha I_{orig}(h, w) + \beta G(h, w, \sigma) * I_{orig}(h, w) + \gamma \tag{6.3}$$

$$I(h,w)_{ms-retinex} = \sum^{N} W_n(log(I(h, w)) - log((I(h, w)) * G(h, w, \sigma_n))) \tag{6.4}$$

Recently, Zamir *et al.* (2020) proposed the MIRNET model that integrates parallel multi-resolution convolution, spatial and channel attention for image enhancement. The pre-trained weights of this model were utilized, to enhance the

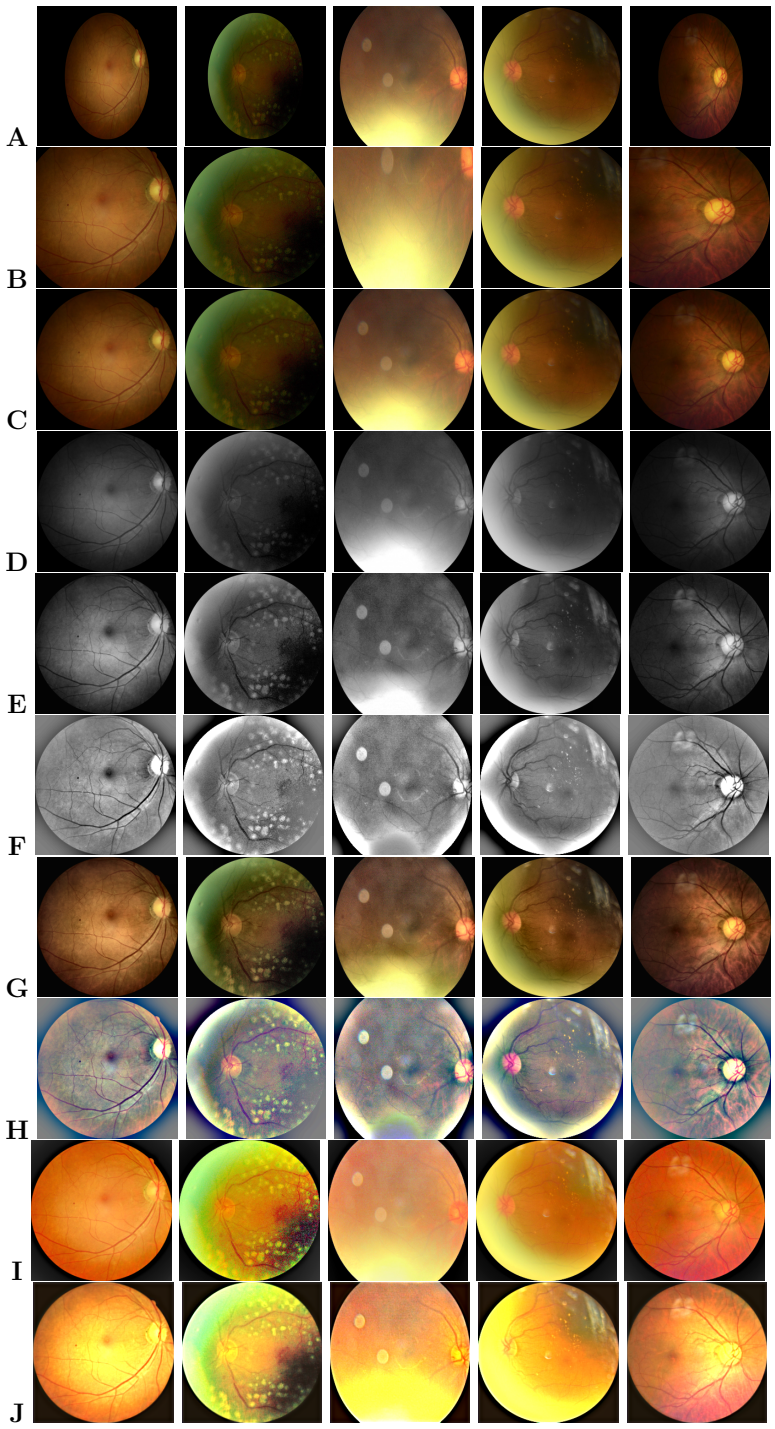


Figure 6.2: Results of preprocessing on input sample images. **A**)Original, **B**)Otsu's thresholding segmentation, **C**)Proposed Hough transform segmentation, **D**)Green channel, **E**)Green channel CLAHE, **F**)Green channel Gaussian convolution, **G**)RGB CLAHE, **H**)RGB Gaussian convolution, **I**)MSR, and **J**)MIRNET images.

contrast of fundus images. From Fig. 6.2, the effect of each preprocessing phase

on the sample input fundus images could be observed. As can be seen from Fig. 6.2, MSR increases the brightness while maintaining/improving the overall visual quality. CLAHE and MIRNET enhance the visual quality of the image but also introduce additional artefacts, while the green channel retains the original image visual quality.

6.4.3 Vessel Segmentation

The progressive changes of retinal vessels are crucial to help detect CODs, for which preprocessing is performed through the segmentation of blood vessel regions. The DRIVE dataset (Staal *et al.*, 2004) was used to train the RetinaNet (Son *et al.*, 2018) model for this purpose. CLAHE was applied to increase the contrast of the color fundus images, and the generative neural model (Son *et al.*, 2018) is re-trained for the segmentation of vessel masks. Regions with intensities less than the threshold (20) are set to zero in the mask. If the number of connected pixels in each connected component is less than 100, the intensities are set to zero. Algorithm 3 details the steps involved in vessel segmentation. To observe the impact of vessel structure on the COD detection performance, a process of background color inpainting is applied to the segmented vessel masks, using a pre-trained generative model (Yu *et al.*, 2018). Fig. 6.3 shows some sample images from various COD classes.

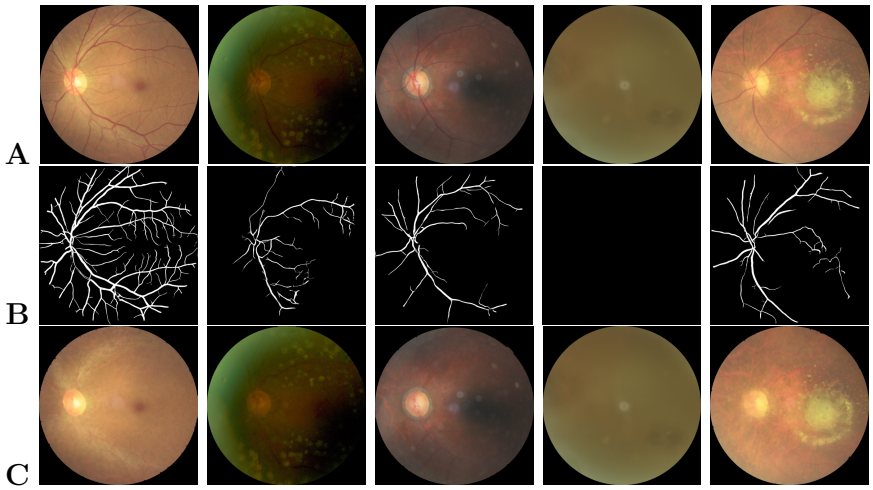


Figure 6.3: Results of segmentation and inpainting of vessel structure on sample input images. **A**)Original (Normal, DR, glaucoma, cataract and AMD), **B**)Vessel segmentation, **C**)Vessel inpainting images.

Algorithm 3 Automated vessel segmentation from input colored fundus image

Input: A sequence of fundus images.
Output: Vessel segmented masks.

Convert DRIVE dataset images to $L*a*b*$ ▷ *$L*$ indicates lightness and $a*$ and $b*$ are chromatic coordinates*
Apply CLAHE to lightness channel
Convert to RGB colorspace
Train for vessel segmentation using RetinaNet
Store the model with best AUC score.
for each input fundus image **do**
    Generate vessel mask using the stored RetinaNet model weights.
    Set zero for the pixels whose intensity is less than 20.
    Find the connected component statistics.
    Sort connection area components in ascending order (in pixels)
    for each connected area **do**
        if the area has less than 100 pixels **then**
            Mask the region (with Zero). ▷ *fully disconnected noisy area*
        end if
    end for
    Save the generated mask image ▷ *refer Fig. 6.3(b).*
end for

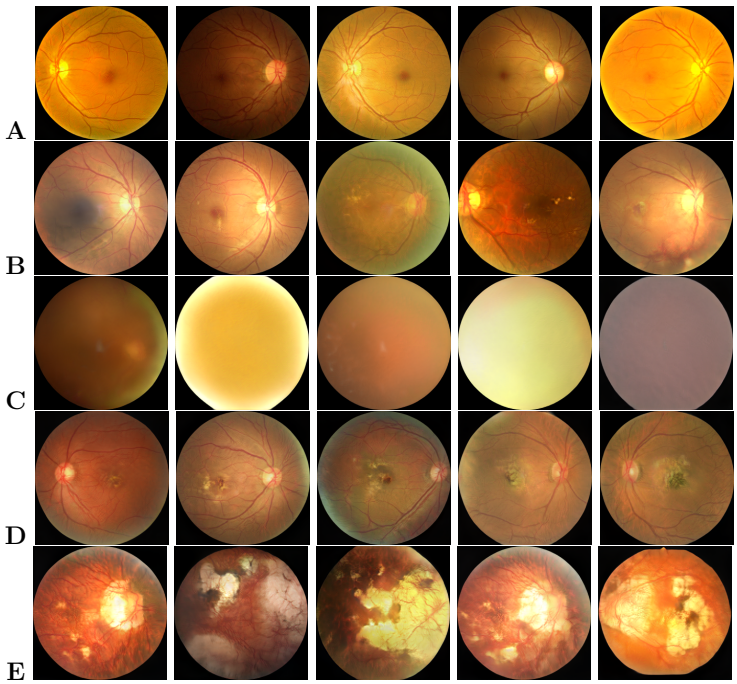


Figure 6.4: Sample of generated fundus images using StyleGAN2 for the given ocular conditions. **A**)Normal **B**)DR **C**)Cataract **D**)AMD **E**)Myopia.

6.4.4 Data Augmentation

For effective learning performance, generalizable deep neural models require a large number of labelled images. A variety of data augmentation techniques are widely used to deal with the limited number of training images. Batch-level and condition-level data augmentation techniques were incorporated for increasing the number of images used for training the neural models. In batch-level augmentation, horizontal and vertical flipped and random angle rotated images were generated and added to the training set. In condition-level augmentation, training images were augmented conditionally using the StyleGAN2 model (Karras *et al.*, 2020*a*,*b*). The StyleGAN2 was applied to the training dataset, using the textual descriptions of the ocular diseases. The diseases with fewer than five training images were excluded, which resulted in 48 conditions. The network was trained for 25 million training images. Fig. 6.4 presents a representative sample of fundus images generated for the given COD. As a result of the data augmentation, over 21,000 training images were obtained by augmenting 300 images for each ocular condition.

6.5 Convolutional Neural Models

CNNs have shown exceptional performance in various computer vision tasks, including the classification of COD. Experimentation was carried-out with eleven prominent CNN models that have demonstrated high performance on the ImageNet (Russakovsky *et al.*, 2015) challenge dataset. This included DenseNet (Huang *et al.*, 2017), EfficientNet (Tan and Le, 2019), Inception (Szegedy *et al.*, 2015), MobileNet (Howard *et al.*, 2017), ResNet (He *et al.*, 2016), ResNeXt (Xie *et al.*, 2017*a*), WideResNet (Zagoruyko and Komodakis, 2017), VGG16 (Liu and Deng, 2015), and SqueezeNet (Iandola *et al.*, 2017) models using both cropped and non-cropped fundus images.

COD classification is a multi-label classification problem that requires binary classification of multiple diagnostic labels, each label indicating a specific COD. As a result, binary predictions were used as target scores, with actual and predicted values compared pairwise. A multi-label one-versus-all loss based on max-entropy was used to train the CNN, as illustrated in eq. (6.5) (where $i \in \{0 \ldots, N\}$, $y[i] \in \{0, 1\}$). In this study, Y represents the actual target labels, while $\hat{y}$ represents the predicted labels of dimension (N, C) (where N is the batch size and C is the number of classes=8).

$$\begin{aligned} \text{loss}(\hat{y},y) = -\frac{1}{C} * \sum_{i} y[i] * log((1 + exp(-\hat{y}[i]))^{-1}) \\ +(1 - y[i]) * log\left(\frac{exp(-\hat{y}[i])}{(1 + exp(-\hat{y}[i]))}\right) \end{aligned} \tag{6.5}$$

Table 6.1 lists the number of training parameters used for the DL models in ascending order. The model parameters were initialized using the ImageNet pre-trained weights. After observing the outcomes of the experiments, the model with the highest F_1 score was utilized as a baseline for examining the effect of various preprocessing techniques on CNN performance. Fundus images were preprocessed individually and fed into the CNN models for inference. The maximum score is regarded as the final screening result at the patient level.

Table 6.1: Details of model training parameters.

Model	**Parameters** ***(in millions)***
SqueezeNet	**0.726600**
MobileNetv2	2.234120
GoogleNet	5.608104
DenseNet121	6.962056
EfficientNet-B3	10.708528
ResNeXt50	22.996296
ResNet50	23.524424
Inceptionv3	25.128656
EfficientNet-B7	63.807448
WideResNet50	66.850632
VGG16	**134.29332**

6.6 Experimental Results and Discussion

For the experimental evaluation of the proposed approaches, the ODIR-5K (Ocular Disease Intelligent Recognition) challenge dataset (ODIR, 2019) consisting of a total of 5,000 patients' data was employed. The training dataset comprises 3,500 patient records, with 7,000 fundus images, and the final label is based on both eye conditions. The dataset is highly imbalanced. It contains over 3,000 normal images but only 190 hypertensive retinopathy images. About 500 images

have multiple labels, and the "*Others*" category consists of more than 20 distinct eye diseases. The statistics of the training data are summarized in Table 6.2. Automatic relabeling was carried out based on the textual information available for each eye condition, and the validity was checked by comparing the union of the labels with the final available labels. For example: *0_left.jpg* has the text *cataract*, so it is labelled *00010000*, and *0_right.jpg* contains the *normal fundus*, which is labelled *10000000*, and the union 00010000 is compared to the final available label *00010000*. Normal labels were considered only when both eyes had the text *normal fundus*, ignored otherwise. All the images were resized to 256×256.

Table 6.2: Details of ODIR training data.

Type of COD (Class)	**Training data**
Normal	3,098
DR	1,801
Others	1,200
Glaucoma	326
Cataract	313
AMD	280
Myopia	268
Hypertension	193

Several standard metrics were used for validation purpose. F_1 score was used as a primary metric for validating the output of preprocessing techniques. F_1 ($F_{\beta=1}$) is a weighted harmonic mean of precision and recall that seeks to balance the two (see eq. (6.6)). As a result, models with higher F_1 scores are expected to improve the system's predictability. The F_1 metric is also thought to be more indicative than the standard accuracy score because it accounts for true and false positives (TP and FP) as well as true and false negatives (TN and FN). The precision and recall for a neural system learned to predict over C classes is computed using eq. (6.7). In this study, the precision and recall values are macro-averaged over the target output classes.

$$F_{\beta=1} = (1+\beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}} \tag{6.6}$$

$$\text{precision} = \frac{1}{C} \sum_{c=1}^{C} \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}; \text{recall} = \frac{1}{C} \sum_{c=1}^{C} \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} \tag{6.7}$$

In addition to F_1 scores, the models' performance was reported using the area under the curve (AUC) and Kappa score (average of Cohen's kappa for each label) multi-label classification metrics. The field under the ROC curve is referred to as the AUC. The model's classification accuracy improves as it gets closer to 1. It is often used to determine the model's stability. Cohen's Kappa (McHugh, 2012) (see eq. 6.8) is a quantitative measure of reliability; a score of 0 indicates that there is a random match, while a score of 1 means that true and predicted labels are fully in the agreement.

$$Kappa_{score} = \frac{P_o - P_e}{1 - P_e};$$

$$P_o = \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C} TP_c + FN_c}; \tag{6.8}$$

$$P_e = \frac{\sum_{c=1}^{C} TP_c * (TP_c + FN_c)}{N^2}$$

During inference, the left and right eye images were considered, and the final ocular condition was determined by a label-wise maximum score among two output predictions. The ODIR test set contains 500 unlabeled patient records (1,000 images) that can be labelled in any of the eight possible ways. Thus, if a single label is correctly predicted, an F_1 score of 0.00025 is achieved, demonstrating the significance of prediction scores. Table 6.3 summarizes the results obtained with original color and cropped images obtained using state-of-the-art neural models with test ODIR data[1]. The experimentation was conducted using a batch-level augmentation method (please refer Section 6.4.4). The performance of the top scoring neural model (ResNeXt50) with RoI crop is benchmarked using several preprocessing methods. As stated earlier, nine preprocessing methods were employed, and the results are shown in Table 6.4.

The ResNeXt50 model achieved the best F_1, Kappa and AUC scores on the cropped images obtained using the proposed RoI segmentation algorithm. The observed results are tabulated in Table 6.5. It can be observed that batch-level augmentation achieved the best performance. Therefore, for the rest of the experiments, it was utilized for training the models. The majority rule voting approach

[1]Submitted to: `https://odir2019.grand-challenge.org/evaluation/challenge/submissions/create/`

was used to ensemble the predictions of the top three DL models (ResNeXt50, EfficientNetB7 and VGG16) with only RoI cropped color images and the top three preprocessing approaches (RoI cropped, green channel, and MSR) trained with the ResNeXt50 model. The results of ensemble models' performance are shown in Table 6.5 and the results of benchmarking experiments with respect to state-of-the-art DL models are summarized in Table 6.6.

Table 6.3: Observed performance for state-of-the-art DL models on the testset.

	With original image			**With cropped image**		
	Kappa	**AUC**	**F_1**	**Kappa**	**AUC**	**F_1**
DenseNet	0.4659	0.7888	0.8698	0.5195	0.8210	0.8804
EfficientNetB3	0.4691	0.7922	0.8695	0.5090	0.8199	0.8803
EfficientNetB7	0.4677	0.7265	0.872	0.5260	0.8317	0.8845
GoogleNet	0.4124	0.7746	0.8553	0.4733	0.8021	0.8718
InceptionNet	0.3131	0.7201	0.8365	0.3882	0.7671	0.8535
MobileNet	0.4664	**0.7984**	0.8703	0.4852	0.8191	0.8740
ResNet50	0.4110	0.7602	0.8575	0.5022	0.8093	0.8782
ResNeXt50	0.4733	0.7921	0.8717	**0.5680**	**0.8606**	**0.8953**
WideResNet	0.4222	0.7634	0.8618	0.4734	0.8343	0.8743
SqueezeNet	0.1347	0.6178	0.7838	0.1594	0.6412	0.7873
VGG16	**0.5092**	0.7828	**0.8790**	0.5268	0.8248	0.8813

During the extensive experiments conducted to evaluate the effectiveness of the proposed approaches, it was observed that the models trained using cropped fundus images always outperformed those trained on non-cropped images by an average percentage difference of 15% Kappa score as shown in Table 6.3. This can be attributed to the enhanced predictability afforded due to the proposed RoI segmentation algorithm. Gradient-weighted class activation mapping (Grad-CAM) (Selvaraju *et al.*, 2019) was used to visualize the dominant features learned by the proposed model for identifying a particular type of COD. Fig. 6.5 shows the Grad-CAM visualization for the original (non-cropped) and cropped images trained with ResNeXt50. The last convolution layer's coarse localization map (before AdaptiveAvgPool2d) reflects the important regions in the input image to detect a particular type of COD. The obtained Grad-CAM is normalized and resized to the original image size. A mask image is generated from Grad-CAM with a threshold of 100. The contours are drawn using the mask image and visualized on the input fundus image as shown in Fig. 6.5 (iv & viii). Contours

Table 6.4: Observations w.r.t best performing model, ResNeXt50, when used with proposed preprocessing pipeline.

Preprocessing method	**Observed performance**		
	Kappa	**AUC**	**F_1**
Original image	0.4733	0.7921	0.8717
Cropped image	0.5680	0.8606	0.8953
Green channel	0.5336	0.8186	0.8882
Green channel+CLAHE	0.5198	0.8189	0.8840
Green channel+Gaussian	0.4898	0.8082	0.8777
RGB+CLAHE	0.5308	0.8368	0.8860
RGB+Gaussian	0.5260	0.8206	0.8840
Multiscale Retinex (MSR)	0.5330	0.8418	0.8865
MIRNET	0.4438	0.8403	0.8550
Vessel segmentation	0.3097	0.7212	0.8325
Vessel inpaint	0.4865	0.8500	0.8765

Table 6.5: Performance of proposed augmentation and ensemble techniques.

Method	**Observed performance**		
	Kappa	**AUC**	**F_1**
No augmentation	0.5246	0.8323	0.850
Batch-level *(Flip + Rotation)*	0.5680	0.8606	0.8953
Condition-level *(GAN)*	0.4228	0.8534	0.8710
Ensemble 1 *(ResNeXt50, EfficientNetB7 & VGG16)*	0.5815	0.8532	0.9008
Ensemble 2 *(RoI cropped, green channel & MSR)*	0.6081	0.8806	0.9070

are indicated in green when the prediction score is greater than or equal to 0.5; otherwise, they are highlighted in red. Owing to the difficulties in identifying minute lesions (e.g., microaneurysms, drusens, cup to disc ratio, etc.), DL models trained on non-cropped images failed to detect a majority of early-stage ocular diseases (refer Fig. 6.5 ii-iv). In contrast, the proposed approach performed well in accurately identifying the majority of lesions, thus aiding in the generation of explainable predictions.

The proposed ensemble model outperformed several state-of-the-art models (Gour and Khanna, 2020; Islam *et al.*, 2019; Li *et al.*, 2021*a*) in terms of Kappa and F_1 scores. AUC is insensitive to class imbalance, i.e., when the labels include many

Table 6.6: Comparative performance of proposed approaches against state-of-the-art techniques.

No.	Models	Dataset	Performance		
			Kappa	AUC	F_1
1	ResNet-101 backbone (Li *et al.*, 2020; He *et al.*, 2021*a*)	1166 patients data (ODIR train set)	0.6370	0.9300	0.9130
2	ResNet-101 + Textual features (He *et al.*, 2021*b*)	1166 patients data (ODIR train set)	0.5800	0.9280	0.9080
3	EfficientNet-B3 (Wang *et al.*, 2020)	40 images from DRIVE (Staal *et al.*, 2004)	0.4900	0.7300	0.8900
4	Graph convolutional network (Lin *et al.*, 2021)	996 images of 498 patients (ODIR)	0.5765	0.7816	0.8966
5	Shallow CNN (Islam *et al.*, 2019)	ODIR offline challenge test set	0.3100	0.8050	-
6	Two input VGG16 (Gour and Khanna, 2020)	ODIR offline challenge test set	-	-	0.8557
7	VGG-16 (Li *et al.*, 2021*a*)	ODIR offline challenge test set	0.4494	0.8881	0.8730
8	Proposed pipeline	ODIR offline challenge test set	0.5680	0.8606	0.8953
9	Proposed DL ensemble	ODIR offline challenge test set	0.5891	0.8610	0.9025
10	Proposed preprocessing ensemble	ODIR offline challenge test set	0.6081	0.8806	0.9070

zeros, correctly detecting them may also lead to high AUC. As a result, a high F_1 score is more significant than a high AUC in cases with a high class imbalance. For the studies using ODIR training using 1166 patients' data, the models put forth by He *et al.* (2021*b*) and Li *et al.* (2020) showed better results. Due to the lack of precise patient IDs for each split, the proposed method could not be evaluated and compared to these models. Additionally, owing to the fusion of features, obtaining evidence for the output predictions is difficult with these models. With the proposed method, the evidence can be visualized for each predicted label (true or false). Moreover, He *et al.* (2021*b*) utilized diagnostic keywords indicative of

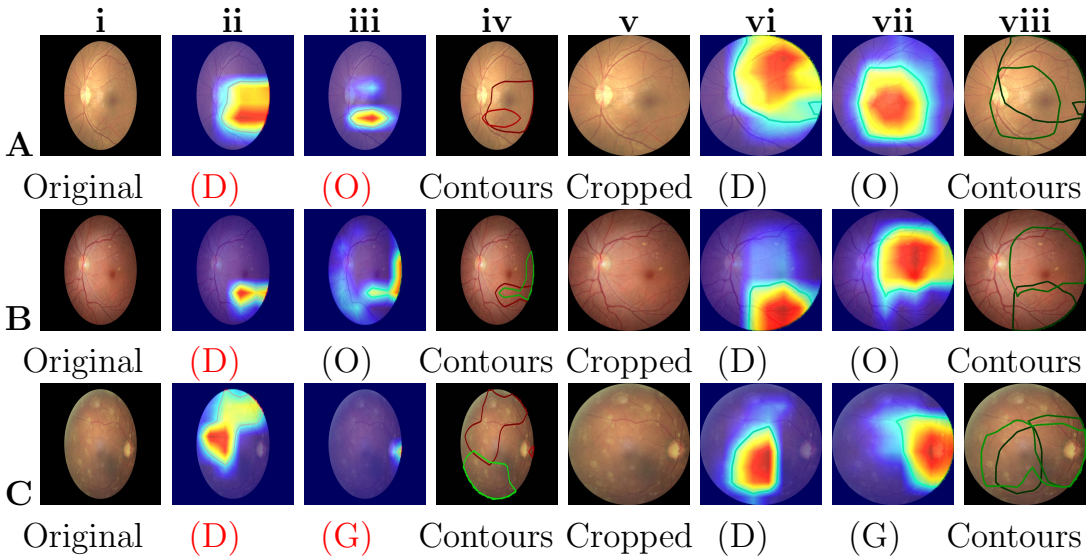


Figure 6.5: Visualization of Grad-CAM heatmap on the original input images. Columns **i-iv** show the original images, where as, columns **v-viii** are cropped versions. The annotated labels are **A**) Mild DR (D), epiretinal membrane (O); **B**) Mild DR (D), drusen (O); and **C**) Mild DR (D), glaucoma (G), vitreous degeneration (O).

actual diagnosis along with fundus images to augment their accuracy. Due to this, there is a dependency on expert generated diagnosis reports, which imposes an additional load on the learning models.

All other preprocessing strategies, except for the RoI cropping method, had no significant impact on the prediction performance of DL models (please see Table 6.4). Though the images created with CLAHE and the MIRNET seemed to improve contrast visually, the techniques did not significantly boost the CNN performance. The efficiency is similar to that of the RGB channels when only the green channel is used, but the number of training parameters is decreased by around 10000. Hence, this helps for efficient training and inference. Furthermore, the vessel segmentation technique that was primarily investigated for DR detection (Roshini *et al.*, 2020; Saranya and Prabakaran, 2020; Saranya *et al.*, 2021), had no discernible effect on COD detection. Condition-level augmentation improved prediction accuracy for the *Normal*, *DR*, *Cataract*, *AMD*, and *Myopia* classes. It did not, however, improve prediction for *Glaucoma* or *Other* categories of diseases. This limitation could be addressed by including more representative images for these classes (particularly with minute lesions). To further investigate the impact of augmentation on prediction performance, the highest-scoring DL model (ResNeXt) trained on ODIR is tested on the publicly available DDR test dataset (Li *et al.*, 2019). The DDR dataset makes use of the International DR

Grade Classification (Wilkinson *et al.*, 2003) (ranging from 0 to 4), as well as a special label (5) for low-quality images. These ungradable images were excluded due to their low quality. The remaining dataset (3,759 images) is split into normal (0) and abnormal (1,2,3,4). The foreground region is cropped using the proposed method (Section 6.4) and tested using the ODIR pretrained ResNeXt model. The final score is obtained using the predictions of the "*Normal*" class. The results of the three augmentation techniques for DR screening are presented in Table 6.7. The observations revealed that the condition-level augmented model holds promise for building a generalizable DL model. The inference for DR screening was also achieved using an ensemble of models trained on the ODIR dataset. The proposed ensemble model outperformed a patch-based lesion localization deep network proposed by Zago *et al.* (2020) in terms of AUC and sensitivity scores (Table 6.7).

Table 6.7: Comparative evaluation of augmentation and preprocessing techniques on DDR testset.

Method	**Observed performance**			
	Kappa	F_1	**AUC**	**Sensitivity**
No augmentation	0.5797	0.7898	0.8669	0.9276
Batch-level augmentation	0.5812	0.7906	0.8793	0.9170
Condition-level augmentation	0.6052	0.8026	0.8769	0.9091
Ensemble 1 *(ResNeXt50, EfficientNetB7 & VGG16)*	0.6196	0.8098	0.8749	0.9563
Ensemble 2 *(RoI cropped, green channel & MSR)*	0.6302	0.8151	0.8730	0.9570
Patch-based lesion localization model (Zago *et al.*, 2020)	-	-	0.8480	0.8910

6.7 Summary

Chronic Ocular Diseases (COD) can affect the eye and may even lead to severe vision impairment or blindness. However, if COD is detected early, vision impairment can be avoided by early intervention and cost-effective treatment. Several preprocessing approaches were combined with convolutional neural networks to accurately detect COD in eye fundus images. Experimental results demonstrate that CNNs trained on the region of interest segmented images outperform the

models trained on original input images by a substantial margin. Additionally, an ensemble of three preprocessing techniques outperformed other state-of-the-art approaches by 30% and 3%, in terms of Kappa and F_1 scores, respectively. The developed prototype has been extensively tested and can be evaluated on more comprehensive COD datasets for deployment in the clinical setup.

Publications

(based on works presented in this chapter)

1. Veena Mayya, Sowmya Kamath S., Uma Kulkarni, Divyalakshmi Kaiyoor Surya, U. Rajendra Acharya, "*An empirical study of preprocessing techniques with convolutional neural networks for accurate detection of chronic ocular diseases using fundus images*", Applied Intelligence, Springer. (SCI & Scopus Indexed IF: 5.086) *(Online)*

2. Veena Mayya, Sowmya Kamath S., Uma Kulkarni, "*Automated microaneurysms detection for early diagnosis of diabetic retinopathy: A comprehensive review*", Computer Methods and Programs in Bio-medicine Update, Elsevier, Volume 1, 2021. *(Online)*

Chapter 7

Multi-Scale CNN model for Corneal Segmentation for Early Detection of Fungal Keratitis

7.1 Introduction

The cornea is a transparent layer of tissue covering the front surface of the eye that acts as a window, allowing light to enter the eye. Infection is the most common cause of corneal ulcers (keratitis), and at least 4.2 million people worldwide are reported to suffer from corneal opacities, according to the 2019 World Vision Report (WHO, 2019). Corneal opacity is caused by a variety of conditions that cause the cornea to scar or become opaque. Microbial keratitis (MK) or infectious keratitis (IK) is the primary cause of corneal opacification, and the fifth leading cause of visual impairment in the developing world Ung *et al.* (2019). If such infections are not detected and treated early, they can cause irreversible corneal blindness due to perforation, endophthalmitis and panophthalmitis (Anutarapongpan and Brien, 2014; Schein, 2016; Maharana *et al.*, 2016; Tananuvat *et al.*, 2021).

FK, in particular, is challenging to treat at later stages and may necessitate surgery. The primary causative factors for FK are prolonged contact lens wear, topical steroid usage, trauma caused due to organic or vegetative matter, pre-existing systemic disorders, and other ocular surface issues (Kalkancı and Ozdek, 2011). It is a serious public health problem resulting in significant morbidity, if left untreated. Early interventions and treatment are critical for the recovery and prevention of corneal blindness (Kalkancı and Ozdek, 2011; Lopes *et al.*, 2019). Although ophthalmologists are trained to diagnose FK based on specific clinical signs and symptoms, isolation of fungi in micro-biological culture-based techniques

remains the gold standard for diagnosis. However, these are time-consuming and labor-intensive (Prajna *et al.*, 2017; Ferrer and Alió, 2011). Fungal organisms are slow growing and may not be florid in the early stages of FK. Fungal cultures may also have limited sensitivity due to the scant quantity of material accessible from corneal scrapings, which may, in turn, lead to false-negative results (Ferrer and Alió, 2011).

7.2 Problem Statement

Slit-lamp examination of the ocular surface, particularly the cornea, conjunctiva, and anterior chamber, is widely used in the diagnosis of MK. The most clinical signs based on which ophthalmologists differentiate corneal ulcers are infiltrate location, pattern, depth, epithelial defect size, surrounding stromal haze, and the presence (or absence) of hypopyon. Specifically, FK is associated with the occurrence of an uneven or feathery border, raised profile, deep stromal infiltrates, satellite lesions, endothelial plaques and/or pigmentation (Thomas *et al.*, 2005; Leck and Burton, 2015). However, the findings of corneal staining combined with slit-lamp biomicroscopy are heavily reliant on the grader's clinical knowledge.

As per studies conducted, general ophthalmologists typically differentiate FK from non-FK about 49.3–67.1% of the time, while trained corneal specialists can distinguish FK from non-FK 66.00–75.90% of the time (Dalmon *et al.*, 2012; Thomas *et al.*, 2005; Dahlgren *et al.*, 2007). These statistics are a significant cause for concern and highlight the need for effective automated diagnostic systems to assist doctors in the early detection of FK. Furthermore, certain clinical signs typically attributed to FK may also be of bacterial or protozoal origin, thereby complicating the diagnostic process. Automated grading of FK images could overcome these limitations by lowering physician burden and improving patient prognosis through early diagnosis. These systems can also aid in the timely diagnosis of FK by means of tele-ophthalmology in rural areas where there is a shortage of doctors. Thus, the problem to be addressed here is defined as follows:

> *Given the challenges associated with the differentiation of FK from non-FK, design and develop effective preprocessing, modelling and segmentation approaches for early diagnosis of FK based on slit-lamp biomicroscopy images.*

7.3 Motivating Example

To describe the prevailing conditions that emphasize the need for automated FK detection CDSS based on slit-lamp images, consider scenarios where the retinal images are captured by physicians using their portable devices during routine patient visits or during their mass campaigning in rural areas. Agriculture is a dominant occupation in rural areas, and FK is primarily caused by corneal injuries caused by flying debris from plant/soil materials. Since current diagnosis is primarily through microbial culture, limited access to both trained corneal experts and microbial culture labs in rural areas severely reduces the chances of early diagnosis. Often, minute lesions in the corneal surface may be missed even by trained corneal experts. There is also a possibility of inter- and intra-observational differences, i.e., based on a particular observation, the experts may infer the possibility of fungal, while, after a few hours/days when the same image is observed, the expert may find a very low probability of infection of fungal origin. There is also a possibility of difference of opinion between different experts when diagnosing based on the same images.

These challenges in the manual screening process could be avoided if the FK could be automatically identified using the slit-lamp imaging data. The slit-lamp biomicroscopy images tend to contain noise in the form of surrounding background that may affect the system's prediction. The automated disease detection CDSS could directly preprocess the slit-lamp images and detect the FK possibility, along with highlighting the regions that the system found to be relevant for detecting FK. Thus, avoid surgical complications associated with late diagnosis. Automated systems can aid in the early detection of FK via tele-ophthalmology in rural areas where there is a shortage of corneal specialists.

In this chapter, approaches towards developing effective FK detection models built on slit-lamp imaging data are presented. This study addresses several lacunae in existing approaches by utilizing automated corneal region segmentation to enhance the performance of evidence-based CDSS. In contrast to existing works, a two-stage approach is proposed that can potentially improve the performance of the etiological classification while also providing reliable evidence for predictions. The next section provides a detailed overview of the proposed methodology adopted for the early detection of FK. The performance of the proposed model is compared to that of the state-of-the-art region of interest (RoI) segmentation systems built on slit-lamp images.

7.4 KeratNet - Multi-task CNN for Corneal RoI Segmentation

The dataset provided by Loo *et al.* (2021*a*) consisting of 133 clinically suspected MK images were used for RoI segmentation task. 540 public-domain images were collated for FK detection task. This included FK (250) and non-FK images (290). Of the 290 images, 150 images were of *viral* keratitis, 120 images of *bacterial* keratitis and 20 images of *acanthamoeba* keratitis. These were obtained from various Web sources, and all are microbiologically validated cases which were used for meta-analysis purposes. This study was exempted from the purview of ethical clearance by Yenepoya Ethics Committee-1 (YEC-1/2021/046). The dataset provided by Loo *et al.* (2021*a*) was analysed by two ophthalmologists and labelled as *fungal*(1) or *non-fungal*(0), based on clinical observations. The clinically suspected MK images were assigned to the FK group if at least one of the ophthalmologists who participated in the study identified it as FK. Similarly, when both ophthalmologists labelled the images with non-FK, the images were assigned to the non-FK group. The corneal region annotation was performed using VGG Image annotator (Dutta and Zisserman, 2019)], after which the mask images were formed using the annotated regions.

The collated data was preprocessed and augmented before the RoI segmentation and classification phases. CLAHE algorithm (Contrast Limited Adaptive Histogram Equalization) (Reza, 2004) algorithm was utilized to increase the contrast and highlight the corneal border. All the images and the corneal masks were resized to 512×512. Before classifying into FK and non-FK, the images were scaled to ($width = 384 \times height = 256$) based on the training image normal distribution. The images were augmented by vertical and horizontal flipping of images to prevent the overfitting of the model. Rotated images at random angles ranging from 200° to 360° degrees were also included in each training batch.

The overall workflow of the proposed approach is depicted in Fig. 7.1. The slit-lamp biomicroscopy images tend to contain noise in the form of surrounding background that may affect the classification model's prediction. To enable the classification model to learn the minute changes in the corneal region, the RoI was segmented using the proposed KeratNet. The cropped RoI images were preprocessed and classified using a classification network. To highlight the features of an input slit-lamp image that the classification model considered relevant for prediction, the features were visualized using Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju *et al.*, 2017). Each process depicted is described in more

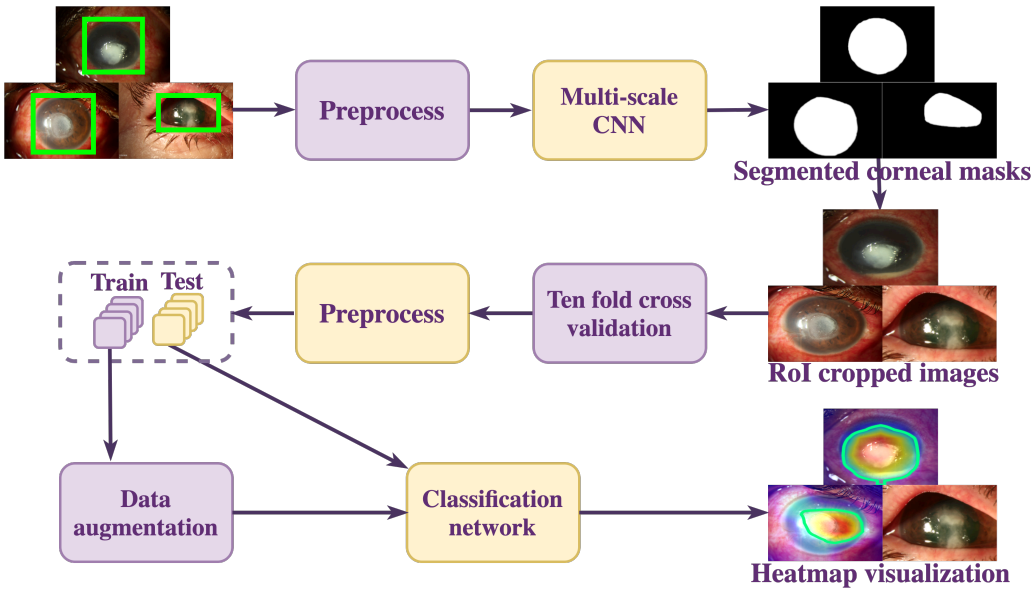


Figure 7.1: Proposed methodology for FK classification.

detail in subsequent sections.

Since the data collated in this study included images of varying dimensions, a KeratNet model is proposed for accurate segmentation of the corneal region. The network architecture is shown in Fig. 7.2 and is based on UNet (Ronneberger *et al.*, 2015*a*) and attention UNet (Abraham and Khan, 2019) architectures, which work well with modest training data. After improving the contrast of the corneal boundaries using CLAHE, the images were passed through a succession of convolution, and max-pooling layers for local feature extraction. The expansion layers were utilized to re-sample the image maps using extracted contextual information. Skip connections were utilized to encourage more semantically relevant outputs and handle varying resolution images to mix high-dimensional local characteristics with low-dimensional global information. The output of each dimension is then up-sampled and concatenated with the output from the first dimension. Ultimately, the resultant concatenation layer was subjected to a sigmoid non-linearity activation function and trained using binary cross-entropy loss to get the final corneal mask. Attention gates aided in learning the semantically important features.This technique increases segmentation accuracy for the dataset where tiny RoI features may be lost in cascading convolutions. Furthermore, the model can learn more location-aware features in relation to the classification objective. The corneal mask generated by KeratNet is used to automatically crop the RoI. The bounding rectangular region around the maximal contour is automatically cropped in the generated mask and used in the classification phase.

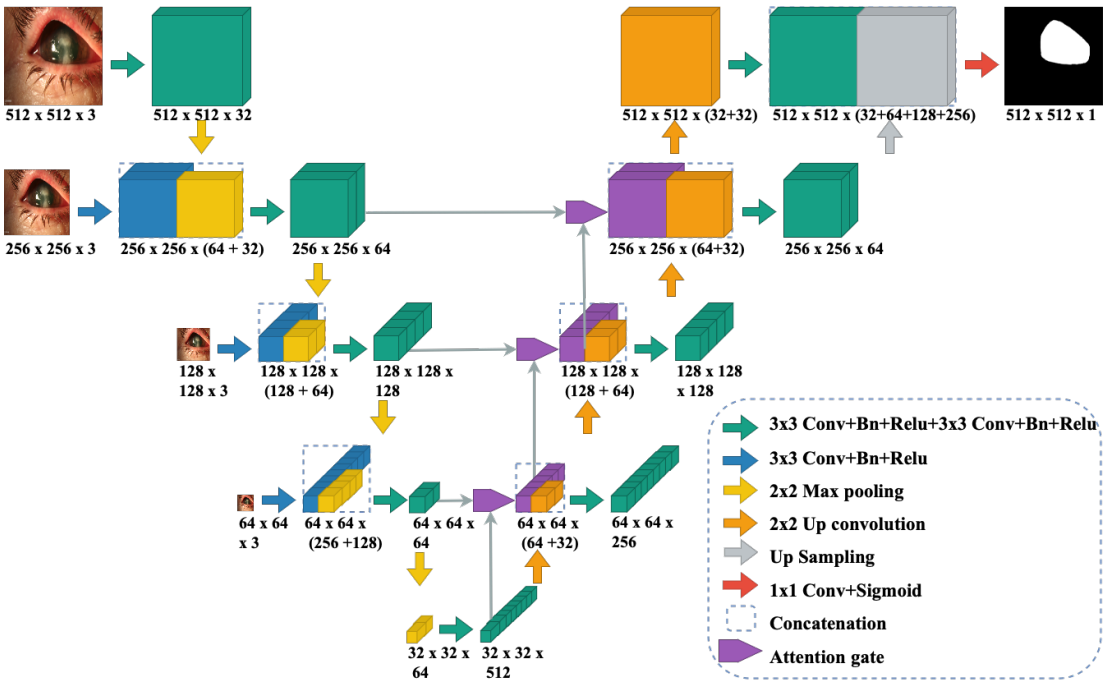


Figure 7.2: KeratNet architecture used for corneal RoI segmentation.

For classifying the RoI cropped images into FK and non-FK classes, transfer learning (with ImageNet pre-trained weights) based on the ResNeXt50 (Xie *et al.*, 2017*b*) architecture has been employed. ResNeXt50 is modularized based on VGG (Liu and Deng, 2015) and ResNet (He *et al.*, 2016). The multiple paths of ResNeXt50 share the same topology, and it has substantially fewer parameters than VGG. The coarse localization map (before *AdaptiveAvgPool2d*) of the last convolution layer represents the essential features in the input image to detect FK. Grad-CAM calculates attention scores based on gradients determined for the FK output. The attention scores are then normalized and resized to the size of the original image.

7.5 Experimental Results and Discussion

A detailed discussion on the experimental evaluation of the proposed methodology and the observations are presented in this section. For the implementation and training of the proposed approach, Python 3.8.8 with the Torch 1.8.0, Keras 2.4.3, and TensorFlow 2.2.0 as backend was utilized, running on Ubuntu OS with four NVIDIA Tesla V100-DGXS 32GB GPUs with CUDA v11.2. Based on the normal distribution of available images, the images were resized to ($width$ = 384 × $height$ = 256) dimensions. As per the available system configuration,

the batch size was set to 32 images. The model was trained for a maximum of 30 epochs, overall ten folds of the hold-out validation (Arlot and Celisse, 2010).

Several standard metrics were used for validating the proposed approach. Dice similarity coefficient (DSC) or F_1 (with configuration parameter $\beta = 1$) score and accuracy (Eq. 7.1) are used as primary metrics for validating the output over C classes. Dice coefficient is a weighted harmonic mean of positive predictive value (PPV) and true positive rate (TPR), and it seeks to strike a balance between the two (see Eq. 7.2). Both true/false positives (TP and FP) and true/false negatives (TN and FN) are accounted for in the dice coefficient/F1 measure. Thus, it is more informative than the conventional accuracy score. The positive and negative predictive values (NPV) are computed using Eq. (7.3). True positive and negative rates are computed as per Eq. (7.4).

$$\text{Accuracy} = \frac{1}{C}\sum_{c=1}^{C}\frac{\text{TP}_c + \text{TN}_c}{\text{TP}_c + \text{TN}_c + \text{FP}_c + \text{FN}_c} \tag{7.1}$$

$$F_{\beta=1} = (1+\beta^2)\cdot\frac{\text{PPV}\cdot\text{TPR}}{(\beta^2\cdot\text{PPV}) + \text{TPR}} \tag{7.2}$$

$$\text{PPV} = \frac{1}{C}\sum_{c=1}^{C}\frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c};\ \text{NPV} = \frac{1}{C}\sum_{c=1}^{C}\frac{\text{TN}_c}{\text{TN}_c + \text{FN}_c} \tag{7.3}$$

$$\text{TPR} = \frac{1}{C}\sum_{c=1}^{C}\frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c};\ \text{TNR} = \frac{1}{C}\sum_{c=1}^{C}\frac{\text{TN}_c}{\text{TN}_c + \text{FP}_c} \tag{7.4}$$

The performance of the proposed KeratNet model is observed using seven-fold cross-validation on all 133 diffuse white light images provided by Loo et al. Loo *et al.* (2021*a*). Table 7.1 lists the average dice similarity coefficient (DSC) values of KeratNet and state-of-the-art corneal limbus segmentation techniques. As is evident from Table 7.1, the proposed KeratNet outperformed the state-of-the-art model, SLIT-Net Loo *et al.* (2021*a*), by a margin of 1.42%. Furthermore, KeratNet requires only 5.67 million training parameters compared to 44.62 million for SLIT-Net, which is a 7× reduction. As a result, the proposed KeratNet is capable of faster training and inference while still enabling more accurate learning of the RoI even with variable sized input images. Fig. 7.3 shows a few samples of actual and predicted corneal region segments for the second test fold. It can be observed that the actual and segmented corneal limbus are in good agreement (see Fig. 7.3D).

The FK detection training loss for each fold and accuracy obtained with each

Table 7.1: Summary of average DSC of the proposed KeratNet and state-of-the-art corneal limbus segmentation methods, using diffuse white light images.

Method	**DSC (%)**	**Confidence Interval** (*with 0.05 Significance Level*)	**Training Parameters** (*in Millions*)
U-Net (Loo *et al.*, 2021*a*)	91	74–100%	34.51
U^2 Net (Qin *et al.*, 2020)	95.1	93.54–96.66%	44.01
nnU-Net (Loo *et al.*, 2021*a*)	96	11	–*
Mask R-CNN (Loo *et al.*, 2021*a*)	95	3	–*
SLIT-Net (Loo *et al.*, 2021*a*)	95	93–97%	44.62
KeratNet *(Proposed)*	96.42	95.65–97.19%	5.67

* Not reported by the authors.

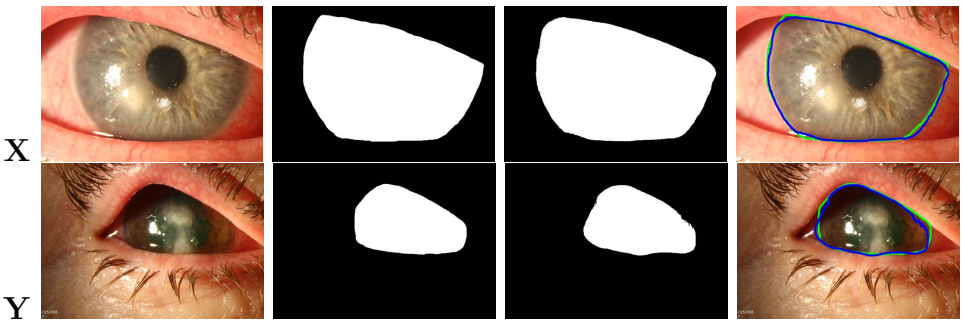


Figure 7.3: Sample of fully-automatic segmentation results by KeratNet on diffuse white light images. (**A**) Original images (source: Loo *et al.* (2021*a*)); (**B**) Actual masks obtained for images in (**A**); (**C**) Predicted masks for images in (**A**) using KeratNet; (**D**) Contour plots of actual (green) and predicted (blue) masks on original images.

test fold are plotted in Fig. 7.4. It can be seen that the loss converges after 12 epochs for all the folds. The accuracy stabilizes after a few initial variations, and the weights with which the model achieved highest precision are saved for each fold. Table 7.2 presents the details of prediction performance in terms of standard metrics. The confusion matrix obtained for all the ten test folds is shown in Fig. 7.5. The dominant features learned by the model to detect FK are visualized using gradient-weighted class activation mapping (Grad-CAM) Selvaraju *et al.* (2017). Fig. 7.6 shows the Grad-CAM visualization for the correctly diagnosed patients having FK. The maximal contour (illustrated using green marking in Fig. 7.6) is drawn using the heatmap mask produced when the threshold value is set to 100.

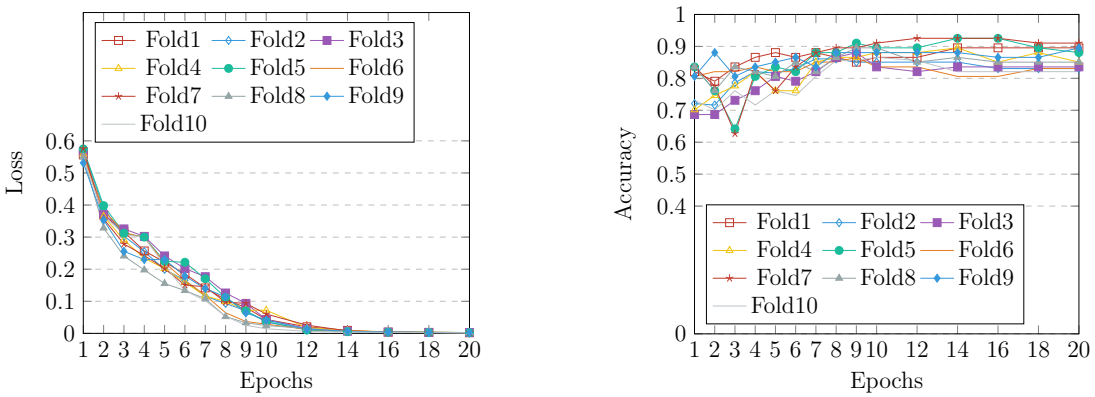


Figure 7.4: Observations w.r.t each of the 10 folds: (**A**) Loss vs. number of epochs (**B**) Accuracy vs. number of epochs.

Table 7.2: Performance evaluation of proposed approach with standard metrics.

Metric	**Mean Value**	**Confidence Interval** (*@0.05 Significance Level*)
Accuracy	88.96%	87.43–90.48%
Sensitivity/Recall/TPR	90.67%	87.95–93.39%
Specificity/TNR	87.57%	85.45–89.69%
Precision/PPV	85.65%	83.59–87.75%
Negative predictive values/NPV	92.18%	90.01–94.33%
F1/Dice coefficient score/DSC	88.01%	86.32–89.70%

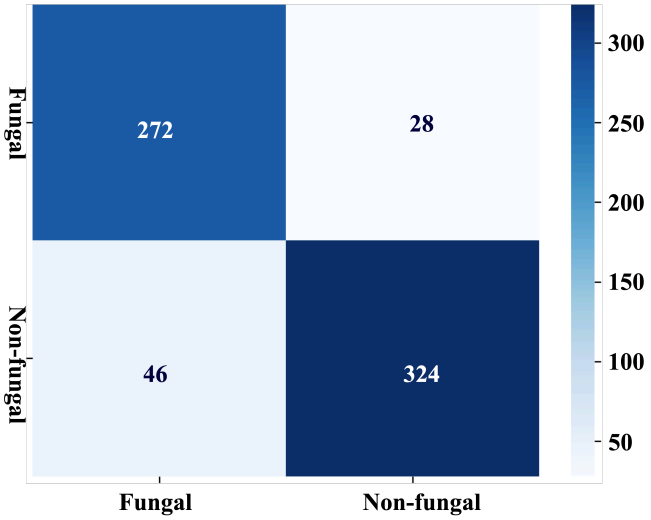


Figure 7.5: Confusion matrix obtained for KeratNet with 10-fold CV.

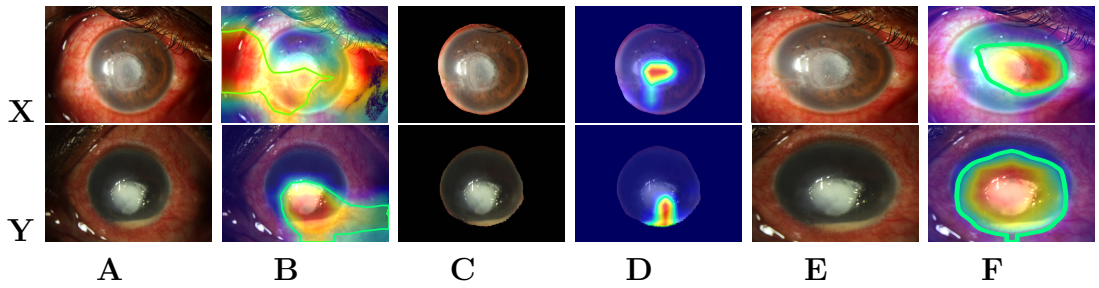


Figure 7.6: Sample Grad-CAM visualizations generated by the proposed model for correctly identified FK images. (**A**) Original keratitis images (source: Loo *et al.* (2021*a*)); (**B**) Grad-CAM visualizations for (**A**) images; (**C**) Automatic segmented corneal images in (**A**) using KeratNet; (**D**) Grad-CAM visualizations for (**C**) images; (**E**) Automatic RoI cropped images in (**A**) using KeratNet; (**F**) Grad-CAM visualizations for (**E**) images.

During the experiments, it was observed that the losses and accuracy between the folds are almost the same (refer Fig. 7.4), proving that the proposed model (with cropping) is generalizable for keratitis diagnosis, despite being trained on varied dimension images. This can be attributed to the effectiveness of the data augmentation and proposed RoI cropping process, which enabled strong focus on the FK lesions while avoiding any over-fitting. To understand the role and importance of the RoI cropping process in the prediction pipeline, model training using the original (non-cropped) images is also experimented with. The results revealed a higher variation in the between-fold mean and confidence interval values, when original images (without cropping) were used. This may be attributed to the fact that the model focuses on non-corneal areas (mainly conjunctiva region) for most of the identified FK images. However with cropping, the model is able to distinctly focus and learn the features from dominant lesions like epithelial defect, immune ring, satellite lesions, feathery margins and deep stromal infiltration for detecting FK (refer Fig. 7.6).

An ablation study was carried out for determining the contributions of various individual modules in the proposed FK detection approach. The results obtained for the first fold with the experimented methods have been summarized in Table 7.3. For obtaining the direct segmented corneal images, an approach similar to that proposed by Hung *et al.* (2021) was used—the pixels within the generated corneal mask were retained, while the remaining pixels in the original input image were set to zero. The segmented masks thus generated had noisy boundaries, clipped cornea regions and many black (zero valued) pixels, resulting in unfocused images. Scaling these images during training data preparation further deteriorated the images, making it difficult for the model to learn the required features. This is

evident from the visualized heatmaps (see Fig. 7.6D). When the direct segmented corneal images were used, the model failed to identify the required corneal lesions. In the proposed approach, RoI was initially cropped by using the bounding box around the generated corneal mask contour. Therefore, it is evident that scaling had minimal impact on the model's performance. The ablation study also revealed that the model's predictive performance degraded significantly with the absence of transfer learning (i.e., pre-trained weight initialization). Due to random initialization of network parameters when transfer learning is not used, inconsistent results were observed during each training run. In order to attain convergence and reliable results, the network must be trained over a larger number of epochs.

Table 7.3: Ablation study results for the proposed approaches.

Model	**Accuracy (%)**	**F1 Score (%)**
Proposed approach	89.55	89.23
Original images	$87.88^{-\theta}$	$87.59^{-\theta}$
Segmented corneal images	$84.62^{-\theta}$	$84.52^{-\theta}$
RoI cropped images	$76.12^{-\eta}$	$76.11^{-\eta}$

* *Proposed RoI cropping process is denoted by θ. The transfer learning is represented using η. Exclusion of a technique is indicated using $-$.*

Furthermore, the proposed approach correctly identified most cases of fungal and non-fungal (viral & bacterial) keratitis (refer confusion matrix shown in Figure 7.5). While false positive instances were noted in cases of acanthamoeba keratitis images, false negative instances appeared to be due to a lack of significant infiltration in the FK images. The latter could be attributed to images of patients who might have been in early stages of the disease or are undergoing medical or surgical treatment, thereby altering the morphology of the infectious infiltrate. In order to address the false negative performance, the poor quality images were removed, and more FK images were included than class-wise non-FK images. This may have improved the proposed model's performance. However, in order to achieve a high degree of precision, variability of the datasets is essential. Therefore, this limitation could be addressed by generating datasets with additional acanthamoeba, bacterial keratitis images, and also high-quality images. The proposed model detects FK primarily through identification of morphological changes in the corneal area, but the challenge of identifying a non-infectious corneal infiltrate, or a species-specific form of keratitis is yet to be addressed. Not just micro-organisms, but different species within a group can cause varied oc-

ular signs based on the presence or absence of polymicrobial association and/or periocular conditions. For example, candidal keratitis can cause a collar-stud like morphology as opposed to Fusarium keratitis that causes feathery branch–like extensions or a ring-shaped infiltrate.

7.6 Summary

Microbial keratitis is an infection of the cornea of the eye that is commonly caused by prolonged contact lens wear, corneal trauma, pre-existing systemic disorders and other ocular surface disorders. It can result in severe visual impairment if improperly managed. In this work, a multi-scale convolutional neural network (KeratNet) was proposed for accurate segmentation of the corneal region to enable early FK diagnosis. The proposed approach consisted of a deep neural pipeline for corneal region segmentation followed by a ResNeXt model to differentiate between FK and non-FK classes. The model trained on the segmented images in the region of interest, achieved a diagnostic accuracy of 88.96%. The features learnt by the model emphasize that it can correctly identify dominant corneal lesions for detecting FK.

Publications

(based on works presented in this chapter)

1. Veena Mayya, Sowmya Kamath S., Uma Kulkarni, Manali Hazarika, Prabal Datta Barua, U. Rajendra Acharya, "*Multi-scale convolutional neural network for accurate corneal segmentation in early detection of fungal keratitis*", Journal of Fungi, MDPI, Volume 10, 2021, Article Number 850, ISSN 2309-608X, (SCI & Scopus, IF: 5.816) *(Published)*

PART IV

AI-based CDSS using Multimodal Healthcare Data

Chapter 8

Learning Lung Disease Representations from Multimodal Radiology Data

8.1 Introduction

The COVID-19 pandemic has already spread on a global scale, affecting more than 280 million people with more than 5 million casualties worldwide. During the first and second waves of the pandemic, large-scale screening efforts were severely affected due to the short supply of COVID-19 test kits and the delay in notifying the test results. Recent works have studied the possibility of a speedy diagnosis of COVID-19 using chest X-ray images due to the wide-scale availability of X-ray machines and the low cost. Over the continued course of the pandemic, a significant volume of expert-written diagnosis reports has also been accumulated that capture a wide variety of symptoms and observations with reference to diagnosed COVID-19 cases. The utility of the rich, latent information embedded in such unstructured expert-written diagnosis reports has been mostly overlooked, and its importance as a source of valuable disease-specific information has been underexploited. A CDSS for predicting the presence or absence of COVID-19 infection using diagnostic scans can be beneficial to healthcare professionals as well as patients. A contactless X-ray scan workflow could be achieved by using cameras for patient monitoring purposes (Scheib, 2009; Forthmann and Pfleiderer, 2019), after a few days of onset of COVID-19-like symptoms, the patient is subjected to an X-ray. Such analysis can also contribute to isolating asymptomatic COVID-19 patients who undergo chest X-rays for other reasons (e.g., pre-operation evaluation, routine medical check-up, rib fractures etc.).

8.2 Problem Statement

The clinical testing period is one of the main reasons contributing to the fast spread of the COVID-19 pandemic. So the main aim is to construct an automated CAD system capable of detecting COVID-19 samples from healthy individuals and shortness of breath (SoB) patients using multimodal radiology data. Machine learning based medical image or text analysis and classification have seen extensive healthcare applications enabling COVID-19 management and improved diagnosis. Lee *et al.* (2021) examined the training techniques used to categorize CT scans into COVID-19 and non-COVID-19 classes, as well as performance disparities amongst 13 international institutions and eight nations. Rangarajan and Ramachandran (2021) tested the effectiveness of a transfer learned chest X-ray image classifier on a smartphone for detecting pneumonia, COVID-19, and normal cases. For categorizing chest X-ray images into pneumonia and COVID-19 pneumonia classes, Alhudhaif *et al.* (2021) used the transfer learning technique. Jain *et al.* (2021) also used the transfer learning technique for classifying chest X-ray images into normal, pneumonia and COVID-19 pneumonia classes. Shakarami *et al.* (2021) used the CNN network to extract the features and categorized the X-ray images into COVID-19 and non-COVID-19 groups. Sait *et al.* (2021) proposed the transfer learning-based CovScanNet, which classifies breathing sound spectrograms and X-ray images into normal, pneumonia, and COVID-19 pneumonia classes. Few researchers experimented with segmenting the lung region for COVID-19 classification and lesion visualization (Wang *et al.*, 2020*a*; Xu *et al.*, 2021). Several gaps were identified after reviewing existing works in this area. It was observed that the associated metadata of patients has not been considered in most works. Also, valuable expert-written diagnosis maintained as natural language text reports after checking a patient's chest radiography images have not been explored for the task of disease prediction. Furthermore, there is ample scope for the development of a complete, easy-to-use diagnostic framework for the use of healthcare professionals. Using such tools, expert opinion & other metadata about patients can also be obtained to incorporate relevance feedback into the prediction model, building accurate CDSSs. Thus, the problem to be addressed here is defined as follows:

> *Given the need for an easy-to-use prescreening and diagnostic framework for fast and accurate detection of COVID-19, design and develop effective comprehensive CDSS powered by ensemble deep learning models (CADNN) using multimodal radiology data.*

8.3 Motivating Example

Although the reverse transcription polymerase chain reaction (RT-PCR) is currently widely used as the leading COVID-19 diagnostic test, the turnaround time and expense of these tests are lengthy, necessitating the development of new rapid and accessible diagnostic techniques. To describe the prevailing conditions that emphasize the need for a comprehensive CDSS that make use of both textual and imaging healthcare data, consider scenarios where the chest X-ray images are captured by physicians during routine patient visits or during their mass campaigning in rural areas. Images captured may include minute changes in lung structure facilitating the COVID-19 diagnosis, which may be missed by human experts (due to the noise created by rib shadow and surrounding regions). There is also a possibility of difference of opinion between different experts when diagnosing based on the same images. The delay in processing and other challenges in COVID-19 detection could be avoided if the prediction scores could be automatically generated using the multimodal radiology data. The automated comprehensive CDSS could directly preprocess the multimodal radiology data and detect the COVID-19 possibility along with highlighting the regions (both in the image and textual report) that the system found to be relevant for detecting COVID-19. Thus, avoid complications associated with late diagnosis.

In this chapter, multiple deep learning models are incorporated for classifying X-ray images as COVID-19 positive or negative. The contributions of image features and the latent information contained in the expert-written diagnosis text reports are modelled for the diagnosis. To alleviate the manual effort required to assess and generate diagnosis reports when a large number of diagnosed cases arrive, a content-based report generation model for automatically generating natural language diagnosis reports is also designed for reducing the cognitive burden of radiologists and other medical personnel involved in medical record management. The complete framework is deployed on the cloud and is made available as a web application for managing patients metadata (from the day of admission till discharge). Functionalities like validity checks for X-ray images, evidence-based diagnosis support through highlighting of important features learnt by the model, and automatic report generation for further processing are incorporated in the proposed framework. A feedback system is provided to verify the prediction and generated reports, which is later utilized for improving the offline training process, for fine-tuning the prediction performance of the CDSS.

8.4 CADNN – Disease Diagnosis based on Multimodal Data

The objective of this work is to adapt and learn disease specific features from multiple modalities of data and utilize them for effective and accurate diagnosis of COVID-19. Fig. 8.1 depicts the overall methodology. A cloud-based application that allows learning disease representations from radiography images and early observations in the form of clinical notes has been deployed. Preprocessing and modelling of clinical notes and X-ray images were carried out separately. During the inference phase, the application visualizes the "heatmap" for both the clinical notes and the X-ray images.

For curating the dataset, a total of 150 confirmed COVID-19 patient cases were collected from publicly available sources. Each X-ray in the collated data has associated metadata – demographics details like age, gender and findings in the form of plain natural language text (reports) as observed by expert radiologists. COVID-19 X-ray images were also collated from available open datasets. In total, the dataset contained 450 chest X-ray images of COVID-19 infected patients, of which 150 images had associated metadata (clinical notes). In addition to this, about 2,000 normal case X-ray images were taken from the Pneumonia Detection Challenge (Radiological Society of North America, 2018). X-ray images and clinical note were prepossessed separately and are elaborated in section 8.4.1. A set of X-ray images along with their expert-generated diagnosis descriptions which were available from the IU dataset provided by Indiana University (Demner-Fushman *et al.*, 2016) was considered, for the shortness of breath (SoB) patients. The dataset was split as per the 70:30 ratio, i.e., 70% of the input records were used for training, and the remaining 30% of the records were utilized as a testset. This results in a total of 820 X-ray images with clinical notes.

8.4.1 X-ray Image and Report Preprocessing

As the X-ray images were captured using different machines, there exists a large variability, mainly in pixel intensities and focus on lung regions. To reduce the change in color intensities across the images, histogram matching (Gonzalez and Woods, 2008) was applied to the dataset, for which an X-ray image was considered as a reference image (R_{img}) and then matched all the other X-ray pixel intensity histograms with R_{img}. To suppress the effect of rib shadows on the classifier's prediction, *RIBDL* (Gusarev *et al.*, 2017) model was used. Then, the lung regions

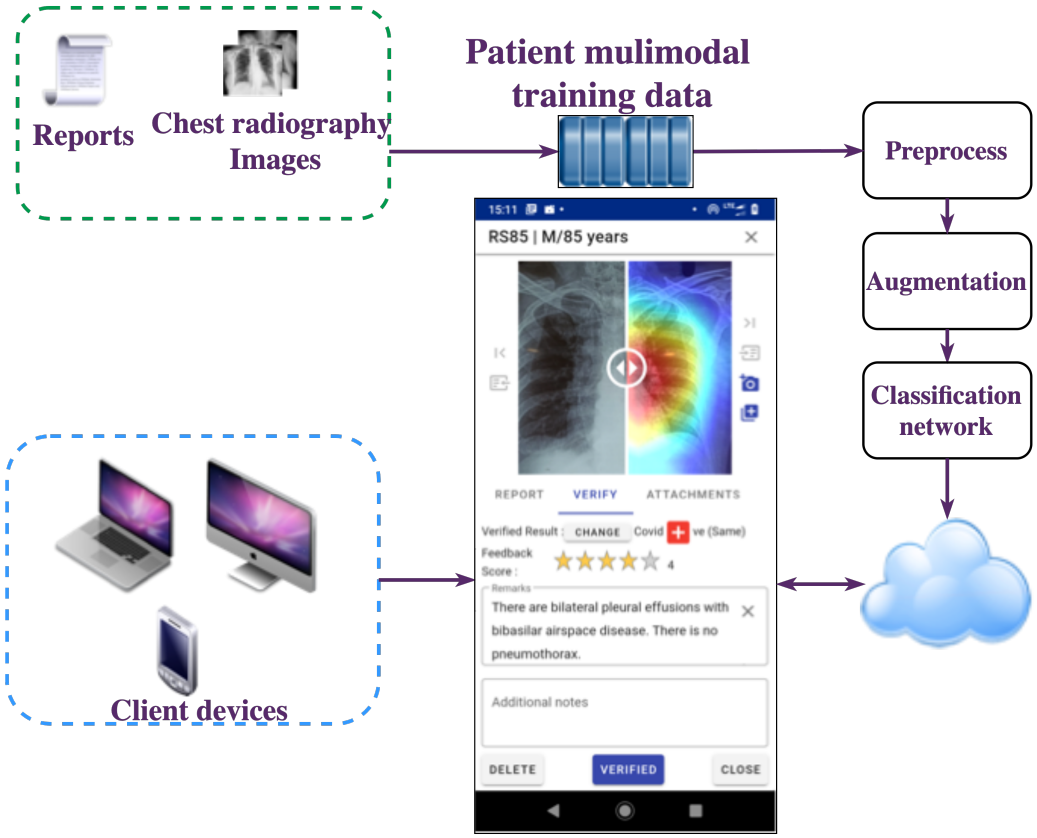


Figure 8.1: High-level design of the proposed CADNN framework

were segmented to reduce the effect of the surrounding background on the model's prediction. This enables the classification models to learn the minute changes in lung structure, which is often missed by human experts (due to the noise created by rib shadow). Algorithm 4 details the process of segmentation.

Initially, the lung region was segmented using *PIXGAN*, then the bounding box around the lung cavity is cropped and resized to the original size. The sample output after each step is depicted in Fig. 8.2. Cropping only the region of interest (RoI) allows the deep neural model to learn important features from within the lung regions. Crop and rotate facilities were also provided in the developed CADNN framework so that users could select only the lung region while uploading new test images. The local contrast of RoI segmented input X-ray grayscale images was further improved by applying Contrast Limited Adaptive Histogram Equalization (CLAHE) (Reza, 2004). Image augmentation was performed by applying rotation with different angles ($10° - 120°$) with an interval of $40°$ to the training images, which resulted in more than 3,000 COVID-19 and non-COVID-19 (a total of 3,474 training case samples).

Algorithm 4 Chest X-ray prepossessing pipeline

Input: Input chest X-ray images
Output: Preprocessed X-ray image

for each $img \in InputImages$ **do**
    Perform histogram matching of img with reference image.
    Perform rib shadow removal using *RIBDL*.
    Perform lung region segmentation using *PIXGAN* & resize img to 512x512
    Find the contours of the generated $MaskImg$
    Remove all small contours (with width < 50 and height < 50) in $MaskImg$
    Dilate $MaskImg$ with kernel size of (5,3) until a single contour is formed.
    Draw the bounding box over the single contoured $MaskImg$ ▷ *Set all other pixel intensities to zero.*
    Remove all black regions from input image and resize to required shape.
    Apply CLAHE.
end for

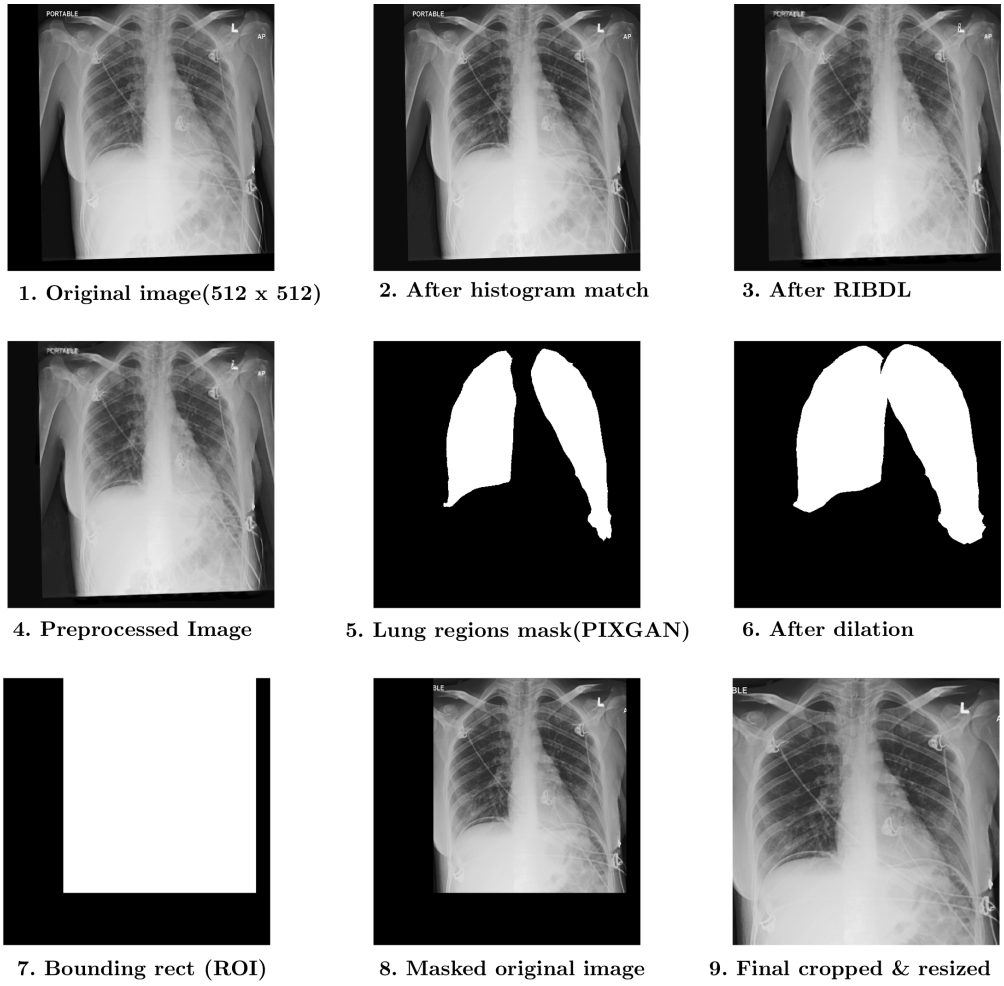


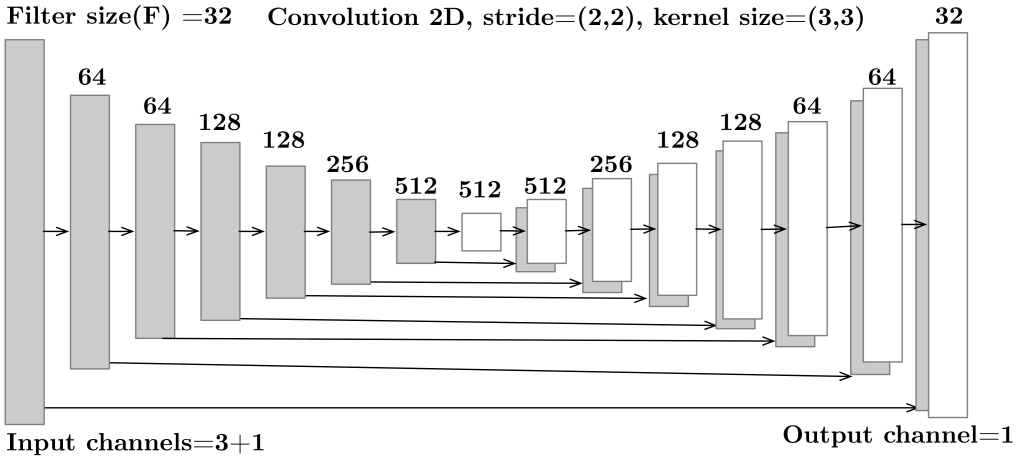
Figure 8.2: Stages of the chest X-ray preprocessing pipeline

As the collated dataset includes the X-ray images captured by different technicians using a variety of X-ray machines, substantial variation is observed in the area of focus. Some images included only lung regions, while many others covered the entire abdominal cavity. Also, several images from the collated dataset included X-ray machine labels in the form of characters/texts/numbers. To overcome variations and to restrict the classifier for effectively learning patterns from the lung region only, PIXGAN (Isola *et al.*, 2018) is trained from scratch to segment the RoI. The number of convolution and de-convolution layers were increased to handle the larger input image size (512, 512) and a modified loss function (Son *et al.*, 2017) that uses a tuning parameter (λ) was incorporated. λ was used while summing up the discriminator's binary cross entropy loss (among *predicted* and *true* labels) and generator's binary cross entropy loss (among *generated* and *true* lung masks). Thus, the generator enables the discriminator to produce outputs that are very similar to the real lung mask.

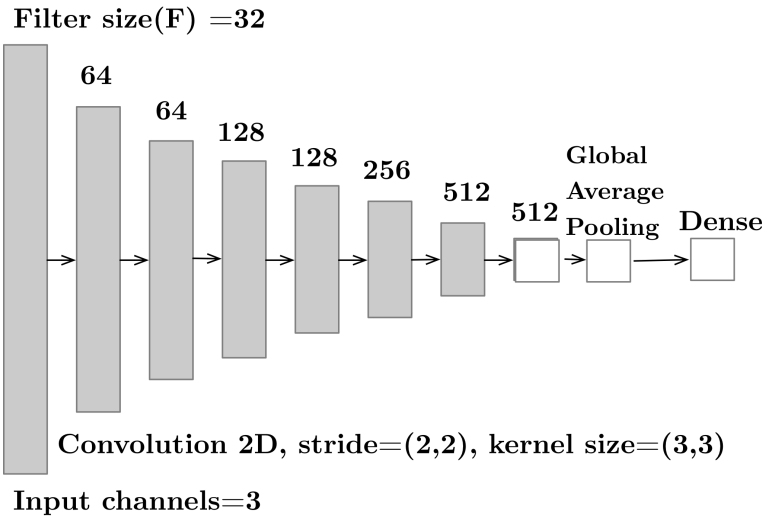
A total of 800 X-ray images and lung region masks obtained from the Kaggle challenge (Jaeger *et al.*, 2014; Candemir *et al.*, 2014)[1] was used to train the *PIXGAN*. The *PIXGAN* discriminator is fed both X-ray ($Image_{xray}$) and lung mask images ($Mask_{lung}$), which must determine whether $Mask_{lung}$ is a plausible transformation of $Image_{xray}$, as local style statistics are efficiently captured by *PIXGAN*. The generator is built on U-Net (Ronneberger *et al.*, 2015*a*) architecture and makes use of convolution and de-convolution layers for learning to generate realistic lungs mask from very few training X-ray images. Fig. 8.3 depicts the configuration of *PIXGAN* model which was used to segment the lung region for a given input X-ray image. The generator was designed using eight encoding and decoding units as shown in Fig. 8.3a, while the discriminator included the complete encoding part of generator network with global average pooling and final dense layers (shown in Fig. 8.3b). Once the lung region mask was generated, the original images were cropped to include only the lung region (as illustrated in Algorithm 4).

The expert reports consisting of physician observations contain a wealth of information regarding the condition, symptoms and other details regarding the patients' status. This rich latent information can be used to model patient representations, which can then be leveraged to potentially screen COVID-19 infected patients. Each report was subjected to preprocessing using standard natural language processing (NLP) techniques. Any punctuation, digits and stop words were removed from the patient's X-ray reports. Out of vocabulary(OOV) terms were

[1]https://www.kaggle.com/nikhilpandey360/chest-xray-masks-and-labels



(a) Generator network configuration



(b) Discriminator network configuration

Figure 8.3: *PIXGAN* for lung region segmentation

handled by including special OOV token, and the maximum allowed document length was fixed to 100. From the preprocessed text, embeddings were generated using the Word2Vec Continuous Bag-of-Words (CBoW) model (Mikolov *et al.*, 2013). The learning rate was fixed to a default value of 0.025 (same as that of Word2Vec model), the number of iterations was set to 10 and embedding size used was 200. Python Gensim library (Řehůřek and Sojka, 2010) was used to generate the word embeddings using the preprocessed X-ray reports.

8.5 Multimodal Data Modelling

As shown in Fig. 8.1, the proposed framework was built on the predictive framework powered by five neural models. Transfer learning was employed to obtain pretrained weights for the initial layers while some of the models were trained from scratch. In the designed application, users were provided with an interface to select and upload X-ray images from those available on their smartphones/system. However, there is a possibility of them knowingly or unknowingly uploading natural photographs. To accept only valid images, a two-layered convolution network *ValidateDL* was trained on the CIFAR10 dataset (Krizhevsky, 2012) (containing 60,000 32×32 color natural images) and 5,000 X-ray images to classify between natural images and X-ray images. For every batch, randomly selected CIFAR10 images were converted to grayscale and copied to all three channels, to ensue that even grayscale images are correctly classified by the model.

The configuration of the network is shown in Fig. 8.4. The network was trained with stochastic gradient descent (SGD) optimizer with learning rate of 0.01. Batch size was set to 32 and trained for a maximum of 20 epochs. Early stopping was used to prevent overfitting problems. The rib cage forms a significant part of the chest X-ray, and the rib shadow in the input images was suppressed using a pretrained autoencoder[2] model, to ensure that the training of the classification network focuses on relevant information within the lungs region instead of the rib structure. The pre-trained model was deployed during the validation phase of the online CDSS, to generate the rib suppressed image for all collated input images.

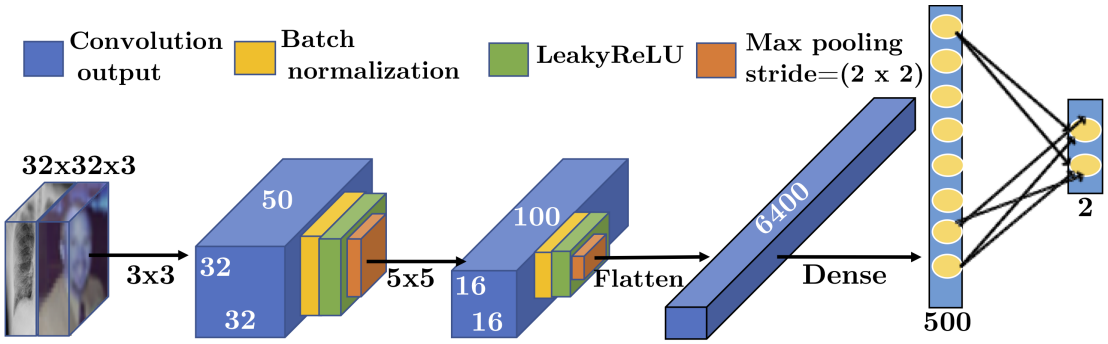


Figure 8.4: *ValidateDL* network architecture

These preprocessed images were trained using deep residual network (ResNet) (He *et al.*, 2015). ResNet-18 was used for training the classifier to distinguish between COVID-19 and non-COVID-19 cases. A content-based technique as de-

[2]https://github.com/hmchuong/ML-BoneSuppression

scribed in Section 8.5.1 was utilized to obtain the findings in the input X-ray image in the form of natural language textual report. The generated reports were pre-filled in the CADNN framework, so that the radiologists could verify and make the changes if necessary. The collated expert X-ray reports were classified into COVID and non-COVID using the proposed explainable report prediction deep learning model R$\mathscr{DX}$. This was mainly performed to highlight the important terms in pre-filled reports of CADNN framework. R$\mathscr{DX}$ model architecture is discussed in Section 8.5.1.1.

8.5.1 Diagnostic Report Generation

For this task, above trained ResNet18 was re-utilized. The last convolution layer output of ResNet18 provides a plausible disease representation of the input X-ray image. A feature vector ($features$) was generated by summing the last convolution layer output from the trained ResNet18. A dictionary ($D_{features}$) of feature vectors and reports indexed by image names was generated for the collated X-ray images for which expert reports were available. $D_{features}$ was also updated with frontal X-ray image features and textual findings obtained from the IU dataset (Demner-Fushman *et al.*, 2015). In total, 820 COVID and non-COVID reports along with corresponding frontal chest X-ray image features were utilized for report generation and report classification. For the given input test X-ray image, the image features ($Test_{feature}$) were extracted during classification along with the predicted label. Cosine similarity between $Test_{feature}$ and features of $D_{features}$ was computed. The report was obtained using index($Isimax$) of $D_{features}$ for which the maximum cosine similarity exists between $Test_{feature}$ and $D_{features}[Isimax]$. The generated reports were further processed by the X-ray report classifier.

8.5.1.1 Report Classification

As the main purpose of NLP classifier was to highlight important terms in the reports, a convolutional attention explainable neural network (R$\mathscr{DX}$) was designed to classify the X-ray reports. The model consists of a 1D convolution layer followed by an attention and a dense layer, which is depicted in Figure 8.5. All the reports were padded up to a maximum length of the reports in a batch, 100 being the maximum allowed length for a report. Given the test report, which is generated for the input X-ray image using the content-based approach, it is classified by the R$\mathscr{DX}$ model. The important terms contributing to the model's prediction are highlighted using color codes. This feature is also used to enable faster

verification features for experts through the functionalities provided through the CADNN interface, and to update/edit findings if necessary.

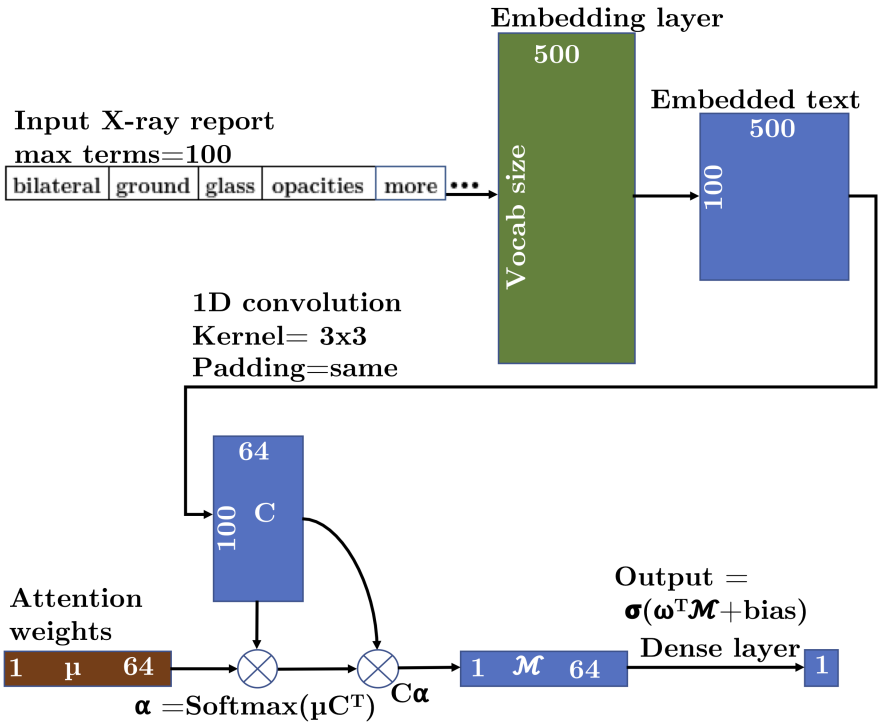


Figure 8.5: R$\mathscr{D}\mathscr{X}$ X-ray report classification DL model

8.6 Prediction Models

To choose an appropriate neural models, a through study was carried out on a smaller dataset. A total of 100 cases were used to put together a balanced dataset containing an equal number of COVID-19 and SoB cases. It is observed that SoB is also a primary symptom of early-stage infection; thus, experiments were conducted to train neural networks to assess and evaluate features for fine-grained differentiation between SoB and COVID-19 cases.

Expert reports based on physician observations provide a wealth of information on the patient's health, symptoms, and other pertinent factors. This wealth of latent information can be utilized to create patient representations, which can subsequently be used to possibly screen individuals infected with COVID-19 and SoB. A comprehensive experiment was conducted to determine the most effective neural model for report-based classification. Each report was subjected to preprocessing using standard natural language processing (NLP) techniques. Any punctuation

and digits were removed from the patient's X-ray reports. Stop words were retained due to the small dataset, and the maximum allowed document length was fixed to 100. From the preprocessed text, embeddings were generated using the Word2Vec Continuous Bag-of-Words (CBoW) model (Mikolov *et al.*, 2013). Table 8.1 lists the parameters used for generating the word embedding. The learning rate is fixed to a default value of 0.025 (same as that of the Word2Vec model), the number of iterations is set to 10, and the embedding size used is 200. For each report, a consolidated vector of the embedding of dimension 200 was generated by summing the embedding of words in the report. These were used to train three supervised classifiers - Logistic Regression, Decision Tree classifier and Support Vector Machine (with the linear kernel) for predicting COVID-19/SoB cases. State-of-art deep learning methods such as TextCNN (Kim, 2014), LSTM (Hochreiter and Schmidhuber, 1997) and GRU (Cho *et al.*, 2014) were applied to the chest X-ray text reports for predicting COVID-19. Values of hyperparameters like the number of hidden units, filter size, output channels, number of layers etc., were chosen through extensive experimentation and are listed in Table 8.2. As the size of the dataset is small, it was observed that increasing the hyperparameter values did not affect the model's prediction performance to a significant extent.

Table 8.1: Hyperparameter values used for the Word2Vec model

Hyperparameter	**Chosen value**
Number of iterations	10
Size of Summarized Vocabulary (post preprocessing)	559
Window size of the context in CBoW	5
Learning rate of neural model	0.025
Sizes of the word embedding	200

Table 8.2: Deep learning methods and hyper-parameters for chest X-ray report classification

Method	**Layer Description**
TextCNN	Embedding, Convolution1D, Linear
LSTM And Bi-LSTM	Embedding, LSTM, Linear
GRU and Bi-GRU	Embedding, GRU, Linear
CAD	Embedding, Conv1d, Linear (Attention), Linear

Interestingly, even though the dataset is small, it could be observed that deep learning methods outperform traditional supervised machine learning methods.

ML models use a fixed set of features per x-ray report, while DL methods learn multiple levels of representation from the input reports. The proposed R$\mathscr{DX}$ model performed better than the existing NLP classification DL models. The inclusion of an attention layer helps intensify focus on the most critical features of each COVID-19 or SoB report, rather than a uniform pooling operation for all codes, as is the case with TextCNN. R$\mathscr{DX}$ can easily be adapted for the larger dataset as it is more computationally efficient than recurrent neural networks like Bi-LSTM and Bi-GRU. Thus, R$\mathscr{DX}$ is utilized for designing the CADNN framework.

For report generation task, features extracted from the X-ray images and the expert-written diagnosis reports were modeled for automatically generating the reports of identified COVID-19 patients. The features from the X-ray images were extracted using the the last average pooling layer of the models, that were previously trained for classification task (Inception v3 and ResNet-18). During the feature extraction process, each image in the training set was used for creating a feature vector. When a test image feature set is given as a query, the pairwise distance measure is used to compute distances that can be used to obtain matching feature sets with the smallest distance from the images in the training set. Initially, eight different distance measures, Cosine, Correlation, Cityblock, Euclidean, Spearman, Minkowski, Euclidean and Chebychev, were used to check the closest distance measure among the test and training feature sets. For each observation in Y (*Test image features*), the pairwise distance method finds the smallest distances by computing and comparing the distance values to all the observations in X (*Training image features*). Based on experimentation performance, Cosine and Euclidean similarity measures are used to find the nearest feature set from the training set and its index is taken as the closest reference.

8.7 Experimental Results and Discussion

The proposed deep neural approaches were developed using Pytorch (Paszke *et al.*, 2019*b*) and Tensorflow (Abadi *et al.*, 2016) deep learning python frameworks and trained on Ubuntu 18.04 system with NVIDIA Tesla M40 and Tesla V100-DGXS. All the test data splits were made before image augmentation. Accuracy is used as an evaluation metric for verification of classification results, which is calculated based on true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) cases predicted by a particular neural model, and is given by Eq. (1). Here, TP is the number of cases that are correctly identified by the

prediction model to be COVID-19 positives, which match with experts' opinion, while FN are incorrectly rejected cases. TN is the number of correctly identified non-COVID-19 cases, and FP gives the total incorrectly identified COVID-19 cases.

$$Accuracy = \frac{TP + TN}{(TP + FP + FN + TN)} \tag{8.1}$$

Four different deep neural models (DenseNet-201, Inception v3 and Resnet-18) were experimented for classifying the chest X-ray images into COVID-19 and Shortness of Breath (SoB) cases. The results of the experiments and the performance achieved by the various models when applied to the test X-ray images are shown in Table 8.3. DenseNet-201 achieved the best overall accuracy on correctly classifying both COVID-19 and SoB cases. In the case of COVID-19, Inception v3 performed well by properly predicting all the test cases as COVID-19. Thus, the sensitivity is 100%. Also, it is noted that the features extracted using ResNet model contributed greatly towards the report generation task and Grad-CAM visualization. Thus, ResNet model is utilized for designing the CADNN framework.

Table 8.3: Performance evaluation of the chest X-ray image classification task

NN Model	**Accuracy**	**Sensitivity**	**Specificity**	**F1-Score**
ResNet-18	0.8730	0.9091	0.8333	0.8621
DenseNet-201	**0.9048**	0.9000	**0.8788**	**0.8850**
Inception v3	0.8889	**1.0**	0.7667	0.8679

To model the text reports in a more intuitive way, R$\mathscr{D}\mathscr{X}$ was designed to classify the X-ray reports. The model consists of a 1D convolution layer followed by an attention and a dense layer. First, all the reports were padded up to a maximum length of the reports in a batch, 100 being the maximum allowed length for a report. Next, a convolution layer with 64 filters and a kernel size of 3 was applied to the input embedding of reports. The resulting matrix was then fed into the attention layer followed by a dense layer resulting in a vector that determines the output class. The results for the test X-ray report generation task are listed in Table 8.4.

BLEU score (BiLingual Evaluation Understudy) proposed by Papineni *et al.* (2002) was utilized for assessing the quality of the generated texts. It evaluates the similarity between a candidate document and a collection of reference documents. As per an ordering from 1 to n, cumulative scores of individual n-grams can be calculated. N-gram is an evaluation of matching terms i.e., single word (1-gram), two word (2-gram or bigram) and so on. Weighing them together gives

Table 8.4: Performance evaluation of chest X-ray report classification task using ML classifiers and proposed CADNN model

NN Model	Accuracy	Sensitivity	Specificity	F1-Score
Logistic Regression	0.868	0.875	0.862	0.857
Decision Tree	0.868	0.958	0.793	0.868
SVM	0.868	0.917	0.828	0.862
TextCNN	0.906	0.792	1.000	0.8837
TextLSTM	0.943	0.917	0.966	0.936
TextLSTM-Bidirectional	0.943	0.875	1.000	0.933
TextGRU	0.887	0.750	1.000	0.857
TextGRU-Bidirectional	0.962	0.917	1.000	0.956
CAD	**0.981**	**0.958**	**1.0**	**0.979**

the geometric mean. In other words, for each i-gram where $i = 1, 2, 3 \ldots N$, the percentage of i-gram tuples in the candidate document that also occur in the reference document represented as BLEU-i is given by Eq. (8.2), where, C(i) is the number of i-gram tuples in the candidate document. In this work, suppose C="the lungs are clear" then C(*1*)=4, C(*2*)=3... C(*4*)=1.

$$BLEU - (i) = \frac{Matched(i)}{C(i)} \tag{8.2}$$

$$Matched(i) = \sum_{i=1} \min \{H_c(t_i), \max_j H_{cj}(t_i)\} \tag{8.3}$$

Here, (t_i) is an i-gram tuple in candidate C; $H_c(t_i)$ is the number of times (t_i) occurs in the candidate; $H_{cj}(t_i)$ is the number of times (t_i) occurs in reference j of this candidate. For calculating the BLEU score, each report in the training dataset is considered as a potential candidate, and the report generated by each ML/DL model for each test image is taken as a reference. The number of distinct sentences generated for the whole set (training and testing cases) is calculated separately. References are then evaluated with each of the different candidate sets and is depicted in Table 8.5. It could be observed that highest BLEU scores are achieved with ResNet. Thus, ResNet is utilized for designing the CADNN framework.

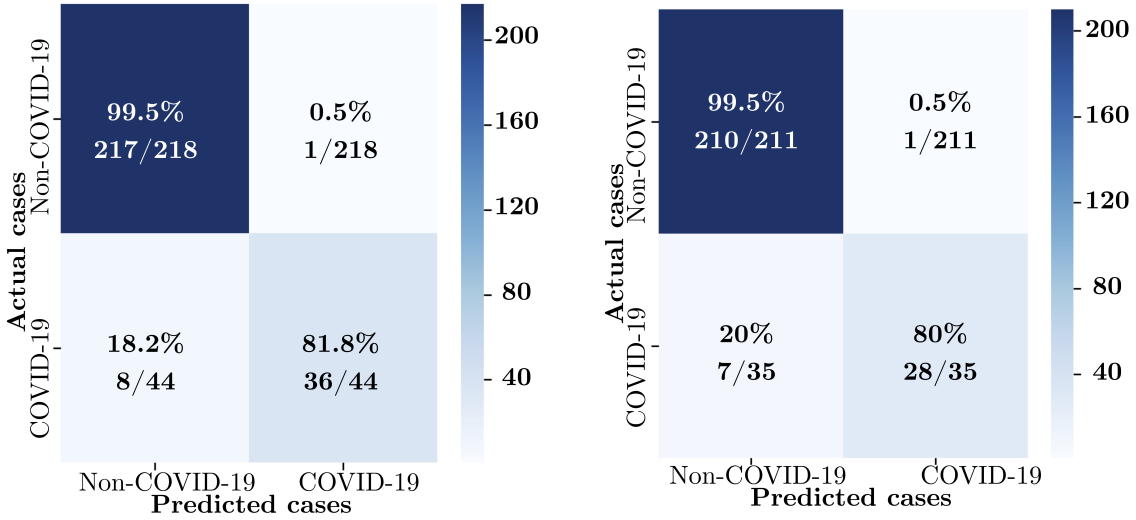
The *ValidateDL* model is evaluated on the CIFAR-10 test set and 450 test X-ray images from RSNA challenge data. The model was able to identify all X-ray images correctly and achieved 97.8% accuracy. Out of the 10,000 CIFAR-10

Table 8.5: Performance evaluation of chest X-ray report generation w.r.t BLEU scores

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
ResNet-18	0.9125	0.8872	0.6251	0.6211
DenseNet-201	0.8879	0.7790	0.6184	0.5867
Inception v3	0.8984	0.7008	0.6304	0.5637

images, 221 images that included only cloudy sky or runway images (which appear similar to X-ray image structure in 32 x 32 dimension) were wrongly classified as X-ray images instead of natural images. For training *PIXGAN*, batch size of 32, λ value of 0.5 and Adam optimizer with 0.0002 learning rate was used. Training was performed for a maximum of 100 epochs, and the *PIXGAN* was evaluated on 50 test images out of 800 X-ray images. The generator model that achieved the highest dice coefficient (Zijdenbos *et al.*, 1994) (obtained at 28^{th} epoch on validation data) was used to extract lung mask regions for the collated data.

For testing the ResNet18 X-ray image classification model, 80% of input data for training, 10% for validation and rest 10% was considered. The predicted and actual class for the X-ray images are summarized in the confusion matrix shown in Figure 8.6a. An accuracy of 97% was achieved using the proposed X-ray image classification model. The R$\mathscr{DX}$ model was evaluated on the 20% test split, and accuracy of 96.74% was observed. The results are summarized in the confusion matrix depicted in Fig. 8.6b.



(a) Chest X-ray dataset

(b) Diagnostic report dataset

Figure 8.6: Confusion matrix for different datasets

8.7.1 Qualitative Evaluation

For enabling evidence-based diagnosis, Class Activation Mapping (Grad-CAM) (Selvaraju *et al.*, 2017) was utilized to highlight the regions of the input X-ray image that the classification model considered relevant to perform the prediction. The visualization is shown in the CADNN framework to aid the clinical decision. The regions in the image, where this gradient is predominant, are shown in Figure 8.7), along with the generated report that shows the highlighted important terms. As can be observed from the attentions, the model has successfully learnt important features in the X-ray image, restricting itself mostly to within the lung region. The important radiography terms are identified (highlighted with white and red colors) by R$\mathscr{DX}$. This will enable CADNN users to validate whether correct regions/terms are learnt by the image/NLP models for the predicted output. In the case of x-ray reports, the significant terms of the precited output will be highlighted using the two colors. Blue shades denote terms that contribute the least to the output (darker shades denote irrelevance, whereas lighter hues influence¡0.5). White shades are the influential terms (=0.5). Red shades are most influential terms for predicting the corresponding output. This allows clinicians to verify if the model is able to learn the correct medical terminology and thus build trust to use the system in the clinical setup. For the X-ray images, the regions learned by the model for the predicted output are marked using contours/heatmaps. Such evaluation also aids in fine tuning the model. So, if due to the poor quality training data, if the model learnt non-relevant regions, more relevant data could be fed into re-train the models.

8.8 Summary

A cost-effective early lung disease screening strategy is crucial to prevent new outbreaks and to curtail the rapid spread. Chest X-ray images have been widely used to diagnose various lung conditions such as pneumonia, emphysema, broken ribs and cancer. In this work, the utility of chest X-ray images and available expert-written diagnosis reports were explored for training neural network models to learn disease representations for the diagnosis of COVID-19. A manually curated dataset consisting of 450 chest X-rays of COVID-19 patients and 2,000 non-COVID cases, along with their diagnosis reports, were collected from reputed online sources. Convolutional neural network models were trained on this multimodal dataset for the prediction of COVID-19 induced pneumonia. A com-

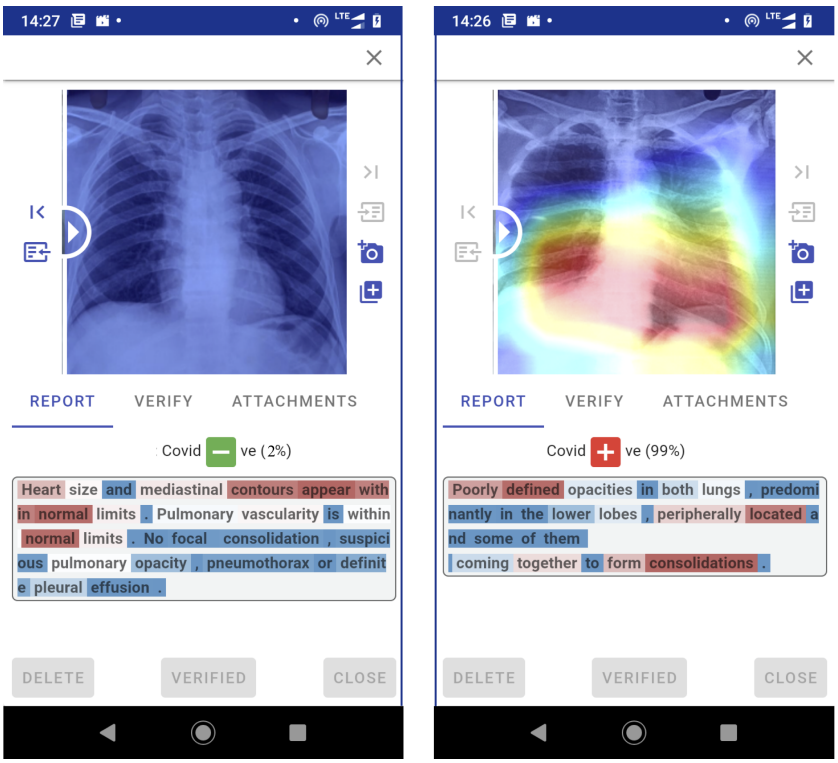


Figure 8.7: Proposed CADNN in action

prehensive clinical decision support system powered by ensemble deep learning models (CADNN) is designed and deployed on the web. The system also provides a relevance feedback mechanism through which it learns multimodal COVID-19 representations for supporting clinical decisions.

Publications

(based on works presented in this chapter)

1. Mayya V, Karthik, K., Kamath S., Karadka K, and Jeganathan J, "COVIDDX: AI-based clinical decision support system for learning COVID-19 disease representations from multimodal patient data", 14th International Joint Conference on Biomedical Engineering Systems and Technologies, HEALTHINF, BIOSTEC 2021, SciTePress, Vienna, Austria (Virtual), 2021. (CORE Ranked) *(Published)*.

2. Mayya V*, K Karthik*, Kamath S., Karadka K, "*Multi-task deep neural network models for learning COVID-19 disease representations from multimodal data*", International Journal of Medical Engineering and Informatics, Inderscience, ISSN 1755-0653, 2021, (Scopus) *(In press) (* Equal contribution.)*

Chapter 9

Conclusion & Future Work

9.1 Conclusion

Healthcare data is a key and essential data source, based on which CDSS can be built for revolutionizing the way in which personal healthcare is delivered and managed. For diagnosis and treatment, a computer-aided CDSS often plays a critical role and provides essential benefits to physicians. A CDSS could serve as an expert for a less experienced physician or as a backup option/opinion for an experienced physician while making the clinical decisions. CDSSs that provide precise diagnostic guidance and recommend cost-effective treatment options may benefit the target patient population. However, designing and developing a CDSS where precision of system performance is critical has been a significant challenge. An extensive literature analysis revealed significant potential for developing AI-based clinical decision support systems that use multimodal healthcare data. Several research gaps have been identified, most notably in the fields of diagnostic code assignment, ocular disease detection, and radiology.

Based on an extensive literature review, ample scope for improving the performance of automated disease code assignment systems, specifically addressing the challenges of utilizing unstructured discharge summaries was observed, which was actively pursued as part of the first research objective. In Chapter 4, *EnCAML*, a multi-channel, variable-sized convolutional attention model, was designed to enable the clinical task of diagnostic code assignment as a multi-label classification problem. It was demonstrated that the proposed model enhances the code predictability by extracting multi-granular text snippets, using which the attention mechanism enables the selection of those segments that are most contributing to the corresponding diagnostic code. Extensive benchmarking against several state-of-the-art models, including convolution-based models, sequence models, single-channel convolutional attention models, models employing transfer learning, and

others, revealed the efficacy of the proposed approach in modelling noisy, unstructured discharge summaries of the MIMIC-III corpus. In part, the reported high performance was attributed to the proposed preprocessing pipeline, which facilitated the effective pruning of irrelevant content in the free-text summaries. Furthermore, to demonstrate the robustness and adaptability of the proposed model, the minimal effect of the choice of initial embedding layer on the overall performance was also established. Finally, the enhanced interpretability of the prediction output of the *EnCAML* model was demonstrated using the learned per-code attention weights, thereby establishing the impact of the proposed model on instigating trust in intelligent healthcare systems.

A novel approach for automatic assignment of ICD-10 codes for the unstructured non-English clinical cases, the $\mathcal{LATA}$ was presented. $\mathcal{LATA}$ supports information aggregation across a patients' case reports via label attention encapsulated with transformer self-attention layers, aimed at extracting textual evidence that maps to the corresponding ICD-10 diagnostic code. It could be observed that $\mathcal{LATA}$ variants consistently outperformed basic BERT counterparts by a huge margin of 33-49%. A qualitative analysis was presented to demonstrate the model's ability to directly capture crucial input tokens contributing to particular ICD-10 diagnostic code through the label attention mechanism, despite the disparity in the length of the case reports.

Automated early COD diagnosis systems that use fundus and slit-lamp bi-microscopy imaging data are a critical requirement in Ophthalmology. A comprehensive study on the effectiveness of preprocessing techniques for automated COD diagnosis was studied. Experiments revealed that ResNeXt was most effective at modelling the very imbalanced and noisy ODIR dataset, when compared to the other state-of-the-art transfer learning approaches considered for the evaluation. It was demonstrated that the models trained on images processed using the proposed RoI segmentation algorithm outperformed those models trained on original non-cropped input images by a significant margin. The interpretability was demonstrated using the CNN learned features, thereby establishing the impact of the proposed RoI segmentation on instigating trust in intelligent healthcare systems. The experimental results showed that except for the RoI segmentation method, the other preprocessing strategies do not impact much on CNN performance. The proposed ensemble approach with batch-level augmentation was found to be superior when compared to state-of-the-art techniques, benchmarked on the ODIR-5K dataset.

Early diagnosis of FK is essential for clinical decision-making and can poten-

tially eliminate vision impairment. Existing manual screening approaches and corneal scraping for microbiological culture-senstivity tests are cumbersome and time-consuming. A multi-scale CNN model for automatic segmentation of corneal region combined with ResNeXt neural model was discussed for automated FK diagnosis. The Grad-CAM learnt features were visualized to illustrate the interpretability of the proposed pipeline, thereby instilling trust in intelligent healthcare systems. Experimental results showed that the proposed MS-CNN trained for segmentation of corneal region achieved superior performance for SLIT-Net dataset, underscoring its effectiveness against state-of-the-art methods.

Additionally, an emerging paradigm with the potential to drastically improve healthcare delivery models, that of leveraging multimodal data was explored. A cost-effective, early-screening strategy for COVID-19 diagnosis based on chest X-ray images and expert-written diagnosis reports was proposed and the framework has been deployed as a web-based CDSS called CADNN. The input images were subjected to extensive validation and preprocessing steps to eliminate variance and ensure effective learning by the prediction model. Preprocessed images were used to train a ResNet model for COVID-19, SoB or non-COVID-19 classification and the findings obtained from the images were used to automatically generate the natural language diagnosis reports, using content-based learning approach. The proposed CADNN also included feedback mechanisms so that the results could be verified by the experts, and feedback from experts can be utilized during offline retraining of the models. CADNN allows users to upload additional documents like CT scan images/DICOM sequences for additional insights into the patient's condition. The proposed framework could be easily adapted for diagnosis of other lung related diseases and provide a comprehensive CDSS support to medical professionals.

9.2 Future Directions

This thesis put forth several approaches towards the design and development of CDSS using multi-modal healthcare data. These approaches could be further extended to improve the existing CDS systems. Alternate sources of patient data could be accommodated for extending the models, especially in cases where the underlying discharge summaries are rather uninformative. Additionally, patient profiling can be extended via automated generation of summarized and well-formatted reports, sourced from multiple patient data sources, including discharge summaries, nursing notes, radiology reports, and various others. These

aggregated, rich semi-structured data sets can also be used to improve the interpretability and predictability of the underlying CDSS. Though the superiority of the $\mathcal{LATA}$ models over basic BERT variants was verified for the ICD-10 coding task, in general, the proposed model can be easily adapted for other domain sentence classification tasks too. This work can be extended to validate $\mathcal{LATA}$ model for procedural ICD-10 codes. Pre-training the encoder module of $\mathcal{LATA}$ on a large corpus for language modelling tasks potentially reduces false positive cases, which is to be experimented upon. Incorporating $\mathcal{LATA}$ for clinical case reports from other languages might potentially be considered for multilingual ICD-10 code assignment.

As part of extended work for diagnostic imaging based CDSSs, the approaches presented can be augmented with a detailed study of the impact of attention layers at various stages of inference using CNNs. Additionally, the possibility of using patient profiling via automatic generation of textual findings while considering both eye conditions can be explored to further improve the COD detection performance. The proposed CDSS makes use of only imaging data for early COD detection. A combination of patient metadata such as demographic factors (e.g., age, gender) and vital signs (blood pressure, blood sugar level) is not explored. There is scope for the development of CDSS that utilize such data from different modalities. The proposed CDSS uses the static images collected during the patient's single visit. Data from follow-up visits of patients can be utilized to identify the lesions that are likely to cause problems within the next few months. Collating more Acanthamoeba Keratitis images may further improve the performance of the proposed FK detection model and reduce the number of false positives. The model's prediction performance for various corneal lesions, such as corneal oedema border, ulcer border, degree of stromal infiltration, and height of hypopyon, needs to be studied further to understand its superiority.

The proposed framework that utilizes multimodal healthcare data can be easily adapted for the diagnosis of other lung-related diseases and provide comprehensive CDSS support to medical professionals. The approaches can be extended to use other imaging modalities, such as CT-scan, for potential improvements in performance. The final prediction can be generated based on an ensemble of these imaging modality classifiers along with the radiography text reports. At present, a content based approach is utilized to obtain the textual reports for the input X-ray images, experimentation with other automated text generation approaches using deep learning approaches can be carried out to further improve the performance and truly alleviate the cognitive burden of radiologists and clinical

personnel involved in medical record management.

The training samples were selected in accordance with the respective state-of-the-art studies to enable accurate benchmarking of the obtained performance. Data selection bias could occur and may affect the model's learning process and performance while validating with data obtained from real-world clinical setup. The development of the imaging-based CDSS utilized vast dataset collected with a variety of devices, and the use of data augmentation techniques at the input layer aided in enhancing the robustness of the proposed models and compensating for image quality-induced training data bias. By collecting multi-modal patient data and implementing suitable multi-modal workflows, the robustness of the proposed models could be enhanced further. While the proposed AI-based CDSS may aid in making clinical decisions during patient care, the ethical implications of its deployment to limit its potential harms, particularly for the most vulnerable, must be examined. Given the lack of practical methodologies or frameworks to evaluate adherence to sustaining ethical norms for AI-based CDSS (Karimian *et al.*, 2022), future research is necessary to build such tools. The interpretability of the proposed models through (Grad-CAM) (Selvaraju *et al.*, 2017) for both true and false positives enables to build confidence and reduce the trust barrier between doctors and automated systems. The interpretability of the proposed models using Grad-CAM for both true and false positives allows clinicians to gain confidence in CDSS and lower the trust barrier, other recent XAI techniques (Vilone and Longo, 2021), such as score deviation maps and recursive division methods, can be investigated.

Publications based on Research Work

Journal Publications

1. Veena Mayya, Sowmya Kamath S., Gokul Krishnan, Tushaar Gangavarapu, "*Multi-channel, convolutional attention based neural model for automated diagnostic coding of unstructured patient discharge summaries.*", Future Generation Computer Systems, Elsevier, Volume 118, 2021, Pages 374-391, ISSN 0167-739X, DOI: 10.1016/j.future.2021.01.013 (SCI & Scopus, IF: 7.187) *(Online)*

2. Veena Mayya, Sowmya Kamath S., Uma Kulkarni, Manali Hazarika, Prabal Datta Barua, U. Rajendra Acharya, "*Multi-scale convolutional neural network for accurate corneal segmentation in early detection of fungal keratitis*", Journal of Fungi, MDPI, Volume 10, 2021, Article Number 850, ISSN 2309-608X, DOI: 10.3390/jof7100850 (SCI & Scopus, IF: 5.816) *(Online)*

3. Veena Mayya, Sowmya Kamath S., Uma Kulkarni, Divyalakshmi Kaiyoor Surya, U. Rajendra Acharya, "*An empirical study of preprocessing techniques with convolutional neural networks for accurate detection of chronic ocular diseases using fundus images*", Applied Intelligence, Springer, ISSN 1573-7497, DOI: 10.1007/s10489-022-03490-8 (SCI & Scopus Indexed IF: 5.086) *(Online)*

4. Veena Mayya*, Karthik K.*, Sowmya Kamath S., Karadka K. P., "*Multi-task deep neural network models for learning COVID-19 disease representations from multimodal data*", International Journal of Medical Engineering and Informatics, Inderscience, ISSN 1755-0653, 2021, DOI: 10.1504/IJMEI .2021.10043617 (Scopus) * Equal contribution *(In press)*

5. Veena Mayya, Sowmya Kamath S., Uma Kulkarni, "*Automated microaneurysms detection for early diagnosis of diabetic retinopathy: A comprehensive review*", Computer Methods and Programs in Bio-medicine Update, Elsevier, ISSN 2666-9900, 2021, DOI: 10.1016/j.cmpbup.2021.100013. *(Online)*

Conference Publications

1. Veena Mayya, Sowmya Kamath S. and Vijayan Sugumaran, "$\mathcal{LATA}$– Label Attention Transformer Architectures for ICD-10 Coding of unstructured clinical Notes", In the proceedings of the 18th International Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE, Australia (Virtual), October 2021, Pages 1-7, DOI: 10.1109/CIBCB49929.2021.9562815. (CORE Ranked) *(Online)*.

2. Veena Mayya V, Karthik, K., Sowmya Kamath S., Karadka K. P., and Jeganathan J, "COVIDDX: AI-based clinical decision support system for learning COVID-19 disease representations from multimodal patient data", In the proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies, HEALTHINF, BIOSTEC 2021, Vienna, Austria (Virtual), February 2021, 659-666. (CORE Ranked) *(Online)*.

References

Abadi, M., P. Barham, J. Chen, *et al.*, Tensorflow: A system for large-scale machine learning. *In 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. 2016.

Abraham, N. and N. Khan (2019). A novel focal tversky loss function with improved attention u-net for lesion segmentation. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 683–687.

Acharya, R., W. Yu, K. Zhu, J. Nayak, T.-C. Lim, and J. Chan (2010). Identification of cataract and post-cataract surgery optical images using artificial intelligence techniques. *Journal of Medical Systems*, 34(4), 619–628. ISSN 01485598.

Acharya, U., S. Dua, X. Du, V. Sree S, and C. Chua (2011). Automated diagnosis of glaucoma using texture and higher order spectra features. *IEEE Transactions on Information Technology in Biomedicine*, 15(3), 449–455. ISSN 10897771.

Agarwal, S., L. Vasudevan, T. Tamrat, C. Glenton, S. Lewin, H. Bergman, N. Henschke, G. Mehl, and M. Fønhus (2018). Digital tracking, provider decision support systems, and targeted client communication via mobile devices to improve primary health care. *Cochrane Database of Systematic Reviews*, 2018.

Akut, R. (2019). Film: finding the location of microaneurysms on the retina. *Biomedical Engineering Letters*, 9(4), 497–506. ISSN 20939868.

Al-Bander, B., B. M. Williams, W. Al-Nuaimy, M. A. Al-Taee, H. Pratt, and Y. Zheng (2018). Dense fully convolutional segmentation of the optic disc and cup in colour fundus for glaucoma diagnosis. *Symmetry*, 10(4). ISSN 2073-8994.

Al-Diri, B., F. Calivá, P. Chudzik, G. Ometto, and M. Habib, Chapter 12 - diabetic retinopathy and maculopathy lesions. *In* E. Trucco, T. MacGillivray, and Y. Xu (eds.), *Computational Retinal Image Analysis*, The Elsevier and MICCAI Society Book Series. Academic Press, 2019. ISBN 978-0-08-102816-2, 223–243.

Algaze, C., M. Wood, N. Pageler, P. Sharek, C. Longhurst, and A. Shin (2016). Use of a checklist and clinical decision support tool reduces laboratory use and improves cost. *Pediatrics*, 137(1). ISSN 00314005.

Alhudhaif, A., K. Polat, and O. Karaman (2021). Determination of covid-19 pneumonia based on generalized convolutional neural network model from chest x-ray images. *Expert Systems with Applications*, 180, 115141.

Almagro, M., R. Martínez-Unanue, V. Fresno-Fernández, S. Montalvo, and H. Tissot, Icd-10 coding based on semantic distance: Lsi_uned at clef ehealth 2020 task 1. *In CLEF*. 2020.

Alsentzer, E., J. Murphy, W. Boag, W.-H. Weng, *et al.*, Publicly available clinical BERT embeddings. *In Proceedings of the 2nd Clinical Natural Language Processing Workshop*. 2019.

Amirfar, S., J. Taverna, S. Anane, and J. Singer (2011). Developing public health clinical decision support systems (cdss) for the outpatient community in new york city: Our experience. *BMC public health*, 11, 753.

Anutarapongpan, O. and T. Brien (2014). Update on management of fungal keratitis. *Clin Microbial*, 3(5).

Aramaki, E., Y. Kano, T. Ohkuma, and M. Morita, MedNLPDoc: Japanese shared task for clinical NLP. *In Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*. The COLING 2016 Organizing Committee, Osaka, Japan, 2016. URL `https://aclanthology.org/W16-4203`.

Arlot, S. and A. Celisse (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40–79.

Awad, A., M. B. Bader-El-Den, and J. McNicholas (2017). Patient length of stay and mortality prediction: A survey. *Health Services Management Research*, 30, 105 – 120.

Bagchi, S. (2008). Growth generates health care challenges in booming india. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, 178(8), 981 – 983. ISSN 14882329.

Baumel, T., J. Nassour-Kassis, R. Cohen, M. Elhadad, and N. Elhadad, Multi-label classification of patient notes: case study on ICD code assignment. *In*

Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence. 2018*a*.

Baumel, T., J. Nassour-Kassis, R. Cohen, M. Elhadad, and N. Elhadad, Multi-label classification of patient notes: Case study on ICD code assignment. *In The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, AAAI Workshops. AAAI Press, 2018*b*.

Berner, E. S., *Clinical Decision Support Systems: Theory and Practice*. Springer Publishing Company, Incorporated, 2010, 2nd edition. ISBN 1441922237.

Birkhead, G. S., M. Klompas, and N. R. Shah (2015). Uses of electronic health records for public health surveillance to advance public health. *Annu Rev Public Health*, 36, 345–359.

Biyani, R. and B. Patre (2018). Algorithms for red lesion detection in diabetic retinopathy: A review. *Biomedicine & Pharmacotherapy*, 107, 681 – 688. ISSN 0753-3322.

Bizopoulos, P. and D. Koutsouris (2019). Deep learning in cardiology. *IEEE Reviews in Biomedical Engineering*, 12, 168–193.

Blackmore, C., R. Mecklenburg, and G. Kaplan (2011). Effectiveness of clinical decision support in controlling inappropriate imaging. *Journal of the American College of Radiology*, 8(1), 19–25. ISSN 15461440.

Blanco, A., A. Pérez, and A. Casillas, Ixa-aaa at clef ehealth 2020 codiesp. automatic classification of medical records with multi-label classifiers and similarity match coders. *In CLEF*. 2020.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146.

Buisson, M., V. Navel, A. Labbé, S. Watson, J. Baker, P. Murtagh, F. Chiambaretta, and F. Dutheil (2021). Deep learning versus ophthalmologists for screening for glaucoma on fundus examination: A systematic review and meta-analysis. *Clinical and Experimental Ophthalmology*, 49(9), 1027–1038. ISSN 14426404.

Burlina, P., N. Joshi, K. Pacheco, T. Liu, and N. Bressler (2019). Assessment of deep generative models for high-resolution synthetic retinal image generation of age-related macular degeneration. *JAMA Ophthalmology*, 137(3), 258–264. ISSN 21686165.

Cabana, M. D., C. S. Rand, N. R. Powe, A. W. Wu, M. H. Wilson, P.-A. C. Abboud, and H. R. Rubin (1999). Why Don't Physicians Follow Clinical Practice Guidelines?A Framework for Improvement. *JAMA*, 282(15), 1458–1465. ISSN 0098-7484.

Cai, Q., H. Wang, Z. Li, and X. Liu (2019). A survey on multimodal data-driven smart healthcare systems: Approaches and applications. *IEEE Access*, 7, 133583–133599. ISSN 21693536.

Caicedo-Torres, W. and J. Gutierrez (2019). Iseeu: Visually interpretable deep learning for mortality prediction inside the icu. *Journal of Biomedical Informatics*, 98, 103269. ISSN 1532-0464.

Calloway, S., H. A. Akilo, and K. W. Bierman (2013). Impact of a clinical decision support system on pharmacy clinical interventions, documentation efforts, and costs. *Hospital Pharmacy*, 48, 744 – 752.

Candemir, S., S. Jaeger, K. Palaniappan, J. P. Musco, R. K. Singh, Z. Xue, A. Karargyris, S. Antani, G. R. Thoma, and C. J. McDonald (2014). Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Transactions on Medical Imaging*, 33, 577–590.

Catling, F., G. P. Spithourakis, and S. Riedel (2018). Towards automated clinical coding. *International journal of medical informatics*, 120, 50–61.

Cennamo, G., D. Montorio, M. Carosielli, M. Romano, and G. Cennamo (2021). Multimodal imaging in choroidal metastasis. *Ophthalmic Research*, 64(3), 411–416. ISSN 00303747.

Chalakkal, R., F. Hafiz, W. Abdulla, and A. Swain (2021). An efficient framework for automated screening of clinically significant macular edema. *Computers in Biology and Medicine*, 130, 104128. ISSN 0010-4825.

Chen, C., J. H. Chuah, R. Ali, and Y. Wang (2021). Retinal vessel segmentation using deep learning: A review. *IEEE Access*, 9, 111985–112004.

Chen, Y., H. Lu, and L. Li (2017). Automatic icd-10 coding algorithm using an improved longest common subsequence based on semantic similarity. *PloS one*, 12(3), e0173410.

Cho, K., B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation. *In Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. 2014.

Chudzik, P., S. Majumdar, F. Calivá, B. Al-Diri, and A. Hunter (2018). Microaneurysm detection using fully convolutional neural networks. *Computer Methods and Programs in Biomedicine*, 158, 185–192. ISSN 01692607.

Civit-Masot, J., M. Dominguez-Morales, S. Vicente-Diaz, and A. Civit (2020). Dual machine-learning system to aid glaucoma diagnosis using disc and cup feature extraction. *IEEE Access*, 8, 519–529.

Clarke, E., J. Evans, and L. Smeeth (2018). Community screening for visual impairment in older people. *Cochrane Database of Systematic Reviews*, 2018(2). ISSN 1469493X.

Coiera, E., *A Guide to Health Informatics*. CRC Press, 2003.

Cossin, S. and V. Jouhet, Iam at clef ehealth 2020: Concept annotation in spanish electronic health records. *In CLEF*. 2020.

Costa, J., I. Lopes, A. Carreiro, D. Ribeiro, and C. Soares, Fraunhofer aicos at clef ehealth 2020 task 1: Clinical code extraction from textual data using fine-tuned bert models. *In CLEF*. 2020.

Costa, P., A. Galdran, M. I. Meyer, M. Niemeijer, M. Abràmoff, A. M. Mendonça, and A. Campilho (2018). End-to-end adversarial retinal image synthesis. *IEEE Transactions on Medical Imaging*, 37(3), 781–791.

Cowie, M. R., J. I. Blomster, L. H. Curtis, S. Duclaux, I. Ford, F. Fritz, S. Goldman, S. Janmohamed, J. Kreuzer, M. Leenay, A. Michel, S. Ong, J. P. Pell, M. R. Southworth, W. G. Stough, M. Thoenes, F. Zannad, and A. Zalewski (2017). Electronic health records to facilitate clinical research. *Clinical research in cardiology : official journal of the German Cardiac Society*, 106(1), 1–9. ISSN 1861-0692.

Cox, J., V. Sadiraj, K. Schnier, and J. Sweeney (2016). Higher quality and lower cost from improving hospital discharge decision making. *Journal of Economic Behavior and Organization*, 131, 1–16. ISSN 01672681.

Dahlgren, M. A., A. Lingappan, and K. R. Wilhelmus (2007). The clinical diagnosis of microbial keratitis. *American Journal of Ophthalmology*, 143(6), 940–944.e1. ISSN 0002-9394.

Dalmon, C. A., T. Porco, T. Lietman, N. Prajna, L. Prajna, M. Das, J. A. Kumar, J. Mascarenhas, T. Margolis, J. Whitcher, B. Jeng, J. Keenan, M. Chan, S. McLeod, and N. Acharya (2012). The clinical differentiation of bacterial and fungal keratitis: a photographic survey. *Investigative ophthalmology & visual science*, 53 4, 1787–91.

Darabi, H. R., D. Tsinis, K. Zecchini, D. Tsinis, K. Liss, F. Whitcomb, W. F. Whitcomb, and A. Liss (2018). Forecasting Mortality Risk for Patients Admitted to Intensive Care Units Using Machine Learning. *Procedia Computer Science*, 140, 306–313. ISSN 1877-0509.

Davis, D. A. and A. L. Taylor-Vaisey (1997). Translating guidelines into practice. a systematic review of theoretic concepts, practical experience and research evidence in the adoption of clinical practice guidelines. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, 157 4, 408–16.

Davoodi, R. and M. H. Moradi (2018). Mortality prediction in intensive care units (ICUs) using a deep rule-based fuzzy classi fi er. *Journal of Biomedical Informatics*, 79(February), 48–59. ISSN 1532-0464.

de la Iglesia, I., M. Martínez-Puente, A. Platas, I. S. Miguel, A. Atutxa, and K. Gojenola, Media team: Clef-2020 ehealth task 1: Multilingual information extraction - codiesp. *In CLEF*. 2020.

de la Torre, J., A. Valls, and D. Puig (2019). A deep learning interpretable classifier for diabetic retinopathy disease grading. *Neurocomputing*. ISSN 0925-2312.

Demner-Fushman, D., M. Kohli, M. Rosenman, S. E. Shooshan, L. M. Rodriguez, S. Antani, G. Thoma, and C. McDonald (2016). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association : JAMIA*, 23(2), 304–10.

Demner-Fushman, D., M. Kohli, *et al.* (2015). Preparing a collection of radiology examinations for distribution and retrieval. *JAMIA*, 23.

Dermouche, M., J. Velcin, R. Flicoteaux, S. Chevret, and N. Taright, Supervised topic models for diagnosis code assignment to discharge summaries. *In International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2016.

Deshmukh, A. and J. Sivaswamy, Synthesis of optical nerve head region of fundus image. *In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. 2019.

Deshmukh, M., Y.-C. Liu, T. H. Rim, A. Venkatraman, M. Davidson, M. Yu, H. S. Kim, G. Lee, I. Jun, J. S. Mehta, and E. K. Kim (2021). Automatic segmentation of corneal deposits from corneal stromal dystrophy images via deep learning. *Computers in Biology and Medicine*, 137, 104675. ISSN 0010-4825.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of NAACL: Human Language Technologies, Volume 1*. 2019*a*.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. Association for Computational Linguistics, Minneapolis, Minnesota, 2019*b*.

Diaz-Pinto, A., A. Colomer, V. Naranjo, S. Morales, Y. Xu, and A. F. Frangi (2019*a*). Retinal image synthesis and semi-supervised learning for glaucoma assessment. *IEEE Transactions on Medical Imaging*, 38(9), 2211–2218.

Diaz-Pinto, A., S. Morales, V. Naranjo, T. Köhler, J. Mossi, and A. Navea (2019*b*). Cnns for automatic glaucoma assessment using fundus images: An extensive validation. *BioMedical Engineering Online*, 18(1). ISSN 1475925X.

Dimagno, M., E.-J. Wamsteker, R. Rizk, J. Spaete, S. Gupta, T. Sahay, J. Costanzo, J. Inadomi, L. Napolitano, R. Hyzy, and J. Desmond (2014). A combined paging alert and web-based instrument alters clinician behavior and shortens hospital length of stay in acute pancreatitis. *American Journal of Gastroenterology*, 109(3), 306–315. ISSN 00029270.

Domingues, I., G. Pereira, P. Martins, H. Duarte, J. Santos, and P. H. Abreu (2019). Using deep learning techniques in medical imaging: a systematic review of applications on ct and pet. *Artificial Intelligence Review*. ISSN 1573-7462.

Dubielzig, R. R., K. Ketring, G. J. McLellan, and D. M. Albert, Chapter 11 - the retina. *In Veterinary Ocular Pathology*. 2010. ISBN 978-0-7020-2797-0, 349 – 397.

Dutta, A. and A. Zisserman, The VIA annotation software for images, audio and video. *In Proceedings of the 27th ACM International Conference on Multimedia*, MM '19. ACM, New York, NY, USA, 2019. ISBN 978-1-4503-6889-6/19/10.

Eftekhari, N., H.-R. Pourreza, M. Masoudi, K. Ghiasi-Shirazi, and E. Saeedi (2019). Microaneurysm detection in fundus images using a two-step convolutional neural network. *BioMedical Engineering OnLine*, 18(1), 67. ISSN 1475-925X.

Erraguntla, M., B. Gopal, S. Ramachandran, and R. Mayer, Inference of missing icd 9 codes using text mining and nearest neighbor techniques. *In 45th Hawaii International Conference on System Sciences*. IEEE, 2012.

Eslami, S., P. Adorján, and C. Meinel, Sehmic: Semi-hierarchical multi-label icd code classification. *In CLEF*. 2020.

Falavigna, G. (2020). Prediction of general medical admission length of stay with natural language processing and deep learning: a pilot study. *Internal and Emergency Medicine*, 15(6), 917–918. ISSN 18280447.

Farkas, R. and G. Szarvas (2008). Automatic construction of rule-based ICD-9-cm coding systems. *BMC Bioinformatics*, 9, S10 – S10.

Ferrao, J. C., F. Janela, M. D. Oliveira, and H. M. Martins, Using structured ehr data and svm to support ICD-9-cm coding. *In IEEE International Conference on Healthcare Informatics*. 2013.

Ferrer, C. and J. Alió (2011). Evaluation of molecular diagnosis in fungal keratitis. ten years of experience. *Journal of Ophthalmic Inflammation and Infection*, 1, 15 – 22.

Figueiredo, I., S. Kumar, C. Oliveira, J. Ramos, and B. Engquist (2015). Automated lesion detectors in retinal fundus images. *Computers in Biology and Medicine*, 66, 47 – 65. ISSN 0010-4825.

Forthmann, P. and G. Pfleiderer (2019). Augmented display device for use in a medical imaging laboratory. US2018197337 Google Patents.

Fraz, M., P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen, and S. A. Barman (2012). Blood vessel segmentation methodologies in retinal images–a survey. *Computer methods and programs in biomedicine*, 108(1), 407–433.

Fu, H., J. Cheng, Y. Xu, C. Zhang, D. Wong, J. Liu, and X. Cao (2018). Disc-aware ensemble network for glaucoma screening from fundus image. *IEEE Transactions on Medical Imaging*, 37(11), 2493–2501. ISSN 02780062.

García-Santa, N. and K. Cetina, FLE at CLEF ehealth 2020: Text mining and semantic knowledge for automated clinical encoding. *In* L. Cappellato, C. Eickhoff, N. Ferro, and A. Névéol (eds.), *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Greece, Sep 22-25, 2020*, CEUR Workshop Proceedings. 2020.

Gargeya, R. and T. Leng (2017). Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, 124(7), 962–969. ISSN 0161-6420.

Gentimis, T., A. J. Alnaser, A. Durante, K. Cook, and R. Steele, Predicting hospital length of stay using neural networks on MIMIC III data. *In IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing*. 2017.

Georgiou, A., M. Prgomet, A. Markewycz, E. Adams, and J. Westbrook (2011). The impact of computerized provider order entry systems on medical-imaging services: A systematic review. *Journal of the American Medical Informatics Association*, 18(3), 335–340. ISSN 10675027.

Goeuriot, L., L. Kelly, H. Suominen, A. Névéol, A. Robert, E. Kanoulas, R. Spijker, J. Palotti, and G. Zuccon, Clef 2017 ehealth evaluation lab overview. 2017. ISBN 978-3-319-65812-4.

Goh, J. H. L., Z. W. Lim, X. L. Fang, A. Anees, S. Nusinovici, T. H. Rim, C.-Y. Cheng, and Y.-C. Tham (2020). Artificial intelligence for cataract detection and management. *Asia-Pacific Journal of Ophthalmology*.

Gonzalez, R. C. and R. E. Woods, *Digital image processing*. Prentice Hall, Upper Saddle River, N.J., 2008.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, *et al.*, Generative adversarial nets. *In* Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014.

Gopal, K. (2019). Strategies for ensuring quality health care in india: Experiences from the field. *Indian Journal of Community Medicine*, 44(1), 1 – 3. ISSN 09700218.

Gour, N. and P. Khanna (2020). Multi-class multi-label ophthalmological disease detection using transfer learning based convolutional neural network. *Biomedical Signal Processing and Control*. ISSN 17468094.

Grewal, P. S., F. Oloumi, U. Rubin, and M. T. Tennant (2018). Deep learning in ophthalmology: a review. *Canadian Journal of Ophthalmology*, 53(4), 309 – 313. ISSN 0008-4182.

Grzybowski, A., P. Brona, G. Lim, P. Ruamviboonsuk, G. Tan, M. Abràmoff, and D. Ting (2019). Artificial intelligence for diabetic retinopathy screening: a review. *Eye*, 34, 451–460.

Guibas, J., T. S. Virdi, and P. S. Li (2017). Synthetic medical images from dual generative adversarial networks. *ArXiv*, abs/1709.01872.

Guo, D., G. Duan, Y. Yu, Y. Li, F.-X. Wu, and M. Li (2019). A disease inference method based on symptom extraction and bidirectional long short term memory networks. *Methods*. ISSN 1046-2023.

Guo, S., C.-C. Huang, W. Zhao, A. C. Yang, C.-P. Lin, T. Nichols, and S.-J. Tsai (2018). Combining multi-modality data for searching biomarkers in schizophrenia. *PLOS ONE*, 13(2), 1–20.

Gusarev, M., R. Kuleev, A. Khan, A. Ramirez Rivera, and A. M. Khattak, Deep learning models for bone suppression in chest radiographs. *In 2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. 2017.

Hagiwara, Y., J. Koh, J. Tan, S. Bhandary, A. Laude, E. Ciaccio, L. Tong, and U. Acharya (2018). Computer-aided diagnosis of glaucoma using fundus images: A review. *Computer Methods and Programs in Biomedicine*, 165, 1–12. ISSN 01692607.

Harutyunyan, H., H. Khachatrian, D. Kale, G. Ver Steeg, and A. Galstyan (2019*a*). Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1). ISSN 20524463.

Harutyunyan, H., H. Khachatrian, D. Kale, G. Ver Steeg, and A. Galstyan (2019*b*). Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1). ISSN 20524463.

He, J., C. Li, J. Ye, Y. Qiao, and L. Gu (2021*a*). Multi-label ocular disease classification with a dense correlation deep neural network. *Biomedical Signal Processing and Control*, 63. ISSN 17468094.

He, J., C. Li, J. Ye, Y. Qiao, and L. Gu (2021*b*). Self-speculation of clinical features based on knowledge distillation for accurate ocular disease classification. *Biomedical Signal Processing and Control*, 67, 102491. ISSN 1746-8094.

He, K., X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition. *In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016.

He, K., X. Zhang, *et al.* (2015). Deep residual learning for image recognition.

He, P., X. Liu, J. Gao, and W. Chen (2020). Deberta: Decoding-enhanced bert with disentangled attention. *ArXiv*, abs/2006.03654.

Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

Howard, A., M. Zhu, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv*, abs/1704.04861.

Hsu, J.-L., T.-J. Hsu, C.-H. Hsieh, and A. Singaravelan (2020). Applying convolutional neural networks to predict the icd-9 codes of medical records. *Sensors*, 20(24). ISSN 1424-8220.

Huang, G., Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, Densely connected convolutional networks. *In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

Huang, J., C. Osorio, and L. W. Sy (2019). An empirical evaluation of deep learning for ICD-9 code assignment using mimic-iii clinical notes. *Computer Methods and Programs in Biomedicine*, 177, 141–153. ISSN 0169-2607.

Huang, K., J. Altosaar, and R. Ranganath (2019). ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv e-prints*, arXiv:1904.05342.

Huang, S.-C., A. Pareek, S. Seyyedi, I. Banerjee, and M. Lungren (2020). Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digital Medicine*, 3(1). ISSN 23986352.

Hung, N., A. K.-Y. Shih, C. Lin, M.-T. Kuo, Y.-S. Hwang, W.-C. Wu, C.-F. Kuo, E. Y.-C. Kang, and C.-H. Hsiao (2021). Using slit-lamp images for deep learning-based identification of bacterial and fungal keratitis: Model development and validation with different convolutional neural networks. *Diagnostics*, 11(7). ISSN 2075-4418. URL https://www.mdpi.com/2075-4418/11/7/1246.

Iandola, F. N., M. W. Moskewicz, K. Ashraf, S. Han, W. Dally, and K. Keutzer (2017). Squeezenet: Alexnet-level accuracy with 50x fewer parameters and ¡1mb model size. *ArXiv*, abs/1602.07360.

Imran, A., J. Li, Y. Pei, F. Akhtar, J.-J. Yang, and Y. Dang (2020). Automated identification of cataract severity using retinal fundus images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*, 8(6), 691–698. ISSN 21681163.

India, G. O. (2020). Report of the artificial intelligence task force. Online. URL "https://dpiit.gov.in/sites/default/files/Report_of_Task_Force_on_ArtificialIntelligence_20March2018_2.pdf".

Institute of Medicine (2015). Improving diagnosis in health care. Online. URL "https://www.nap.edu/resource/21794/DiagnosticError_ReportBrief.pdf".

Islam, M. T., S. A. Imran, A. Arefeen, M. Hasan, and C. Shahnaz, Source and camera independent ophthalmic disease recognition from fundus image using neural network. *In 2019 IEEE International Conference on Signal Processing, Information, Communication Systems (SPICSCON)*. 2019.

Isola, P., J.-Y. Zhu, T. Zhou, and A. A. Efros (2018). Image-to-image translation with conditional adversarial networks.

Jadhav, A., P. Patil, and S. Biradar (2020*a*). Analysis on diagnosing diabetic retinopathy by segmenting blood vessels, optic disc and retinal abnormalities. *Journal of Medical Engineering and Technology*, 44(6), 299–316. ISSN 03091902.

Jadhav, A., P. Patil, and S. Biradar (2020*b*). Analysis on diagnosing diabetic retinopathy by segmenting blood vessels, optic disc and retinal abnormalities. *Journal of Medical Engineering and Technology*, 44(6), 299–316. ISSN 03091902.

Jaeger, S., A. Karargyris, S. Candemir, L. R. Folio, J. Siegelman, F. M. Callaghan, Z. Xue, K. Palaniappan, R. K. Singh, S. Antani, G. R. Thoma, Y. Wang, P.-X. Lu, and C. J. McDonald (2014). Automatic tuberculosis screening using chest radiographs. *IEEE Transactions on Medical Imaging*, 33, 233–245.

Jain, R., M. Gupta, S. Taneja, and D. Hemanth (2021). Deep learning based detection and analysis of covid-19 on chest x-ray images. *Applied Intelligence*, 51(3), 1690–1700.

Jain, S., R. Mohammadi, and B. C. Wallace (2019). An Analysis of Attention over Clinical Notes for Predictive Tasks. *arXiv e-prints*, arXiv:1904.03244.

Jenks, G. F. (1967). The data model concept in statistical mapping. *International Yearbook of Cartography*, 7, 186–190.

Jensen, P. B., L. J. Jensen, and S. Brunak (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6), 395.

Ji, S., E. Cambria, and P. Marttinen, Dilated convolutional attention network for medical code assignment from clinical text. *In Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Online, 2020.

Jiang, J., S. Hewner, and V. Chandola, Tree-based regularization for interpretable readmission prediction. *In AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*. 2019.

Jin, M., M. Taha Bahadori, A. Colak, P. Bhatia, B. Celikkaya, R. Bhakta, S. Senthivel, M. Khalilia, D. Navarro, B. Zhang, T. Doman, A. Ravi, M. Liger, and T. Kass-hout (2018). Improving Hospital Mortality Prediction with Medical Named Entities and Multimodal Learning. *arXiv e-prints*, arXiv:1811.12276.

Jing, B., P. Xie, and E. Xing, On the automatic generation of medical imaging reports. *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 2018.

Jobson, D. J., Z.-u. Rahman, and G. A. Woodell (1997). A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image processing*, 6(7), 965–976.

Johnson, A. E., T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3, 160035.

Joshi, G., J. Sivaswamy, and S. Krishnadas (2011). Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment. *IEEE Transactions on Medical Imaging*, 30(6), 1192–1205. ISSN 02780062.

Juneja, M., S. Singh, N. Agarwal, S. Bali, S. Gupta, N. Thakur, and P. Jindal (2020). Automated detection of glaucoma using deep learning convolution network (g-net). *Multimedia Tools and Applications*, 79(21-22), 15531–15553. ISSN 13807501.

Kalkancı, A. and S. Ozdek (2011). Ocular fungal infections. *Current Eye Research*, 36, 179 – 189.

Kandel, I. and M. Castelli (2020). Transfer learning with convolutional neural networks for diabetic retinopathy image classification. a review. *Applied Sciences (Switzerland)*, 10(6). ISSN 20763417.

Kar, S. and S. Maity (2018). Automatic detection of retinal lesions for screening of diabetic retinopathy. *IEEE Transactions on Biomedical Engineering*, 65(3), 608–618. ISSN 00189294.

Karimian, G., E. Petelos, and S. Evers (2022). The ethical issues of the application of artificial intelligence in healthcare: a systematic scoping review. *AI and Ethics*, 1–13.

Karras, T., M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, Training generative adversarial networks with limited data. *In* H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020*a*.

Karras, T., S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, Analyzing and improving the image quality of stylegan. *In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020*b*.

Kavuluru, R., A. Rios, and Y. Lu (2015). An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial Intelligence in Medicine*, 65.

Kelly, L., H. Suominen, L. Goeuriot, M. Neves, E. Kanoulas, D. Li, L. Azzopardi, R. Spijker, G. Zuccon, H. Scells, and J. Palotti, *Overview of the CLEF eHealth Evaluation Lab 2019*. 2019. ISBN 978-3-030-28576-0, 322–339.

Khan, S., X. Liu, S. Nath, E. Korot, L. Faes, S. Wagner, P. Keane, N. Sebire, M. Burton, and A. Denniston (2021). A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *The Lancet Digital Health*, 3(1), e51–e66. Cited By 15.

Khojasteh, P., B. Aliahmad, and D. K. Kumar (2018). Fundus images analysis using deep features for detection of exudates, hemorrhages and microaneurysms. *BMC Ophthalmology*, 18.

Kim, Y., Convolutional neural networks for sentence classification. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 2014. URL `https://www.aclweb.org/anthology/D14-1181`.

Kingma, D. P. and J. Ba (2015). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Knaus, W., E. Draper, D. Wagner, and J. Zimmerman (1985). Apache ii: A severity of disease classification system. *Critical Care Medicine*, 13(10), 818–829.

Kou, C., W. Li, W. Liang, Z. Yu, and J. Hao (2019). Microaneurysms segmentation with a u-net based on recurrent residual convolutional neural network. *Journal of Medical Imaging*, 6(2).

Krizhevsky, A. (2012). Convolutional deep belief networks on cifar-10. URL `https://www.cs.toronto.edu/~kriz/cifar.html`.

Krizhevsky, A., I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks. *In* F. Pereira, C. J. C. Burges, L. Bottou, and

K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2012.

Kruse, C. S., A. Stein, H. Thomas, and H. Kaur (2018). The use of electronic health records to support population health: A systematic review of the literature. *Journal of Medical Systems*, 42(11), 214. ISSN 1573-689X.

Kuo, M., B. W.-Y. Hsu, Y.-K. Yin, P.-C. Fang, H.-Y. Lai, A. Chen, M.-S. Yu, and V. S. Tseng (2020). A deep learning approach in diagnosing fungal keratitis based on corneal photographs. *Scientific Reports*, 10.

Kuperman, G., A. Bobb, T. Payne, A. Avery, T. Gandhi, G. Burns, D. Classen, and D. Bates (2007). Medication-related clinical decision support in computerized provider order entry systems: A review. *Journal of the American Medical Informatics Association*, 14(1), 29–40. ISSN 10675027.

Kwok, R. C. L., M. M. Dinh, D. Dinh, and M. Chu (2009). Improving adherence to asthma clinical guidelines and discharge documentation from emergency departments: Implementation of a dynamic and integrated electronic decision support system †. *Emergency Medicine Australasia*, 21.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, Albert: A lite bert for self-supervised learning of language representations. *In International Conference on Learning Representations*. 2020.

Land, E. H. (1977). The retinex theory of color vision. *Sci. Amer*, 108–128.

Lang, D. (2007). Consultant report-natural language processing in the health care industry. *Cincinnati Children's Hospital Medical Center, Winter*, 6.

Le Gall, J.-R., P. Loirat, A. Alperovitch, P. Glaser, C. Granthil, D. Mathieu, P. Mercier, R. Thomas, and D. Villers (1984). A simplified acute physiology score for icu patients. *Critical Care Medicine*, 12(11), 975–977.

Leck, A. and M. Burton (2015). Distinguishing fungal and bacterial keratitis on clinical signs. *Community Eye Health*, 28, 6 – 7.

Lee, E., J. Zheng, E. Colak, and Others (2021). Deep covid detect: an international experience on covid-19 lung detection and prognosis using chest ct. *npj Digital Medicine*, 4(1), 11. ISSN 23986352.

Li, C., L. Chen, J. Feng, D. Wu, Z. Wang, J. Liu, and W. Xu (2019*a*). Prediction of length of stay on the intensive care unit based on least absolute shrinkage and selection operator. *IEEE Access*, 7, 110710–110721. ISSN 2169-3536.

Li, C., J. Ye, J. He, S. Wang, Y. Qiao, and L. Gu, Dense correlation network for automated multi-label ocular disease detection with paired color fundus photographs. *In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. 2020.

Li, F. and H. Yu, Icd coding from clinical text using multi-filter residual convolutional neural network. *In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*. 2020.

Li, M., Z. Fei, M. Zeng, F. Wu, Y. Li, Y. Pan, and J. Wang (2019*b*). Automated ICD-9 coding via a deep learning approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(4), 1193–1202.

Li, N., T. Li, C. Hu, K. Wang, and H. Kang, A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection. *In* F. Wolf and W. Gao (eds.), *Benchmarking, Measuring, and Optimizing*. Springer International Publishing, Cham, 2021*a*.

Li, T., W. Bo, C. Hu, H. Kang, H. Liu, K. Wang, and H. Fu (2021*b*). Applications of deep learning in fundus images: A review. *Medical Image Analysis*, 69, 101971. ISSN 1361-8415.

Li, T., Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang (2019). Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, 501, 511–522. ISSN 0020-0255.

Li, W., Y. Yang, K. Zhang, E. Long, L. He, L. Zhang, Y. Zhu, C. Chen, Z. Liu, X. Wu, D. Yun, J. Lv, Y. Liu, X. Liu, and H. Lin (2020). Dense anatomical annotation of slit-lamp images improves the performance of deep learning for the diagnosis of ophthalmic disorders. *Nature Biomedical Engineering*, 1–11.

Lin, J., Q. Cai, and M. Lin (2021). Multi-label classification of fundus images with graph convolutional network and self-supervised learning. *IEEE Signal Processing Letters*, 28, 454–458.

Lin, Y.-W., Y. Zhou, F. Faghri, M. Shaw, and R. H. Campbell (2019). Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. *PLoS ONE*, 14.

Liskowski, P. and K. Krawiec (2016). Segmenting retinal blood vessels with deep neural networks. *IEEE Transactions on Medical Imaging*, 35(11), 2369–2380.

Liu, G., T.-M. H. Hsu, M. McDermott, W. Boag, W.-H. Weng, P. Szolovits, and M. Ghassemi, Clinically accurate chest x-ray report generation. *In* F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, and J. Wiens (eds.), *Proceedings of the 4th Machine Learning for Healthcare Conference*, Proceedings of Machine Learning Research. PMLR, 2019*a*.

Liu, J., Z. Zhang, and N. Razavian, Deep ehr: Chronic disease prediction using medical notes. *In* F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, and J. Wiens (eds.), *Proceedings of the 3rd Machine Learning for Healthcare Conference*, Proceedings of Machine Learning Research. PMLR, 2018.

Liu, S. and W. Deng, Very deep convolutional neural network based image classification using small training sample size. *In 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. 2015.

Liu, X., J. Jiang, K. Zhang, E. Long, J. Cui, M. Zhu, Y. An, J. Zhang, Z. Liu, Z. Lin, X. Li, J. Chen, Q. Cao, J. Li, X. Wu, D. Wang, and H. Lin (2017). Localization and diagnosis framework for pediatric cataracts based on slit-lamp images using deep features of a convolutional neural network. *PLoS ONE*, 12(3). ISSN 19326203.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov (2019*b*). Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Loo, J., M. Kriegel, M. Tuohy, K. Kim, V. Prajna, M. Woodward, and S. Farsiu (2021*a*). Open-source automatic segmentation of ocular structures and biomarkers of microbial keratitis on slit-lamp photography images using deep learning. *IEEE Journal of Biomedical and Health Informatics*, 25(1), 88–99. ISSN 21682194.

Loo, J., M. Woodward, V. Prajna, M. Kriegel, M. Pawar, M. Khan, L. Niziol, and S. Farsiu (2021*b*). Open-source automatic biomarker measurement on slit-lamp photography to estimate visual acuity in microbial keratitis. *Translational Vision Science and Technology*, 10(12). ISSN 21642591.

Lopes, B. T., A. Eliasy, and R. Ambrósio (2019). Artificial intelligence in corneal diagnosis: Where are we? *Current Ophthalmology Reports*, 7, 204 – 211.

Lovelace, J. R., N. C. Hurley, A. D. Haimovich, and B. J. Mortazavi (2019). Explainable Prediction of Adverse Outcomes Using Clinical Notes. *arXiv e-prints*, arXiv:1910.14095.

Lu, W., Y. Tong, Y. Yu, Y. Xing, C. Chen, and Y. Shen (2018). Applications of artificial intelligence in ophthalmology: General overview. *Journal of Ophthalmology*, 2018, 1–15.

Maharana, P., N. Sharma, R. Nagpal, V. Jhanji, S. Das, and R. Vajpayee (2016). Recent advances in diagnosis and management of mycotic keratitis. *Indian Journal of Ophthalmology*, 64, 346 – 357.

Mainor, A., N. Morden, J. Smith, S. Tomlin, and J. Skinner (2019). Icd-10 coding will challenge researchers: Caution and collaboration may reduce measurement error and improve comparability over time. *Medical Care*, 57(7), E42–E46. ISSN 00257079.

Maji, D., A. Santara, P. Mitra, and D. Sheet (2016). Ensemble of deep convolutional neural networks for learning to detect retinal vessels in fundus images. *ArXiv*, abs/1603.04833.

Maqsood, S., R. Damaševičius, and R. Maskeliūnas (2021). Hemorrhage detection based on 3d cnn deep learning framework and feature fusion for evaluating retinal abnormality in diabetic patients. *Sensors*, 21(11). ISSN 14248220.

Marcilly, R., N. Leroy, M. Luyckx, S. Pelayo, C. Riccioli, and M.-C. Beuscart-Zéphir (2011). Medication related computerized decision support system (cdss): make it a clinicians' partner! *Studies in health technology and informatics*, 166, 84–94.

Marshall, J., A. Chahin, and B. Rush, *Review of Clinical Databases*. Springer International Publishing, Cham, 2016, 9–16.

McDonald, R., G. Brokos, and I. Androutsopoulos, Deep relevance ranking using enhanced document-query interactions. *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

McHugh, M. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 22, 276 – 282.

McMullin, S., T. Lonergan, C. Rynearson, T. Doerr, P. Veregge, and E. Scanlan (2004). Impact of an evidence-based computerized decision support system on primary care prescription costs. *Annals of Family Medicine*, 2(5), 494–498. ISSN 15441709.

McNutt, T., S. Benedict, D. Low, K. Moore, I. Shpitser, W. Jiang, P. Lakshminarayanan, Z. Cheng, P. Han, X. Hui, M. Nakatsugawa, J. Lee, J. Moore, S. Robertson, V. Shah, R. Taylor, H. Quon, J. Wong, and T. DeWeese (2018). Using big data analytics to advance precision radiation oncology. *International Journal of Radiation Oncology Biology Physics*, 101(2), 285–291. ISSN 03603016.

Medori, J. and C. Fairon, Machine learning and features selection for semi-automatic ICD-9-cm encoding. *In Proceedings of the NAACL HLT Second Louhi Workshop on Text and Data Mining of Health Documents*. 2010*a*.

Medori, J. and C. Fairon, Machine learning and features selection for semi-automatic ICD-9-cm encoding. *In Proceedings of the NAACL HLT Second Louhi Workshop on Text and Data Mining of Health Documents*. Association for Computational Linguistics, 2010*b*.

Messina, P., P. Pino, D. Parra, Á. Soto, C. Besa, S. Uribe, M. and'ia, C. Tejos, C. Prieto, and D. Capurro (2020). A survey on deep learning and explainability for automatic image-based medical report generation. *ArXiv*, abs/2010.10563.

Metsing, I., W. Jacobs, and R. Hansraj (2018). A review of vision screening methods for children. *African Vision and Eye Health*, 77(1). ISSN 24133183.

Mikolov, T., K. Chen, G. Corrado, and J. Dean, Efficient estimation of word representations in vector space. *In 1st International Conference on Learning Representations, ICLR*. 2013.

Miranda-Escalada, A., A. Gonzalez-Agirre, *et al.*, Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of clef ehealth 2020. *In Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*. 2020.

Mohaimenul, M., H.-C. Yang, *et al.* (2020). Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis. *Computer Methods and Programs in Biomedicine*, 191, 105320.

Monshi, M., J. Poon, and V. Chung (2020). Deep learning in generating radiology reports: A survey. *Artificial Intelligence in Medicine*, 106. ISSN 09333657.

Mookiah, M., S. Hogg, T. MacGillivray, V. Prathiba, R. Pradeepa, V. Mohan, R. Anjana, A. Doney, C. Palmer, and E. Trucco (2021*a*). A review of machine learning methods for retinal blood vessel segmentation and artery/vein classification. *Medical Image Analysis*, 68. ISSN 13618415.

Mookiah, M. R. K., S. Hogg, T. J. MacGillivray, V. Prathiba, R. Pradeepa, V. Mohan, R. M. Anjana, A. S. Doney, C. N. Palmer, and E. Trucco (2021*b*). A review of machine learning methods for retinal blood vessel segmentation and artery/vein classification. *Medical Image Analysis*, 68, 101905. ISSN 1361-8415. URL `https://www.sciencedirect.com/science/article/pii/S1361841520302693`.

Moons, E. and M.-F. Moens, Convolutional attention models with post-processing heuristics at clef ehealth 2020. *In CLEF*. 2020.

Moreno, R., P. Metnitz, E. Almeida, B. Jordan, P. Bauer, R. Campos, G. Iapichino, D. Edbrooke, M. Capuzzo, and J.-R. Le Gall (2005). Saps 3 - from evaluation of the patient to evaluation of the intensive care unit. part 2: Development of a prognostic model for hospital mortality at icu admission. *Intensive Care Medicine*, 31(10), 1345–1355.

Mullenbach, J., S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein, Explainable prediction of medical codes from clinical text. *In Proceedings of the 2018 Conference of NAACL: Human Language Technologies, Volume 1*. 2018.

Mumtaz, R., M. Hussain, S. Sarwar, K. Khan, S. Mumtaz, and M. Mumtaz (2018). Automatic detection of retinal hemorrhages by exploiting image processing techniques for screening retinal diseases in diabetic patients. *International Journal of Diabetes in Developing Countries*, 38(1), 80–87. ISSN 09733930.

Murugan, R. (2019). An automatic detection of hemorrhages in retinal fundus images by motion pattern generation. *Biomedical and Pharmacology Journal*, 12(3), 1433–1440. ISSN 09746242.

Mykowiecka, A., M. Marciniak, and A. Kupść (2009). Rule-based information extraction from patients' clinical data. *Journal of Biomedical Informatics*, 42(5), 923 – 936. ISSN 1532-0464. Biomedical Natural Language Processing.

Njie, G., K. Proia, A. Thota, R. Finnie, D. Hopkins, S. Banks, D. Callahan, N. Pronk, K. Rask, D. Lackland, T. Kottke, and C. P. S. T. Force (2015). Clinical decision support systems and prevention: A community guide cardiovascular disease systematic review. *American Journal of Preventive Medicine*, 49(5), 784–795. ISSN 07493797.

Nunzio, G. M. D., As simple as possible: Using the r tidyverse for multilingual information extraction. ims unipd ad clef ehealth 2020 task 1. *In CLEF*. 2020.

ODIR (2019). Ocular Disease Intelligent Recognition, ODIR-5K dataset. URL `https://odir2019.grand-challenge.org/`.

Orlando, J., E. Prokofyeva, M. del Fresno, and M. Blaschko (2018). An ensemble deep learning based approach for red lesion detection in fundus images. *Computer Methods and Programs in Biomedicine*, 153, 115–127. ISSN 01692607.

Ortega, J., P. López-Úbeda, M. C. Díaz-Galiano, M. T. M. Valdivia, and L. López, Sinai at clef ehealth 2020: Testing different pre-trained word embeddings for clinical coding in spanish. *In CLEF*. 2020.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1), 62–66.

Pakhomov, S., J. Buntrock, and C. Chute (2006*a*). Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association : JAMIA*, 13, 516–25.

Pakhomov, S. V., J. D. Buntrock, and C. G. Chute (2006*b*). Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association*, 13(5), 516–525.

Pandey, B., D. Kumar Pandey, B. Pratap Mishra, and W. Rhmann (2021). A comprehensive survey of deep learning in the field of medical imaging and medical natural language processing: Challenges and research directions. *Journal of King Saud University - Computer and Information Sciences*. ISSN 1319-1578.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation. *In Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002.

Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, *et al.*, Pytorch: An imperative style, high-performance deep learning library. *In Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019*a*, 8024–8035.

Paszke, A., S. Gross, F. Massa, *et al.*, Pytorch: An imperative style, high-performance deep learning library. *In Advances in Neural Information Processing Systems 32*. 2019*b*.

Pathan, S., P. Kumar, R. Pai, and S. Bhandary (2021). Automated segmentation and classifcation of retinal features for glaucoma diagnosis. *Biomedical Signal Processing and Control*, 63. ISSN 17468094.

Payrovnaziri, S. N., L. A. Barrett, D. Bis, J. Bian, and Z. He (2019*a*). Enhancing prediction models for one-year mortality in patients with acute myocardial infarction and post myocardial infarction syndrome. *Studies in health technology and informatics*, 264, 273–277.

Payrovnaziri, S. N., L. A. Barrett, D. Bis, J. Bian, and Z. He (2019*b*). Enhancing prediction models for one-year mortality in patients with acute myocardial infarction and post myocardial infarction syndrome. *Studies in health technology and informatics*, 264, 273–277.

Peng, Y., S. Yan, and Z. Lu, Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. *In Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*. 2019.

Perotte, A., R. Pivovarov, K. Natarajan, N. Weiskopf, F. Wood, and N. Elhadad (2013). Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, 21(2), 231–237.

Perotte, A. J., F. Wood, N. Elhadad, and N. Bartlett, Hierarchically supervised latent dirichlet allocation. *In Advances in Neural Information Processing Systems*. 2011.

Perreault, L. E. and J. B. Metzger (1999). A pragmatic framework for understanding clinical decision support. *Journal of Healthcare Information Management*, 13, 5–22.

Petro, A. B., C. Sbert, and J.-M. Morel (2014). Multiscale Retinex. *Image Processing On Line*, 71–88.

Pham, Q., S. Ahn, S. Song, and J. Shin (2020). Automatic drusen segmentation for age-related macular degeneration in fundus images using deep learning. *Electronics (Switzerland)*, 9(10), 1–11. ISSN 20799292.

Polignano, M., V. Suriano, P. Lops, M. Degemmis, and G. Semeraro, A study of machine learning models for clinical coding of medical reports at codiesp 2020. *In CLEF*. 2020.

Prajna, V., L. Prajna, and S. Muthiah (2017). Fungal keratitis: The aravind experience. *Indian Journal of Ophthalmology*, 65, 912 – 919.

Pratap, T. and P. Kokil (2019). Computer-aided diagnosis of cataract using deep transfer learning. *Biomedical Signal Processing and Control*, 53, 101533. ISSN 1746-8094.

Procop, G., C. Keating, P. Stagno, K. Kottke-Marchant, M. Partin, R. Tuttle, and R. Wyllie (2015). Reducing duplicate testing: A comparison of two clinical decision support tools. *American journal of clinical pathology*, 143, 623–6.

Pruszydlo, M., S. Walk-Fritz, T. Hoppe-Tichy, J. Kaltschmidt, and W. Haefeli (2012). Development and evaluation of a computerised clinical decision support system for switching drugs at the interface between primary and tertiary care. *BMC Medical Informatics and Decision Making*, 12(1). ISSN 14726947.

Purushotham, S., C. Meng, Z. Che, and Y. Liu (2018). Benchmarking deep learning models on large healthcare datasets. *Journal of Biomedical Informatics*, 83, 112 – 134. ISSN 1532-0464.

Qin, X., Z. Zhang, C. Huang, M. Dehghan, O. R. Zaïane, and M. Jägersand (2020). U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognit.*, 106, 107404.

Queipo-Álvarez, P. and I. González-Carrasco, Classifying clinical case studies with icd-10 atcodiesp clef ehealth 2020 task 1-diagnostics. *In CLEF*. 2020.

Radiological Society of North America (2018). Rsna pneumonia detection challenge. `https://www.kaggle.com/c/rsna-pneumonia-detection-challenge`. Accessed: 2020-08-11.

Rajkomar, A., E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang, Y. Zhang, G. Flores, G. E. Duggan, J. Irvine, Q. V. Le, K. Litsch, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S. L. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N. H. Shah, A. J. Butte, M. D. Howell, C. Cui, G. Corrado, and J. Dean (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1.

Ramaswamy, A., N. R. Gowda, H. Vikas, M. Prabhu, D. Sharma, P. R. Gowda, D. Mohan, and A. Kumar (2022). It's the data, stupid: Inflection point for artificial intelligence in indian healthcare. *Artificial Intelligence in Medicine*, 128, 102300. ISSN 0933-3657.

Rampat, R., R. Deshmukh, X. Chen, D. Ting, D. Said, H. Dua, and D. Ting (2021). Artificial intelligence in cornea, refractive surgery, and cataract: Basic principles, clinical applications, and future directions. *Asia-Pacific journal of ophthalmology (Philadelphia, Pa.)*, 10(3), 268–281. ISSN 21620989.

Rangarajan, A. and H. Ramachandran (2021). A preliminary analysis of ai based smartphone application for diagnosis of covid-19 using chest x-ray images. *Expert Systems with Applications*, 183. ISSN 09574174.

Rashidian, S., J. Hajagos, R. Moffitt, F. Wang, *et al.* (2018). Disease phenotyping using deep learning: A diabetes case study. *arXiv e-prints*, arXiv:1811.11818.

Reddy, K. S., V. Patel, P. Jha, V. K. Paul, A. Kumar, and L. Dandona (2011). Towards achievement of universal health care in india by 2020: a call to action. *The Lancet*, 377, 760–768.

Řehůřek, R. and P. Sojka, Software Framework for Topic Modelling with Large Corpora. *In Proceedings of the LREC Workshop on New Challenges for NLP Frameworks*. ELRA, 2010.

Ren, S., K. He, R. Girshick, and J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks. *In* C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015.

Ren, X., X. Zheng, X. Dong, and X. Cui (2020). Deep feature extraction via adaptive collaborative learning for drusen segmentation from fundus images. *Signal, Image and Video Processing*. ISSN 18631703.

Reza, A. (2004). Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *VLSI Signal Processing*, 38, 35–44.

Rios, A. and R. Kavuluru (2019). Neural transfer learning for assigning diagnosis codes to EMRs. *Artificial Intelligence In Medicine*, 96(December 2018), 116–122.

Rishivardhan, K., S. Kayalvizhi, D. Thenmozhi, S. Krishan, and C. Aravindan, Transformers in semantic indexing of clinical codes. *In CLEF*. 2020.

Ronneberger, O., P. Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation. *In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer, 2015*a*. ISBN 978-3-319-24574-4.

Ronneberger, O., P.Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation. *In Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS. Springer, 2015*b*.

Roshini, T., R. Ravi, A. Reema Mathew, A. Kadan, and P. Subbian (2020). Automatic diagnosis of diabetic retinopathy with the aid of adaptive average filtering with optimized deep convolutional neural network. *International Journal of Imaging Systems and Technology*, 30(4), 1173–1193. ISSN 08999457.

Rouzbahman, M., A. Jovicic, and M. Chignell (2017). Can cluster-boosted regression improve prediction of death and length of stay in the icu? *IEEE Journal of Biomedical and Health Informatics*, 21(3), 851–858. ISSN 2168-2208.

Ruch, P., J. Gobeill, I. Tbahriti, and A. Geissbühler, From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. *In AMIA Annual Symposium Proceedings*. American Medical Informatics Association, 2008.

Rudnicka, A. R., C. G. Owen, R. A. Welikala, *et al.* (2020). Retinal vasculometry associations with glaucoma: Findings from the european prospective investigation of cancer–norfolk eye study. *American Journal of Ophthalmology*, 220, 140–151. ISSN 0002-9394.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115, 211–252.

Sadeghi, R., T. Banerjee, and W. Romine (2018). Smart Health Early hospital mortality prediction using vital signals. *Smart Health*, 9-10, 265–274. ISSN 2352-6483.

Sait, U., G. L. K.V., S. Shivakumar, T. Kumar, R. Bhaumik, S. Prajapati, K. Bhalla, and A. Chakrapani (2021). A deep-learning based multimodal system for covid-19 diagnosis using breathing sounds and chest x-ray images. *Applied Soft Computing*, 109, 107522. ISSN 15684946.

Samonte, M. J. C., B. D. Gerardo, A. C. Fajardo, and R. P. Medina, ICD-9 tagging of clinical notes using topical word embedding. *In Proceedings of the 2018 International Conference on Internet and e-Business*. ACM, 2018.

Samuel, P. and T. Veeramalai (2020). Review on retinal blood vessel segmentation - an algorithmic perspective. *International Journal of Biomedical Engineering and Technology*, 34(1), 75–105. ISSN 17526418.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Saranya, P. and S. Prabakaran (2020). Automatic detection of non-proliferative diabetic retinopathy in retinal fundus images using convolution neural network. *Journal of Ambient Intelligence and Humanized Computing*. ISSN 18685137.

Saranya, P., S. Prabakaran, R. Kumar, and E. Das (2021). Blood vessel segmentation in retinal fundus images for proliferative diabetic retinopathy screening using deep learning. *Visual Computer*. ISSN 01782789.

Schäfer, H. and C. Friedrich, Multilingual icd-10 code assignment with transformer architectures using mimic-iii discharge summaries. *In CLEF*. 2020.

Scheib, S. (2009). Dosimetric end-to-end verification devices, systems, and methods. US 20150085993 Google Patents.

Schein, O. D. (2016). Evidence-Based Treatment of Fungal Keratitis. *JAMA Ophthalmology*, 134(12), 1372–1373. ISSN 2168-6165.

Scherpf, M., F. Gräßer, H. Malberg, and S. Zaunseder (2019). Predicting sepsis with a recurrent neural network using the MIMIC III database. *Computers in Biology and Medicine*, 113(August), 103395. ISSN 0010-4825.

Scheurwegs, E., B. Cule, K. Luyckx, L. Luyten, and W. Daelemans (2017). Selecting relevant features from the electronic health record for clinical code prediction. *Journal of Biomedical Informatics*, 74, 92 – 103. ISSN 1532-0464.

Schivre, G. (2021). Multiscale retinex. Online. URL "https://www.mathworks.com/matlabcentral/fileexchange/71386-multiscale-retinex".

Schmidt-Erfurth, U., A. Sadeghipour, B. S. Gerendas, S. M. Waldstein, and H. Bogunović (2018). Artificial intelligence in retina. *Progress in Retinal and Eye Research*, 67, 1 – 29. ISSN 1350-9462.

Selvaraju, R., M. Cogswell, *et al.*, Grad-cam: Visual explanations from deep networks via gradient-based localization. *In IEEE international conference on computer vision*. 2017.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization. *In IEEE International Conference on Computer Vision (ICCV)*. 2017.

Selvaraju, R. R., A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra (2019). Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128, 336–359.

Sengupta, S., A. Singh, H. Leopold, T. Gulati, and V. Lakshminarayanan (2020). Ophthalmic diagnosis using deep learning with fundus images – a critical review. *Artificial Intelligence in Medicine*, 102. ISSN 09333657.

Sengür, A., Y. Guo, Ü. Budak, and L. J. Vespa (2017). A retinal vessel detection approach using convolution neural network. *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 1–4.

Shakarami, A., M. Menhaj, and H. Tarrah (2021). Diagnosing covid-19 disease using an efficient cad system. *Optik*, 241, 167199. ISSN 00304026.

Sheikhalishahi, S., V. Balaraman, and V. Osmani (2020). Benchmarking machine learning models on multi-centre eicu critical care dataset. *PLoS ONE*, 15(7). ISSN 19326203.

Shickel, B., P. J. Tighe, A. Bihorac, and P. Rashidi (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5), 1589–1604. ISSN 2168-2208.

Si, Y. and K. Roberts (2019). Deep patient representation of clinical notes via multi-task learning for mortality prediction. *AMIA Summits on Translational Science Proceedings*, 2019, 779.

Singh, H., A. Meyer, and E. Thomas (2014). The frequency of diagnostic errors in outpatient care: Estimations from three large observational studies involving us adult populations. *BMJ Quality and Safety*, 23(9), 727–731. ISSN 20445415.

Singh, N., D. Bansal, and D. Nagpal (2020). Deep learning based retinal vessel segmentation: A review. *Advances in Mathematics: Scientific Journal*, 9(6), 3827–3837. ISSN 18578365.

Son, J., S. Park, and K.-H. Jung (2018). Towards accurate segmentation of retinal vessels and the optic disc in fundoscopic images with generative adversarial networks. *Journal of Digital Imaging*, 32, 499–512.

Son, J., S. J. Park, and K.-H. Jung (2017). Retinal Vessel Segmentation in Fundoscopic Images with Generative Adversarial Networks. *arXiv e-prints*, arXiv:1706.09318.

Sreng, S., N. Maneerat, K. Hamamoto, and K. Win (2020*a*). Deep learning for optic disc segmentation and glaucoma diagnosis on retinal images. *Applied Sciences (Switzerland)*, 10(14). ISSN 20763417.

Sreng, S., N. Maneerat, K. Hamamoto, and K. Y. Win (2020*b*). Deep learning for optic disc segmentation and glaucoma diagnosis on retinal images. *Applied Sciences*, 10(14). ISSN 2076-3417.

Staal, J., M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken (2004). Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23(4), 501–509.

Stella Mary, M. C. V., E. B. Rajsingh, and G. R. Naik (2016). Retinal fundus image analysis for diagnosis of glaucoma: A comprehensive survey. *IEEE Access*, 4, 4327–4354.

Su, T.-Y., P.-J. Ting, S. Chang, and D.-Y. Chen (2020). Superficial punctate keratitis grading for dry eye screening using deep convolutional neural networks. *IEEE Sensors Journal*, 20, 1672–1678.

Sudha, V. and T. Ganeshbabu (2021). A convolutional neural network classifier vgg-19 architecture for lesion detection and grading in diabetic retinopathy based on deep learning. *Computers, Materials and Continua*, 66(1), 827–842. ISSN 15462218.

Sun, Z., H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, MobileBERT: a compact task-agnostic BERT for resource-limited devices. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2020.

Suresh, H., J. J. Gong, and J. V. Guttag, Learning tasks for multitask learning: Heterogeneous patient populations in the icu. *In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018.

Szegedy, C., Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, Going deeper with convolutions. *In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

Tagawa, Y., N. Nakano, R. Ozaki, T. Taniguchi, and T. Ohkuma, Teamx at clef ehealth 2020: Icd coding with n-gram encoder and code-filtering strategy. *In CLEF*. 2020.

Tan, J. H., H. Fujita, S. Sivaprasad, S. V. Bhandary, A. K. Rao, K. C. Chua, and U. R. Acharya (2017). Automated segmentation of exudates, haemorrhages, microaneurysms using single convolutional neural network. *Information Sciences*, 420, 66 – 76. ISSN 0020-0255.

Tan, M. and Q. Le, EfficientNet: Rethinking model scaling for convolutional neural networks. *In Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2019.

Tananuvat, N., P. Upaphong, C. Tangmonkongvoragul, M. Niparugs, W. Chaidaroon, and M. Pongpom (2021). Fungal keratitis at a tertiary eye care in Northern Thailand: Etiology and prognostic factors for treatment outcomes. *Journal of Infection*, 83(1), 112–118. ISSN 0163-4453.

Teng, F., W. Yang, L. Chen, L. Huang, and Q. Xu (2020). Explainable prediction of medical codes with knowledge graphs. *Frontiers in Bioengineering and Biotechnology*, 8, 867. ISSN 2296-4185.

Thakur, N. and M. Juneja (2018). Survey on segmentation and classification approaches of optic cup and optic disc for diagnosis of glaucoma. *Biomedical Signal Processing and Control*, 42, 162–189. ISSN 17468094.

Thomas, P., A. Leck, and M. Myatt (2005). Characteristic clinical features as an aid to the diagnosis of suppurative keratitis caused by filamentous fungi. *British Journal of Ophthalmology*, 89, 1554 – 1558.

Ting, D. S. W., C. Y.-L. Cheung, G. Lim, G. S. W. Tan, N. D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I. Y. San Yeo, S. Y. Lee, E. Y. M. Wong, C. Sabanayagam, M. Baskaran, F. Ibrahim, N. C. Tan, E. A. Finkelstein, E. L. Lamoureux, I. Y. Wong, N. M. Bressler, S. Sivaprasad, R. Varma, J. B. Jonas, M. G. He, C.-Y. Cheng, G. C. M. Cheung, T. Aung, W. Hsu, M. L. Lee, and T. Y. Wong (2017). Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA*, 318(22), 2211–2223. ISSN 0098-7484.

Tognetto, D., R. Giglio, A. L. Vinciguerra, S. Milan, R. Rejdak, M. Rejdak, K. Zaluska-Ogryzek, S. Zweifel, and M. D. Toro (2021). Artificial intelligence applications and cataract management: A systematic review. *Survey of Ophthalmology*. ISSN 0039-6257.

Torio, C. M. and B. J. Moore (2016). National inpatient hospital costs: the most expensive conditions by payer, 2013: statistical brief# 204.

Tsoumakas, G., I. Katakis, and I. Vlahavas, Mining multi-label data. *In In Data Mining and Knowledge Discovery Handbook*. 2010.

Ung, L., P. J. M. Bispo, S. Shanbhag, M. Gilmore, and J. Chodosh (2019). The persistent dilemma of microbial keratitis: Global burden, diagnosis, and antimicrobial resistance. *Survey of ophthalmology*, 64 3, 255–271.

Valdes, G., I. Simone, C.B., J. Chen, A. Lin, S. Yom, A. Pattison, C. Carpenter, and T. Solberg (2017). Clinical decision support of radiotherapy treatment

planning: A data-driven machine learning strategy for patient-specific dosimetric decision making. *Radiotherapy and Oncology*, 125(3), 392–397. ISSN 01678140.

Veena, H., A. Muruganandham, and T. Kumaran (2020). A review on the optic disc and optic cup segmentation and classification approaches over retinal fundus images for detection of glaucoma. *SN Applied Sciences*, 2(9). ISSN 25233971.

Vilone, G. and L. Longo (2021). Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction*, 3(3), 615–661. ISSN 2504-4990.

Vincent, J.-L., R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. Reinhart, P. Suter, and L. Thijs (1996). The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22(7), 707–710.

Vu, T., D. Q. Nguyen, and A. Nguyen, A label attention model for icd coding from clinical text. *In Proceedings of 29th International Joint Conference on Artificial Intelligence, IJCAI-20*. 2020.

Wagner, R. A. and M. J. Fischer (1974). The string-to-string correction problem. *J. ACM*, 21(1), 168–173. ISSN 0004-5411.

Wang, D., L. Wang, Z. Zhang, D. Wang, H. Zhu, Y. Gao, X. Fan, and F. Tian, *"Brilliant AI Doctor" in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment*. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450380966.

Wang, J., L. Yang, Z. Huo, W. He, and J. Luo (2020). Multi-label classification of fundus images with efficientnet. *IEEE Access*, 8.

Wang, N., M. CHEN, and K. P. Subbalakshmi (2020*a*). Explainable cnn-attention networks (c-attention network) for automated detection of alzheimer's disease. *CoRR*, abs/2006.14135.

Wang, S., X. Li, L. Yao, Q. Z. Sheng, G. Long, *et al.* (2017). Learning multiple diagnosis codes for icu patients with local disease correlation mining. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 11(3), 31.

Wang, S.-M., Y.-H. Chang, L.-C. Kuo, F. Lai1, *et al.* (2020*b*). Using deep learning for automated ICD-10 classification from free text data. *EJBI*, 1–10.

Wang, Z. and J. Yang, Diabetic retinopathy detection via deep convolutional networks for discriminative localization and visual explanation. *In The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, AAAI Workshops. AAAI Press, 2018.

WHO (2019). World health organization report on vision. Online. URL "https://www.who.int/publications/i/item/world-report-on-vision".

Wilkinson, C., F. L. Ferris, R. E. Klein, P. P. Lee, C. D. Agardh, M. Davis, D. Dills, A. Kampik, R. Pararajasegaram, and J. T. Verdaguer (2003). Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, 110(9), 1677–1682. ISSN 0161-6420.

Wintergerst, M., T. Schultz, J. Birtel, A. Schuster, N. Pfeiffer, S. Schmitz-Valckenberg, F. Holz, and R. Finger (2017). Algorithms for the automated analysis of age-related macular degeneration biomarkers on optical coherence tomography: A systematic review. *Translational Vision Science and Technology*, 6(4). ISSN 21642591.

Wisely, C., D. Wang, R. Henao, D. Grewal, A. Thompson, C. Robbins, S. Yoon, S. Soundararajan, B. Polascik, J. Burke, A. Liu, L. Carin, and S. Fekrat (2020). Convolutional neural network to identify symptomatic alzheimer's disease using multimodal retinal imaging. *British Journal of Ophthalmology*. ISSN 00071161.

World Health Organization, *ICD-10 : International statistical lassification of iseases and related health problems / World Health Organization*. World Health Organization Geneva, 2004, 10th revision, 2nd ed. edition.

Wu, X., Y. Huang, Z. Liu, W. Lai, E. Long, K. Zhang, J. Jiang, D. Lin, K. Chen, T. Yu, D. Wu, C. Li, Y. Chen, M. Zou, C. Chen, Y. Zhu, C. Guo, X. Zhang, R. Wang, Y. Yang, Y. Xiang, L. Chen, C. Liu, J. Xiong, Z. Ge, D. Wang, G. Xu, S. Du, C. Xiao, J. Wu, K. Zhu, D. Nie, F. Xu, J. Lv, W. Chen, Y. Liu, and H. Lin (2019). Universal artificial intelligence platform for collaborative management of cataracts. *British Journal of Ophthalmology*, 103(11), 1553–1560. ISSN 00071161.

Xie, P. and E. Xing, A neural architecture for automated icd coding. *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018.

Xie, S., R. B. Girshick, P. Dollár, Z. Tu, and K. He (2017*a*). Aggregated residual transformations for deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5987–5995.

Xie, S., R. B. Girshick, P. Dollár, Z. Tu, and K. He (2017*b*). Aggregated residual transformations for deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5987–5995.

Xu, C., X. Zhu, W. He, Y. Lu, X. He, Z. Shang, J. Wu, K. Zhang, Y. Zhang, X. Rong, Z. Zhao, L. Cai, D. Ding, and X. Li (2019*a*). Fully deep learning for slit-lamp photo based nuclear cataract grading. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11767 LNCS, 513–521. ISSN 03029743.

Xu, K., M. Lam, J. Pang, X. Gao, C. Band, *et al.*, Multimodal machine learning for automated ICD coding. *In Proceedings of the 4th Machine Learning for Healthcare Conference*, Proceedings of Machine Learning Research. PMLR, 2019*b*.

Xu, X., L. Zhang, J. Li, Y. Guan, and L. Zhang (2020*a*). A hybrid global-local representation cnn model for automatic cataract grading. *IEEE Journal of Biomedical and Health Informatics*, 24(2), 556–567. ISSN 21682194.

Xu, Y., M. Kong, W. Xie, R. Duan, Z. Fang, Y. Lin, Q. Zhu, S. Tang, F. Wu, and Y.-F. Yao (2020*b*). Deep sequential feature learning in clinical image classification of infectious keratitis. *Engineering*. ISSN 20958099.

Xu, Y., H.-K. Lam, and G. Jia (2021). Manet: A two-stage deep learning method for classification of covid-19 from chest x-ray images. *Neurocomputing*, 443, 96–105. ISSN 0925-2312.

Xue, J., S. Yan, J. Qu, *et al.* (2019). Deep membrane systems for multitask segmentation in diabetic retinopathy. *Knowledge-Based Systems*, 183, 104887. ISSN 0950-7051.

Xue, Y., T. Xu, L. Rodney Long, Z. Xue, S. Antani, G. R. Thoma, and X. Huang, Multimodal recurrent model with attention for automated radiology report generation. *In* A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and

G. Fichtinger (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Springer International Publishing, Cham, 2018. ISBN 978-3-030-00928-1.

Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding. *In Advances in Neural Information Processing Systems*. 2019.

Young, J., M. Modat, M. J. Cardoso, A. Mendelson, D. Cash, and S. Ourselin (2013). Accurate multimodal probabilistic prediction of conversion to alzheimer's disease in patients with mild cognitive impairment. *NeuroImage: Clinical*, 2, 735 – 745. ISSN 2213-1582.

Yu, J., Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, Generative image inpainting with contextual attention. *In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018.

Yu, K., Z. Yang, C. Wu, Y. Huang, and X. Xie (2021). In-hospital resource utilization prediction from electronic medical records with deep learning. *Knowledge-Based Systems*, 223, 107052. ISSN 0950-7051.

Yu, S., D. Xiao, S. Frost, and Y. Kanagasingam (2019*a*). Robust optic disc and cup segmentation with deep learning for glaucoma detection. *Computerized Medical Imaging and Graphics*, 74, 61–71. ISSN 0895-6111.

Yu, Y., M. Li, L. Liu, Z. Fei, F.-X. Wu, and J. Wang (2019*b*). Automatic icd code assignment of chinese clinical notes based on multilayer attention birnn. *Journal of Biomedical Informatics*, 91, 103114. ISSN 1532-0464.

Yuan, J., H. Liao, R. Luo, and J. Luo (2019). Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11769 LNCS, 721–729. ISSN 03029743.

Zago, G., R. Andreão, B. Dorizzi, and E. Teatini Salles (2020). Diabetic retinopathy detection using red lesion localization and convolutional neural networks. *Computers in Biology and Medicine*, 116. ISSN 00104825.

Zagoruyko, S. and N. Komodakis (2017). Wide residual networks.

Zamir, S. W., A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, Learning enriched features for real image restoration and enhancement. *In* A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm (eds.), *Computer Vision – ECCV 2020*. Springer International Publishing, Cham, 2020. ISBN 978-3-030-58595-2.

Zebin, T. and T. J. Chaussalet, Design and implementation of a deep recurrent model for prediction of readmission in urgent care using electronic health records. *In 2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. 2019. ISSN null.

Zebin, T., S. Rezvy, and T. J. Chaussalet, A deep learning approach for length of stay prediction in clinical settings from medical records. *In 2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. 2019. ISSN null.

Zeng, M., M. Li, Z. Fei, Y. Yu, Y. Pan, and J. Wang (2019*a*). Neurocomputing Automatic ICD-9 coding via deep transfer learning. *Neurocomputing*, 324, 43–50. ISSN 0925-2312.

Zeng, M., M. Li, Z. Fei, Y. Yu, and J. Wang (2019*b*). Automatic ICD-9 coding via deep transfer learning. *Neurocomputing*, 324, 43–50.

Zhang, D., C. Yin, J. Zeng, X. Yuan, and P. Zhang (2020). Combining structured and unstructured data for predictive models: a deep learning approach. *BMC Medical Informatics and Decision Making*, 20(1). ISSN 14726947.

Zhang, H., K. Niu, Y. Xiong, W. Yang, Z. He, and H. Song (2019). Automatic cataract grading methods based on deep learning. *Computer Methods and Programs in Biomedicine*, 182, 104978. ISSN 0169-2607.

Zhang, Z., Y. Xie, F. Xing, M. McGough, and L. Yang, Mdnet: A semantically and visually interpretable medical image diagnosis network. *In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 2017. ISSN 1063-6919.

Zhao, H., H. Li, S. Maurer-Stroh, and L. Cheng (2018). Synthesizing retinal and neuronal images with generative adversarial nets. *Medical Image Analysis*, 49, 14–26. ISSN 1361-8415.

Zhao, R., X. Chen, X. Liu, Z. Chen, F. Guo, and S. Li (2020). Direct cup-to-disc ratio estimation for glaucoma screening via semi-supervised learning. *IEEE Journal of Biomedical and Health Informatics*, 24(4), 1104–1113. ISSN 21682194.

Zheng, R., L. Liu, S. Zhang, C. Zheng, F. Bunyak, R. Xu, B. Li, and M. Sun (2018). Detection of exudates in fundus photographs with imbalanced learning using conditional generative adversarial network. *Biomedical Optics Express*, 9(10), 4863. ISSN 21567085.

Zhou, Y., B. Wang, X. He, S. Cui, and L. Shao (2020). Dr-gan: Conditional generative adversarial network for fine-grained lesion synthesis on diabetic retinopathy images. *IEEE Journal of Biomedical and Health Informatics*. ISSN 21682194.

Zijdenbos, A., B. Dawant, R. Margolin, and A. C. Palmer (1994). Morphometric analysis of white matter lesions in mr images: method and validation. *IEEE transactions on medical imaging*, 13 4, 716–24.

Zimmerman, J., A. Kramer, D. McNair, and F. Malila (2006). Acute physiology and chronic health evaluation (apache) iv: Hospital mortality assessment for today's critically ill patients. *Critical Care Medicine*, 34(5), 1297–1310.

Zuiderveld, K., *Contrast Limited Adaptive Histogram Equalization*. Academic Press Professional, Inc., USA, 1994. ISBN 0123361559, 474–485.

Bio-data

Name:	Veena Mayya
Current Address:	Research Scholar, Department of Information Technology, NITK Surathkal Mangaluru, Karnataka India - 575025.
Permanent Address:	Samrudhi, Vigneshwar Road, Adiudupi, Udupi, Karnataka India - 576103.
Email:	mayya.veena@gmail.com
Mobile No:	+91 94487 70333

Qualification:

Ph.D. in Information Technology
Department of Information Technology
National Institute of Technology Karnataka, Surathkal
Mangalore, India.

M.Tech in Software Engineering
MIT, MAHE, Manipal University, Karnataka, India.

B.Tech in Computer Science & Engineering
Visvesvaraya Technological University (VTU)
Karnataka, India.

Research Area: Healthcare Analytics, Natural Language Processing, Deep Learning, Computer Vision