PREDICTIVE ANALYTICS BASED INTEGRATED FRAMEWORK FOR INTELLIGENT HEALTHCARE APPLICATIONS

THESIS

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

DOCTOR OF PHILOSOPHY

by

GOKUL S KRISHNAN
(Reg. No.: 165034IT16F03)

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

SEPTEMBER 2020

DECLARATION

I hereby declare that the Research Thesis entitled "**PREDICTIVE ANALYTICS BASED INTEGRATED FRAMEWORK FOR INTELLIGENT HEALTHCARE APPLICATIONS**", 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 :

Gokul S Krishnan
Reg.No.: 165034IT16F03
Department of IT,
NITK Surathkal.

CERTIFICATE

This is to certify that the Research Thesis entitled, "**PREDICTIVE ANALYTICS BASED INTEGRATED FRAMEWORK FOR INTELLIGENT HEALTHCARE APPLICATIONS**", submitted by **GOKUL S KRISHNAN (Reg. No. 165034IT16F03)**, as the record of research work carried out by him, *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 :

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

Dr. Biju R Mohan
Chairman - DRPC,
Department of IT,
NITK Surathkal.

Dedicated to

My Mother, Father, Grandmother

&

Those Special Ones who were 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. Shashidhar G Koolagudi, Dept. of CSE and Dr. Geetha V, Dept. of IT, for all the constructive feedback they have given me which has definitely enhanced my research.

Next, I thank all of Department of Information Technology, NITK – all faculties and staff for numerous opportunities of learning and all the support 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 of attaining my Ph.D.

I thank my wonderful Parents – C Lalitha and K Subrahmony, for the love, patience and moral support they have showered on me throughout my life, even more importantly, during the period of my research. I take this opportunity to offer prayers to my grandmother, Lakshmi Chandrashekhar, whom I lost recently, for her love, support and blessings. I also thank the rest of my super cool family, who has always prayed for me and supported me.

I extend my thanks to my very Special Friends – Manjunath, Sanjay, Karthik, Archana, Sangeetha, Ashwin, Shridhar, Ranjit, Sreenath, Anoop, Praneetha and Deepa, pillars of my life, who have inspired me for moving ahead and more importantly, supported me and made me smile even during hard times. I also thank all my labmates and fellow research scholars for the many productive discussions and fun moments that will be in my memories forever.

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

GOKUL S KRISHNAN

Abstract

Healthcare analytics is a field that deals with the examination of underlying patterns in healthcare data in order to determine ways in which clinical care can be improved - in terms of patient care, hospital management and cost optimization. Towards this end, health information technology systems such as Clinical Decision Support Systems (CDSSs) have received extensive research attention over the years. A CDSS is designed to provide physicians and other health professionals assistance with clinical decision-making tasks, based on automated analysis of patient data and other knowledge sources. Recent advancements in Big Data and Healthcare Analytics have seen an emerging trend in the application of Artificial Intelligence techniques to healthcare data for supporting essential applications like disease prediction, mortality prediction, symptom analysis, epidemic prediction etc. Despite such major advantages offered by CDSSs, there are several issues that need to be overcome to achieve their full potential. There is scope for significant improvements in terms of patient data modeling strategies and prediction models, especially with respect to clinical data of unstructured nature.

In this research thesis, various approaches for building decision support systems towards patient-centric and population-centric predictive analytics on large healthcare data of both structured and unstructured nature are presented. For structured data, an empirical study was performed to observe the effect of feature modeling on mortality prediction performance, which revealed the need for extensive study on the relative relevance of features contributing to mortality risk prediction. Towards this, a Genetic Algorithm based wrapper feature selection method was proposed, for determining the most relevant lab events that help in effective patient-specific mortality prediction.

Clinical data in the form of unstructured text, being rich in patient-specific information sources has remained largely unexplored, and could be potentially used to leverage effective CDSS development. Towards this, an Extreme Learning Machine based patient-specific mortality prediction model built on ECG text reports of cardiac patients was proposed. The approach, which involved word embedding based feature modeling and an unsupervised data cleansing technique to filter out anomalous data, underscored the importance of effective word embeddings. Therefore, our next objective was to study the word embedding models and their role in feature modeling for building effective CDSSs. A benchmarking study on performance of word representation models for patient specific mortality prediction using unstructured clinical notes was performed.

Our next objective involved analyzing and utilizing the unstructured clinical notes for building effective disease prediction models. An ontology-driven feature modeling approach was proposed, for designing a disease group prediction model built on unstructured radiology reports. In order to solve the problems of sparsity and high dimensionality of this approach, another feature modeling approach based on Particle Swarm Optimization (PSO) and neural networks was proposed to further enhance the performance of disease group prediction models using unstructured radiology reports. With the objective of analyzing physician notes, a hybrid feature modeling approach was proposed to leverage the latent information embedded in unstructured patient records for disease group prediction. Towards addressing the incremental and redundant nature of unstructured clinical notes, aggregation of nursing notes using *TAGS* and *FarSight* approaches were also explored for effective disease group prediction, which demonstrated significant potential towards enabling early disease diagnosis.

For population health analysis (flu vaccine hesitancy, flu vaccine behaviour and depression detection), a generic model called Multi-task Deep Social Health Analyzer (MDSHA) was proposed which uses a PSO based topic modeling approach for effective feature representation and predictive modeling. All proposed approaches were compared to existing state-of-the-art approaches for respective prediction tasks using standard datasets. The promising results achieved underscore the superior performance of the approaches designed in this research, and reveal much scope for adaptation in the healthcare field for improving quality of healthcare.

KEYWORDS: Healthcare Informatics, Clinical Decision Support Systems, Predictive Analytics, Machine Learning, Natural Language Processing, Evolutionary Computing

Contents

Part II - Building CDSSs for Structured Clinical Data

Part III - Mortality Risk Prediction CDSSs using Unstructured Clinical Data

Part IV - Disease Prediction CDSSs using Unstructured Clinical Data

Part V - Towards Population Analytics based CDSSs

List of Figures

List of Tables

Abbreviations

ADE	Adverse Drug Events
ADR	Adverse Drug Reaction
AI	Artificial Intelligence
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
APACHE	Acute Physiology And Chronic Health Evaluation
AUPRC	Area Under Precision Recall Curve
AUROC	Area Under Receiver Operating Characteristic Curve
BP	Backpropagation
CART	Classification And Regression Tree
CBOW	Continuous Bag-Of-Words
CDC	Centre for Disease Control and Prevention
CDS	Clinical Decision Support
CDSS	Clinical Decision Support System
CNN	Convolutional Neural Network
DBN	Deep Belief Network
DNN	Deep Neural Network
DPM	Disease Prediction Model
DS	Dataset
ECG	Electrocardiogram
EHR	Electronic Health Record
ELM	Extreme Learning Machine
EMR	Electronic Medical Record
FFNN	Feed Forward Neural Network
FN	False Negatives
FP	False Positives
FPR	False Positive Rate
FS	Feature Selection

GA	Genetic Algorithm
GAWFS	Genetic Algorithm based Wrapper Feature Selection
GloVe	Global Vectors
GRU	Gated Recurrent Unit
HADM_ID	Hospital Admission Identifier
HDP	Hierarchical Dirichlet Process
HIMS	Hospital Information Management System
HSLDA	Hierarchically Supervised Latent Dirichlet Allocation
HVE	Hard Voting Ensemble
ICD9	International Classification of Diseases, 9th Version
ICU	Intensive Care Unit
ILI	Influenza like Illness
IR	Information Retrieval
KNN	K-Nearest Neighbours
LASSO	Least Absolute Shrinkage & Selection Operator
LDA	Latent Dirichlet Allocation
LR	Logistic Regression
LSTM	Long Short Term Memory
MCC	Matthews Correlation Coefficient
MDSHA	Multi-task Deep Social Health Analyzer
ME	Monge-Elkan
MeSH	Medical Sub Headings
MI	Mutual Information
MICU	Medical Intensive Care Unit
MIMIC	Medical Information Mart for Intensive Care
ML	Machine Learning
MLP	Multilayer Perceptron
MMDL	Multi-modal Deep Learning
MPM	Mortality Risk Prediction Model
MRD	Medical Records Department
MSE	Mean Squared Error
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NMF	Nonnegative Matrix Factorization
NN	Neural Network
NPMI	Normalized Pointwise Mutual Information
OASIS	Oxford Acute Severity of Illness Score

OSN	Open Social Network
OvR	One vs Rest
PMC	PubMed Central
PSO	Particle Swarm Optimization
PV	Paragraph Vectors
PV-DBoW	Distributed Bag-of-Words Model of Paragraph Vectors
PV-DM	Distributed Memory Model of Paragraph Vectors
RBF	Radial Basis Function
RBM	Restricted Boltzmann Machine
ReLU	Rectified Linear Unit
RFE	Recursive Feature Elimination
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SAPS	Simplified Acute Physiological Score
SC	Semantic Coherence
SE	Stacking Ensemble
SFS	Sequential Feature Selection
SGD	Stochastic Gradient Descent
SICU	Surgical Intensive Care Unit
Sim	Similarity
SLFNN	Single hidden Layer Feedforward Neural Network
SNEFT	Social Network Enabled Flu Trends
SOFA	Simplified Organ Failure Assessment
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SVR	Support Vector Regression
TAGS	Term weighting of unstructured notes AGgregated using fuzzy Similarity
TC	Topic Coherence
Tf-Idf	Term frequency-Inverse docment frequency
TN	True Negatives
TP	True Positives
TPR	True Positive Rate
UMLS	Unified Medical Language System
WHO	World Health Organization

PART I

Introduction & Background

Chapter 1

Introduction

The healthcare delivery process encompasses "*the maintenance or improvement of health via prevention, diagnosis, treatment, recovery or cure of disease, illness, injury and other physical and mental impairments in people*". Quality of healthcare is an important aspect in the promotion of health and well-being of people around the world. As per World Health Organization (WHO) reports[1], a well-functioning healthcare system requires a financing mechanism, a well trained & adequately paid workforce, reliable information to base decisions & policies and well maintained health facilities to deliver quality medicines and technologies. The United States alone spent $3.5 Trillion in 2017 to maintain their healthcare systems, a whopping 18% of their GDP (White, 2007), and their per capita spending on healthcare stood at $10,348 according to 2016 reports, which exceeds the world average by more than eight times[2]. The US healthcare system has extensively invested in the use of Information Technology in healthcare services delivery.

India, the world's second most populous country, has made vast strides towards the implementation of nation-wide healthcare systems for its people. Infectious diseases like Smallpox and Polio have been eradicated through successful large-scale public awareness and education programs, doubling the life expectancy of citizens (Reddy *et al.*, 2011). However, the health outcomes remain inadequate when India is compared with other countries that were at similar economic stages of development at the time of independence; preventable disease burden remains a significant challenge. The budget allocation for healthcare in India rose 16% in 2019-2020 over the previous year budget, reserving an amount of ₹ 61,398 crores, with an allocation of ₹ 6,400 Crores for the exclusive healthcare initiative – Ayushman Bharat Pradhan Mantri Jan Aroghya Yojana (AB-PMJAY) for making

[1] Health Systems Governance,
https://www.who.int/health-topics/health-systems-governance#tab=tab_1

[2] Global Health Expenditure Database, http://apps.who.int/nha/database

necessary interventions and additions in various government healthcare systems across the country[3]. An efficient and effective healthcare system can significantly contribute to a country's economy, development and even industrialization.

Application of computers and information technology in healthcare, termed as Health Information Technology, has proven to be effective in the betterment of healthcare management in various aspects. Health Information Technology is "*the application of information technology involving computer hardware and software to deal with the storage, retrieval, sharing, and use of healthcare data, information and knowledge for better communication and decision making*". Computerization of medical records has been in practice in the healthcare industry over the past 2-3 decades in developed countries. Early research showed that paper based medical records have limited impact (Pollak, 1983), while their usage covers day-to-day requirements of recording clinical events, these records are not suitable for patients with long-term illnesses or extensive medical history as future procedures depend on the past diagnoses, procedures and medication that they were subjected to.

The availability of patients' medical records creates a positive impact on patient management decisions, only when the records are in an organized, standardized and retrievable format (Stead and Hammond, 1983). In 1983, a medical record system called The Medical Record (TMR) was designed by Stead and Hammond (1983) who reported that using computerized medical records significantly prevented accidental oversights by doctors, while it also improved the communication between doctors and patients. They also reported on the new possibilities for data analysis and review made possible due to the organized nature of the medical records and ways in which computerized medical records can be time-oriented, displayed in a useful fashion for doctors, analyzed and put to use for quality improvement and administrative purposes. McDonald and Tierney (1988) demonstrated how computerized medical records can help in tasks such as, organization and retrieval of patient data, decision-making and diagnoses, retrieving past or similar clinical cases/experiences for care, administration or even research purposes. Investing in and using an electronic medical records system in primary care units has been shown to result in positive returns for the healthcare provider organization (Wang *et al.*, 2003). They reported that the estimated net benefit from using an electronic medical record for a 5-year period was $86,400 per provider, which can be contributed to savings due to optimized drug management, improved utilization of radiology tests, better capture of charges, and decreased billing errors. In one-way sensitivity analyses, the model was most sensitive to

[3] Ayushman Bharat, https://www.pmjay.gov.in/

the proportion of patients whose care was long-term, as larger data is available for designing decision-making tools. In view of this, several prominent hospitals and healthcare organizations have actively invested in implementation of computerized medical records, formatted as per the standards put forth by their respective countries, in the form of Electronic Health Records (EHRs) or Electronic Medical Records (EMRs).

1.1 Electronic Health Records

Gunter and Terry (2005) defined an Electronic Health Record (EHR) as *"the systematized collection of patient or population health information, which are electronically stored in a digital format"*. Electronic Health Records (EHRs) are seen as an significant step towards streamlining the storage, management and dissemination of patient data in hospitals. EHRs are real-time, patient-centered records which provide data in a standardized format so that the patient data can be accessed securely by all authorized stakeholders - doctors, specialists, hospitals, insurance providers, and others. EHRs contain vital patient-specific information like medical history, diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, and laboratory test results. They form the basis for the development of knowledge-enabled healthcare IT systems that support evidence-based clinical decision making, in addition to facilitating automating and streamlining healthcare delivery workflows.

Herland *et al.* (2014) categorized healthcare data available for informatics into four levels, where analytics research and application developments are underway.

1. *Micro-level Data:* Performed on genes, molecules, etc. to predict diseases like cancer and other genetic disorders.

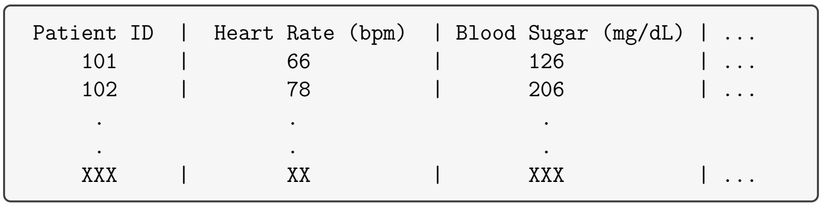
2. *Tissue-level Data:* Performed on tissues in plants and animals for purposes of brain research and the other human-scale biological factors.

3. *Patient-level Data:* Performed on mostly EHRs (text and images) to make predictions on general diseases or specific ones, ICU mortality, etc.

4. *Population-level Data:* Performed on data of a population in form of EHRs or even open social media data to make epidemic predictions.

Patient-level data in terms of EHR can be structured or unstructured with respect to the nature of the data itself.

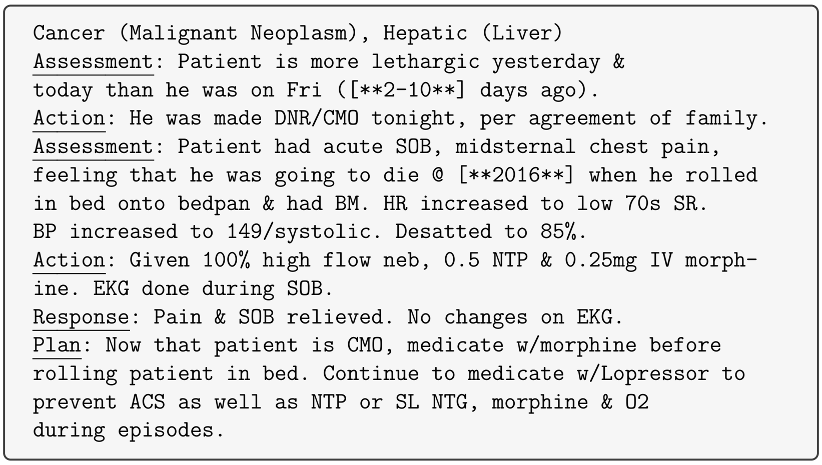
Structured patient data: Patient data stored in a consistent and an organized manner, mostly in the form of rows and columns, with keywords to identify and analyze the data values. Such structured patient data can be easily validated against allowed ranges of values and rules. These include numerical values such as age, gender, height, weight, lab test values, etc. These data can be mostly directly and easily used for various analysis and training ML based prediction models as they are mostly numerical or categorical values. A very basic example of structured patient data in the form of rows and columns is shown in Figure 1.1.



```
Patient ID   |  Heart Rate (bpm)   | Blood Sugar (mg/dL) | ...
    101      |        66           |        126          | ...
    102      |        78           |        206          | ...
     .                .                      .
     .                .                      .
    XXX      |        XX           |        XXX          | ...
```

Figure 1.1: An example of structured patient data – A sample table indicating lab values

Unstructured patient data: Unstructured patient data as the name suggests is the opposite of structured data – there is no organisation of the data. The most common kinds of unstructured data available in hospital scenarios are free text clinical notes narrated by doctors or nurses regarding the condition of patients undergoing treatment and discharge summaries; medical images such as X-rays or MRI scans; etc. More importantly, unstructured patient data requires manual analysis and processing by health personnel, before it can be used for ML applications. Unstructured data cannot be directly consumed for any automated analysis and has to be therefore brought to some form of structured representation for further processing for analysis. A sample nursing note shown in Figure 1.2 is an apt example of unstructured patient data in the form free text.



```
Cancer (Malignant Neoplasm), Hepatic (Liver)
Assessment: Patient is more lethargic yesterday &
today than he was on Fri ([**2-10**] days ago).
Action: He was made DNR/CMO tonight, per agreement of family.
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. HR increased to low 70s SR.
BP increased to 149/systolic. Desatted to 85%.
Action: Given 100% high flow neb, 0.5 NTP & 0.25mg IV morph-
ine. EKG done during SOB.
Response: Pain & SOB relieved. No changes on EKG.
Plan: Now that patient is CMO, medicate w/morphine before
rolling patient in bed. Continue to medicate w/Lopressor to
prevent ACS as well as NTP or SL NTG, morphine & O2
during episodes.
```

Figure 1.2: An example of unstructured patient data – A note recorded by a nurse (Source: MIMIC-III Dataset (Johnson *et al.*, 2016))

Given the wide variety and volume of healthcare data available, traditional approaches to managing it will no longer suffice. Also, such data is continuously generated over time, at every patient contact episode, thus exhibiting Big data characteristics. Hence, emerging computational technologies like Big Data Analytics and Machine Learning can be applied to large-scale patient-specific data, for improving healthcare systems and managing effective care delivery. Such applications encompass Predictive modeling, Preventive modeling, Intelligent Retrieval, Automatic Information/Concept Extraction, Recommendations for Doctors and Patients, etc, which have the potential for revolutionizing and creating a huge positive impact on the way healthcare is provided.

1.2 Clinical Decision Support Systems

Big Data Analytics in Healthcare is an emerging field that has the potential for significant improvement in areas of clinical operations, research & development, public health policies, evidence based medicine, genomic analytics and patient profile analytics. Predictive Analytics based applications built on EHRs that demonstrate superior performance over traditional rule based systems helped caregivers in a lot of aspects such as diagnosis and intervention decisions (Simpao *et al.*, 2014). Therefore, such systems came to be known as Clinical Decision Support Systems (CDSSs). Wikipedia defines CDSS as "*a health information technology system that is designed to provide physicians and other health professionals with Clinical Decision Support (CDS), that is, assistance with clinical decision-making tasks.*" Dr. Robert Hayward has provided a working definition for CDSS – "*a link between health observations and health-related knowledge that influences treatment choices by clinicians improved healthcare.*"

Perreault and Metzger (1999) defined four key functions for electronic CDSSs, which are as follows:

1. *Administrative:* Supporting clinical coding and documentation, authorization of procedures, and referrals.

2. *Managing clinical complexity and details:* Keeping patients on prescribed protocols; tracking orders, referrals follow-up, and preventive care.

3. *Cost control:* Monitoring medication orders; avoiding duplicate or unnecessary tests.

4. *Decision support:* Supporting clinical diagnosis and treatment plan processes; and promoting use of best practices, condition-specific guidelines, and population-based management.

Several studies have shown that computerized CDSSs can improve clinical performance and patient outcomes (Kennebeck *et al.*, 2012; Nachtigall *et al.*, 2014; Moja *et al.*, 2014). CDSS systems, with their potential to minimize practice variation and improve patient care, have been adapted in practice by many healthcare verticals and practitioners on large and small-scale (Trivedi *et al.*, 2002; Dorr *et al.*, 2007). With the emergence of artificial intelligence, computational tools and expert systems designed to capture and encode the knowledge of subject experts, CDSSs can be critical in revolutionizing clinical care. This is especially true in the era of Big Data Analytics, where large-scale data and knowledge is available for pattern analysis and knowledge discovery (Kawamoto *et al.*, 2005; Black *et al.*, 2011). Analytics on various types of EHRs – Computer Vision and Machine Learning based analytics on medical images, Signal analytics on physiological signals and Big Data and Machine Learning based analytics on Gene/Protein data; have the potential to revolutionise healthcare industry and healthcare delivery (Belle *et al.*, 2015).

1.2.1 Knowledge-based vs. Non-knowledge based CDSSs

As per Berner (2007), CDSSs are classified into two – Knowledge-based CDSSs and Non-knowledge based CDSSs. Knowledge-based CDSSs are rule-based systems that provide a means of interaction to the user (patient/caregivers), providing important information that can assist a physician for decision making. It mostly consists of three parts – a knowledge base, a reasoning/inference engine and finally, an interface to communicate with the user. The knowledge base consists of rules defined by healthcare domain experts and the inference or reasoning engine contains the formulas or logic methods to combine the rules in the knowledge base and a real patient's data. Generally, a knowledge-based CDSS accepts an input as patient data mostly in electronic form and then performs the inference or reasoning using the defined logic based on rules in the knowledge base. Finally, the system provides physicians or other caregivers with an output in the form of recommendation, alerts or even diagnosis probabilities that can help in his/her decision-making in the clinical diagnosis. A generic model of a knowledge-based CDSS is as depicted in Figure 1.3.

Non-knowledge based CDSSs are systems that incorporate Artificial Intelli-

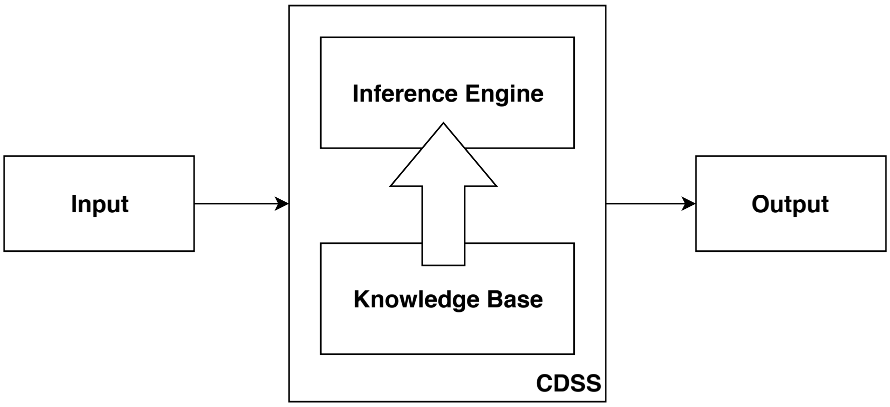


Figure 1.3: General Model of a Knowledge-based CDSS (Berner, 2007)

gence (AI) techniques like Machine Learning, that enable the systems to learn from historic clinical data through pattern recognition (Berner, 2007). A critical shortcoming of Knowledge-based CDSSs is that a human expert should always put the knowledge into the system directly, which are explicit for various cases, diagnoses and scenarios. This meant that development of CDSSs required a huge number of rules to be entered into the knowledge base by a large number of domain experts, leading to bulkiness and inconsistency. Hence, the requirement of systems that can 'learn' to perform inference or reasoning for incoming patient data based on experience or historic data led to application of Data Mining and AI based techniques in the domain of healthcare. Machine Learning techniques and neural network models, using supervised or unsupervised methods, are applied to data in the form of EHRs or even unstructured clinical data to come up with predictions that can aid in the decision making or timely interventions of physicians and other caregivers. A generic model of a non-knowledge based CDSS that uses supervised or unsupervised machine learning approaches is illustrated in Figure 1.4.

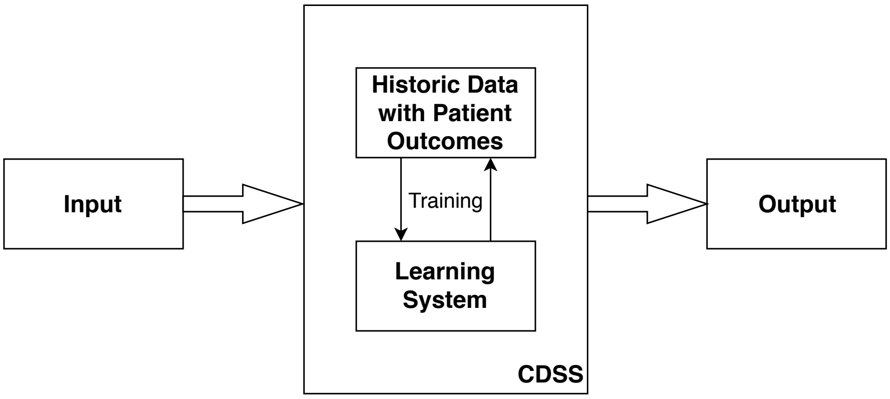


Figure 1.4: A General Model of a Non-knowledge based CDSS (Berner, 2007)

A typical non-knowledge based CDSS pipeline, that makes use of Machine Learning techniques to assist physicians, mainly involves three tasks – *Data Preparation/Preprocessing*, *Feature Modeling/Engineering* and *Learning/Prediction Modeling*. The patient data used in the pipeline may be in the form of structured EHRs or unstructured clinical data, which need adequate preprocessing, using various techniques for making it standard and machine readable. The data is then modeled as features by vectorization and similar extraction strategies to bring the data to a form that can be fed into a machine learning based prediction model for training. Finally, the features, along with the patient outcomes of historic data or labels, are trained using machine learning techniques to predict the outcomes for new incoming patient data. The overall workflow for a typical ML based CDSS is as illustrated in Figure 1.5.

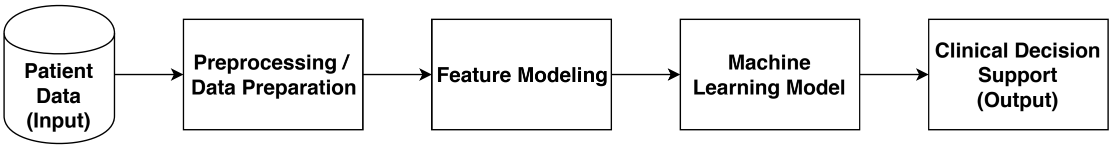

Patient Data (Input)
Preprocessing / Data Preparation
Feature Modeling
Machine Learning Model
Clinical Decision Support (Output)

Figure 1.5: A Typical Machine Learning based CDSS – Workflow

1.2.2 Learnable CDSS Models - Need vs. Impact

Over the past 30 decades, development of systems that offer clinical decision support with the objective of improving healthcare quality and enhance the medical decision making process have seen active research interest. Three important factors that underscore the need for incorporating computer algorithms and IT systems in CDSSs, as put forth by Musen *et al.* (2014), are listed below.

1. *Information Needs and Data Management:* Modern clinical decision making is characterized and has evolved based on an ever-increasing knowledge base and also depends on growing datasets that include patient characteristics from phenotype to genotype. Due to exponential increase in patient volume, countries like the US have seen a steady decline in the average time available for a typical physician-patient encounter, increasing physician fatigue and often causing clinicians to miss out critical information regarding the patients' condition (Baron, 2010). In a study conducted on ambulatory cases, it was observed that 81% of the time, clinicians missed at least four items with respect to patients (Tang *et al.*, 1996). It has also been observed that clinicians and healthcare providers face challenges in acquiring detailed

and relevant information regarding a patient's condition or even a summary of the patient's history, which can be helpful in making decisions regarding further tests or treatment procedures. Studies suggest that 18% of medical errors may be caused due to unavailability of relevant or crucial patient information at the opportune time (Leape, 1994). Therefore, there is a rising demand for improved decision support systems based on clinical information of patients.

2. *Personalized Medicine:* An emerging era of genomics and biomedicine demands the need for personalized medicine and the need to tailor healthcare to individualized factors for improving healthcare delivery (Ginsburg and Willard, 2009). Personalized medicine aims to facilitate customized decision making by taking into account various patient-specific factors (data) like – family history, social and environmental factors along with health (genomics/physiological) data and even patient preferences of care, wherever applicable. This requires clinical practitioners to master an additional-level of knowledge that can add to their ever-increasing cognitive fatigue. As personalized medicine is well on its way to becoming a practical norm, clinicians will be hugely benefited with the unintrusive assistance that computers and information technology can provide, thus underscoring the need for intelligent CDSSs. Such data-driven CDSSs not only help the clinical decision support process to be personalized to tune in on patient needs, but also help justify the decision by providing evidence-based explanations in terms of various physiological variables or even genetic-level data.

3. *Cost-Benefit Tradeoff:* Computer based learnable CDSSs have the potential to influence hospitals and clinicians to better optimize their resources in terms of money and other aspects like lab machinery, ICU beds, ventilators, etc. While the hospital can make use of intelligent CDSSs for various applications to deliver better healthcare to more number of patients in lesser time, the patients can also save resources in terms of cost and time for only required treatment procedures and reduced waiting time for appointments and medicines. In a country like India, where patients bear a considerable share of healthcare costs, it is imperative to adopt practices that not only alleviate the burden, but also significantly enhance the quality of care delivery.

In view of these three aspects, a fundamental change in the way hospitals and clinicians create, store and manage patient-specific data is evidenced, to enable the use of intelligent CDSSs for applications like diagnosis support, severity/risk prediction, procedure recommendation, patient history summary generation, appointment prioritizing, etc. Patients too are benefitted through better access to their own health-related data, made possible by providing a secure profile through which they can manage appointments, optimized appointment schedules and treatment procedures, prescriptions, recommendations etc. Therefore, learnable CDSSs with patient-centric personalization and better patient information management capabilities that can also save resources are in demand and can be a boon to the healthcare industry.

1.2.3 EHR Adoption and CDSS Development: Current Scenario

The vast majority of hospitals in developed countries like USA, UK, Germany and Australia are well on their way towards adopting standardised and structured EHRs. In some countries, almost 90% of hospital records are now digital (Coorevits *et al.*, 2013). If exploited in the right direction, EHR adoption for analytics is key to solving problems related to Clinical Decision Support, clinical care quality and reliable information flow among individuals and organisations participating in healthcare. This requires good regulations and standards proposed by organisations, respective governments etc. for full scale adoption and realization (Coorevits *et al.*, 2013).

However, the fact remains that most developing countries are far from moving away from paper-based medical records. Many of these developing countries, including India, face complex challenges in management of healthcare data and delivery of healthcare services (Braa *et al.*, 2004). The implementation of technology based healthcare solutions such as CDSSs encounter challenges like inadequate funding, lack of resources and weak infrastructures (Sood *et al.*, 2008). Other kinds of challenges identified by Sood and Tech (2004) include computer technology related illiteracy and inadequate numbers of trained caregivers that can utilize such expert systems. These challenges have resulted in a low adoption rate of EHR systems in developing countries like India. However, over the past few years, through the increased usage of computers, training of caregivers, infrastructure development and stricter government regulations and laws, the number of hospitals and clinics that use computer systems to store patient information are on the rise. But, these computerized patient records are not in the form of structured EHRs and are mostly in the form of plaintext records and medical images, which are stored in raw format and are used mostly for manual reference by doctors and nurses. The unstructured nature of textual medical data and images, in such scenarios, presents many challenges in exploiting their full potential.

Hospitals that have adopted standard EHR formats can provide structured EHR data as a service through respective hospital warehouses. Few such examples are Mount Sinai Data Warehouse used by Miotto *et al.* (2016) and Sutter Health hospital data used by Choi *et al.* (2016) in developing their respective CDSSs. Other than that, there are open datasets as well that are made available for coding challenges and research purposes such as i2b2 (Sun *et al.*, 2013), MIMIC II (Saeed *et al.*, 2002) and MIMIC III (Johnson *et al.*, 2016). Research studies are being performed on many such datasets (both structured and unstructured) for development of not only CDSSs like ICU mortality risk prediction models, disease prediction models, etc., but also for other important tasks of the healthcare management workflows like hospital patient management, effective hospital finance management, designing insurance models for both clients as well as companies, etc. Along with traditional techniques of Big Data Analytics, Data Mining and Natural Language Processing, concepts of Artificial Intelligence, Machine Learning, Neural Networks and Deep Learning applied to the field of healthcare have shown great potential in creation of predictive and analytical models with a very good accuracy and precision. Moreover, the problem of developing and extracting structured patient representation from raw unstructured data itself is an open research problem. With the use of Text Mining, Natural Language Processing and Machine Learning techniques on de-identified patient records, such challenges can be addressed through automatic concept extraction, categorization and modelling for development of effective CDSSs. This highlights a major area of work, based on under-utilized unstructured clinical text sources, which provides ample scope for exploring avenues for impactful research towards development of CDSSs.

1.2.4 A Motivating Example

To highlight the prevalent situation in practical hospital scenarios, we consider an example. This scenario will be treated as a running example throughout this thesis, wherever real-world context to the research problems addressed are to be provided. Let us consider three physicians, *Dr. Alice*, *Dr. Bob* and *Dr. Charlie* working in hospitals A, B and C, respectively. The hospitals have incorporated various levels of IT infrastructure and medical record management systems, as highlighted in the list below. These three types of hospitals effectively cover the various stages at which real-world hospital IT systems are at, and we aim to describe how CDSSs fit into the mix, given the inherent challenges evoked by this highly heterogeneous ecosystem.

1. *Hospital A* is a non-computerized hospital that still employs a paper-based medical records system. Examples of such hospitals include primary care centres and rural hospitals where computerization and other information technology services are yet to be implemented.

2. *Hospital B* has a full-fledged implementation of a standardized EHR system based on relational databases, forming a well-designed Hospital Information Management System (HIMS). The EHR system of *Hospital B* consists of only structured data, strictly adhering to defined EHR standards, and therefore, data entry is undertaken mostly in the form of readings and values. Hospitals in major cities of developed countries can be categorized into this group.

3. *Hospital C* has a 'semi-EHR' system, i.e., computerization of records is implemented, however, they still do not have a sophisticated EHR system similar to *Hospital B*. For each patient in *Hospital C*, various types of notes are maintained, typically in text format, thus making their patient records semi-structured. Examples include hospitals in tier-2 and tier-3 cities and towns in developing countries where computerization is available, however, patient records are still not stored in standardized EHR standards. The hospitals in this category can be also mapped to those organizations which are in the transition phase towards becoming a full-fledged EHR-based hospital.

Given this background, we now consider three use cases from a normal routine work day in a hospital. Firstly, we consider the process from the perspective of the out-patient who visits the hospital for consultation with doctors, typically on pre-fixed appointments. Second, the case of an in-patient admitted to the Intensive care unit (ICU) of the hospital, requiring significantly more medical attention is discussed. Thirdly, we also consider the scenario from the point of view of the Hospitals' Medical Records Department, which plays a critical role in the daily operations of a hospital.

1.2.4.1 Case 1: Out-patient visits.

Let us consider a scenario where an out-patient visits the respective hospitals. A patient arriving at the reception of *Hospital A* to seek the medical services of *Dr.*

Alice will have to provide a card or a patient ID (along with payment of fees) so that the staff can go to a particular rack in the MRD department to fetch the record. The patient meanwhile waits at the front desk or outside the doctor's room. The patient may have to wait for a long time as the file has to reach the doctor's room from the MRD once the staff finds the file. Based on the queuing system, the patient awaits their turn and consults with *Dr. Alice*. During consultation, *Dr. Alice* asks the patient's problems, symptoms, family history, etc and reaches a diagnosis decision and provides the patient with a prescription (while making a copy for the hospital file as well). The patient then has to visit the pharmacy for purchasing medicines, which creates another paper-based record, which may or may not be mapped to the patient's record.

Now, patients arriving at the reception of computerized *Hospital B* or *Hospital C* have the option of providing their hospital card (embedded with an unique identifier in the form of barcode, QR code or RFID[4]) or even just their mobile number, as the hospital is equipped with a full-fledged (or intermediate-level) EHR based system. Preliminary collection of symptoms or problems faced by the patient may also be performed, which are fed into the EHR system, to update the patient record. *Dr. Bob*/*Dr. Charlie*'s patient list on his desktop computer gets updated dynamically with all details given by the patient during the preliminary evaluation. Now, a CDSS consuming data from the patient records can process the symptoms/problems reported by the patient and the order of consultation gets dynamically updated based on his/her severity. During the consultation, *Dr. Bob*/*Dr. Charlie* already has access to all necessary details of the patient generated by the hospital's CDSS in time for the consultation, including a summary of the patient's medical history, and insights with reference to relevant details like his/her allergy information. The doctor and patient during the consultation time can spend on result-oriented discussion on personalizing the treatment plan, which is once again mapped to the patient records. After the consultation, the prescription will be automatically dispatched through the system to the pharmacy. By the time the patient moves towards the pharmacy, his order will be ready and he/she just has to collect it from the desk. At the end of the visit, the patient will have to pay the overall amount and will have various payment options for the same (cash, card or even automatically deducted from the wallet/bank account attached to the hospital record).

The difference between the circumstances faced by both patients and doctors in the two categories of hospitals are predominantly noticeable and it is clear how

[4] Quick Response codes, Radio Frequency IDentification

hospitals equipped with well-designed IT systems providing support to operate CDSSs can enhance hospital-centric experience of patients significantly, while also optimizing doctors' time and hospital resources. The overall time and efforts spent by a patient in *Hospital B* and *Hospital C* are significantly lesser than that spent by a patient in *Hospital A*. Meanwhile, *Dr. Bob* and *Dr. Charlie* in *Hospital B* and *Hospital C* respectively, can potentially treat more patients in a day than *Dr. Alice* in *Hospital A*. Moreover, *Dr. Bob* and *Dr. Charlie* will be able to deliver treatment to the patients with a personal touch as the decision support system provides them with good insights of complete background data and medical history of the patients. In contrast to this, *Dr. Alice* will have to spend a lot of time reading the file of patients to arrive at a diagnosis, and may therefore choose to adhere to a more generic treatment approach.

1.2.4.2 Case 2: In-patient hospital stay and treatment.

Next, we consider the case of a patient admitted to the ICU, which are specialized critical care units meant for patients suffering from life-threatening diseases. When a patient is admitted to the ICU, critical care personnel are trained to perform standard routines to assess their vital signs and signs of deterioration. For instance, periodic check-up by nurses, periodic visit by the on-duty ICU doctors and also the visit by the doctor assigned to the patient are the norm in the ICU. In *Hospital A*, the nurses are required to note down readings in a paper-based format provided by the hospital, which will be frequently monitored by the on-duty doctors during their periodic visits (or during emergency visits, if any abnormalities are reported by the nurses). The duty roster of the critical care personnel keeps changing over the day and each of them have only this continually updated sheet to refer to for taking care of the patient during their duty time. In case of emergencies, the senior doctor who is in charge of the patient will be notified by the on-duty doctors regarding the patient's conditions. The diagnosis of patients may change over time depending on the variation in the patient's monitored signs.

Now, in case of *Hospital B*, the ICU patients' vital signs are still governed periodically, however, as per standard policy set up, the critical care personnel have to enter these observations directly into the EHR system as readings. The CDSS built on a robust IT infrastructure is designed to consume this streaming data for generating actionable insights for the critical care personnel, and also alert on-duty doctors and the doctor-in-charge when an abnormality is observed in the patient's condition. The CDSS may also provide them with diagnostic assistance providing statistics (mostly probabilistic) on prospective diagnosis of the patient based on the variations recorded in the EHR of the patient. In case of *Hospital C*, the system may be equipped to take such reading partially, while for the most part, paper-based readings may be recorded, hence CDSSs built on this data, may have very rudimentary capabilities. Hence, it is a critical requirement to build CDSSs that are able to process text-based records directly to enable prediction capabilities.

A detailed analysis from the stakeholders' perspective in the above scenario, in the context of *Hospital A* reveals the considerable effort required on the part of critical care personnel, in order to discharge their duties of effectively monitoring an ICU patient. It is also to be noted that the doctor-in-charge, who visits the patient mostly once a day has to go through a lot of paperwork to understand the patient's condition. Moreover, during the other times, he/she is unaware of his/her patient's condition unless and until an on-duty doctor contacts him/her with a situation. In contrast to this, the computerized *Hospital B* and *Hospital C* will be able to provide information and even alerts to the on-duty doctors as well as the doctor-in-charge as soon as the nurses feed in the readings/notes into the system. In this way, the ICU patient is assured of prompt and timely intervention in case of emergencies in addition to continuous observation by not only the nurses, but the doctors as well.

1.2.4.3 Case 3: Medical Records Management.

Let us now consider the case of a significant stakeholder in the day-to-day operations of a hospital, that of the overworked and often under-staffed Medical Records Departments (MRDs). It is easy to imagine a scenario with reference to the MRD of *Hospital A*, that of a room full of file cabinets storing patient records and other hospital records in paper format. The responsibility of the MRD staff is mainly limited to search and retrieve patient records whenever required, and refile them once the patient is treated/discharged. The MRD staff will have huge responsibilities of securing the paper based records and also will have to work relentlessly fetching and replacing medical records of patients everyday, in order to ensure a suitable categorization and cataloging system.

The MRD of *Hospital B* uses minimal paper records and the main job of the MRD staff is to deal with the extensive data entry required to continually update the patient records (whenever the healthcare personnel themselves are not doing this). The medical personnel such as doctors and nurses also are responsible to record event-generated patient data using suitable interfaces to the HIMS, either as readings or as notes. The MRD staff are trained to further process and code the data into defined formats, and generate the EHR record as required for CDSS consumption. This task is to be performed by highly trained medical coders, with domain expertise and are equipped to handle challenges like extensive use of medical jargon, acronyms, shorthand, notations etc. It is also a laborious process, that is cost-prohibitive, time-consuming, & error-prone and has been reported to be a significant burden on hospitals. It has been reported that the additional costs incurred due to inaccurate coding and the financial outlay towards improving diagnostic coding efficacy is estimated to be approximately $25 billion per year, in the United States alone (Lang, 2007; Farkas and Szarvas, 2008).

The MRD of *Hospital C* may have both paper-based records and EHR based formats and therefore, the main jobs of MRD staff will be to manage records and also data entry. Their data entry jobs include only those cases, where the EHR system has to be updated from old records or if notes for a patient are presented to them in paper form. Further, if the CDSS system for the hospital is designed to be able to directly process text-based clinical data, then, the MRD staff are relieved of the responsibility of performing the tedious task of manually coding medical records into formats mandated by traditional CDSS built on structured EHR data. Instead their effort can be focused on digitizing older records (if any) for use in CDSS and for improving the implementation of the EHR system.

The scenarios discussed in the context of the MRD and its staff in each hospital not only highlights the importance and need for CDSSs, but also stresses that CDSS systems can be most helpful when patient data is easily processable manually, if not automated processing mechanisms are incorporated. The need for well-designed HIMS equipped with effective CDSSs that can provide decision-making assistance to medical personnel based on structured, unstructured and semi-structured medical data is critical. In case of *Hospital A*, there is a critical requirement of an effective HIMS as the hospital is not in a state to provide any decision support to medical personnel due to its paper-based records management. Several technical challenges exist, including the steep costs of implementing a full-fledged HIMS equipped with CDSS capability for the hospital.

Hospital B and *Hospital C* are far ahead in terms of technology implementation, however require effective CDSSs implementations to ensure their smooth functioning. As CDSS systems in *Hospital B* can be built on structured EHR data, due to their complete adoption of EHR systems, the implementation can be quite straightforward. However, the significant costs and manual effort required for generating structured data from unstructured clinical data can be a significant burden, compounded by the scarcity of trained medical coders. In *Hospital C*, some limited capability to support CDSSs is available, however, a significant volume of their patient data is unstructured. Hence, specialized CDSSs that can directly consume semi-structured and unstructured raw clinical data for facilitating actionable insights are seen as a crucial solution for *Hospital C*. However, there are several challenges in developing such CDSSs, which are discussed in detail in the next section.

1.3 Prevalent Challenges

With the application of computation techniques like Machine Learning and Big Data Analytics, Healthcare Information Technology has evolved towards adaptable CDSSs. Despite the vast strides in research in the area of CDSSs over the past 30 areas, numerous challenges still exist (Miotto *et al.*, 2016). We discuss some of these challenges, that are sees as significant roadblocks that need to be overcome for designing next generation CDSSs.

1. **Data Volume & Variety:** In the field of biomedicine and healthcare, large volumes of data gets generated in various modalities, such as – text, images, signals, etc. In addition to this volume, the heterogeneous nature of the health data is an additional aspect that compounds the issues. Moreover, understanding diseases and their variability is another complicating aspect that needs to be tackled. However, the availability of clinical/health data in huge volumes can be often seen as an advantage from a data science perspective as artificial intelligence and machine learning based systems as built on the thumb rule “more the data, the better”.

2. **Data Quality:** In addition to being voluminous and heterogeneous, healthcare/biomedical data can be highly noisy, incomplete, ambiguous and unstructured. Moreover, data can be sparse or redundant too. This means that extensive preprocessing mechanisms need to be incorporated to prepare and structure the data into a representation that can be utilized effectively for predictive analytics based applications and CDSSs.

3. **Temporality:** The temporal nature of clinical data might be one of the hardest challenges as machine learning based models are not very well-versed to deal with the time aspect. Disease progression and changes over time

have always been non-deterministic and hence, it is critical that the data be modeled and analyzed with appropriate techniques that can capture and analyze the temporal nature of the data.

4. **Domain Complexity:** With respect to other domains, the problems in healthcare and biomedicine can be much more complicated and critical. The disease progression, their causes and changes are extremely difficult to analyze and the domain knowledge pertaining to the same are limited. Moreover, for certain disease cases, the number of patients can be extremely rare, resulting in unbalanced data, a classic case where most machine learning models fail to perform well. This makes it even more difficult to perform analytics on the data, thus adding to the complexity for developing predictive analytics based CDSSs.

5. **Interpretability:** 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 performs well. Additionally, this interpretability is also crucial in making the medical personnel understand about the outputs of the predictive analytics based CDSSs.

1.4 Summary

In this chapter, issues and challenges pertaining to the healthcare domain and improving healthcare delivery were discussed. The need and motivation for CDSSs and the support offered by the availability of standardized EHRs in the current healthcare scenario were presented. The need for CDSSs has become a matter of critical importance, to improve healthcare delivery and the advantages of the same were highlighted with help of example scenarios from healthcare practice. Enumerating the prevalent challenges affecting the development of intelligent, non-knowledge based CDSSs, a significant scope for designing predictive analytics and machine learning based CDSSs is observed, for overcoming the challenges associated with the voluminosity & variety of the healthcare data, and to create value. This value can be then used for generating actionable insights for healthcare personnel, ultimately improving the healthcare delivery process for patients and all other stakeholders.

1.5 Thesis Organization

The rest of this thesis is organized as follows.

- In chapter 2, an extensive literature review on CDSSs 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 is formally defined. The scope of this research and a brief description of the proposed methodologies are also provided in Chapter 3.
- In Chapter 4, proposed approaches for predictive analytics based patient-specific CDSSs using structured patient data are discussed.
- Chapters 5 and 6 cover the details of the proposed approaches for predictive analytics based patient-specific CDSSs using unstructured clinical notes.
- In chapter 7, aggregation based unstructured clinical text modeling strategies to build patient-centric CDSSs are presented in detail.
- Chapter 8 discusses the proposed approach for population based predictive analytics using social network data, is presented.
- Chapter 9 presents concluding remarks about the extensive research work carried out and prospects of future research in the area.

Chapter 2

Literature Review

2.1 Background

Clinical health analytics and informatics research is an emerging field in the development of intelligent systems for the medical field, for realizing personalised healthcare and improved understanding into disease dynamics. The advent of Electronic Health Records (EHRs) have paved the way to development of intelligent decision-making systems and have made a significant impact in the domain of healthcare analytics and informatics, due to the availability of health data in a more usable structured format (Coorevits *et al.*, 2013). This forms an area of active research interest with real-world implications in the form of CDSSs for diagnosis prediction (Choi *et al.*, 2016; Miotto *et al.*, 2016), ICU mortality prediction (Silva *et al.*, 2012) and patient risk prediction (Cheng *et al.*, 2016). Other applications include ICU patient care recommendation (Saeed *et al.*, 2011), efficient hospital management (Lovis, 2011), data quality measurement of EHR (Weiskopf and Weng, 2013), patient feature representations derived from EHRs (Miotto *et al.*, 2016; Choi *et al.*, 2016), automatic concept extraction (Xu *et al.*, 2010), and many more. EHRs also store temporal data which may be exceptionally helpful for time-oriented representations and predictions of future events, diseases, etc (Wu *et al.*, 2010).

In view of these foreseeable benefits, healthcare analytics has generated much interest among researchers and the healthcare community alike. However, the critical challenges to be overcome include dealing with several issues ranging from extensive availability of well-defined EHR data to generating actionable insights from it, when available. In this section, we present a comprehensive review on the wide-ranging spectrum of healthcare analytics applications leading to the development of intelligent decision-making systems.

2.2 Related Work

Existing research in the domain of Healthcare Analytics and Clinical Decision Support Systems (CDSSs) can be broadly classified into four main categories based on the methodologies employed by these systems. These are listed below. In the subsequent sections, we discuss each of these categories and the ongoing research in these areas in detail.

1. Information Retrieval (IR) based Systems
2. Natural Language Processing (NLP) based Systems
3. Data Mining and Learning based Systems
4. Population based Healthcare Systems

2.2.1 Information Retrieval based Systems

One of the first problems addressed by early works in clinical data management, was to categorize and retrieve medical documents based on patients or conditions. Later, the challenge was designing effective ranking for the retrieved documents, improving keyword based searches to retrieve medical data from the web, to support context-sensitive querying etc. Such methods used techniques like document similarity, knowledge bases, ontologies and semantic web concepts in the latest works.

In 1996, a system named *Medical World Search*, was developed and made operational by Suarez *et al.* (1997). It was a medical search engine that performed information retrieval based on the knowledge base Unified Medical Language System (UMLS) (Lindberg *et al.*, 1993). It accepted queries from users and returned ranked medical documents from the internet based on relevance of the query with respect to the documents retrieved (Suarez *et al.*, 1997). However, the method lacked a strong term or concept matching algorithm with the UMLS knowledge base and moreover, the work was performed on a very small database.

Malet *et al.* (1999) proposed an approach to enhance dynamic and online retrieval of medical documents. The approach used the MeSH vocabulary (Medical Sub Headings) (Lowe and Barnett, 1994) standardized by the US National Library of Medicine (NLM) and also MEDLINE (Greenhalgh, 1997) type descriptions for the purpose of referencing during retrieval. They also extended Dublin Core Metadata (Weibel *et al.*, 1998) to form the Medical Core Metadata (Malet *et al.*, 1999). These together enable the medically represented documents (using

Clinical Decision Support Systems

- **IR based**
  - Ranked Retrieval (Suarez *et al.* (1997), Malet *et al.* (1999), Brown and Sönksen (2000))
  - Query Enhancement (Göbel *et al.* (2001), Leroy and Chen (2001) Bayegan (2002), Jain and Huimin Zhao (2005))
  - Ontologies Used (UMLS Lindberg *et al.* (1993), MeSH Lowe and Barnett (1994))
- **NLP based**
  - Reminder/Alerts (McDonald *et al.* (1999), Evans *et al.* (1998))
  - Decision Support (Fiszman *et al.* (2000), Haug *et al.* (2007))
  - Knowledge Base Related (Baud *et al.* (2001), Fontelo *et al.* (2005))
- **Data Mining and Learning based**
  - Mortality/Severity Prediction
    - Parametric/Scoring based (Knaus *et al.* (1981, 1985, 1991), Zimmerman *et al.* (2006) Gall *et al.* (1984, 1993), Moreno *et al.* (2005) Johnson *et al.* (2013)
    - Non Parametric based (Kim *et al.* (2011), Pirracchio *et al.* (2015) Calvert *et al.* (2016*a*), Harutyunyan *et al.* (2017), Che *et al.* (2018))
  - Disease Prediction
    - Disease Specific (Himes *et al.* (2009), Michelson *et al.* (2014), Lipton *et al.* (2015))
    - Generic Disease
      - Disease Group/Conditions (Zhang *et al.* (2012), Miotto *et al.* (2016) Choi *et al.* (2016), Cheng *et al.* (2016), Nguyen *et al.* (2017) Che *et al.* (2018), Purushotham *et al.* (2018))
      - Disease Code (Perotte *et al.* (2011), Ferrao *et al.* (2013) Dermouche *et al.* (2016), Wang *et al.* (2017), Baumel *et al.* (2018) Li *et al.* (2018), Mullenbach *et al.* (2018), Xie and Xing (2018) Huang *et al.* (2019), Zeng *et al.* (2019))
  - Others
    - Length of Stay (Gentimis *et al.* (2017), Zebin *et al.* (2019), Li *et al.* (2019))
    - Readmission Prediction (Campbell *et al.* (2008), Fialho *et al.* (2012))
- **Population based Healthcare Systems**
  - Epidemic Prediction (Ginsberg *et al.* (2009), Signorini *et al.* (2011) Aramaki *et al.* (2011), Achrekar *et al.* (2011) Yuan *et al.* (2013), Santillana *et al.* (2015))
  - Adverse Drug Reactions/Events (Nikfarjam *et al.* (2015) Sarker *et al.* (2015), Cocos *et al.* (2017))
  - Vaccine Sentiment (Huang *et al.* (2017), Joshi *et al.* (2018))
  - Depression Detection (McManus *et al.* (2015), Shen *et al.* (2017), Orabi *et al.* (2018))

Figure 2.1: Categorization of Clinical Decision Support Systems

Medical Core Metadata) to serve as a base which is then used to support enhanced medical information retrieval using keyword search and web crawlers. Brown and Sönksen (2000) presented a method to evaluate Information Retrieval performance in a computerized patient database using a semantic terminological model built within Clinical Terms Version 3. However, the model remained incomplete and worked well only for a small domain.

Several researchers focused on incorporating standard vocabularies and taxonomies to boost document-matching performance, given the diversity in health data terminologies and parlance. Göbel *et al.* (2001) designed an intermediary model for consumer health information systems that performed query enhancement on queries input by the users (or patients). The system used the controlled vocabulary of MeSH for the purpose of query enhancement (Jain and Huimin Zhao, 2005; Göbel *et al.*, 2001). A similar approach was put forth by Leroy and Chen (2001), in the form of a tool named Medical Concept Mapper, which eased access to medical sources online by recommending users with appropriate search terms for their input medical queries, using the UMLS vocabulary. Both these systems were meant for those customers or patients who were not familiar with medical terms. Although the above methods provided satisfactory results, they lacked the ability to consider patient data and recommend queries accordingly.

In an evaluation study, Plovnick and Zeng (2004) reported that, query enhancement or reformulation using the vocabulary of UMLS improved the precision of clinical information retrieval. Liu and Chu (2005) developed a UMLS based system that performed scenario-specific retrieval of medical free text, by expanding queries. They also proved with results that domain specific retrieval performs better than generic retrieval without any domain specification (Plovnick and Zeng, 2004). Although it accomplishes its objectives, it would have been better if this was based on patient records or information as well. This was taken care of by Bayegan and Tu (2002) and Bayegan (2002) proposed a knowledge based medical record system that could perform problem oriented view of patients data. The system was meant to be helpful for physicians and ranked the relevance of patient information in a particular context using the physicians' work processes as knowledge base. Jain and Huimin Zhao (2005) proposed a method to semantically retrieve medical records related to patient symptoms. It was implemented using various techniques in information retrieval, domain ontologies and a body of domain knowledge created by healthcare experts. The major drawback was that the same domain knowledge base cannot be used as a generic one as it was manually created by healthcare experts specifically for this objective.

Table 2.1: Summary of Information Retrieval based CDSSs

Work	**Concept/Method**	**Explanation/Remarks**
Suarez *et al.* (1997)	UMLS based	Medical web search engine with ranking
Malet *et al.* (1999)	MeSH vocabulary & MEDLINE type definitions	Enhance the online retrieval of medical documents
Brown and Sönksen (2000)	Semantic terminological model built within CTV 3	Evaluate IR performance in a computerised patient database
Göbel *et al.* (2001)	Query enhancement using MeSH vocabulary	Intermediary model for consumer health information systems
Leroy and Chen (2001)	UMLS based Medical Concept Mapper	Recommending users with search terms for their input medical queries
Bayegan (2002)	Knowledge based medical record system using multiple ontologies	For physicians to rank patient info
Liu and Chu (2005)	UMLS based scenario-specific retrieval of medical free text using VSM	Query expansion and retrieval using similarity measure
Jain and Huimin Zhao (2005)	Information retrieval, domain ontologies and KB by healthcare experts	Semantically retrieve medical records related to patient symptoms

2.2.2 Natural Language Processing based Systems

Generating actionable insights from clinical data is a complex and challenging task due to the huge volume and variety of natural language textual data generated, such as, narrations or notes by doctors and nurses, discharge summaries, prescriptions given to patients, etc. Clinical Data Analytics is a field that leverages Natural Language Processing (NLP) techniques for processing unstructured free text and for extracting latent knowledge like symptoms, diseases, patient history, etc. Several existing works used basic NLP techniques like tokenization, stopping, stemming and n-gram extraction, which are performed on the text data and then used in association with systems like UMLS, MESH or ontologies like SNOMED-CT (Donnelly, 2006) or machine/deep learning techniques for the development of intelligent predictive analytics based healthcare applications.

In early 1972, McDonald *et al.* (1999) of the Regenstrief Institute at Indianapolis, USA, developed the Regenstrief Medical Record System (RMRS) – a system that provided a protocol-driven reminder service for physicians which reminded them to perform certain clinical tests on their patients (Demner-Fushman *et al.*, 2009; Mamlin *et al.*, 2007). Antibiotic Assistant was a system developed by Evans *et al.* (1998) at LDS Hospital, Utah, USA, which could identify patients with potential infections and then alert the physicians about anti-infection therapy, with a suggestion regarding the dosage of antibiotic for the same. Although physicians could save time using this, the system could never measure the quality of the anti-infection treatment or even more importantly, failed to detect drug events and reactions. Fiszman *et al.* (2000) developed an enhanced system that used the Antibiotic Assistant but could detect acute bacterial pneumonia from chest X-ray reports using NLP techniques. They proved that applying NLP techniques is more suitable for prediction than simple keyword based techniques. Although its accuracy was satisfactory, the time taken for prediction was quite long, thus failing to generate fast actionable insights for early decision making.

Baud *et al.* (2001) proposed a light knowledge model for medical texts. The medical documents or records were tokenized, disambiguated, parsed, performed semantic tagging of words and finally, syntax-driven modeling in which semantic relationships between two words are tagged, thus resulting in a light knowledge model for medical texts. Fontelo *et al.* (2005) developed a system named askMEDLINE, a free-text, natural language search tool for the medical literature system MEDLINE[1]/PubMed[2] without using any domain specific vocabularies. HELP (Health Evaluation through Logical Processing) (Haug *et al.*, 2007) was the first hospital information system to integrate clinical data accumulation and clinical decision support. The HELP system provided diagnostic decision support to help diagnose Adverse Drug Events (ADE). It was basically a rule based system that generated predictions based on chemical tests, drug tests, drug prescriptions and orders, symptoms, etc. available in patients' EHRs. The system would have per-

[1] https://www.nlm.nih.gov/bsd/medline.html

[2] https://www.ncbi.nlm.nih.gov/pubmed/

formed better if better preprocessing and data mining techniques were applied on the hospital EHR data.

Table 2.2: Summary of Natural Language Processing based Systems

Work	**Concept/Method**	**Explanation/Remarks**
McDonald *et al.* (1999)	NLP and rule-based	Protocol-driven reminder system for physicians to perform timely clinical tests on patients
Evans *et al.* (1998)	NLP on structured medical text records & MEDLINE type definitions	Antibiotic Assistant for identifying patients with potential infections and alert physicians for anti-infection therapy
Fiszman *et al.* (2000)	NLP grammar and keywords on chest X-ray reports	System able to detect acute bacterial pneumonia
Baud *et al.* (2001)	NLP and term relationship semantics	A light knowledge model for medical texts
Fontelo *et al.* (2005)	NLP and keyword search	askMEDLINE: NL search tool for MEDLINE/PubMed
Haug *et al.* (2007)	HELP, NLP, Rule-based (using EHRs)	Diagnostic decision support to diagnose Adverse Drug Events

2.2.3 Data Mining and Learning based Systems

The amount of data generated in the field of healthcare is very high in volume and also is highly complex to be processed and analyzed by traditional methods (Koh *et al.*, 2011). Over the years, Data mining techniques like pattern analysis, clustering etc., have been used by many researchers to transform these huge amounts of data into useful information for decision making and predictive modeling. With the massive explosion of multimodal data availability in the field of Healthcare, the role of data analytics in making meaningful use of it has grown more vital (Ravì *et al.*, 2017). This has resulted in a new generation of analytical, data-driven and predictive models based on Artificial Intelligence through Machine Learning and Deep Learning techniques. As this is the primary area of focus and is an extensive area of research, we have classified the works further based on the common prediction tasks performed by these CDSSs.

2.2.3.1 Mortality/Severity Prediction

In critical care applications, the process of taking practical decisions on managing the care of intensive care patients can help augment the efficiency of caregivers, through the use of predictive data analysis on the large amounts of data generated while monitoring these patients. The most important aspect of a CDSS in the ICU is undoubtedly, its ability to accurately predict in advance the mortality or severity risk of a patient, so that doctors and other healthcare personnel can be prepared to intervene in time, with the resources available in ICU. Apart from measuring the severity of illness, mortality prediction can also play a crucial role in the assessment of treatment and critical care policies of a hospital. Hence, ICU mortality prediction has remained a well-researched problem over the years.

Before delving into the ML based mortality prediction CDSSs, we discuss a few important traditional parametric severity scoring based Mortality Prediction Models (MPMs). Parametric scoring based MPMs typically use the perceived relevance of the clinical measurements of an ICU patient, to calculate a score in a particular range, as per a model derived by clinical experts. Knaus *et al.* (1981) proposed a physiological scoring system APACHE (Acute Physiology And Chronic Health Evaluation) that measures severity of illness using 34 physiological variables in patients from critical conditions in ICUs. Gall *et al.* (1984) proposed another scoring system Simplified Acute Physiological Score (SAPS), which used measurements of 14 physiological variables to group patients in various probabilities of death risk. SAPS claimed to be simpler and less time-consuming that APACHE in calculation of score. Knaus *et al.* (1985) put forth a revision of APACHE, APACHE-II, that used 12 physiological variables and two additional diagnosis related variables to calculate severity of illness and risk of death in ICU patients.

To improve the accuracy of determining mortality risk, APACHE-III was introduced later by Knaus *et al.* (1991), which required measurement of eight additional physiological variables along with those used in APACHE-II. The next version of SAPS, SAPS-II which was proposed by Gall *et al.* (1993), was reported to have performed better in comparison to APACHE-II. It included easy-to-measure 12

Table 2.3: Summary of Traditional Mortality/Severity Scores

Work	**Concept/Method**	**Explanation/Remarks**
APACHE (Knaus *et al.*, 1981, 1985, 1991; Zimmerman *et al.*, 2006)	Parametric method based on clinical variables	Acute Physiology And Chronic Health Evaluation
SAPS (Gall *et al.*, 1984, 1993; Moreno *et al.*, 2005)	Parametric method based on clinical variables	Simplified Acute Physiological Score
SOFA Vincent *et al.* (1996)	Parametric method based on clinical variables	Sepsis-related Organ Failure Assessment
OASIS (Johnson *et al.*, 2013)	Parametric method based on clinical variables	OASIS (Oxford Acute Severity of Illness Score)

physiological variables to calculate severity of illness and mortality risk of ICU patients. Another score, SOFA (Sequential Organ Failure Assessment) (Vincent *et al.*, 1996), determines severity and mortality risk of ICU patients and was considered advantageous for its simplicity of measuring only six variables (related to Respiration, Central Nervous System, Cardiovascular, Renal, Coagulation and Comorbidity) to calculate the score. SAPS-III, introduced by Moreno *et al.* (2005), was meant to supplement the SAPS-II scoring system for better determining mortality and severity risk in ICU patients. The fourth version of APACHE, APACHE-IV Zimmerman *et al.* (2006) used multivariate logistic regression for measuring severity of illness and mortality risk in ICU patients. OASIS (Oxford Acute Severity of Illness Score), a recent scoring system proposed by Johnson *et al.* (2013), uses a subset of APACHE-IV variables along with others like age, length-of-stay and elective surgery prior to ICU admission, to predict mortality of ICU patients. According to the authors, its performance is at par with that of APACHE-IV and is considered superior to APACHE-IV as it requires lesser features. Despite this, validation studies carried out by various researchers have shown that these scores can be further fine-tuned for better performance (Pirracchio *et al.*, 2015; Awad *et al.*, 2017). Any such fine-tuning can help in reducing the time taken in collecting patient data, thus enabling earlier predictions with better accuracy than that achieved by traditional scoring systems.

In recent years, researchers have focused on designing non-parametric CDSSs built using data mining and machine learning techniques to enable higher accuracy for applications like ICU mortality prediction, length-of-stay prediction, readmission prediction etc. Dybowski *et al.* (1996) proposed the use of an Artificial Neural Network (ANN) optimized by Genetic Algorithm (GA) to effectively predict mortality risk in ICU patients. They compared the performances of their approach against Logistic Regression based MPM. Wong and Young (1998) compared the performances of an APACHE-II model and an ANN based mortality prediction model for ICU patients and reported that ANN based MPM outperformed the APACHE-II score, underscoring the significant potential of learnability in effective prediction of mortality risk. Clermont *et al.* (2001) reported similar performances of Logistic Regression based MPMs and ANN based MPMs with adequate samples of data.

Nimgaonkar *et al.* (2004) compared the performance of APACHE-II based MPMs and ANN based MPMs on Indian ICU data and reported that ANN based MPMs outperformed APACHE-II based MPMs. Kim *et al.* (2011) compared the predictive accuracy of MPMs based on ANN, Support Vector Machine (SVM) and decision trees with APACHE-III scoring system based MPM using ICU patient data collected at University of Kentucky Hospital, where, the C5.0 decision tree algorithm outperformed APACHE-III and ANN based MPMs. Celi *et al.* (2012) developed customized ML based MPMs for various categories of ICU patients, like, patients suffering from acute kidney injury (for which ANN performed better than SAPS), patients suffering from subarchanoid hemorrhage (for which Bayesian Networks outperformed SAPS) and elderly patients who had undergone open heart surgery (for which ANN performed better compared to SAPS). Their work underscores the very generic nature of standard and traditional severity scoring systems due to which they often fail to predict well for locally customized models (Awad *et al.*, 2017).

Pirracchio *et al.* (2015) developed the Super ICU Learner Algorithm (SICULA), a MPM which uses a ML cascade trained on the MIMIC-II critical care database, and its performance significantly outperformed that of SAPS-II and SOFA based MPMs. Calvert *et al.* (2016*a*) proposed a MPM that uses Logistic Regression and binning to predict mortality risk using 12 hours of patient data after admission into the ICU, outperformed performances of SAPS-II and SOFA based MPMs. Calvert *et al.* (2016*b*) showed the applicability of similar methodology in developing a MPM specifically for critically ill patients suffering from alcohol disorder patients. Harutyunyan *et al.* (2017) proposed and benchmarked multitask Long Short Term Memory (LSTM) neural network based learning models that perform prediction tasks to enable better clinical decision support, out of which one task is in-hospital mortality. Che *et al.* (2018) proposed and benchmarked a modified Gated Recurrent Unit (GRU) neural network based CDSS models including a MPM that included novel preprocessing strategies for missing values in patient data. Purushotham *et al.* (2018) benchmarked performances of deep learning models and proposed a Multi-modal Deep Learning (MMDL) based prediction model, an ensemble of Feed Forward Neural Network (FFNN) and LSTM, for various tasks, one of them being mortality prediction of ICU patients.

Table 2.4: Summary of Works on ML based Mortality/Severity Prediction Models

Work	**Concept/Method**	**Explanation/Remarks**
Dybowski *et al.* (1996)	ANN with GA optimization	Compared with Logistic Regression based MPM
Wong and Young (1998)	ANN based MPM	Comparison with APACHE-II
Clermont *et al.* (2001)	Logistic Regression and ANN based MPMs	Compared performances of LR and ANN based MPMs
Kim *et al.* (2011)	Various ML classifiers - comparison study	ICU Mortality prediction using private hospital dataset
Celi *et al.* (2012)	ANN and Bayesian Network based MPMs	customized MPMs for various conditions
Pirracchio *et al.* (2015)	Logistic Regression	ICU mortality prediction using MIMIC-II dataset
Grnarova *et al.* (2016)	Neural document embeddings	ICU Mortality prediction
Calvert *et al.* (2016*a*)	Logistic Regression with binning feature modeling	ICU mortality prediction using MIMIC-III dataset
Calvert *et al.* (2016*b*)	Logistic Regression with binning feature modeling	Mortality prediction in alcoholic patients
Harutyunyan *et al.* (2017)	Time series analysis	Benchmarking CDSS predictions - ICU Mortality, readmission, length of stay
Che *et al.* (2018)	Recurrent Neural Networks & multivariate time series	Mortality prediction with missing values handling

2.2.3.2 Disease Prediction

Disease prediction models can be categorized into three types – (i) disease specific models for predicting risks of developing a particular disease (such as diabetes prediction, heart disease prediction, etc.), (ii) Generic disease categories and group prediction models and (iii) Generic disease coding models. We discuss various works in each of these categories in the next section.

Disease-specific Prediction Models. Himes *et al.* (2009) proposed a prediction model to predict Chronic Obstructive Pulmonary Disease (COPD) in asthma patients using data extracted from EMRs. Their system used a Bayesian Network classifier (Friedman *et al.*, 1997) and K2 algorithm (Cooper and Herskovits, 1992) for the purpose of prediction. However, their model lacked good preprocessing and feature modeling techniques due to which the Bayesian Classifier did not have enough or quality features to learn. Lipton *et al.* (2015) demonstrated the effectiveness of gradient based learning algorithms like Recurrent Neural Networks (RNN) (Williams and Zipser, 1989) in making diagnosis using EHRs. The system is intended for multi-label classification for patient phenotyping and predicting diagnoses. While the proposed system performed well, the evauation experiments were performed on a small set of data from a hospital, which was found to be inadequate for training a deep learning model and for optimal learning. Moreover, it lacked effective preprocessing and patient representations, which further degraded the system's performance. Michelson *et al.* (2014) proposed a system that can detect surgical site infections, using text mining and NLP for the purpose of retrieval and extraction of information from structured, semi structured and unstructured medical text. The system lacked good preprocessing techniques, especially in the context of unstructured clinical text, and thus was not very effective.

Generic Disease Prediction Models. This category includes models built on patient data for predicting the existence of specific diseases such as heart diseases, lung diseases, sepsis, kidney diseases, etc. Zhang *et al.* (2012) developed a system that used historical datasets along with streaming data for predicting diagnosis in real-time, based on the concept of data stream mining (Domingos and Hulten, 2000) and classification using Very Fast Decision Trees (Li *et al.*, 2009). Miotto *et al.* (2016) designed a system named Deep Patient, a model for disease prediction for clinical decision making. It used an unsupervised feature learning method over EHRs using autoencoders to generate effective patient representations. They used a random forest classifier (Rodriguez *et al.*, 2006) to train and predict the probability of occurrence of all the diseases in the vocabulary. Even though the system achieves pretty good results, the supervised random forest classifier at the end layer could have been another deep learning architecture which would have refined the results to a better extent. The system proposed by Choi *et al.* (2016), named Doctor AI, is a generic predictive model that considers medical conditions that were previously observed in patients. Doctor AI uses a temporal model implemented using RNNs which was then applied over longitudinal timestamped EHRs to make multi-label predictions (for diagnosis, medication and time of patient's next visit). The advantages of a derived patient representation and unsupervised multi-label classification are significant. However, the generated patient representations need improvement, as very rudimentary feature engineering has been utilized. Moreover, both DeepPatient and DoctorAI are designed to consume structured EHR data and are custom made for their own hospital EHRs, thus have not been benchmarked on standard open datasets. Also, they do not consider the case of unstructured text data, which forms a substantial percentage of clinical data volume generated by hospitals.

Deepr is a system proposed by Nguyen *et al.* (2017) which uses a deep learning approach that performs risk prediction for patients. It is an end-to-end system that takes EHRs as input, learns to extract features from them and predicts risk of patients in an automated way. It uses a Convolution Neural Network (CNN) (LeCun *et al.*, 1998) for understanding clinical patterns to predict risk of patients. Another system put forward by Cheng *et al.* (2016) is also a deep learning approach for risk prediction of patients, that uses a four layer CNN model – EHR Matrices, a phenotyping layer, a max pooling layer to remove insignificant features and finally, a softmax classification layer. The authors stress that feature engineering is a major bottleneck in predictive systems based on EHRs, and deep learning models which achieve better feature learning are critical for optimal performance (Nguyen *et al.*, 2017). Moreover, these prediction models could have performed with a better performance with the use of derived patient representations from EHRs.

Che *et al.* (2018) proposed and benchmarked GRU based models for mortality and disease prediction of ICU patients. Purushotham *et al.* (2018) benchmarked performances of deep learning models and proposed a Multi-modal Deep Learning (MMDL) based prediction model, an ensemble of FFNN and LSTM, for various tasks. Both these works (Che *et al.*, 2018; Purushotham *et al.*, 2018) proposed models that perform ICD9 disease group prediction, utilizing structured patient data and deep learning architectures. While both report promising results, they failed to effectively generate and utilize derived patient data representations for the prediction modeling, which is a significant limitation.

Generic Disease Code Prediction Models. In hospitals, the ICD9[3] coding taxonomy is widely employed to describe patient conditions and the associated diagnoses. ICD9 is essentially a hierarchical classification, with unique codes for patient conditions, diseases, infections, symptoms, causes of injury, and others, maintained by the World Health Organization. These unique diagnostic codes are assigned to patient records to facilitate the clinical and financial decisions made by the hospital management for billing, insurance claims, and reimbursements (Jensen *et al.*, 2012; Li *et al.*, 2018). Based on clinicians' free-text notes and other patient records such as discharge summaries, trained professional medical coders employed by the Hospitals' MRD transcribe the patient records into a set of appropriate medical diagnostic codes (from a potentially large set of codes – around 13,000 codes in case of ICD9). These medical coders utilize their expertise in the field of medicine, coding rules, and terminologies, to facilitate the note-to-code mapping, thereby making such manual coding process expensive, inexact, time consuming, and error-prone (Chen *et al.*, 2017; Zeng *et al.*, 2019). It is riveting to note that the additional cost incurred as a result of incorrect coding and the financial investment spent in improving the efficacy of coding, estimates to approximately $25 billion per year (in the United States alone) (Lang, 2007; Farkas and Szarvas, 2008). This emphasizes the crucial need for an automated computational system for ICD9 coding of patient records.

ICD9 coding can be considered as a multi-label classification task (binary classification of multiple labels), where each label pertains to a patient diagnosis or condition and each label has a binary value of zero or one, indicating absence or presence of the condition. Using the concept of multi-label classification, several Machine Learning (ML) approaches for automated ICD9 coding have been proposed over the years. Pakhomov *et al.* (2006) proposed the use of a Bag-of-Words model and a Naive Bayes classifier for automated assignment of diagnosis codes to patient encounters. Medori and Fairon (2010) proposed a Naive Bayes classifier based semi-automatic method to assign ICD9 codes for patient records and the paper also identified some important attributes such as diseases, symptoms, factors, procedures, etc. through the process of feature selection. Perotte *et al.* (2011) proposed a Hierarchically Supervised Latent Dirichlet Allocation (HSLDA)

[3] International Classification of Diseases, 9th revision. Online: `https://www.cdc.gov/nchs/icd/icd9.htm`

as an improved version of Latent Dirichlet Allocation (LDA) topic modeling technique. The authors validated the efficacy of the proposed topic modeling method on various tasks, one of them being ICD9 coding of free text clinical records. Ferrao *et al.* (2013) used SVM classifiers for assignment of ICD9 codes to admission episodes using structured patient data. Dermouche *et al.* (2016) used LDA for feature modeling and used SVM as a classifier for assigning ICD9 codes to patient records. Wang *et al.* (2017) proposed an automatic ICD9 coding model that uses an ensemble of linear classifiers with the help of disease correlation information at local level of patient data for improved prediction of disease codes.

Recent ML based approaches such as (Baumel *et al.*, 2018; Li *et al.*, 2018; Xie and Xing, 2018; Mullenbach *et al.*, 2018; Huang *et al.*, 2019; Zeng *et al.*, 2019) have been put forward to automate ICD9 coding using unstructured clinical text through machine learning. These used the standard and openly available MIMIC-III (Johnson *et al.*, 2016) dataset for benchmarking their works. Baumel *et al.* (2018) proposed a hierarchical attention model to identify relevant sentences for ICD9 labels and used Gated Recurrent Unit (GRU) for the ICD9 coding. Li *et al.* (2018) proposed an approach called 'DeepLabeler' to assign ICD9 codes to discharge summaries. They used Doc2Vec and Convolutional Neural Networks (CNNs) for ICD9 coding. Mullenbach *et al.* (2018) proposed Convolutional Attention for Multi-Label classification(CAML) approach for coding of patient records and benchmarked the results for 8921 unique ICD9 codes that includes 6,918 diagnosis codes and 2,003 procedure codes. Xie and Xing (2018) proposed an approach in which the 'Diagnosis Description' section of the discharge summary documents was taken as the input. They used LSTM to encode hierarchy of ICD9 codes and also used attention models to perform the coding of the discharge summaries. Huang *et al.* (2019) evaluate and benchmark the performance of state-of-the-art deep learning models like RNNs (LSTM and GRU), feed forward neural networks and CNNs for ICD9 coding of discharge summaries. They also compare the performances of deep learning models with that of machine learning classifiers such as Logistic Regression and Random Forest. Zeng *et al.* (2019) used the concept of transfer learning from the task of Medical SubHeadings (MeSH) indexing for ICD9 coding as both are binary classification of multiple labels. They used CNN for their prediction model and they compared their performances against SVM and flat-SVM models.

Table 2.5: Summary of Existing works in the area of Disease Prediction Systems

Work	**Concept/Method**	**Explanation/Remarks**
Pakhomov *et al.* (2006)	BoW model & Naive Bayes classifier	Automatic assignment of diagnosis codes to patient encounters
Medori and Fairon (2010)	Naive Bayes classifier	Identification of important attributes for semi-automatic method to assign ICD9 codes for patient records
Himes *et al.* (2009)	Bayesian Network classifier & K2 algorithm	Prediction of COPD in asthma patients
Perotte *et al.* (2011)	Hierarchically Supervised LDA	ICD9 coding of free text clinical records
Zhang *et al.* (2012)	Very Fast Decision Trees	Real time diagnosis prediction
Ferrao *et al.* (2013)	SVM	Assignment of ICD9 codes to admission episodes using structured patient data
Michelson *et al.* (2014)	Text Mining and NLP techniques	Surgical site infections detection using structured, semi structured and unstructured medical text
Lipton *et al.* (2015)	RNN	predicting diagnosis with the usage of EHRs
Miotto *et al.* (2016)	Autoencoders and Random Forest	predict the probability of occurrence of diseases using patient data warehouse
Choi *et al.* (2016)	RNN	Predict diagnosis using longitudinal timestamped EHRs
Cheng *et al.* (2016)	CNN	Predict risks of patient conditions
Nguyen *et al.* (2017)	CNN	Predict risks of patient conditions

Table 2.5: Summary of Existing works in the area of Disease Prediction Systems

Work	Concept/Method	Explanation/Remarks
Baumel *et al.* (2018)	GRU & attention model	Automated ICD9 coding
Li *et al.* (2018)	Doc2Vec and CNN	Automated ICD9 coding
Mullenbach *et al.* (2018)	CAML (CNN and attention model)	Automated ICD9 coding – diagnosis and procedures
Xie and Xing (2018)	LSTM & attention model using Diagnosis Descriptions	Automated ICD9 coding
Che *et al.* (2018)	GRU-D (GRU)	Disease group prediction
Purushotham *et al.* (2018)	MMDL (FFNN and GRU)	Disease group prediction with benchmarks performances of other models
Huang *et al.* (2019)	RNN, CNN, FFNN	Automated ICD9 coding with benchmarks performances of other models
Zeng *et al.* (2019)	Transfer Learning from MeSH indexing task	Automated ICD9 coding

2.2.3.3 Other CDSSs

Efforts for automatically predicting the outcomes of several other clinical tasks have been attempted over the past decade. Campbell *et al.* (2008) proposed a system that predicted probability of death/readmission of ICU patients using their discharge summaries and other patients' historical clinical records, built on multivariate logistic regression. Their system used a small collection of manually collected data, that failed to intuitively model disease severity and incremental risk of the patient. A similar system for ICU readmission prediction was put forward by Fialho *et al.* (2012), that used data mining and fuzzy modelling on MIMIC-II database (Saeed *et al.*, 2002) to make the predictions.

Bennett and Doub (2010) proposed a predictive and 'adaptive' model for CDS, that helps medical personnel select the optimal clinical treatment for a patient by modeling the outcomes of all treatment options, based on the patient's EHR data. The system was implemented by performing frequent pattern mining and predictive analysis on EHRs. Liang *et al.* (2014) proposed a system that applies Deep Belief Networks (DBN) (Hinton *et al.*, 2006) on EHRs for supporting clinical decision making. The system also used a Restricted Boltzmann Machine (RBM) (Salakhutdinov and Hinton, 2009) for implementing layer-wise training for the implemented DBN, and was tested for generic disease diagnosis assistance as well as hypertension prediction. The system had an advantage of designing an improved patient representation format by incorporating efficient preprocessing and unsupervised feature extraction techniques.

Gentimis *et al.* (2017) proposed an approach that used neural networks to predict the number of days a patient could remain admitted in the ICU (length of stay). This was a multilayer perceptron based classification approach that categorized the length of stay of ICU patients in the MIMIC-III dataset into *long* (more than 5 days) and *short* (5 or lesser days). Zebin *et al.* (2019) proposed an autoencoder based deep neural network approach to determine the length of stay of ICU patients and benchmarked their system on the MIMIC-III dataset. They too used a deep neural network to classify the length of stay into long (more than 7 days) and short (7 or lesser days). Li *et al.* (2019) proposed an approach based on exploratory data analysis and Least Absolute Shrinkage & Selection Operator (LASSO) regression technique to predict length of stay of ICU patients in the MIMIC-III dataset and expressed the results in terms of Root Mean Squared Error (RMSE). Table 2.6 presents a summary of CDSSs for varied clinical tasks.

2.2.4 Population Analytics based Healthcare Systems

Automating population health surveillance based on the confluence of technologies like Big Data Analytics, Data Mining and Machine Learning (ML) has emerged as a significant solution for extracting latent patterns and gaining potentially actionable insights to help govern the health of a population and also drive public health policies (Darcy *et al.*, 2016; Krumholz, 2014). Patient level data available in hospitals can provide government agencies with systematic data on instances of disease or virus outbreak, which can help put effective prevention and quarantine procedures in place, also provide long-term data for future prediction of similar outbreaks even before the symptoms manifest themselves. Statuses and posts on

Table 2.6: Summary of Other CDSS Models

Work	**Concept/Method**	**Explanation/Remarks**
Campbell *et al.* (2008)	Logistic Regression	ICU mortality and readmission prediction
Bennett and Doub (2010)	Frequent Pattern Mining	Optimal treatment selection through patient outcome prediction
Fialho *et al.* (2012)	Data Mining and Fuzzy Modelling	ICU readmission prediction
Liang *et al.* (2014)	DBN and RBM	Generic Healthcare decision making
Gentimis *et al.* (2017)	Multilayer Perceptron	ICU length-of-stay categorization
Zebin *et al.* (2019)	Autoencoder and DNN	ICU length-of-stay categorization
Li *et al.* (2019)	LASSO regression	ICU length-of-stay prediction

Online Social Network (OSN) sites such as Twitter, Facebook, etc. have proven to be an abundant source of useful information for such population based analytics and several research works that use Big Data Analytics and Machine Learning have been proposed over the years that prove the same.

Computational techniques like NLP and ML are extensively employed for performing predictive analytics on social media data. Several works in the areas of influenza or flu monitoring/detection (Alshammari and Nielsen, 2018; Aramaki *et al.*, 2011; Byrd *et al.*, 2016; Santillana *et al.*, 2015; Wakamiya *et al.*, 2018), adverse drug event detection (Cocos *et al.*, 2017; Sarker *et al.*, 2015), vaccine sentiment (Huang *et al.*, 2017), vaccine behaviour/vaccine shot status (whether vaccine shot was received or not) (Huang *et al.*, 2017; Joshi *et al.*, 2018), and vaccine hesitancy/vaccine intent (whether vaccine is intended to be taken or not) (Huang *et al.*, 2017) have been proposed over the past decade. Computational models for depression detection (whether a person is suffering from or prone to mental illness or depression problems) (McManus *et al.*, 2015; Orabi *et al.*, 2018; Shen *et al.*, 2017) have also been attempted through effective social media analysis.

Basak *et al.* (2007) used a supervised learning cum regression method based on Support Vector Regression (SVR) (Suykens and Vandewalle, 1999), to classify and measure the contribution of each influenza related term (feature) in Twitter data. Ginsberg *et al.* (2009) proposed a system that detects influenza epidemic with the help of search engine query data. The system uses data mining to analyze the data, linear regression for automatic query selection and univariate models to make predictions (Ginsberg *et al.*, 2009). Signorini *et al.* (2011) designed a Twitter data based system for tracking and measuring the disease activity of Influenza A H1N1 (or commonly referred to as Swine flu) using SVM. Aramaki *et al.* (2011) proposed an approach named TWEET-SVM, that uses NLP techniques to mine Twitter data and use it for detecting influenza epidemics using SVM classifier, which performed better than Google Flu Trends based detection. Achrekar *et al.* (2011) proposed a system named Social Network Enabled Flu Trends (SNEFT), which used data from US Centres of Disease Control and Prevention (CDC), as well as Twitter data to analyze flu trends. The system used mining and correlation techniques for analysis and prediction of influenza epidemics which also proved that including Twitter data along with CDC data improves accuracy of prediction (Achrekar *et al.*, 2011).

Yuan *et al.* (2013) designed a system to analyze, monitor and predict influenza epidemic in China, using the internet search query data from Baidu [4]. Using text mining, keywords or terms related to influenza are retained and the rest are ignored. The terms were weighted in terms of importance/relevance and then, a time series model and a regression process was used to make the prediction. Santillana *et al.* (2015) proposed an approach called 'Nowcast' that used Stacked linear regression, Adaboost regression with decision trees and SVR based on combined data from various sources such as Google searches, Twitter microblogs, nearly real-time hospital visit records and data from a participatory surveillance system to forecast estimates of multiple Influenza like Illness (ILI) diseases in the US. Byrd *et al.* (2016) proposed an approach that showed that Twitter data can not only be used for influenza epidemics detection, but also to predict the spread of the disease and monitor it in real-time. The approach involved usage of Twitter data and all its attributes including geolocation codes, along with a sentiment analysis module using NLP techniques and Naive Bayes classifier for enabling the detection and surveillance of influenza epidemic.

Huang *et al.* (2017) presented a study that uses NLP and several machine learning classifiers to analyze Twitter users' behaviour towards influenza vaccination, out of which Logistic Regression performed best. They performed prediction tasks such as vaccine relevance, vaccine shot detection, vaccine intent detection and vaccine sentiment. Joshi *et al.* (2018) designed and benchmarked NLP

[4] Chinese Web Search Engine - `https://www.baidu.com/`

based approaches (rule based, statistical based and deep learning based) for detecting whether a Twitter user received a flu vaccination shot or not, built on a LSTM based language model. Alshammari and Nielsen (2018) used NLP and machine learning techniques (SVM and Random Forest classifiers) for detecting self-reported flu cases using Twitter data. The authors reported that a tweet consisting of 280 characters, along with other aspects of the user such as followers, retweet, likes, replies, etc, is indeed a valuable source of information that supports accurate detection of self-reported flu cases. Wakamiya *et al.* (2018) proposed an approach based on NLP for influenza detection using direct and indirect information in Twitter data, for both urban and rural areas. The approach proved that influenza detection can be performed using tweets with not just direct information like announcements or reported cases, but also using tweets that indirectly points to the flu (maybe as a reason for some event cancellation).

Nikfarjam *et al.* (2015) proposed ADRMine, a machine learning based concept extraction system, based on NLP techniques and clustering, to detect and extract Adverse Drug Reactions (ADRs) from Twitter and DailyStrength[5] data. Sarker and Gonzalez (2015) presented a dynamic approach for ADR detection based on multiple corpus data – Twitter, DailyStrength and an openly available Adverse Drug Events (ADE) corpus using NLP techniques and SVM classifier. Cocos *et al.* (2017) proposed a deep learning based approach that used RNNs and tweet word embeddings to detect ADR in Twitter data. The authors also presented an RNN based model learned from minimal examples to label the words in the data with ADR membership tags, therefore minimizing manual labeling effort and making the model a bit scalable.

McManus *et al.* (2015) proposed an approach to detect Schizophrenia (a mental disorder) in Twitter users using NLP techniques and a SVM classifier. Shen *et al.* (2017) constructed a well-labelled depression and non-depression dataset consisting of twitter users and also presented and benchmarked various classifiers for depression detection. The authors proposed a multi-modal depressive dictionary learning model (MDL) based on dictionary learning and a custom gradient descent based binary classifier for detecting depression related characteristics in social media users. Orabi *et al.* (2018) proposed an approach based on word embeddings and deep learning (CNN model) for depression detection based on tweets by users. The authors came up with an approach to optimize the weights of word embeddings and benchmarked performance of several deep learning models on Twitter data. A summary of the works discussed in this section is presented in

[5] Health based Social network, `https://www.dailystrength.org/`

Table 2.7.

Table 2.7: Summary of Population Analytics based Healthcare Systems

Work	**Concept/Method**	**Explanation/Remarks**
Basak *et al.* (2007)	SVR	Classify and measure contribution of influenza related terms
Ginsberg *et al.* (2009)	Linear Regression	Influenza epidemic detection using search queries
Signorini *et al.* (2011)	SVM & NLP	Influenza H1N1 detection OSN data
Aramaki *et al.* (2011)	SVM & NLP	Influenza epidemics detection using OSN data
Achrekar *et al.* (2011)	Mining & Correlation	Influenza epidemics detection using OSN and CDC data
Yuan *et al.* (2013)	Time series and Regression	Influenza detection & surveillance using Baidu data
Santillana *et al.* (2015)	Regression techniques	Influenza surveillance based on combined data from various sources
Nikfarjam *et al.* (2015)	ADRMine - based on NLP and clustering	ADR extraction using OSN data
Sarker and Gonzalez (2015)	SVM & NLP	ADR detection using multiple data sources
McManus *et al.* (2015)	SVM & NLP	Depression detection (Schizophrenia) using OSN data
Byrd *et al.* (2016)	Naive Bayes & NLP	Real time influenza detection and monitoring
Cocos *et al.* (2017)	RNN & NLP	scalable ADR detection using OSN data
Huang *et al.* (2017)	Logistic Regression & NLP	Flu vaccine intend recognition, sentiment recognition using OSN data

Table 2.7: Summary of Population Analytics based Healthcare Systems

Work	Concept/Method	Explanation/Remarks
Shen *et al.* (2017)	Dictionary Learning & binary classification	Depression detection using OSN data
Joshi *et al.* (2018)	LSTM & NLP	Flu vaccine shot detection using OSN data
Alshammari and Nielsen (2018)	Random Forest & NLP	Self-reported flu cases detection using OSN data
Wakamiya *et al.* (2018)	TRAP & NLP	Influenza detection using indirect information in OSN data
Orabi *et al.* (2018)	CNN & NLP	Depression detection using OSN data

2.3 Outcome of Literature Review

After an extensive survey of existing literature, several research gaps were identified, specifically in the area of CDSSs using predictive models trained on text based clinical data. An important limitation of most works discussed earlier is that they focus on structured patient data and processed EHRs which are mostly standardized and in extensive use in western countries. However, currently in India (and other developing countries), 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 CDS to doctors and hospital personnel can be a huge contribution, given the prevalent conditions in the Indian healthcare ecosystem. Based on this observation, we focus on this problem with an intent to explore individual-centric healthcare dynamics and additionally, population-centric health analytics as well.

Another observation gleaned from our review was concerning the techniques adopted for preprocessing and feature modeling of the clinical data for usage in the prediction models based CDSSs. We observed that this is given minimal importance in most works, limiting to just the standard techniques. With unstructured clinical data, few basic NLP and vectorization techniques are usually applied and then fed on to a training algorithm, without an effective feature modeling strategy to derive effective patient data representations. In case of structured patient data and processed EHRs, most existing works have not employed effective preprocessing techniques to derive effective patient representations. Some recent works like DeepPatient (Miotto *et al.*, 2016) and DoctorAI (Choi *et al.*, 2016) tried to address this problem to an extent, by using extensive preprocessing and feature modeling approaches in generating better patient data representations for better training of the learning algorithms. Other works (Cheng *et al.*, 2016; Nguyen *et al.*, 2017) discussed their intent towards developing innovative methodologies to develop well-rounded patient data representations. In summary, more effective the patient representation, better the prediction model will be. In this context, we intend to focus on designing better and more effective techniques for capturing useful representative knowledge from raw clinical data, beyond the capabilities of the basic text/data processing pipeline to come up with patient data representations that can ultimately give rise to better CDSSs.

CDSSs have been an area of active research interest over the past two decades, and the methodologies used have been evolving over time. Machine Learning and Deep Learning models have proved to be the most accurate and effective ones so far. CDSSs that perform personalized prediction for a patient are most necessary in the field of clinical healthcare. These prediction systems not only help the patient with predictions of diagnoses and reminders, but also ensures that healthcare personnel, especially doctors are able to get a comprehensive view of the patients' medical history at a glance and can utilize the trained system's suggestions regarding the diagnoses and medical tests that ought to be performed, to make better treatment decisions. Recent works by (Miotto *et al.*, 2016; Choi *et al.*, 2016; Nguyen *et al.*, 2017; Purushotham *et al.*, 2018) have put forward deep learning based methodologies to provide CDS by predicting diseases and risks for a patient. Although some of these methods (Miotto *et al.*, 2016; Choi *et al.*, 2016; Purushotham *et al.*, 2018) reported good results, there is definite scope for further improvement in terms of both patient data representations and neural network architectures. Hence, better systems that provide effective personalised individual-centric predictions for clinical decision over historic patient data is an avenue we wish to explore further.

Population health management is an extremely important responsibility for a country's government. To analyze a target population's health statistics is a challenging process due to scale of the data and also due to its continuously streaming nature. An intelligent system for automated population health analytics will be extremely useful for organizations such as health departments or ministries, with functionalities like periodic trend analysis and timely reports on the population's health dynamics. In case of any unanticipated events such as disease outbreaks and adverse reactions to newly introduced medications, such organizations can be better prepared to efficiently handle such situations or to take positive action to avert them. Some existing works like (Achrekar *et al.*, 2011; Yuan *et al.*, 2013; Wakamiya *et al.*, 2018) proposed methodologies to model epidemics and tried to predict them based on various factors. It is also important to know the opinions of public regarding vaccine policies, adverse drug events among new drugs, how much of the population actually received a vaccine shot and may be even detect signs of bio-war through such information. Some of the existing works (Dredze *et al.*, 2016; Huang *et al.*, 2017; Joshi *et al.*, 2018) proposed approaches for understanding population intent towards flu vaccines and vaccine shot detection. Detecting signs of suicidal behavior and depression-related illness among people are also an important task for health based organizations. Some recent works (Shen *et al.*, 2017; Orabi *et al.*, 2018) proposed approaches for depression detection in social media users. The performance of the existing approaches that use OSN data for population analytics can be enhanced further by designing improved textual feature modeling techniques and prediction model architectures. For an effective healthcare system (for a city, state or country), population health analytics applications are critical and hence, this also is one of the issues we plan to address in our work.

2.4 Summary

In this chapter, the various approaches and models that have been proposed as part of CDS research were discussed. The existing approaches of building CDSSs were grouped into four categories – IR based Systems, NLP based Systems, Data Mining and Learning based Systems and Population based Healthcare Systems. The extensive review of the existing literature revealed that there is definite requirement for approaches in developing CDSSs, for both individual as well as population-centric predictive health analytics, based on structured clinical data, along with ample scope for introducing better preprocessing and feature modeling strategies for the unstructured data.

In Chapter 3, we formally define the research problem addressed in this thesis, based on the identified research gaps in the existing literature. We also briefly discuss the proposed methodologies designed to address the observed research gaps, the details of which are presented in subsequent chapters of this thesis. We believe that, the domain of healthcare analytics being a critical emerging field in the healthcare domain, the work presented in this thesis will make significant positive contributions to the ongoing research in this area.

Chapter 3

Problem Description

3.1 Background

In the previous chapter, an extensive review of existing approaches focusing on designing CDSS applications for augmenting healthcare delivery was presented. The prevalent issues and requirements for enabling improved CDSSs were also summarized. In this chapter, the identified research gaps are formally presented and defined as a problem statement. In addition, the scope of the proposed research work presented in this thesis and a brief overview of the approaches designed for solving the formally defined problems are also discussed.

3.2 Scope of the Work

An extensive review of the existing research in the domain of healthcare informatics and CDSSs was presented in Chapter 2. The research issues or gaps identified are summarized in the previous chapter. From the review of existing approaches for building CDSSs, it is clear that non-knowledge based CDSSs that involve techniques such as Data Analytics and Machine Learning techniques are the most effective till date. It was also gleaned that effectiveness of developed CDSSs depend heavily on how the clinical data are modeled and represented, as this forms the basis for prediction models. With this objective and with an aim of bridging the observed gaps, the research work presented in this thesis has contributions in five major aspects, as listed below:

1. Design and develop approaches for optimal feature modeling of structured EHR data for enabling CDSS development.

2. Designing patient-centric CDSS based on machine learning models built on structured EHRs, for improved prediction accuracy in end-user applications.

3. Designing approaches for handling unstructured and semi-structured clinical data and extracting latent knowledge for generating patient-specific representations.

4. Developing effective patient-centric predictive analytics based CDSS applications built on unstructured or semi-structured clinical data with learning-based models trained on generated patient data representations.

5. Development of population analytics based predictive analytics CDSSs using OSN data, using intelligent approaches for deriving data representations and machine learning models.

3.2.1 Problem Statement

Based on the understanding of the gaps identified from the review of existing literature in the domain of healthcare analytics and informatics, the research problem addressed by the work presented in this thesis is defined as below.

> *"To design and develop approaches for individual-centric and population-centric predictive healthcare analytics applications using unstructured and structured clinical data."*

3.2.2 Research Objectives

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

1. To design and develop effective preprocessing, feature modeling and representation techniques for structured clinical data for individual-centric predictive analytics.

2. To design and develop effective preprocessing, feature modeling and representation techniques for unstructured clinical data for individual-centric predictive analytics.

3. To design and develop a system for population-centric predictive analytics performed over population data.

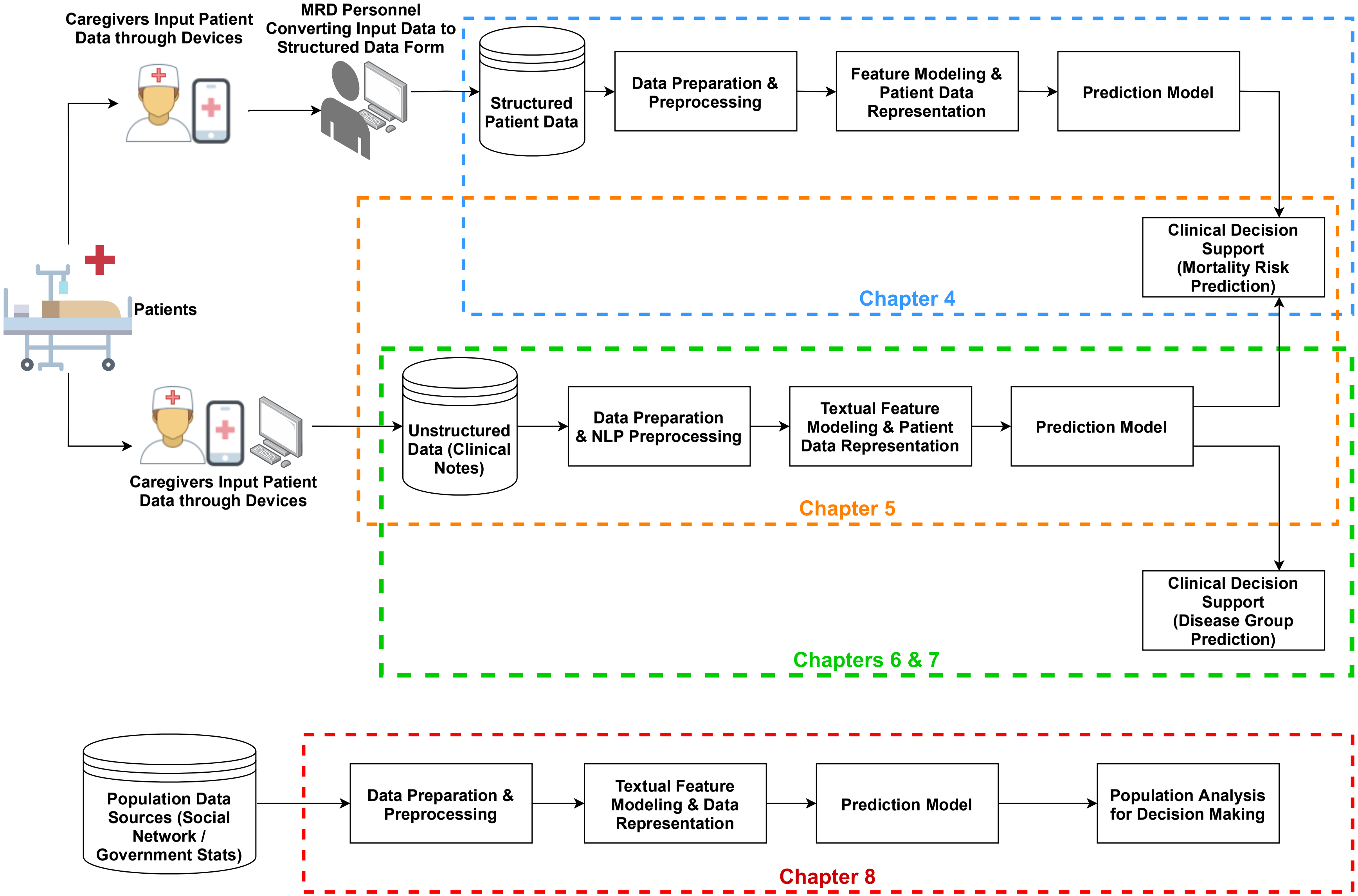

Caregivers Input Patient Data through Devices
MRD Personnel Converting Input Data to Structured Data Form
Structured Patient Data
Data Preparation & Preprocessing
Feature Modeling & Patient Data Representation
Prediction Model
Clinical Decision Support (Mortality Risk Prediction)
Chapter 4
Patients
Caregivers Input Patient Data through Devices
Unstructured Data (Clinical Notes)
Data Preparation & NLP Preprocessing
Textual Feature Modeling & Patient Data Representation
Prediction Model
Chapter 5
Clinical Decision Support (Disease Group Prediction)
Chapters 6 & 7
Population Data Sources (Social Network / Government Stats)
Data Preparation & Preprocessing
Textual Feature Modeling & Data Representation
Prediction Model
Population Analysis for Decision Making
Chapter 8

Figure 3.1: Overall workflow of the proposed integrated framework for predictive analytics of healthcare data - Task-specific view

3.3 Brief Overview of Proposed Methodology

The overall system architecture of the proposed Integrated Predictive Analytics based Framework for Intelligent Healthcare Applications is depicted in Fig. 3.1. The various contributions made towards the defined research objectives are indicated with respect to the individual thesis chapters in which they are presented in more detail. Here, a brief outline of the overall research work presented in this thesis is discussed.

3.3.1 Individual-centric Predictive Analytics for Structured Clinical Data

Hospitals in developed countries record data in the form of structured patient data which mostly consists of readings of labevents and other test reports that can be put to use directly for building CDSS applications. However, review of existing literature showed that there is huge scope for designing non-knowledge based CDSS for overcoming the deficiencies of traditional scoring systems. We address this by proposing efficient feature modeling and patient data representation strategies, which can be generalized better by classification models. A generic workflow for a patient centric CDSS application that make use of structured data is as depicted in Figure 3.2. The research contributions towards feature modeling and deriving patient data representation, along with how they are put to use for mortality risk prediction systems of ICU patients using machine learning, are explained in detail in Chapter 4.

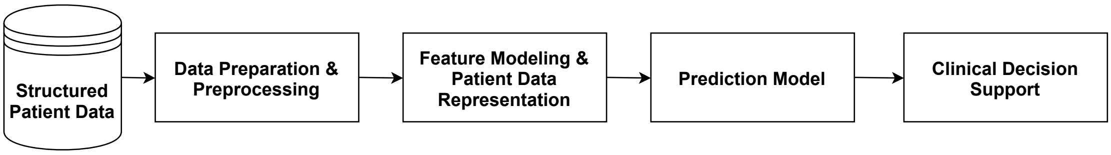

Structured Patient Data
Data Preparation & Preprocessing
Feature Modeling & Patient Data Representation
Prediction Model
Clinical Decision Support

Figure 3.2: Individual-centric Predictive Analytics for Structured Data

3.3.2 Individual-centric Predictive Analytics for Unstructured Patient Data

The EHR adoption rate in most hospitals in developing countries is very low, and they mainly make use of clinical notes which are in the form of unstructured text. From existing literature, a huge scope for developing CDSS applications that are capable of consuming unstructured clinical notes directly for predictive applications was observed. To address this, we incorporate effective textual feature modeling strategies, extracting effective patient data representation which can be used for effective training and prediction using machine learning and deep learning architectures.

Figure 3.3 depicts a generic workflow of a patient centric CDSS application that makes use of unstructured clinical notes. The various textual feature modeling strategies proposed for deriving patient data representations, along with details on how they are put to use for mortality risk prediction systems and disease prediction systems of ICU patients using various machine learning models are presented in Chapters 5 and 6 respectively. We also explored the possibilities of building disease prediction systems through aggregation strategies of multiple clinical records that pertain to a patient's admission and these are discussed in detail in Chapter 7.

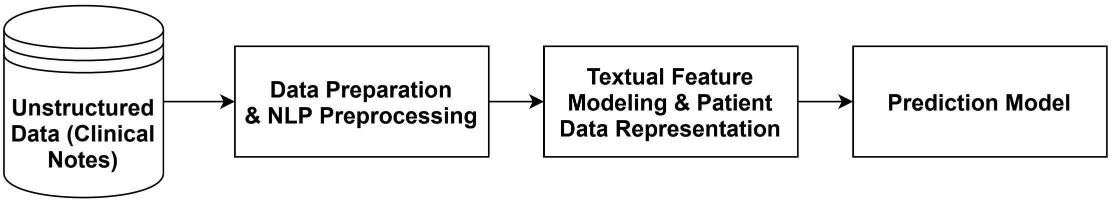

Unstructured Data (Clinical Notes)
Data Preparation & NLP Preprocessing
Textual Feature Modeling & Patient Data Representation
Prediction Model

Figure 3.3: Individual-centric Predictive Analytics for Unstructured Data

3.3.3 Population-centric Predictive Analytics

Gaining actionable insights into a population's health, their intent towards vaccine policies, their mental health analysis is an important task for a country's health organizations and for framing effective public health policies. Social media is a rich platform where users share information on their views towards these aspects. This can be used largely by the health organizations to make numerous decisions and hence, used as a source for designing populations analytics based decision support systems. Towards this, our contributions are in the form of designing effective textual feature modeling strategies for deriving rich data representation that can be used to develop effective prediction models. A generic workflow of a population-centric predictive analytics based decision support system is as depicted in Fig. 3.4. The research contribution towards designing a prediction model, that achieves multiple prediction tasks using novel feature modeling and data representation derivation strategies, is presented in Chapter 8.

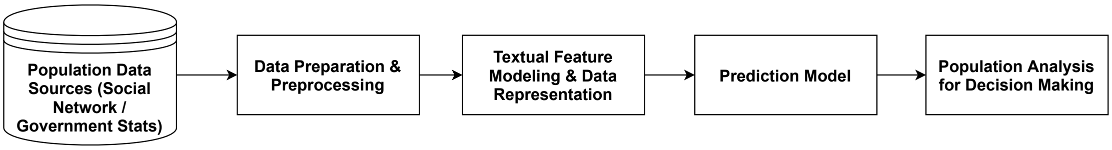

Population Data Sources (Social Network / Government Stats)
Data Preparation & Preprocessing
Textual Feature Modeling & Data Representation
Prediction Model
Population Analysis for Decision Making

Figure 3.4: Population-centric Predictive Analytics

3.4 Research Contributions

In this research thesis, a framework for enabling the design and development of evidence based CDSSs built on both structured and unstructured patient data is presented. The objectives are to design patient-centric and population-centric healthcare analytics methodologies and systems, to provide insights into the patients' health outcomes, thus affording intelligent decision-making capabilities to medical/health personnel. With regards to the outcomes gleaned from the literature review and the scope of work presented, the major contributions of our research work presented in the subsequent chapters of this thesis are as follows:

- An empirical study to understand the effect of feature modeling and selection on patient-specific mortality prediction performance using structured clinical data.

- A feature modeling approach using Genetic Algorithm (GA) and Extreme Learning Machine (ELM) for determining most relevant lab events for effective patient-specific mortality prediction using large-scale structured patient data.

- ELM based patient-specific mortality prediction for cardiac patients using unstructured clinical notes.

- Benchmarking study of word representation models for patient-specific mortality prediction using unstructured clinical notes.

- Ontology-driven feature modeling approach for ICD9 disease group prediction using unstructured clinical notes.

- Two-stage feature modeling approach using Particle Swarm Optimization (PSO) and neural networks for ICD9 disease group prediction using unstructured clinical notes.

- Hybrid feature modeling approach for effective ICD9 disease group prediction using unstructured clinical text records.

- Aggregation strategies for multiple clinical text records pertaining to a patient's hospital admission – *TAGS* and *FarSight*, for effective ICD9 disease group prediction and early prediction of disease onset.

- A Multi-task Deep Social Health Analyzer for performing population based predictive analytics on OSN data for effective monitoring of population health and intent towards policy-making.

3.5 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. We also discussed briefly the proposed approaches towards solving the defined problem, which are explained in detail in subsequent chapters.

PART II

Building CDSSs using Structured Clinical Data

Chapter 4

Individual-Centric Predictive Analytics for Structured Data

4.1 Introduction

In most developed countries, hospitals are typically equipped with advanced hospital information management systems (HIMS), that store patient data in structured formats like relational databases or spreadsheets. The objective is to impose a strict scheme on the patient data, such that it can be maintained as name-value pairs with reference to each patient, i.e., the attributes and values corresponding to various patient-specific data like demographic details like age, gender, lab test results, medication, allergies etc. This structured patient data is manually generated by collating various unstructured data sources, through manual conversion processes, after which they are stored.

A variety of applications are built on the availability of such structured data, the most important being consumption of such data for facilitating analytics and inference engines. These structured data sources are amenable for building machine learning models that are designed to identify underlying patterns, with appropriate feature modeling as per the requirements of the end-user application. Hospitals with advanced health information technology systems use large number of trained personnel in the Medical Records Department (MRD) for converting unstructured patient data sources like demographic data collection forms, physicians notes, nurses notes, discharge summaries, lab reports etc to structured form, for making them suitable for use as training data in prediction models. In this chapter, we aim to explore avenues of leveraging such structured patient data sources for the development of CDSSs. We present experimental studies on the role of structured data in an important CDSS system extensively in use in modern hospitals - ICU mortality risk prediction. We explore the suitability of structured patient data for enabling individual-centric, i.e., patient-specific prediction in estimating risk of death given multi-dimensional feature spaces, and leveraging it for fast and accurate estimation and prediction of patient's risk profile.

4.1.1 Problem Definition

While most parametric scoring systems like APACHE-II, SAPS-II and SOFA are now considered standard for ICU mortality measurement in practice, the accuracy achieved by them is low in comparison to non-parametric methods. Moreover, the patient-specific data points (examples include - results of specific lab test, blood sugar, urine output etc) considered as features by each scoring system is different and often, significantly large in number. Due to this, all such score-specific features, i.e. the required lab tests have to be performed for each ICU patient before a mortality risk can be assessed. This contributes to an additional delay in making time-critical mortality decisions, while also adversely affecting cost and resource usage. It is therefore important to optimize the number of such features required for predicting mortality risk at the earliest possible time, with high precision and accuracy, thereby reducing the number of clinical variables or lab events (features) that need to be collected. To the best of our knowledge, such an investigation into mortality risk prediction has not been conducted on large-scale patient data, with an exclusive focus on the contribution of individual or group of lab event features.

The problem to be addressed here is defined as follows:

> *Given the known issues arising due to the large number of lab events considered by traditional mortality risk assessment scores, design and develop approaches for optimal feature modeling of multi-dimensional patient data for effective ICU mortality risk prediction, based on structured clinical data.*

4.1.2 Motivating Example

We again take up the example scenario introduced in Chapter 1, Section 1.2.4, for underscoring the challenges faced in real-world hospitals. As mentioned earlier, *Dr. Bob* works in *Hospital B*, which has adopted structured data standards, and employs CDSSs built on structured EHRs. When a patient is admitted to the ICU, *Dr. Bob* orders the required X lab tests and readings to be performed as per requirements of one of the traditional severity scoring based systems (such as

APACHE-II, SAPS-II, etc.), in order to measure the severity/mortality risk of the patient. Naturally, this requires some time to be completed, say, T time units, after which the results of the prescribed lab tests are available for scrutiny. Only after the elapsed time T, the CDSS can leverage the X lab test values as features for assessing the mortality risk of the patient, which is then provided to *Dr. Bob* and his team, to enable them to make informed decisions about the next course of action with reference to the patient's treatment.

In this scenario, let us consider that the same prediction could be potentially determined faster and more economically based on a newer risk prediction model, without compromising on prediction performance, with x lab test values/readings that take t time units for completion and result generation, where $x < X$ and $t < T$. That would be highly advantageous to all direct stakeholders like *Dr. Bob* and his team, and the concerned patient, and also the hospital, due to the significant savings in medical and human resources alike. The doctor can make decisions faster, and the patient would receive the required medical intervention faster, lessening the chances of deterioration in his condition. Due to these advantages, optimization in prescribed tests and associated data required to build mortality risk prediction CDSSs commonly used in ICUs are of critical importance.

In this chapter, we address this challenge in a two-fold manner. Firstly, we present extensive empirical studies on the nature of feature variables used by traditional severity scoring based Mortality Risk Prediction Models (MPMs) used in practice. Based on the insights obtained, we employ this for training Machine learning classifiers for designing ML based MPMs. Secondly, we focus on designing a novel learning based model that facilitates time-sensitive ICU mortality risk prediction with improved accuracy. The proposed model uses Genetic Algorithm (GA) based optimization and Extreme Learning Machine (ELM) neural models for deriving optimal subset of lab events, that achieves significant performance improvement and cost/time savings.

4.2 Traditional Mortality Risk Scoring Models - An Empirical Study

In this study, an investigation into the working of traditional severity scoring based MPMs and their constituent feature variables was undertaken. As discussed in Chapter 2 (In Section 2.2.3.1 and Table 2.4), some popular MPMs being used in practice in hospitals include APACHE (version I to IV), SAPS (version I to

III), SOFA, OASIS etc. Most such scores use the values from tests performed on patients for computing their mortality scores. Several such tests are often required to predict mortality risks, in most traditional scoring systems. SAPS-II requires 17 test values for mortality prediction while SOFA uses just 6 and OASIS is dependent on 10 physiological variables. The main objectives of our study are three-fold.

1. To assess the effect of features used by traditional ICU mortality scores.
2. To observe how Machine Learning (ML) based MPMs work in comparison to the traditional severity scoring based MPMs.
3. To examine the effects of feature modeling on mortality prediction performance and compare its performance against the conventional severity scoring based MPMs.

The overall methodology adopted for the empirical study is shown in Figure 4.1. The different processes defined as part of this empirical study are described in detail next.

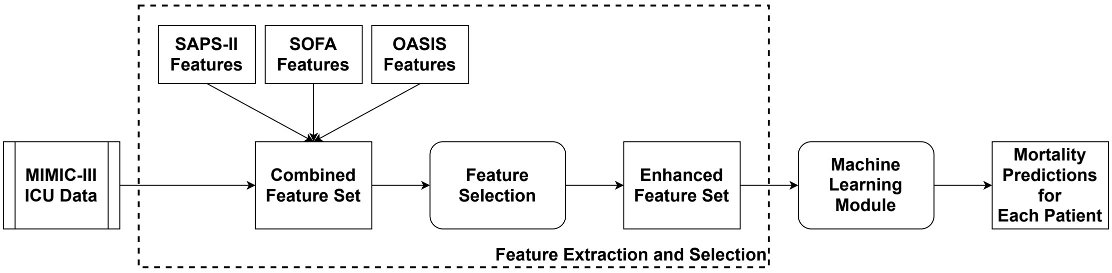


Figure 4.1: Proposed Approach for the Empirical Study

4.2.1 Patient Cohort Selection and Preprocessing

For the proposed empirical study, the MIMIC-III[1] dataset (Johnson *et al.*, 2016) was used, which contains de-identified clinical data of 46,520 critical care patients. The data is quite extensive and we designed certain criteria for selecting specific data. Using these criteria, a subset of 32,622 patient records were selected for training and validation of the proposed approach. We list the patient cohort selection criteria below.

1. Only the records of adult patients were considered ($age \geq 15$), as pediatric patients are treated with infants and adolescent-specific procedures.

[1] Medical Information Mart for Intensive Care – https://mimic.physionet.org/

2. To ensure that the prediction model had enough data to make predictions on, only those patient records, where the patient admitted in the ICU stayed there for at least one day (length of stay - $los \geq 1$), were selected.

Missing Data Handling. A major issue that we had to deal with at this stage was the large number of missing values for clinical features in MIMIC-III. Directly ignoring records with missing values adversely affected the number of patients selected for the study, hence we used a specific way for handling missing data. For the 32,622 patient records, those records with missing values were filled with the statistical median values of respective columns. Filling with median values ensures that the statistics of the model does not deviate from or be biased towards a particular clinical feature.

4.2.2 Clinical Variable Extraction and Feature Selection

During this phase, we experimented with the different features used by traditional mortality risk assessment methods. For this, popular scores - SOFA, SAPS-II and OASIS scoring systems were implemented for each patient in the selected patient cohort, using which a combined feature set was generated. Furthermore, we also considered some additional features like patient demographics (*gender*) and *first care unit* (type of ICU to which patient was first admitted to, e.g. surgical, medical, trauma, cardiac care etc). We also considered the ICD9 code of the first disease diagnosis for each patient. These two additional features, i.e., *first care unit* and ICD9 code, were considered on the basis that they can help determine the severity of the patient condition, which may contribute to the correct prediction of mortality risk. With the inclusion of the standard severity score features and those we considered additionally, the final number of features obtained was 45.

To derive the most relevant features from this set, we used a feature selection technique called Recursive Feature Elimination (RFE) (Guyon *et al.*, 2002). RFE is a wrapper feature selection technique, which uses an estimator to repeatedly create a model based on a feature subset, and then prune features with low importance based on the accuracy of the estimation. The RFE algorithm is a 3-step process which is performed iteratively. Firstly, the estimator is trained on the features by optimizing feature weights, next, the features are ranked based on a cost function, based on which finally, the features with the least rank are removed during each iteration. These steps are performed for a number of iterations till an optimal features set with the required number of features remain. We used the

RFE algorithm with Logistic Regression as the estimator with its performance as the cost function, which uses the categorical variable (mortality prediction in this case) as the dependent variable. Logistic Regression is a statistical model that uses a logistic function to model a dependent variable, which has a certain probability of belonging to a specific class. Logistic Regression being a statistical probabilistic prediction model based on features, works well for binary classification and hence is well-suited for selecting features that contribute most to the mortality prediction. Algorithm 1 depicts the process in detail.

Algorithm 1 Optimal Feature Subset Selection using RFE

Input: The whole set of features and mortality labels and n, the required number of features to be selected.

Output: Optimal subset of n features.

```
Set the whole set of features as the current feature set
while current feature set size $> n$, do
    Train Logistic Regression model with the set of features and corresponding
    labels, thereby tuning feature weights
    Calculate cross-entropy loss function values
    Rank the set of features based on cross-entropy loss values
    Remove feature with the least rank from the current set
Generated optimal subset of $n$ features (Training dataset)
```

After the feature selection process, only 8 features were selected as the most relevant out of the original feature set of 45. These included scores that capture the correct functioning of the liver, renal output, cardiovascular activity, age score, blood urea nitrogen (BUN) score, sodium score, comorbidity score and first care unit. Interestingly, features were selected only from SAPS-II and SOFA and the additional variable, *first care unit*. Intuitively, this makes very good sense, as the first care unit in which the patient was admitted to does play an important role in determining the severity or mortality risk of the patients newly admitted. The score mostly says how critical the patient is, and hence contributes a lot of information towards predicting mortality. For predicting the mortality, a label called *expire flag* is used, which is binary in nature (i.e., 0 for alive and 1 for expired). The final set of optimal features along with the labels are then fed into the ML module for obtaining patient-specific ICU mortality prediction.

4.2.3 Supervised Learning Process

The optimal feature set can now be used as a representation of patient-specific clinical data, based on which their personalized profile can be modeled. The optimal feature set along with the associated mortality labels in MIMIC-III data were used for training a suite of machine learning classifiers (the ML module), for mortality risk predictions namely, Naive Bayes (NB) (Gaussian, Multinomial and Bernoulli), Support Vector Machine (SVM-linear kernel), Decision Tree (CART - Classification and Regression Trees) and Random Forest classifiers. NB Classifiers are probability based classifiers which work based on application of Bayes' Theorem with an assumption that features are independent. SVM classifies data points into potential classes (alive and expired patients in this case) by representing them in vector space. The SVM algorithm constructs a hyperplane or a set of hyperplanes that divide the space so that data is classified into potential classes. The decision trees classifier employs a tree-like structure where the internal nodes represent the features and the leaf nodes represent the labels (the mortality labels - alive(0) or expired(1) in this case). Various branches represent various feature values leading to corresponding mortality labels, which are used to perform classification. Finally, the random forest classifier is an ensemble classifier which works by generating numerous decision trees the votes of which are used to predict the label. The mortality class which is voted or predicted by most decision trees decides the prediction of the random forest classifier.

The patient representations modeled as per the proposed feature extraction and selection approach are used as training data and are fed into the various classifiers for observing mortality prediction performance. We used 10-fold cross validation on the data for training and testing. We used standard metrics like average accuracy, average precision, average F-score and average Area Under Receiver Operating Characteristic Curve (AUROC) for evaluating the performance of the trained model.

4.2.4 Experimental Results and Discussion

The experiments were performed on a workstation running Ubuntu 17.04 with 3.5 GHz Intel Core i7 Processor, 16GB RAM and 2TB Hard Drive. The proposed prediction model was developed in Python and packages like pandas[2], Scikit-Learn (sklearn[3]), and matplotlib[4] were used for performing dataset operations, Machine Learning algorithms and plotting ROC curves. For validating the proposed approach, both the proposed model and the standard traditional severity score based MPMs (SAPS-II, SOFA and OASIS) were applied to the MIMIC-III subset of

[2] https://pandas.pydata.org/

[3] https://scikit-learn.org/stable/

[4] https://matplotlib.org/

32,622 patients. For each patient in the selected cohort, we implemented each traditional severity score based MPM and based on generated scores, the mortality prediction results were obtained. For SAPS-II, the probability of mortality for each patient was calculated as per Eq. 4.1 (Pirracchio *et al.*, 2015; Gall *et al.*, 1993).

$$\log(P_m/1 - P_m) = -7.7631 + 0.0737 * S + 0.09971 * \log(1 + S) \tag{4.1}$$

where, P_m is the required mortality probability of a patient and S is the SAPS-II score of the patient. The threshold of classification for SAPS-II based mortality probability was taken as 0.5 as done by Patel and Grant (1999).

In the case of SOFA, the mortality prediction of each patient was obtained by regressing the mortality on the SOFA score using a main-term logistic regression model as per Pirracchio *et al.* (2015).

The probability of mortality for each patient as per OASIS scoring system is given by the in-hospital mortality score calculation, given by Eq. 4.2 (Johnson *et al.*, 2013).

$$\log(P_m/1 - P_m) = -6.1746 + 0.1275 * OASIS \tag{4.2}$$

where, P_m is the required mortality probability of a patient and $OASIS$ is the OASIS score of the patient. The threshold of classification for OASIS based mortality probabilities were also considered to be 0.5.

Our observations on the performance of the various ML models and the standard severity scores (SAPS-II, SOFA and OASIS) based MPMs on the MIMIC-III dataset are tabulated in Table 4.1. From the values of the various metrics, it can be seen that the Random Forest classifier performed best at an average accuracy of 0.71, average AUROC of 0.77, average precision of 0.71 and average F-score of 0.71, while SVM and Decision Tree classifiers were a close second. Random Forests being an ensemble classifier, predicts a label based on the voting of multiple decision trees, hence performed the best. It can be observed that, among the standard severity scoring systems, SAPS-II outperformed SOFA and OASIS. However, the proposed ML based model with Random Forests classifier achieved significant improvement over SAPS-II in all metrics. A plot of the Receiver Operating Characteristic (ROC) curve is shown in Fig. 4.2, which highlights the superior performance of the proposed (random forest based) MPM over existing severity score based MPMs for MIMIC-III data.

From Table 4.1, it can be observed that among the classifiers used, Random Forest achieved the best accuracy of 0.71 and an average AUROC of 0.77. Its

Table 4.1: Benchmarking ML approaches against Traditional severity scores (SAPS-II, SOFA and OASIS)

Classifier	**Accuracy**	**AUROC**	**Precision**	**F-Score**
Gaussian NB	0.69	0.75	0.68	0.68
Multinomial NB	0.69	0.68	0.68	0.67
Bernoulli NB	0.68	0.71	0.69	0.68
Decision Tree	0.70	0.73	0.70	0.69
SVM	0.70	0.76	0.70	0.69
Random Forest	0.71	0.77	0.71	0.71
SAPS-II	0.63	0.72	0.65	0.57
SOFA	0.62	0.61	0.60	0.57
OASIS	0.61	0.64	0.67	0.50

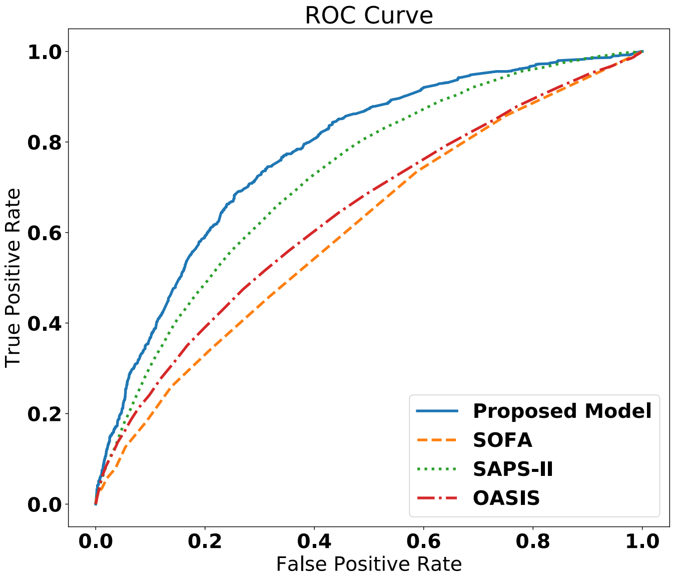


Figure 4.2: Area under ROC Curve for various models (The best performing model, Random Forest, was considered the *Proposed Model*)

performance when compared to that of SAPS-II, SOFA and OASIS, was significantly higher by a factor of 12–16% in terms of prediction accuracy. It can also be inferred that ML based models can effectively predict patient-specific mortality using lesser number of feature variables than traditional severity scoring based MPMs SAPS-II and OASIS. Even though SOFA uses a lower number of feature variables, the Random Forest based MPM performs far better in terms of all metrics even though it considers just two additional feature variables. This indicates an ample scope for development of CDSSs that dynamically determine the most important feature variables and predict patient-specific mortality effectively, thereby reducing time and hospital sources.

4.3 Predicting ICU Mortality using Large-scale Lab Events Data

From the previous work, a significant scope for effective approaches that can determine the most important lab events or physiological tests that contribute most towards mortality risk and using them for mortality prediction applications for ICU patients was inferred. Existing non-parametric based models use a multitude of features as input data to train machine learning models to predict mortality risk of ICU patients. It was also understood from discussion with experts and doctors that, reducing unimportant lab tests not only saves time, but also optimizes cost for patients and resource allocation in hospitals. Moreover, in practice, hospitals often follow their own customized versions of MPMs, where a significant number of lab events are to be performed for each patient. Hence, there is a requirement for approaches that can determine the most important lab events to be performed so that mortality risk prediction can be as accurate as possible with fewer lab events. The processes defined as part of the proposed approach for ICU mortality prediction based on a patient cohort's clinical data is depicted in Fig. 4.3. Each of these processes are discussed in further detail in this section.

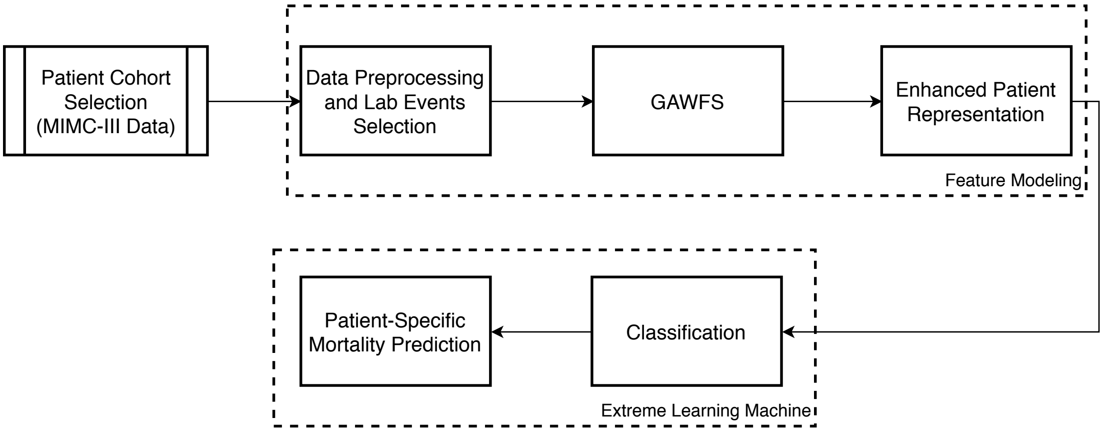

Patient Cohort Selection (MIMC-III Data)
Data Preprocessing and Lab Events Selection
GAWFS
Enhanced Patient Representation
Feature Modeling
Patient-Specific Mortality Prediction
Classification
Extreme Learning Machine

Figure 4.3: Workflow of the proposed GA-ELM Model

4.3.1 Patient Cohort Selection and Data Preprocessing

For validating the proposed methodology, we used the openly available standard dataset, MIMIC-III (v1.4) (Johnson *et al.*, 2016) for our experiments., as similar to the previous work. From this data, a patient cohort was selected based on the following criteria:

1. Clinical data of only adult patients (age>15) was selected for the cohort, in accordance with previous studies. This is important as the procedures used for pediatric patients are highly specific in nature (Pirracchio *et al.*, 2015).
2. Only the first ICU admission of each patient, in cases where a patient was admitted to ICU multiple times, was considered for the study. This helps ensure the CDSS nature of a mortality prediction model which helps in predicting mortality risk with respect to earliest available data on a patient's condition.

Accordingly, a subset of 31,691 eligible patients was chosen as the patient cohort. For these patients, the results of a total of 573 lab tests performed are available in a MIMIC-III table called 'labevents'. These lab test values are extracted and modeled into a representation, where each row represents a patient and each column represents a lab test. However, there are several missing values in some rows, as not all tests are necessarily performed on all patients. If the rows (patients) with such missing column values are directly removed from the cohort, then a large number of patients will need to be excluded from the cohort. To overcome this, we separately calculated the statistical median values of each column for all alive and expired patients in the selected cohort and filled these median values in place of any missing values in that particular column. Along with these 573 features, other demographic features like *age* and *gender* were also added to the feature set. Additionally, the *ICD9* disease code of a patient's first diagnosis, *length of stay* and also the *first_careunit* (type of ICU to which the patient was first admitted to) of the patients were also considered as features. After the preprocessing tasks are applied, the final patient cohort consisted of 31,691 patients (rows) and the 578 features (columns) representing them. The outcome labels are the '*expire_flag*' of each patient (0 for alive and 1 for expired). The statistics of the selected cohort is tabulated in Table 4.2.

4.3.2 Optimally Modeling Lab Events

Laboratory tests or events help medical personnel in continuously monitoring a patient's condition. Often, medical personnel are prone to order various lab eval-

Table 4.2: Number of Expired/Alive Patients in Initial and Selected Cohorts

Cohort	**Alive**	**Expired**	**Total**
MIMIC-III data	30,761	15,759	46,520
Selected Cohort	19,225	12,466	31,691

uation procedures for patients, some of which may be unnecessary or redundant in actually understanding the patient's condition. Eliminating such unnecessary and wasteful lab evaluations is of significant importance, given the rapidly escalating healthcare and insurance costs as well as excessive overuse of laboratories and equipment (Chaudhry *et al.*, 2006). Moreover, the extra time taken to perform the unnecessary tests might worsen the patient's condition. For the problem of mortality prediction for ICU patients, it is critical to predict mortality risk at the earliest possible patient condition and hence, reducing the number of lab events required to predict mortality effectively, is a matter of significant importance. Therefore, we attempt to model patient-specific lab event requirements for the two-fold objective of reducing prediction time as well as improving prediction accuracy.

To determine the optimal representation for each patient in the chosen cohort, a Genetic Algorithm based Wrapper Feature Selection (GAWFS) technique is proposed. GAWFS is used to find the most-optimal subset of feature variables of a patient (i.e. lab events) to predict mortality risk of ICU patients. This optimal feature set is used for training a learning based risk prediction model. While feature selection techniques have been extensively used for deriving the optimal feature set, feature extraction can also be used to reduce dimensionality and increase the efficacy of the model. However, feature extraction techniques use a statistical combination of feature values to generate new features which makes it impossible to track which features (in our case, lab events are features) were contributed in the prediction. As the purpose of our work is also to identify the most crucial lab events, we used feature selection and not feature extraction.

4.3.3 Feature Selection

Feature selection (FS) is the "*process of selecting an optimal subset of features as per certain predefined criteria*". Essentially, FS methods can help in reducing the dimensionality of the dataset by ignoring the unimportant or noisy features, so the prediction process can be more accurate and computationally efficient (Sánchez-

Maroño *et al.*, 2007). In this case, if a real world CDSS application is able to make accurate predictions based on a lower number of features (e.g. lab event measurements), then it can potentially save lives, time and cost, and consequently, is more effective and valuable.

FS techniques can be mainly sub-categorized into *filter* and *wrapper* methods. Filter methods are suitable for quick feature selection based on the threshold of general characteristics of the data, such as statistical dependencies, without the need for any induction or classification algorithms (Sánchez-Maroño *et al.*, 2007; Hira and Gillies, 2015). Some popular examples are ANOVA F-test (Analysis of Variance) and Mutual Information (MI) test. Wrapper methods generate an optimal feature subset by evaluating the quality of each feature subset, based on some classification or induction algorithm, regardless of the chosen learning method (Kohavi and John, 1997). Recursive Feature Elimination (RFE) and Sequential Feature Selection (SFS) are popular examples of wrapper based methods. Although wrapper methods are computationally more expensive in comparison to filter methods, the quality of the derived feature subset is better ensured as performance evaluation is performed with respect to a classifier model during the feature selection process.

To determine the optimal set of features, i.e. the reduced set of lab events contributing the most towards mortality risk prediction, a Genetic Algorithm based Wrapper Feature Selection (GAWFS) technique is proposed. Genetic Algorithm (GA) (Yang and Honavar, 1998) is an evolutionary meta-heuristic algorithm inspired by the biological process of natural selection and the idea of "survival of the fittest". GA is known to offer high quality solutions to optimization and search problems by using the operations – Selection, Crossover and Mutation as in the process of natural selection and hence, GA is appropriate for the feature selection process for removing redundant lab events for improved mortality prediction.

The GAWFS process is depicted in Fig. 4.4. We make use of concepts of GA for calculation of fitness of a population (a set of individuals and chromosomes, i.e. a subset of features or lab events) and based on the fitness, a particular feature is selected if it is fit. From the original feature set consisting of 578 features, the initial population was selected as a random subset of lab events for all patients (analogous to 'individuals' in the 'population'). For the selected patient cohort, the feature subset and the associated patient-specific mortality labels are fed into an estimator/classifier, whose classification performance is then measured.

Algorithm 2 illustrates the process of deriving the optimized lab event subset using GAWFS. We use an Extreme Learning Machine (ELM) based neural network

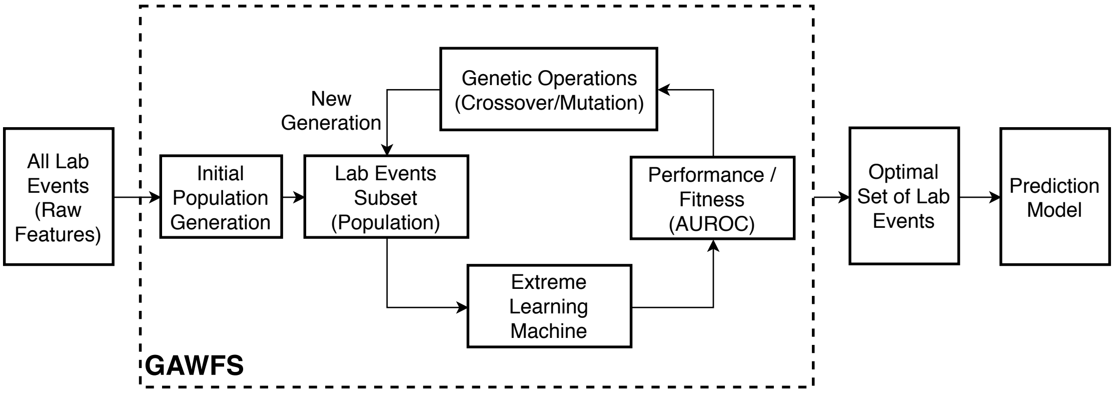

All Lab Events (Raw Features)
Initial Population Generation
Lab Events Subset (Population)
New Generation
Genetic Operations (Crossover/Mutation)
Performance / Fitness (AUROC)
Extreme Learning Machine
GAWFS
Optimal Set of Lab Events
Prediction Model

Figure 4.4: Genetic Algorithm based Wrapper Feature Selection process

Algorithm 2 Optimal Lab Events Subset Selection using GAWFS

Input: Set of all lab events & patient-specific mortality labels

Output: Optimal lab events subset of, say, n features *(Best solution)*

```
while iterations $\leq$ 100 do    ▷ No.of generations=100
    Generate randomly a feature set (lab events) for all patients    ▷ Each feature set represents an individual chromosome, and patients represent the initial population
    Select parents and perform genetic operations    ▷ Single point crossover and mutation with probabilities of 0.5 and 0.2 respectively are used
    Create new generation
    Calculate fitness of new generation    ▷ AUROC performance of ELM
    if new-fitness > old-fitness then    ▷ New generation's fitness is better than that achieved with previous subset of features
        Replace current generation with new generation
    else
        Retain the current generation
    end if
end while
```

based architecture as a classifier or estimator model for the GA technique, thereby making GAWFS a wrapper based feature selection technique. ELM is a training method for a single hidden layer neural network based classifier, for which only the weights between hidden and output layer need to be learned. The ELM model is described in detail in Section 4.3.4. As the fitness function, the metric *Area Under the Receiver Operating Characteristic Curve (AUROC)*, as shown in Eq. 4.3, was used in GAWFS for calculating the fitness value associated with a particular feature subset. AUROC measures the overall quality of a classifier by varying the threshold parameter (say i), which biases the classes and returns a value between 0 and 1 (where a value of 1 indicates best classification performance possible). The number of thresholds varied is determined by the unique number of predicted probabilities of the ELM classifier. As AUROC measures how well a classifier has learned to classify between the majority and minority classes in the presence of class imbalance, it is apt for our problem of mortality prediction, and therefore, it was chosen as the fitness function in GAWFS and is calculated as per Eq.(1).

$$Fitness, f(x) = \sum_{1}^{N-1}(TPR_{i+1} - TPR_i)(FPR_{i+1} - FPR_i) \tag{4.3}$$

where FPR is the False Positive Rate, TPR is the True Positive Rate and i refers to the varying threshold parameter for which at each point FPR and TPR are determined and N is the number of thresholds which was found to be 1062 in our experiment. Eq. 4.3 sums all the area of all the small rectangles in the ROC curve between two FPR and TPR points for adjacent thresholds.

During the FS process using GAWFS, the GA operations of single point crossover and mutation were performed using empirically determined probability values of 0.5 and 0.2 respectively. During this iteration, a new generation gets generated, where, least-fit individuals in the population will be replaced, if their fitness is not better compared to the ones in the population. In order to enable comparative benchmarking of the proposed GAWFS model with state-of-the-art FS techniques, GAWFS was configured to select the most important 10 lab events or features, similar to the other works.

4.3.4 Building the prediction model

The design of the proposed prediction model is driven by the two principal requirements of a mortality prediction CDSS. Firstly, to eliminate false negative mortality predictions, i.e., a wrong low mortality risk prediction for a patient who is actually at high mortality risk should never occur, and, secondly, to ensure learnability after deployment as a real-world CDSS. To address these two major aspects, we propose an architecture built on an Extreme Learning Machine (ELM) neural network, with Rectified Linear Unit (ReLU) as the hidden layer activation function.

ELM is a learning technique for training Single hidden Layer Feedforward Neural Networks (SLFNN) (Huang *et al.*, 2015) that are trained in finite training sets. Initially, the hidden nodes in ELM are randomly fired with random weights and learning is carried out without iterative tuning. By design, only one parameter needs to be learned in ELM, i.e., the set of weights between the hidden layer and output layer. Thus ELMs are extremely fast when compared to traditional SLFNNs and also very well-suited for further re-training for future learning (Huang *et al.*, 2004). Moreover, ELMs can be trained to converge to the smallest possible error with minimal magnitude of weights, due to which the generalization performance of ELMs far exceeds that of traditional feedforward neural networks (as per Bartlett's theory). Another advantage of ELM is its ability to reach solutions in a straightforward manner avoiding problems like local minima, overfitting and improper learning rate (Huang *et al.*, 2004). Based on these observations, we experimented with ELM as an estimator in the proposed mortality prediction model, for exploiting its advantages for the development of real-world CDSSs. After the FS process, the feature subset with the best fitness value along with the class labels, is used for training the ELM model, for predicting patient-specific mortality risk.

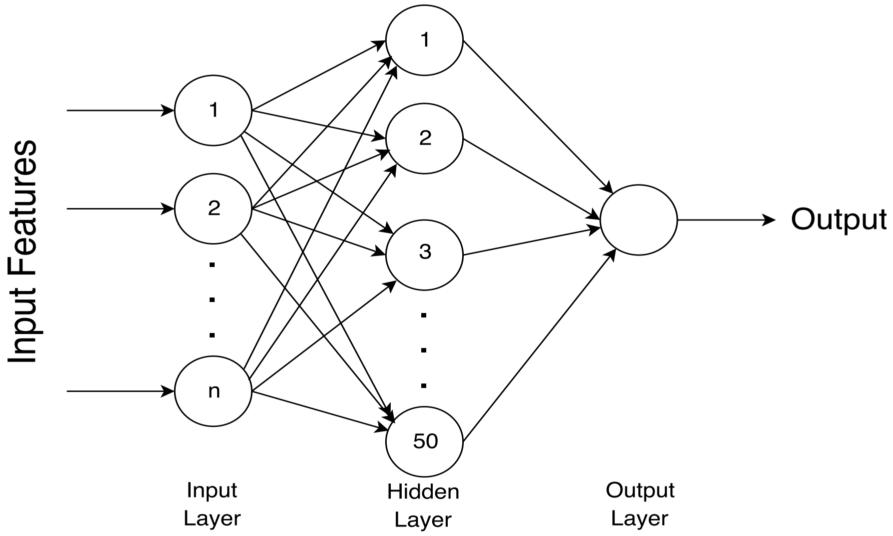


Figure 4.5: Architecture of the ELM Model

Fig. 4.5 illustrates the ELM architecture used in the proposed model. It is primarily a SLFNN architecture, where the number of input nodes is governed by the number of features used for training (as described in 4.3.3). The hidden layer consists of 50 nodes and one nod at the output layer, which predicts the mortality risk. We used ReLU as the hidden layer activation function in the proposed ELM architecture. ReLU, being a ramp function given by $f(x) = max(0, x)$, can trigger for any non-zero input and therefore, helps the ELM classifier to predict even the slightest chance of mortality, thereby completely eliminating as many cases of false negative predictions. The output layer uses a sigmoid activation function as it is a binary classification.

Algorithm 3 Process of training the ELM as an estimator for proposed GAWFS

```
Input: A training set with N samples (from GAWFS) consisting of lab event features and mortality labels $(x_i, y_i)|x_i, y_i \in \mathbb{R}, i = 1, 2, ...N$    Activation function $f(x)$ (ReLU)   Output: ICU Mortality Risk
Randomly assign weights $w_i$ and bias $b_i$, $i = 1, 2...N$ for N training samples
Compute output matrix based on the input lab event feature set, say $H$
Compute output weights using input mortality labels & output matrix $\beta$
Train the ELM network using the Least Square Solution $\beta'$ to the linear system $H\beta = T$
Analytically tune output weights
Perform prediction for test set and observe prediction performance
```

Algorithm 3 depicts the process of training the ELM network as an estimator for the proposed feature selection model, which also works as the final prediction model. The various feature sets generated by the GAWFS technique and the final optimal feature set obtained after the GAWFS process with the associated patient-specific mortality labels, are used for training the ELM model. The parameters and weights are initialized randomly and the output matrix is calculated based on the given lab events or features set and patient-specific mortality labels. During training, the weights between the hidden and output layer are iteratively optimized and finally, patient-specific mortality prediction performance is observed. The details of the experimental validation are reported in Section 4.3.5.

4.3.5 Experimental Results and Discussion

The proposed mortality prediction approach was evaluated by a series of experiments designed to benchmark it against both traditional scoring methods and state-of-the-art learning based approaches. The experiments were performed on a setup consisting of a high-end server running Ubuntu Server OS with 56 cores of Intel Xeon processors, 128GB RAM, 3TB Hard Drive and two NVIDIA Tesla M40 GPUs. For all experiments involving training and testing, 10-fold cross validation was performed. We used Python libraries – sklearn and matplotlib libraries as part of execution. The proposed GAWFS technique was configured to select the top-10 lab events that most contributed to mortality risk prediction, from the original 578 lab events (raw features). For the selected cohort of 31,691 patients, the lab events (features) that were found to be of high importance by proposed GAWFS technique were – *Platelet Count, Red Blood Cells, Hematocrit, Sodium, Chloride, Bicarbonate, Base Excess, Urea Nitrogen, Anion Gap,* and *Partial Thromboplastin Time (PTT)*. These features, along with the corresponding patient-specific mortality labels for the selected cohort were then used for training and validation of the designed prediction model.

4.3.5.1 Evaluation of the Proposed Feature Selection Model

In this phase, two different experiments were conducted. The first experiment was designed for observing the performance of the proposed GAWFS+ELM model, against that of conventional filter and wrapper based methods. Two conventional filter FS methods, ANOVA F-test and MI, and two wrapper FS techniques (RFE and SFS) were selected for the comparison, and were applied to raw features (578 in total) for deriving the respective optimal feature sets. The ELM model was used as an estimator/classifier model for each FS method. After the top-10 lab events were generated by each FS technique, the respective feature sets generated by each technique were used for training a base ELM model. The performance of each of these models were compared against the proposed GAWFS+ELM model. Standard metrics like accuracy, precision, recall, F-score and AUROC were used for comparative evaluation of performance. The validated results for the ELM model trained with feature sets generated by GAWFS, ANOVA, MI, RFE and SFS feature selection techniques are shown in Table 4.3.

Next, in the second experiment, our objective was to observe the performance of the proposed model when no feature selection is used. Towards this, an ELM architecture was trained on the original feature set (578 raw features) without using any FS technique. The prediction performance was compared with that of the proposed GAWFS+ELM model. The results of this experiment are tabulated in Table 4.4.

Table 4.3: Comparison of ICU mortality prediction performance of the proposed GAWFS model and other Traditional FS techniques, using ELM as the estimator

Metric	*ANOVA*	*MI*	*RFE*	*SFS*	***GAWFS***
Accuracy	0.70	0.70	0.72	0.71	**0.75**
Precision	0.74	0.73	0.74	0.74	**0.81**
Recall	0.70	0.70	0.71	0.71	**0.74**
F-Score	0.71	0.71	0.72	0.72	**0.76**
AUROC	0.75	0.74	0.76	0.76	**0.80**

From Tables 4.3 and 4.4, it is clear that the feature selection using the proposed GAWFS technique was most effective for prediction and achieved the best performance with respect to all metrics in contrast to other FS techniques as well

Table 4.4: Comparison of ICU Mortality Prediction Performance - *Base ELM+original feature set* vs. *GAWFS+ELM*

Metric	*Base ELM+OriginalFS*	***GAWFS+ELM***
Accuracy	0.71	**0.75**
Precision	0.70	**0.81**
Recall	0.71	**0.74**
F-Score	0.70	**0.76**
AUROC	0.74	**0.80**

as over the model that used only raw features for prediction. Hence, we conclude that the reduced feature set selected by the proposed GAWFS technique consists of very relevant features or lab events that contribute the most significant patient-specific information, due to which a mortality prediction model trained on it can be effective in real world scenarios too. More importantly, in the case of a real-world CDSS application, a major advantage is foreseen as only 10 features (lab tests/events) need to be measured for each patient thus eliminating wasteful or insignificant lab tests. This can result in significant reduction in costs and unnecessary hospital resource consumption, in addition to making predictions comparatively faster, with better accuracy.

4.3.5.2 Benchmarking against Traditional Mortality Scoring Systems

Several traditional scoring methods are already in use in real-world ICUs, which are primarily parametric mortality scores. To evaluate the effectiveness of the proposed GAWFS+ELM model, we benchmarked the performance of the proposed model against that of four traditional scoring systems, SAPS-II, SOFA, APS-III and OASIS. For each patient in the selected cohort, we implemented each traditional score based MPM using scores generated based on the lab event data from the MIMIC dataset (Johnson *et al.*, 2016) and the mortality risk prediction results were obtained.

For SAPS-II, SOFA and OASIS scores, we calculated the mortality risk or the probabilities of mortality as explained in the previous work, in Section 4.2.4. For SAPS-II and OASIS, the probabilities of mortality were calculated as per Eq. 4.1 and 4.2 respectively, and for SOFA, it was done as per method explained by Pirracchio *et al.* (2015) (see Section 4.2.4).

The threshold of classification for APS-III based mortality probabilities (as similar to others) was considered to be 0.5. Similarly, for APACHE-III (APS III), the mortality probability for each patient is calculated as per Eq. 4.4 (Knaus *et al.*, 1991), where, P_m is the required mortality probability of a patient and APS is the APS-III score of the patient.

$$\log(P_m/1 - P_m) = -4.4360 + 0.04726 * APS \tag{4.4}$$

The results of this experiment are summarized in Table 4.5. It can be observed that the proposed GAWFS+ELM model outperformed all the traditional scoring systems considered for the comparison – SAPS-II, SOFA, OASIS and APS-III, by 15-20% in terms of accuracy, while the observed AUROC improvement was about 11-29%. The superiority of the proposed prediction models trained on a highly relevant feature set is evident from the tabulated results in terms of all other metrics considered. A plot of ROC curves for the proposed model and also the standard scoring systems is shown in Fig.4.6. It can be observed in the plot that the area under Receiver Operating Characteristic (ROC) curve for the proposed GAWFS+ELM model is significantly higher than the standard scoring models.

Table 4.5: Comparison of ICU mortality prediction performance of the proposed GAWFS+ELM model with traditional severity scores – SAPS-II, SOFA, OASIS and APS-III

Metric	***GAWFS+ELM***	*SAPS-II*	*SOFA*	*OASIS*	*APS-III*
Accuracy	**0.75**	0.65	0.63	0.62	0.62
Precision	**0.81**	0.66	0.62	0.67	0.67
Recall	**0.74**	0.65	0.63	0.62	0.61
F-Score	**0.76**	0.59	0.57	0.51	0.49
AUROC	**0.80**	0.72	0.62	0.64	0.67

4.3.5.3 Comparison with State-of-the-art ML based MPMs

We conducted other experiments to benchmark the performance of the proposed GAWFS+ELM model against the current state-of-the-art works in this domain, like the models proposed by Calvert *et al.* (2016*b*), Calvert *et al.* (2016*a*), Grnarova *et al.* (2016), Harutyunyan *et al.* (2017) and Che *et al.* (2018). These state-of-the-art ML based models were developed and benchmarked on the MIMIC-III dataset. For each of these models, we re-generated the cohorts as depicted in the respective

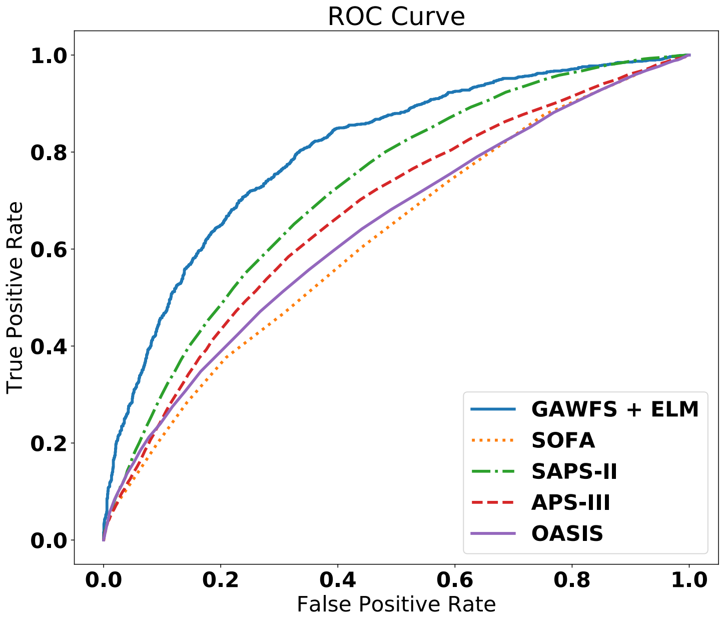


Figure 4.6: Observed AUROC performance of proposed GAWFS+ELM model and traditional severity scores – SAPS-II, SOFA, OASIS and APS-III

models to the highest precision possible. Cohort generation and comparison was carried out in a manner similar to that of Johnson *et al.* (2018). The proposed GAWFS+ELM model was then applied to the patient cohorts used by each of these works. The metrics, prediction accuracy and AUROC were considered for experimental evaluation, the results of which are tabulated in Table 4.6.

It can be observed that the proposed GAWFS+ELM model outperformed all the state-of-the-art models in terms of both prediction accuracy and AUROC. It is to be noted that, some of the state-of-the-art models – Grnarova *et al.* (2016); Harutyunyan *et al.* (2017) and Che *et al.* (2018) did not report their model's prediction accuracy values, due to which we are unable to provide these values in Table 4.6. We conclude that the proposed model was effective in identifying the most optimal set of lab events to be performed for each patient, in order to achieve cost, time and performance improvements over state-of-the-art models.

4.3.5.4 Statistical Significance Testing of the Proposed Model

To further validate the proposed model's improved performance in comparison to both traditional scoring systems and the state-of-the-art machine learning models, the GAWFS+ELM model was subjected to statistical significance testing. Each model under evaluation, including the proposed as well as the state-of-the-art,

Table 4.6: Comparison of ICU mortality prediction performance of proposed GAWFS+ELM model with state-of-the-art ML based models

Study Cohort	*Study Accuracy*	***GAWFS+ELM Accuracy***	*Study AUROC*	***GAWFS+ELM AUROC***
Calvert *et al.* (2016*a*)	0.80	**0.90**	0.88	**0.90**
Calvert *et al.* (2016*b*)	0.81	**0.92**	0.93	**0.94**
Grnarova *et al.* (2016)	–*	**0.96**	0.96	**0.97**
Harutyunyan *et al.* (2017)	–*	**0.92**	0.86	**0.90**
Che *et al.* (2018)	–*	**0.98**	0.84	**0.96**

**Note: The authors of this study reported only AUROC performance in their paper, due to which we are unable to provide prediction accuracy values in Table 4.6.*

was executed for a predefined number of rounds (10 rounds), and a standard-size sample of each model's results during each round, with reference to all evaluation metrics used was collected. Interestingly, it was observed that the result samples were also normally distributed. Therefore, to check if there is a statistically significant difference between the proposed model and the models under comparison, we performed the Student's t-test (Efron, 1969).

The Student's t-test is a statistical hypothesis testing technique that can be used when the samples taken for a normally distributed dataset are small and its standard deviation is not known. To start with, a null hypothesis H_0 was considered, which indicates that there is no statistically significant difference between the result samples of the proposed and existing models. The Student's t-test was performed for the proposed model against each of the existing standard scoring systems, with a significance level of 5% and it was found that the p-value was lesser than 0.04 for all the metrics. Due to this, the null hypothesis H_0, was rejected for all the cases, which means that, there is a statistically significant difference between the performance metrics of the proposed model and that of the existing models. It is also to be noted that the significance level is 5%, i.e., for 95% of the times, the performance of the proposed model is significantly different from that of the existing models. The results of the test of the proposed model against traditional severity scoring models are tabulated in Table 4.7 and that against state-of-the-art machine learning based models in Table 4.8.

Table 4.7: Student's t-test results for Statistical Significant Difference measurement between metric samples of proposed GAWFS+ELM model and traditional severity score based mortality prediction systems - SAPS-II, SOFA, OASIS and APS-III

Metrics →		*Accuracy, Precision, Recall, F-score, AUROC*			
Method →		*SAPS-II*	*SOFA*	*OASIS*	*APS-III*
GAWFS + ELM	**P Value**	<.00001	<.00001	<.00001	<.00001
	Decision[*]	Reject	Reject	Reject	Reject
	Significant Diff	Yes	Yes	Yes	Yes

[*] *Significance level = 0.05*

Table 4.8: Student's t-test Results for Statistical Significant Difference measurement for Accuracy and AUROC metrics of proposed GAWFS+ELM model and different state-of-the-art ML based Models

Metrics →		*Accuracy, AUROC*				
Method →		Calvert *et al.* (2016*a*)	Calvert *et al.* (2016*b*)	Che *et al.* (2018)	Grnarova *et al.* (2016)	Harutyunyan *et al.* (2017)
GAWFS + ELM	**P Value**	<.02	<.04	<.04	<.01	<.01
	Decision[*]	Reject	Reject	Reject	Reject	Reject
	Significant Diff	Yes	Yes	Yes	Yes	Yes

[*] *Significance level = 0.05*

4.3.5.5 Discussion

Based on the results of the validation experiments, several interesting observations can be made. Firstly, the proposed approach ensures that only a reduced set of features or lab events, selected by the proposed GAWFS technique, need to be measured for effectively predicting the mortality risk of a patient. For supporting this claim, we consider the patient with SUBJECT_ID: 22 in the MIMIC III dataset. The patient has spent only a single day in ICU, but the number of labevents measured amounts to 80. Although this includes several labevents pertaining to the condition he or she is suffering from, the mortality risk estimation, being one of the first tasks performed for a patient in ICU will get delayed due to the wait time associated with other lab tests. Most existing ML based CDSS systems require a large number of features or labevents to be available to make predictions with reasonable accuracy. However, in our proposed approach, only the labevents selected by the GAWFS (10 in this case) need to be measured and input to the CDSS for mortality prediction, while still ensuring a good prediction performance (AUROC of 0.80).

Secondly, from Table 4.2, it is evident that the chosen patient cohort exhibits significant class imbalance. Due to the availability of lower number of samples with positive mortality labels (*expire_flag = 1*), the F-score and AUROC metrics are of crucial relevance as they actively measure the model's precision in true positive mortality prediction, i.e., patients at high mortality risk actually, predicted correctly as having high mortality risk. The high values of F-score and AUROC of the proposed model in comparison to that of traditional severity scores currently in popular use (SAPS-II, SOFA, APS-III & OASIS), means that our model can effectively capture latent relationships of features and lab events to predict mortality even in case of class imbalance exhibited by the data. Our model outperformed state-of-the-art machine learning models (Calvert *et al.*, 2016*a*,*b*; Grnarova *et al.*, 2016; Harutyunyan *et al.*, 2017; Che *et al.*, 2018) by a significant margin, thus underscoring its superior performance in making precise predictions for patients at higher mortality risk. Based on observed experimental results (Tables 4.5 & 4.6) and the statistical significance test results (Tables 4.7 & 4.8), it can be thus be conclusively stated that the proposed mortality prediction model can be very effective as a real-world CDSS, and can also help in effective decision towards reduced lab events. Thus, it can contribute positively to patient care and aid in making intelligent decisions in a more effective and productive way. In summary, the experimental and statistical significance test results highlight the suitability of the proposed model for use in real-world ICU mortality prediction CDSSs due to its ability to reduce lab events or features to be considered for early mortality risk prediction. Also of significant importance is the efficacy of the ELM neural network architecture that enables higher prediction accuracy while lowering medical resource consumption footprint.

4.4 Summary

In this chapter, as part of solving the defined problem, i.e., to derive an optimal subset of features for effective ICU mortality risk prediction, two studies were performed. The first one was an empirical study on the effect of feature modeling on features of traditional parametric severity scoring based MPMs on ICU mortality risk prediction performance. The proposed model was built on the Recursive Feature Elimination technique (RFE) for deriving optimal clinical variables/features and Random Forest classifier (best performing) was used to predict mortality of ICU patients. The proposed model was benchmarked against traditional severity scoring based MPMs and it outperformed them by a margin of 12-16% in terms of prediction accuracy. In the second study, a novel GA-ELM feature modeling approach was proposed for capturing the most representative lab events for each individual patient, so that the performance of patient-specific mortality risk prediction can be improved. An ELM based SLFNN architecture was designed to build the prediction model. The proposed GAWFS+ELM model outperformed traditional severity scoring based models as well as state-of-the-art ML based models by a margin of 11-29% and upto 14% in terms of AUROC performance, thereby proving that the modeled patient feature representation was effective in capturing patient-specific nuances. The results of the studies show that it has the capability of improving patient care in ICUs by helping the medical personnel more by providing quicker insights, help the hospitals in judicious use of resources and infrastructure and also help the patients in reducing their hospital costs.

Publications

(based on works presented in this chapter)

1. Gokul S Krishnan and Sowmya Kamath S, "*A Supervised Approach for Patient-specific ICU Mortality Prediction using Feature Modeling*", 7th International Conference on Frontier Computing (FC 2018), Springer, Kuala Lumpur, Malaysia, July 2018. *(Published)*

2. Gokul S. Krishnan, Sowmya Kamath S., "*A novel GA-ELM model for patient-specific mortality prediction over large-scale lab event data*", Applied Soft Computing, Elsevier, Volume 80, 2019, Pages 525-533, ISSN 1568-4946, (SCIE & Scopus, IF: 5.472) *(Published)*

PART III

Mortality Risk Prediction CDSSs using Unstructured Clinical Data

Chapter 5

Individual-centric Mortality Prediction Models for Unstructured Clinical Data

5.1 Introduction

With the rise of EHR adoption rates in developed countries, the availability of structured patient data in the form of EHRs has become abundant. Hence, most existing CDSSs assume EHR availability and are built on EHR data. However, in developing countries like India, clinical experts and caregivers still rely on clinical text notes for decision making. Such clinical notes (e.g. physician notes, nursing notes, discharge summaries, etc.) are primarily unstructured, but contain abundant information on patients' health conditions like status, physiological values, diagnoses and treatments. This forms a significant pool of patient-specific data, which has been explored to a very limited extent for enabling predictive analytics applications like mortality risk prediction and disease prediction. So far, unstructured clinical data which represents a significant volume of clinical data has remained largely unexploited for building predictive analysis models. Big data analytics, NLP and ML can help in developing better CDSSs with these rich clinical information sources leading to significant man-hour and medical resource savings (Belle *et al.*, 2015). In this chapter, two research contributions towards developing mortality prediction based CDSSs that make use of unstructured clinical notes are presented.

5.1.1 Problem Definition

As highlighted in Chapter 4, most parametric scoring systems used in practice, like APACHE-II, SAPS-II, SOFA, OASIS, etc. are considered standard for ICU mortality risk measurement, however, their accuracy is quite low when compared to that achieved by non-parametric methods employing machine learning and other approaches. Although existing non-parametric based approaches like ML based CDSSs have been proven to be better than these traditional systems, the performances of ML based MPMs can be improved to a greater extent by incorporating the ability to process and use clinical data directly. Unstructured clinical data has been shown to contain abundant patient-specific information, which requires effective techniques for extracting such latent information and for leveraging it for the development of CDSSs. Thus, the problem to be addressed here is defined as follows:

> *"Given the low rate adoption of structured EHR in most developing countries, and the availability of abundant unstructured clinical data, design and develop effective preprocessing, feature modeling and prediction modeling approaches for ICU mortality risk prediction based on unstructured clinical notes."*

5.1.2 Motivating Example

Let us consider the example scenario introduced in Chapter 1, Section 1.2.4 once again. As mentioned there, *Hospital B* has a full-fledged EHR system, while *Hospital C* follows a 'semi-EHR' system. During a normal work day, nurses and doctors of both hospitals feed in clinical notes to the hospital information management system. MRD Staff of *Hospital B* manually read, extract and transcribe the relevant patient data like readings, lab test values, vital signs etc into a well-defined structure, which is a highly time and labour-intensive process. The coded data is then stored and is consumed for supporting intelligent applications like mortality risk prediction CDSS implemented in the hospital. Thus, there exists a significant delay in the process of mortality risk prediction, during which the patient condition may also worsen.

In view of this, if clinical data could be processed primarily in its unstructured form, as it is generated, this delay could be avoided, in addition to enormous reduction in cost and labour involved in manual structured EHR conversion processes that would be mandatorily followed in organizations like *Hospital B*. Thus, in *Hospital C*, the raw clinical notes fed in by the medical personnel in the form of unstructured text could be processed by CDSSs equipped with NLP and textual processing capabilities, and valuable patient-specific insights extracted can be consumed to support predictive analytics applications. Furthermore, the manual process of converting the clinical notes into structured form could be potentially eliminated, resulting in substantial savings for hospital management. Hence, there is a critical need for such intelligent CDSSs that can take raw unstructured clinical data as input for enabling fast and accurate decision-making.

In this chapter, our work towards designing two CDSS approaches that focus on utilizing the rich source of information available in latent form in unstructured clinical notes for effective ICU mortality risk prediction is presented. The first approach makes use of Electrocardiogram (ECG) text reports for predicting mortality risk of cardiac patients and its performance was compared to that of traditional parametric severity scoring based MPMs. The second approach is a study on the performance of various word embedding models that can be used for feature modeling, for building MPMs based on clinical text.

5.2 Mortality Risk Prediction using Unstructured Electrocardiogram Text Reports

In this section, an ICU Mortality risk prediction model that utilizes patients' unstructured ECG text reports is presented. The methodology adopted for the design of the proposed MPM is composed of several processes, which are depicted in Figure 5.1.

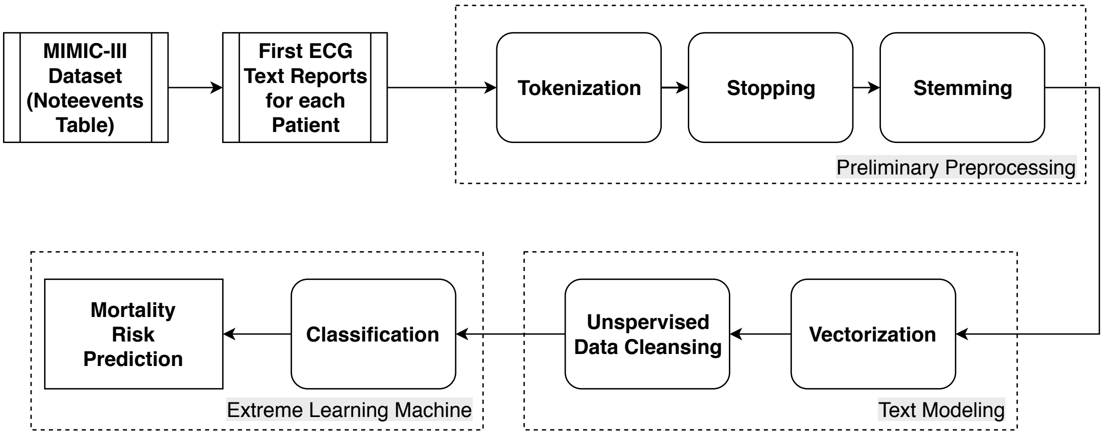

MIMIC-III Dataset (Noteevents Table)
First ECG Text Reports for each Patient
Tokenization
Stopping
Stemming
Preliminary Preprocessing
Vectorization
Unspervised Data Cleansing
Text Modeling
Classification
Mortality Risk Prediction
Extreme Learning Machine

Figure 5.1: Proposed methodology for ICU Mortality Prediction using Unstructured Clinical Notes

5.2.1 Dataset & Cohort Selection

For validating the proposed model, unstructured text data from the open and standard dataset MIMIC-III (Johnson *et al.* (2016)) was used. As stated earlier, MIMIC-III consists of de-identified health data of 46,520 critical care patients. The unstructured clinical text records of these patients are extracted from the 'noteevents' table in the MIMIC-III dataset, from which only the ECG text reports are selected. We have considered only the first ECG report of each patient, based on the requirement that a CDSS must predict risk with the earliest detected condition, enabling fast diagnosis and intervention. The dataset containing the first ECG text reports of 34,159 patients and the corresponding mortality labels, was then subjected to preprocessing using NLP techniques (Details of ECG text corpus summarized in Table 5.1a).

Table 5.1: Dataset Statistics

(a) ECG Text Corpus Statistics

Feature	Total
ECG Reports	34,159
Sentences	108,417
Total Words	802,902
Unique Words	33,748

(b) Statistics of the selected patient cohorts

Data	Total	Alive	Expired
Full ECG Text corpus	34,159	30,464	3,695
Cluster C_1	22,974	20,372	2,602
Cluster C_2	11,185	10,092	1,093
Final Cohort	21,465	20,372	1,093
Training & Test sets	10,155	8,068	2,087
Validation set	2,539	2,024	515

5.2.2 Preliminary Preprocessing

In the next phase, the ECG text corpus is subjected to processing via a NLP pipeline consisting of tokenization, stopping and stemming. During tokenization, the clinical natural language text is split into smaller units called tokens. Generated tokens are filtered to remove unimportant terms (stop words) and finally, stemming is performed on the remaining tokens for suffix stripping. After the initial preprocessing, the tokens are next processed for modeling any latent clinical concepts effectively, during the Text Modeling phase.

5.2.3 Text Modeling

The Text Modeling phase consists of two additional levels of processing - Vectorization and Unsupervised Data Cleansing, which are discussed in detail next.

Vectorization. NLP techniques are critically important in a prediction system based on unstructured data, for generating machine processable representations of the underlying text corpus. Traditional rule and dictionary based NLP techniques, though perform well for certain applications, are not automated and require significant manual effort in tailoring them for various domains. Recent trends in

ML and Deep Learning models and their usage in addition to traditional NLP techniques provide a good avenue for exploiting their performance for improved prediction. However, the effectiveness and performance of such models depend heavily on the optimized vector representations of the underlying text corpus.

Several approaches have been developed for creating meaningful vector representations from text corpus, the prominent ones being Document Term Frequency vectorization and Term frequency-Inverse document frequency (Tf-Idf) Vectorization (Salton and Buckley, 1988). Mikolov *et al.* (2013)'s Word2Vec model is based on two shallow neural network models that are trained on large text corpora, with the objective of mapping each word to a particular dimension in vector space based on its semantic and context similarity, thus, resulting in word vectors of several hundred dimensions. Word2Vec models comprise two distinct approaches – Skipgram and Continuous Bag-of-Words (CBOW). The CBOW model predicts a word from a given window of words (or rather, a context) while the Skipgram model tries to predict the context words, given a target word.

For modeling such latent concepts in the ECG text report corpus, we employed Word2Vec to generate a word embeddings matrix, which consists of the syntactic and semantic textual features obtained from the unstructured ECG corpus. The skip-gram model of Word2Vec was chosen over Continuous Bag-Of-Words (CBOW), due to its effectiveness with infrequent words and also as the order of words is important in the case of clinical reports (Mikolov *et al.*, 2013). We used a standard dimension size of 100, i.e., each ECG report is represented using a 1 x 100 vector, thus resulting in a final matrix of dimension 34159 x 100, each row representing the latent concepts in the ECG report of a specific patient.

Unsupervised Data Cleansing. The vectorized ECG text corpus data is next subjected to an additional process of data cleansing, for identifying special case data points and conflicting records. For this, K-Means Clustering was applied on the vectorized data to cluster the data into two clusters (k=2, as the proposed prediction model is a two-class prediction, '*alive*' and '*expired*' patients) after which a significant overlap was observed in the two clusters. Cluster C_1 contained records of 20,372 alive and 2,602 expired patients while cluster C_2 had 10,092 alive and 1,093 expired patients. As a significant number of the data points representing '*alive*' patients were in cluster C_1, we derived a reduced patient cohort that consists of all '*alive*' patients from cluster C_1 and all '*expired*' patients from cluster C_2, which were then considered for building the prediction model. The remaining patient data points exhibited anomalies due to the existence of patients who might have expired due to causes not related to heart. After this processing, the final patient cohort now consisted of 20,372 alive and 1,093 expired patients (tabulated in Table 5.1b).

5.2.4 Language Modeling based Mortality Risk Prediction

The patient cohort obtained after the data cleansing process is now considered for building the prediction model. Towards this, we designed a neural network model that is built on a fast learning architecture Extreme Learning Machine (ELM) (Huang *et al.*, 2015), as explained in the previous chapter (see Section 4.3.4). ELM is a single hidden layer Feedforward Neural Networks (SLFNN), where the parameters that fire the hidden layer neurons don't require tuning (Huang *et al.*, 2015). The hidden nodes used in ELM fire randomly and learning can be carried out without any iterative tuning. Essentially, the weight between the hidden and output layers of the neural network is the only entity that needs to be learned, thus resulting in an extremely fast learning model. Different implementations of ELMs have been used for tasks like supervised and unsupervised learning, feature learning etc, but to the best of our knowledge, ELMs have not been applied to unstructured clinical text based prediction models. In this SLFNN architecture, we set the number of nodes in the input layer to 100 as the feature vectors obtained after Word2Vec modeling are of similar dimensions. The hidden layer consists of 50 nodes and a single node is used at the output layer, to generate the predicted mortality risk of a patient. The Rectified Linear Unit (ReLU) activation function was used in the layers of the proposed ELM architecture as it is a step function and predicts the slightest chance of mortality. The architecture of ELM, being similar to as described in the previous chapter, is illustrated in Fig. 4.5. During training, the weights between the hidden and output layers are iteratively learned and optimized. Finally, the patient-specific mortality prediction is obtained at the output layer.

5.2.5 Experimental Results and Discussion

For validating the proposed prediction model, an extensive benchmarking exercise was carried out. The experiments were performed using a server running Ubuntu Server OS with 56 cores of Intel Xeon processors, 128 GB RAM, 3 TB Hard Drive and two NVIDIA Tesla M40 GPUs. All implementations were done using Python using packages such as sklearn, gensim[1] (for Word2Vec implementation), Natural

[1] https://radimrehurek.com/gensim/

Language Toolkit (NLTK[2]) and matplotlib. The patient cohort was split into training, test and validation sets (as shown in Table 5.1b). The vectorized feature vectors and the respective mortality labels in the training dataset were used for training the ELM model. We used 10-fold cross validation for all experiments and standard metrics like Accuracy, Precision, Recall, F-score and AUROC (Area under Receiver Operating Characteristic) were used for performance evaluation of the proposed model. Additionally, Matthews Correlation Coefficient (MCC) was also used as a metric, as it takes into account true positives, false positives and false negatives, therefore, regarded as a balanced measure even in presence of class imbalance (Boughorbel *et al.*, 2017).

We also benchmarked the performance of the proposed prediction model against well established, traditional parametric severity scoring methods. The four popular scoring systems used for the study used in Chapter 4 – SAPS-II, SOFA, APS-III and OASIS, were again chosen for this comparison. We implemented and generated the respective scores for each patient in the validation set. For SAPS-II, the mortality probability was calculated as per the process proposed by Gall *et al.* (1993) and Eq. 4.1. For SOFA, the mortality prediction of each patient was obtained by regressing the mortality on the SOFA score using a main-term logistic regression model similar to Pirracchio *et al.* (2015), whereas for APACHE-III (APS III), it was calculated for each patient as per the Knaus *et al.* (1991)'s method based on Eq. 4.4. The mortality probability for each patient as per OASIS scoring system was calculated using the in-hospital mortality score calculation as per Eq. 4.2 defined by Johnson *et al.* (2013). A classification threshold of 0.5 was considered for SAPS-II, APS-III and OASIS.

The validation patient data is fed to the trained model for prediction and its performance was compared to that of traditional scoring methods. The results are tabulated in Table 5.2, from where, it is apparent that the proposed model achieved a significant improvement in performance over all four traditional scores. The proposed model predicted high mortality risk (label 1) correctly for most patients belonging to '*expired*' class, which is a desirable outcome expected out of this CDSS, which is also evident in the high precision values achieved. AUROC, F-Score and MCC are very relevant metrics for this experiment as the data exhibits class imbalance (number of patients in '*alive*' class much greater those in '*expired*' class; see Table 5.1b). MCC, which measures the correlation between the actual and predicted binary classifications, ranges between -1 and +1, where +1 represents perfect prediction, 0 indicates random prediction and -1 indicates total

[2] http://www.nltk.org/

disagreement between actual and predicted values. The high values of F-Score and MCC for the proposed model in contrast to the others, indicates that, regardless of class imbalance in the data, the proposed model was able to achieve a good quality classification for both alive (0) and expired (1) labels.

Table 5.2: Benchmarking proposed ELM model against Traditional severity methods SAPS-II, SOFA, APS-III and OASIS

Models	**Accuracy**	**Precision**	**Recall**	**F-Score**	**AUROC**	**MCC**
Proposed	0.98	0.98	0.98	0.98	0.99	0.84
SAPS-II	0.86	0.87	0.86	0.86	0.80	0.34
SOFA	0.88	0.86	0.88	0.85	0.73	0.22
APS-III	0.89	0.86	0.89	0.86	0.79	0.26
OASIS	0.88	0.86	0.89	0.86	0.77	0.26

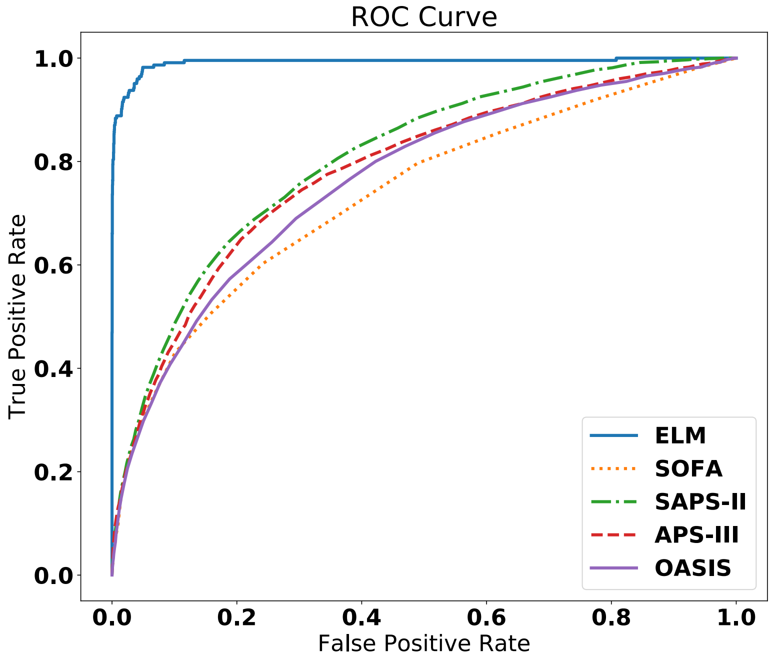


Figure 5.2: Comparison of AUROC performance of the various models

The plot of Receiver Operating Characteristic (ROC) curves generated for all models considered for comparison is shown in Fig. 5.2. The proposed model showed a substantial improvement of nearly 19% in AUROC in comparison to the best performing traditional model, SAPS-II. This indicates that the Word2Vec feature modeling generated quality features and the proposed unsupervised data cleansing approach was successful in filtering out special and anomalous data from the patient data representation, making the features more discriminating and thereby enabling the ELM classifier to generalize and predict the '*alive*' and '*expired*' classes effectively.

5.3 Benchmarking Word Embedding Models for Unstructured Clinical Text based ICU Mortality Risk Prediction

Word Embeddings are typically employed during preliminary NLP processing, for mapping words in a text corpus to a real number in n-dimensional vector space. Word embedding models are built on neural network architectures and have been proven to outperform traditional n-gram based models. Most state-of-the-art word embedding models are inspired by the idea that "*a word is characterized by the company it keeps*" (Firth, 1957). In a nutshell, words are modeled such that, semantically and contextually similar words are mapped closer to each other in the vector space.

Word2Vec (Mikolov *et al.*, 2013), a word embedding model as explained in Section 5.2.3, is an effective approach for generating semantic word embeddings (features) from unstructured text corpus. The generated vectors may be of several hundred dimensions, where unique terms in the text corpus are represented as a vector in the feature space such that corpus terms of similar context are closer to each other (Mikolov *et al.*, 2013). As explained before, Word2Vec models comprise two distinct approaches – Skipgram and Continuous Bag-of-Words (CBOW). The CBOW model predicts a word from a given window of words (or rather, a context) while the Skipgram model tries to predict the context words, given a target word. Pennington *et al.* (2014)'s approach for generating vector representation of words, called Global Vectors (GloVe), uses an unsupervised algorithm trained over word to word co-occurrence statistics of a large text corpus. The model achieved better performance over several other word embedding models for specific corpus data in terms of word analogy, word similarity and named entity recognition tasks (Pennington *et al.*, 2014). Joulin *et al.* (2016) developed an approach called FastText, which uses its own word representation model similar to Word2Vec, with additional rank constraint and fast loss approximation ensuring its faster training capacity for large text corpora. The FastText model claims better quality representation over several state-of-the-art approaches for specific tasks like tag prediction and sentiment analysis (Joulin *et al.*, 2016). Similar to Word2Vec, FastText also provides the Skipgram and CBOW models.

With reference to the document information captured by word embeddings, our aim is to evaluate the applicability of various word embedding models for building effective CDSS applications, when unstructured clinical text notes are available for use. In this section, a study for evaluating quality of word representation models for ICU mortality prediction based on unstructured nursing notes is presented. The workflow adopted for the study is as depicted in Fig. 5.3.

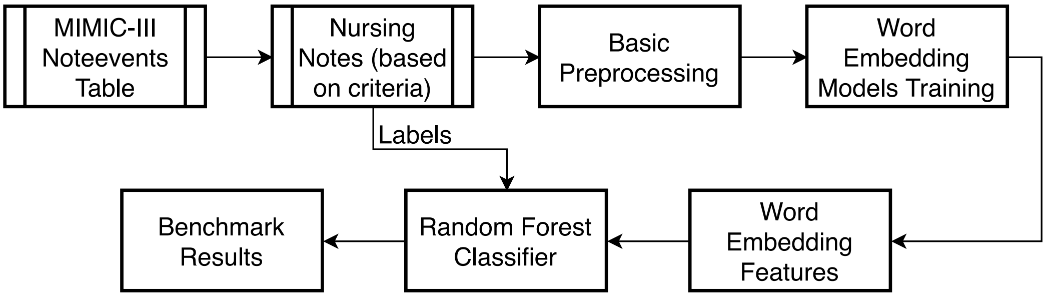

MIMIC-III Noteevents Table
Nursing Notes (based on criteria)
Basic Preprocessing
Word Embedding Models Training
Labels
Benchmark Results
Random Forest Classifier
Word Embedding Features

Figure 5.3: Overall Workflow of Benchmarking Experiment

From the ‘noteevents’ table of MIMIC-III dataset, the nursing reports of all patients were extracted. The patient cohort for this study was selected based on some predefined criteria, similar to the process adopted by existing literature (Pirracchio *et al.*, 2015) (listed below).

1. Only the patients aged above 15 years were selected as paediatrics consists of specific treatment methods.

2. To ensure sufficient data for the system to learn, we included only those patients who spent more than 1 day in the hospital (length of stay > 1).

3. Finally, in case of multiple admissions for a patient, only the first admission of those patients was considered as the system is expected to learn and predict using the patients’ earliest condition possible.

As per the defined criteria, 223,556 text reports pertaining to 5,376 (3,593 alive and 1,783 expired) patients were selected for the study. The characteristics of the nursing notes text corpus of the selected patient cohort are tabulated in Table 5.3. Next, the clinical reports of the selected cohort are preprocessed using various NLP techniques like tokenization, by which the text corpus is broken down into smallest constituent units, i.e., tokens. All special characters except spaces and single quote (’) are ignored (the same process described in Section 5.2.2 is adopted).

After the basic preprocessing, the resultant tokens are fed into corresponding word embedding models for generating a patient representation, which is in turn used to train a classifier to enable patient-specific mortality predictions. Word Embedding models ensure dense representation of the corpus with semantics captured as well, so that words in similar context are mapped and placed in vector space in close proximity. Thus, the final representation of the text corpus can be used as features, for training machine learning classifiers.

Table 5.3: Nursing Notes Text Corpus - Characteristics

Characteristic	**Number**
Nursing Reports	223,556
Sentences	4,774,147
Total Words	80,211,620
Unique Words	434,797

The preprocessed tokens are input to the different word embedding models for training and for semantically mapping words to vector space. Parameters like vector dimension size, initial learning rate, word window size, minimum word count and number of thread workers are most important which are input to the word embedding models. Prediction models built on different popular word embedding models – Word2Vec (Mikolov *et al.*, 2013), FastText (Joulin *et al.*, 2017) and GloVe (Pennington *et al.*, 2014). For Word2Vec and FastText, both Skipgram and CBOW models were trained separately.

Various experiments were performed for observing the effects of variations in dimension size and learning rates on the prediction models and their performance was measured in terms of accuracy. Also, the training time required for various dimension sizes was noted. In the first experiment, for each word embedding model, the training was performed with various dimension sizes – 10, 25, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000. The textual features were generated using each embedding model, and then used to train the classifier. Next, for the same training configuration, the variations in training time with reference to dimension sizes were observed. Based on these two experiments, an optimal dimension size was determined and used in the final experiment. Here, the performance of Random Forest classifier for various initial learning rates of the word embedding models – 0.01, 0.025, 0.05, 0.075, 0.1, 0.125, 0.15, 0.175, 0.2 was monitored, to determine the best initial learning rates of the respective models. This initial rate is configured to later linearly converge to a minimal learning rate of 0.0001 as training progresses.

The mortality prediction model was built on Random Forest classifier (explained in Section 4.2.3), where a standard number of 100 decision trees was used to perform the training for each set of features. Each set of textual features and respective mortality labels were fed into a Random Forest classifier for training. We applied 5-fold cross validation for each model and each model's effectiveness in predicting mortality were measured in terms of standard metrics like – Accuracy, Precision, Recall, F-score and Angle Under Receiver Operating Characteristic (AUROC). Finally, the model which performed the best was identified and benchmarked against four popular traditional severity scores commonly used in hospitals currently – SAPS-II, SOFA, APS-III and OASIS.

5.3.1 Experimental Results and Discussion

As part of experimental validation, the prediction models built on different types of word embeddings were compared, using standard metrics like accuracy, dimension size, learning rate and training time. Each of these studies are described in detail next. The objective was to determine the optimal dimension and learning rate for highest prediction accuracy, so that the best-performing model can be benchmarked against traditional mortality scores. The experiments were performed on a server running Ubuntu 16.04 LTS with 56 cores of Intel Xeon Processors, 128 GB RAM, 3 TB HDD memory and 2 Tesla M40 GPUs. All implementations were carried out using Python and the Gensim implementations of Word2Vec[3] and FastText[4] models were used. A Python implementation of Glove called glove-python[5] was used for its implementation for the experiments.

5.3.1.1 Determining Optimal Dimension & Learning Rate

Different experiments were performed for analyzing the quality of each word embedding model, as discussed below -

1. *Accuracy vs Dimension.* To study the effect of dimension on performance, the vector dimension size was varied, while training the word embedding models to generate word embedding vectors of the corpus. The training was performed with different dimension sizes – 10, 25, 50, 100, 200, 300, 400,

[3] Gensim Word2Vec – `https://radimrehurek.com/gensim/models/word2vec.html`

[4] Gensim FastText – `https://radimrehurek.com/gensim/models/fasttext.html`

[5] GloVe-Python – `https://github.com/maciejkula/glove-python`

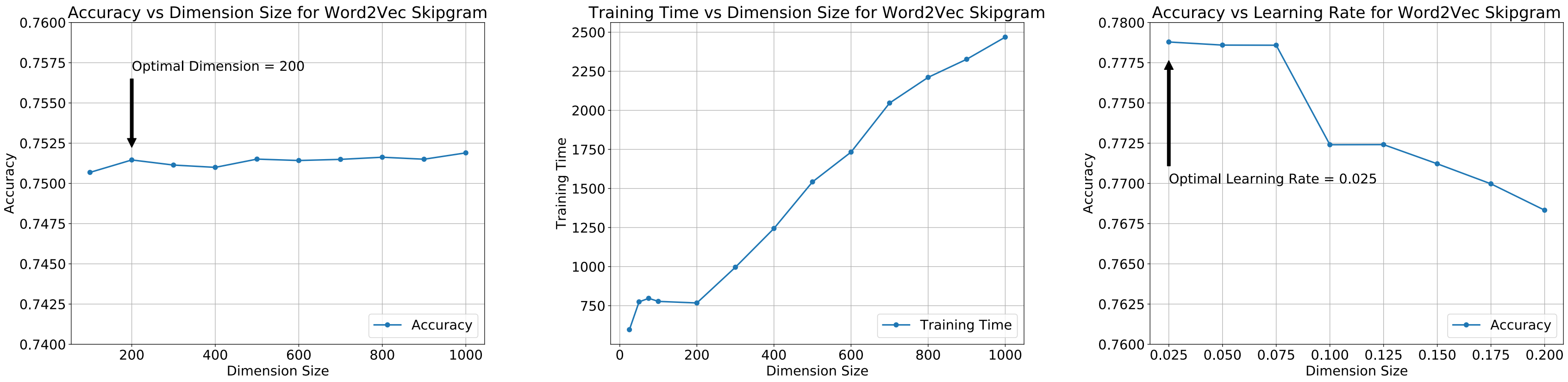
500, 600, 700, 800, 900, 1000; with an initial learning rate of 0.025 and the number of epochs fixed at 5. Next, each set of vectors generated were used as features for training a Random Forest Classifier and the performance of each classification model was observed.

2. *Training Time vs Dimension.* We also performed an analysis on how the training time varies for each word embedding model with variation in dimension sizes. The training was performed with the same conditions as in Experiment 1. It is also to be noted that a default fixed number of 3 worker threads were used for training the respective word embedding models.

3. *Accuracy vs Learning Rate.* After the optimal dimension size was determined (in Experiment 1 & 2), the initial learning rate used for training the word embedding models was varied, while keeping the dimension constant. The training was performed at learning rates – 0.01, 0.025, 0.05, 0.075, 0.1, 0.125, 0.15, 0.175, 0.2; for each, at a fixed epoch of 5. Then, word embedding vectors generated for the corpus were used to train one Random Forest Classifier respectively and the performance was observed.

The results of the experiments conducted for the Word2Vec (Skipgram and CBOW), FastText (Skipgram and CBOW) and GloVe word embedding models are illustrated in Fig. 5.4, 5.5, 5.6, 5.7 and 5.8 respectively. It can be observed that the Word2Vec Skipgram model performed the best at three dimensional sizes – 300, 500 and 800 with a difference of less than 0.01 in terms of accuracy. Moreover, from dimension size 200 to 1000, the change in training time is almost nearly linear. Hence, considering time and dimension, we chose the optimal dimension to be 300. Next, for this chosen dimension size of 300, we varied the initial learning rates and identified the best initial learning rate as 0.05. The performance at this point (size = 300, initial learning rate = 0.05) was chosen to be the best performance of Word2Vec Skipgram model.

Similarly, for Word2Vec CBOW, the model initially performed best at sizes – 500, 800 and 1000 with negligible performance differences and so, 500 was chosen to be the optimal dimension size, taking training time also into account. For dimension size of 500, the optimal learning rate for the same model was 0.025. The accuracy at this point (size=500, initial learning rate = 0.025) was observed to be the highest for Word2Vec CBOW model. In case of FastText Skipgram and CBOW models, the optimal dimension size was 400 (amongst 400, 600 and 800) and optimal learning rate at size 400, was 0.025 and 0.05 respectively. For

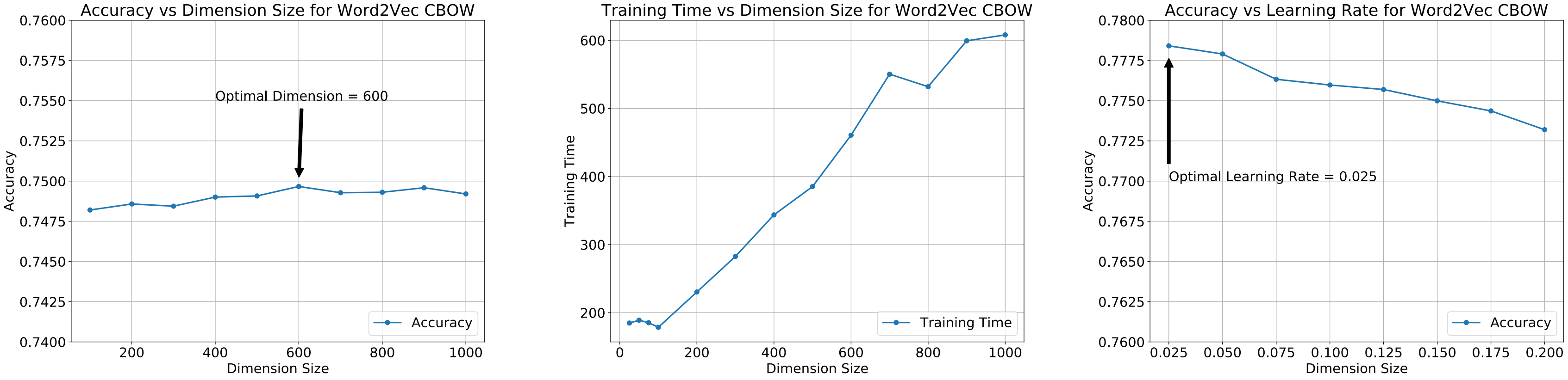


(a) Accuracy vs Dimension Size

(b) Training Time vs Dimension Size

(c) Accuracy vs Learning Rate

Figure 5.4: Performance of Word2Vec Skipgram Word Embedding Model

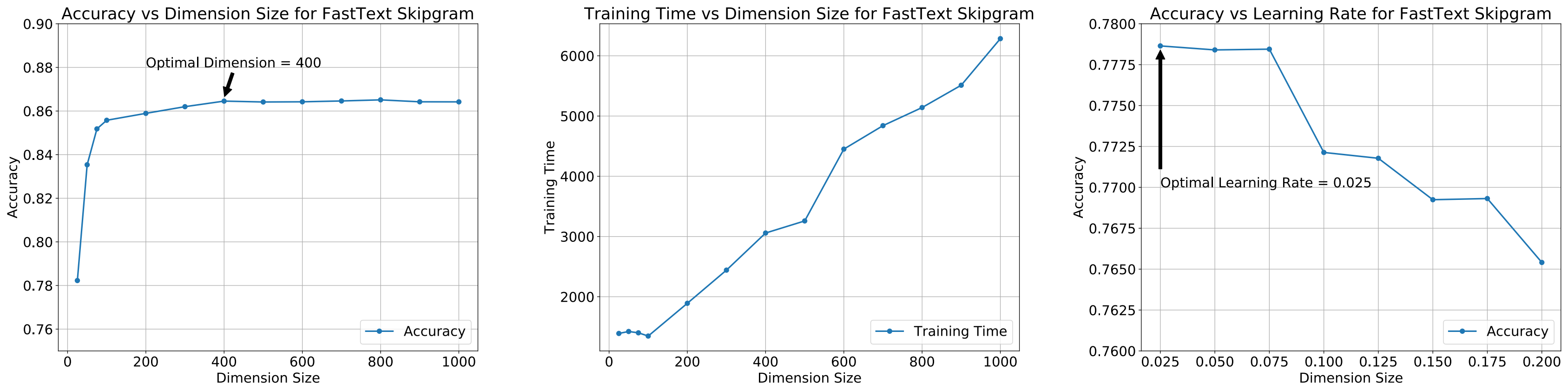


(a) Accuracy vs Dimension Size

(b) Training Time vs Dimension Size

(c) Accuracy vs Learning Rate

Figure 5.5: Performance of Word2Vec CBOW Word Embedding Model

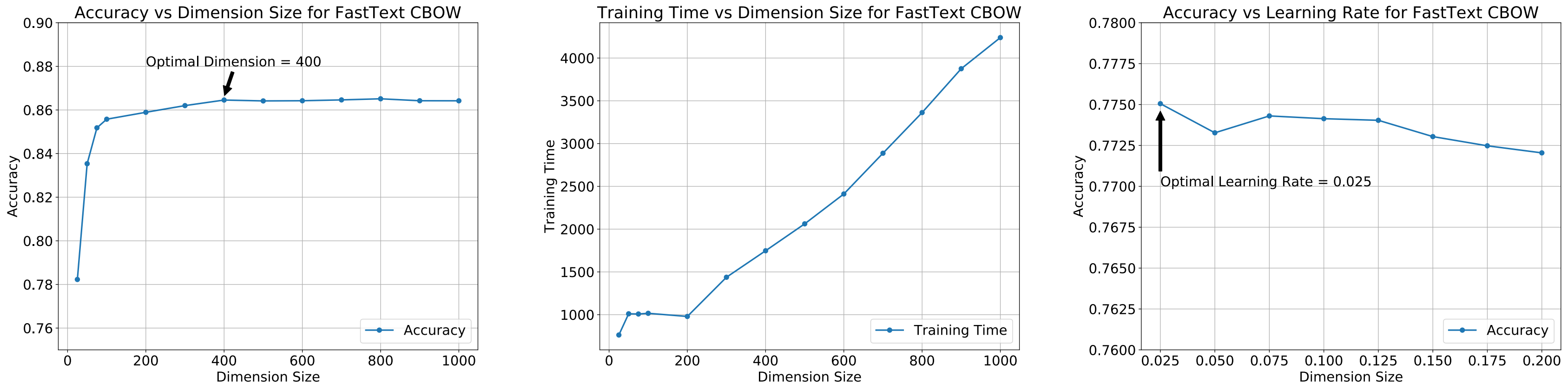


(a) Accuracy vs Dimension Size

(b) Training Time vs Dimension Size

(c) Accuracy vs Learning Rate

Figure 5.6: Performance of FastText Skipgram Word Embedding Model

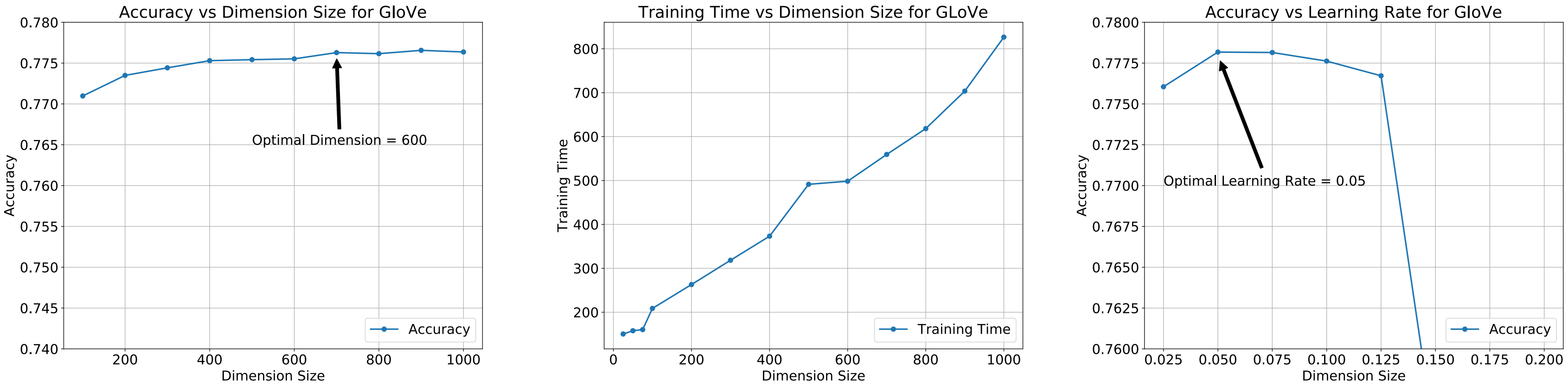


(a) Accuracy vs Dimension Size

(b) Training Time vs Dimension Size

(c) Accuracy vs Learning Rate

Figure 5.7: Performance of FastText CBOW Word Embedding Model



(a) Accuracy vs Dimension Size

(b) Training Time vs Dimension Size

(c) Accuracy vs Learning Rate

Figure 5.8: Performance of the GloVe Word Embedding Model

Table 5.4: Experimental results for various Word Embedding Models considered for the evaluation study using Random Forest Classifier

Model	**Accuracy**	**AUROC**	**TT**	**DS**	**LR**
W2V Skipgram (*W2V_Skip_RF*)	0.864	0.872	1220	300	0.05
W2V CBOW (*W2V_Cbow_RF*)	0.853	0.861	443	500	0.025
FT Skipgram (*FT_Skip_RF*)	0.865	0.873	3309	400	0.025
FT CBOW (*FT_Cbow_RF*)	0.843	0.856	2241	400	0.05
GloVe (*Glove_RF*)	0.851	0.860	659	800	0.125

TT: Training Time (in seconds)
DS: Vector Dimension Size
LR: Learning Rate

GloVe, the optimal dimension size was 800 (amongst 800 and 1000) and optimal learning rate at size 800, was found to be 0.125. Thus, their accuracy at these points were chosen to be the best for FastText and GloVe models respectively. The optimal performances of each word embedding model are tabulated in Table 5.4. Both Word2Vec Skipgram and FastText Skipgram performed equally well, with a negligible difference of 0.1%. As CDSS and other predictive applications generally require retraining as new patient data is available continuously, the Word2Vec Skipgram model was chosen as the best model, due to its lower training time and dimension size.

5.3.1.2 Benchmarking against Traditional Mortality Scores

As can be observed from the experimental results tabulated in Table 5.4, the best performing mortality prediction model was the one built on Word2Vec Skipgram word embeddings and trained using Random Forest classifier, i.e., *W2V_Skip_RF*. The final objective of our work is to benchmark its prediction performance against the four popular traditional mortality scores considered in the previous chapters, SAPS-II, SOFA, APS-III and OASIS. The scores for SOFA, SAPS-II, APS-III and OASIS are calculated for all the 5376 patients using the methods and formulae provided by Gall *et al.* (1993), Knaus *et al.* (1991) and Johnson *et al.* (2013) respectively. The observations of the comparative evaluation are tabulated in Table 5.5.

From Table 5.5, it can be seen that the *W2V_Skip_RF* mortality prediction model built utilizing unstructured nursing notes for capturing patient specifics,

Table 5.5: Comparison of *W2V_Skip_RF* with Traditional Mortality Scores

Model	**Accuracy**	**Precision**	**Recall**	**F-Score**	**AUROC**
W2V_Skip_RF	0.864	0.874	0.863	0.865	0.872
SAPS-II	0.608	0.618	0.608	0.569	0.691
SOFA	0.60	0.598	0.60	0.577	0.629
APS-III	0.565	0.603	0.565	0.448	0.672
OASIS	0.581	0.626	0.581	0.486	0.639

W2V_Skip_RF : Word2Vec Skipgram based Random Forest Classifier

performs significantly better than traditional formulae based severity scores. This emphasises the *W2V_Skip_RF* model's ability to capture the nuances of the varied patient-specific information available in unstructured nursing text notes. The Word2Vec Skipgram model was able to effectively capture these as discriminating features so that the classifier could be trained with high quality training data. Moreover, this shows that leveraging unstructured clinical text notes in the development of CDSS is possible and viable in real-life hospital scenarios and also ensure that conversion to a structured EHR format is not required, thereby saving time, man hours and scarce medical resources.

Various experiments on studying the effect of vector dimension size, training time and initial learning rate on prediction accuracy were performed. The best model identified from Tables 5.4 and 5.5, the Word2Vec Skipgram model based Random Forest Classifier (*W2V_Skip_RF*) when benchmarked against four traditional severity scores currently in use in hospitals, outperformed them by over 43-52% in terms of prediction accuracy. This significant margin highlights the fact that the nuances of patient-specific information present in the unstructured nursing notes was best captured by the Word2Vec Skipgram model, making it suitable for development of better CDSS applications.

5.4 Summary

In this chapter, two CDSS approaches for ICU mortality risk prediction, that are modeled using the latent patient-specific information automatically extracted from unstructured clinical notes, were discussed. In the first approach, ECG text reports available in a standard open dataset were processed and modeled for assessing and predicting risk of mortality of cardiac patients admitted to cardiac ICUs. The second approach was a benchmark study to understand the contribution of various word embedding models for effectively modeling information present in the natural language used in the clinical notes and experimenting with feature modeling strategies for building MPMs based on clinical text. The performances of proposed MPMs were compared against that of traditional severity scoring based MPMs and it was observed the unstructured text based MPMs significantly outperformed traditional mortality scores by a large margin.

Publications

(based on work presented in this chapter)

1. Krishnan, Gokul. S., & Kamath, Sowmya. S., "*A Supervised Learning Approach for ICU Mortality Prediction Based on Unstructured Electrocardiogram Text Reports*", In the proceedings of the 23rd International Conference on Applications of Natural Language to Information Systems (NLDB 2018), Springer, Paris, France, June 2018. (CORE Ranked) *(Published)*

2. Gokul S. Krishnan and Sowmya Kamath S. "*Evaluating the quality of word representation models for unstructured clinical Text based ICU mortality prediction*". In Proceedings of the 20th International Conference on Distributed Computing and Networking (ICDCN '19), ACM, IISc Bangalore, January 2019. (CORE Ranked) *(Published)*

PART IV

Disease Prediction CDSSs using Unstructured Clinical Data

Chapter 6

Individual-Centric Disease Prediction Models for Unstructured Clinical Data

6.1 Introduction

Based on our previous work on designing MPMs using unstructured patient records (presented in Chapter 5), it was observed that there is a huge potential for reducing dependency of hospitals on structured/processed patient records. Additionally, the richness of the patient data that lies latent in unstructured clinical notes is eminently suitable for use towards the development of CDSSs, resulting in significant improvement in accuracy, along with time and cost savings. We explored this avenue further and our next focus is to design generic Disease Prediction Models (DPMs) built on unstructured clinical notes/text.

Disease diagnoses depend on various tests on patients such as labevents, readings from frequent monitoring, patient's history of illness, etc. Most existing disease prediction approaches assume the availability of structured EHRs/patient data. The downside of this is that, in a real-world scenario, the dependency on first having to measure all values of labevents/readings, manually code them to a structured form, only after which prediction is performed, introduces an inevitable delay, which might result in deterioration in the patient's condition. Thus, automated disease prediction models with low latency that can predict with high accuracy even with minimal patient data and no dependency on structured data availability are the need of the hour.

6.1.1 Problem Definition

ICD9 disease coding is an important task in a hospital, as part of which a trained medical coder with the required domain knowledge assigns disease-specific, standardized ICD9 codes to a patient's admission record. As hospital billing and insurance claims are based on the assigned ICD9 codes, the coding task requires high precision but is often prone to human error. This has manifested in an annual spending of more than $25 billion in the US towards efforts to improve coding efficacy (Xie and Xing, 2018; Farkas and Szarvas, 2008). Hence, automated disease coding approaches are seen as a notable solution to this problem. Currently, this is an area of active research, but the performance that has been recorded so far is below par, underscoring a huge scope for improvement and making automated ICD9 coding an open research problem.

For ICD9 disease coding to be effective, the correct determination of generic disease categories or groups prior to specific coding is very crucial as information regarding generic disease groups is prerequisite knowledge for disease-specific ICD9 coding. Furthermore, existing ICD9 coding methods utilize discharge summaries for coding, just like the human medical coders do. However, as the records are digitized, other clinical notes recorded by the caregivers during the same admission can potentially provide additional patient-specific insights pertaining to the diagnosed diseases. Hence, prediction of ICD9 disease groups not only acts as an additional data source for automated ICD9 coding, but can also be a disease risk estimation model that can provide further insights into the comorbidity aspects and even mortality risk prediction. Hence, the problem to be addressed is defined as follows:

> "*Given the rich latent patient-specific information available in unstructured clinical notes, to design and develop effective preprocessing, feature modeling and prediction approaches for effective ICD9 disease group prediction, to enable disease prediction CDSSs*".

6.1.2 Motivating Example

To describe the prevailing conditions that emphasize the need for disease prediction CDSS based on unstructured clinical notes, we consider scenarios from the running example introduced in Chapter 1, Section 1.2.4. Recall that *Hospital B* has a full-fledged EHR system, while *Hospital C* follows a 'semi-EHR' system. In *Hospital B*, the labevents and reading values are recorded by doctors and nurses as notes. The MRD staff perform the conversion from the unstructured to the structured form as required by the disease prediction CDSS implemented in the hospital, after which provides with probability risks of diseases for a particular patient. This delay in processing the data and generating the predictions can be critical as the patient's condition may worsen and can be avoided if the predictions can be directly generated using the clinical notes. The disease prediction CDSS implemented in *Hospital C* can directly process the unstructured clinical notes recorded by the doctors and nurses, and provide them with instantaneous disease risk predictions for a particular patient. In this way, the MRD staff need not perform conversion to any predefined structure and hence, *Hospital C* has the advantage of significant savings in person-hours and scarce medical resources, not to mention precious lives.

In this chapter, various approaches towards developing effective individual-centric disease group prediction models for ICU patients built on unstructured clinical notes are presented. Our contributions towards the defined problem are in the context of designing processes that can automatically process a variety of unstructured clinical text like physicians' notes, nurses notes, radiology notes etc, with their difference 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 ICD9 code group prediction. The performance of the proposed DPMs are compared to that of state-of-the-art DPMs built on structured patient data.

6.2 Ontology-driven Feature Modeling for Disease Prediction

In this section, a generic disease group prediction model is presented that uses ontology-driven text feature modeling and neural networks for prediction. The overall workflow of the proposed model is as depicted in Fig. 6.1. Radiology reports in unstructured text format from the open and standard MIMIC-III (Johnson *et al.*, 2016) dataset were used for this study. From the 'noteevents' table, only the Radiology notes were extracted for this study. Overall, 194,744 radiology text reports generated during 45,512 admissions of 36,447 patients were included for the study. Often, a patient may be diagnosed with multiple diseases in the same admission, hence, it is necessary for the prediction to be a multi-label prediction task. Therefore, for each radiology report, all disease groups were considered as labels and given binary values - 0 (if the disease was not present) and 1 (if the disease was present).

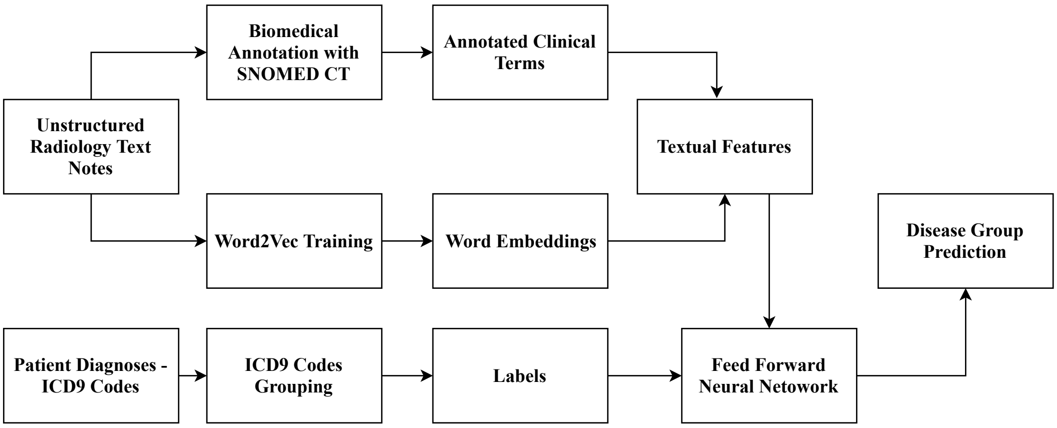

Biomedical Annotation with SNOMED CT
Annotated Clinical Terms
Unstructured Radiology Text Notes
Textual Features
Word2Vec Training
Word Embeddings
Disease Group Prediction
Patient Diagnoses - ICD9 Codes
ICD9 Codes Grouping
Labels
Feed Forward Neural Netowork

Figure 6.1: Proposed Approach

Table 6.1: Charateristics of the Cohort used for Ontology based Clinical Text Modeling

Feature	Total Records
Patients	36,447
Admissions	45,512
Radiology Reports	194,744
Sentences	539,466
Words	45,755,992
Average word Length of Report	235
Unique Diseases	2,593
Disease Groups	21

6.2.1 Preprocessing & Textual Feature Modeling

The radiology reports text corpus were first subjected to a basic NLP pipeline consisting of tokenization and stopping as similar to works in the previous chapter. The tokenization process breaks down the clinical text corpus into tokens and the stopping process filters out unimportant words (stop words) from the corpus. The preprocessed tokens corpus is then fed into a SNOMED-CT ontology based annotator to annotate and extract clinical and biological terms. SNOMED-CT ontology (Snomed, 2011) is an ontology that provides a vocabulary of clinical/biomedical terms and helps extract associated concepts from the preprocessed radiology report corpus. We used the Open BioMedical Annotator (Jonquet *et al.*, 2009) for this purpose, after which 4,366 unique clinical/biological terms were obtained. The presence or absence of each extracted clinical/biomedical term, represented as binary values, is considered as a textual feature representation.

The preprocessed corpus is also used to train a Word2Vec (Mikolov *et al.*, 2013) word embedding model (explained in Section 5.2.3) to extract the word embedding features from the corpus. The Skipgram model of Word2Vec was used for training the corpus as this model takes word ordering into consideration and is effective with infrequent words as well. The Word2Vec Skipgram model is trained with a dimension size of 500 and initial learning rate of 0.01. The word embeddings were extracted such that each report is represented as 1 x 500 vector. The word embedding features were further concatenated with the extracted clinical and biological term features with binary values for each report indicating its presence (1) or absence (0) in the respective reports and the feature matrix was then standardized to values between -1 and 1. These features are used for training the neural network model for disease prediction.

6.2.2 ICD9 Disease Code Grouping

The ICD9 disease codes of patients' diagnoses were retrieved from the 'DIAGNOSES_ICD' table of MIMIC-III dataset and the labels were grouped as per available standards[1] and as previously followed by state-of-the-art work (Purushotham *et al.* (2018)). A total of 2,593 unique ICD9 disease codes were accordingly grouped into 21 ICD9 disease groups (as shown in Table 6.2). As a patient can suffer from multiple diseases, we consider the ICD9 group prediction task as a binary classification of multiple labels. Therefore, 21 different labels (disease groups) were considered with possible binary values: 0 (for absence of the disease) and 1 (for presence of the disease). The 194744 x 4966 feature matrix, 21 ICD9 disease groups were considered as labels to train the neural network model, which is described in the next sub section.

6.2.3 Disease Prediction Model

The feature matrix with both word embedding features and ontologically extracted term-presence features, along with ICD9 group labels are next used for training a Neural network based prediction model. A Feed Forward Neural Network (FFNN) architecture was used to build the prediction model, which is depicted in Figure 6.2. The input layer consists of 2048 neurons with input dimension as 4966 (number of input features); 4 hidden layers with 1024, 512, 256 and 128 neurons respectively and finally an output layer with 21 neurons, each representing an ICD9

[1] Available online `http://tdrdata.com/ipd/ipd_SearchForICD9CodesAndDescriptions.aspx`

Table 6.2: Grouping of ICD9 Codes and Statistics of ICD9 Disease Groups in the MIMIC-III Radiology Corpus Subset

ICD9 Group (Label)	*ICD9 Code Range*	*Description*	*Occurrences in Study Corpus*	*Occurrence Percentage (%)*
1	001 - 139	Infectious & Parasitic Diseases	68,734	35.29
2	140 - 239	Neoplasms	34,668	17.80
3	240 - 279	Endocrine, Nutritional, Metabolic, Immunity	128,357	65.90
4	280 - 289	Blood & Blood-Forming Organs	80,337	41.25
5	290 - 319	Mental Disorders	62,963	32.33
6	320 - 389	Nervous System & Sense Organs	65,149	33.45
7	390 - 459	Circulatory System	152,159	78.13
8	460 - 519	Respiratory System	107,656	55.28
9	520 - 579	Digestive System	84,346	43.31
10	580 - 629	Genitourinary System	89,305	45.85
11	630 - 677	Pregnancy, Childbirth, & Puerperium	601	0.31
12	680 - 709	Skin & Subcutaneous Tissue	29,046	14.91
13	710 - 739	Musculoskeletal System & Connective Tissue	39,703	20.39
14	740 - 759	Congenital Anomalies	10,115	5.19
15	760 - 779	Conditions Originating in Perinatal Period	7,289	3.74
16	780 - 789	Symptoms	71,784	36.86
17	790 - 796	Nonspecific Abnormal Findings	20,803	10.68
18	797 - 799	Ill-defined/Unknown Causes of Morbidity & Mortality	6,664	3.42
19	800 - 999	Injury & Poisoning	108,867	55.90
20	V Codes	Supplementary Factors	100,310	51.50
21	E Codes	External Causes of Injury	79,138	40.63

disease group. To prevent overfitting, two dropout layers, with a dropout rate of 20% was also added to the FFNN model (see Figure 6.2). As this is a binary classification for multiple labels, the loss function used for the FFNN was binary cross entropy, and Stochastic Gradient Descent (SGD) was used as the optimizer and a learning rate of 0.01 was used. The *tanh* activation function was used as the input and hidden layer activation functions as the feature matrix values are standardized to the range -1 and 1. The major hyperparameters for the FFNN model – the optimizer, learning rate of the optimizer and the activation function, were tuned empirically over several experiments using the GridSearchCV function in Python sklearn library. Finally, the output layer activation function was a sigmoid function, again as the classification was binary for each of the 21 labels. Training was performed for 50 epochs and then the model was applied to the validation set to predict disease groups after which the results were observed and analyzed.

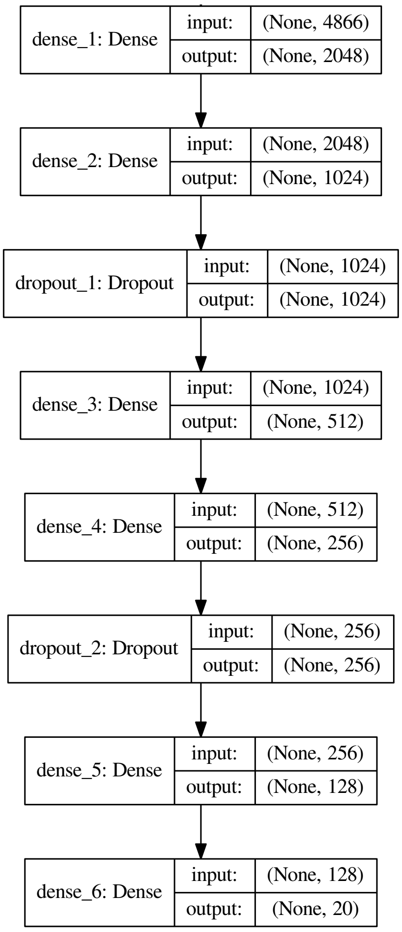


Figure 6.2: Feed Forward Neural Network Model for ICD9 Group Prediction

6.2.4 Experimental Results and Discussion

All experiments were performed on a server running Ubuntu 16.04 LTS with 56 cores of Intel Xeon Processors, 128 GB RAM, 3 TB HDD memory and 2 Tesla M40

GPUs. All implementations were carried out using Python packages – Tensorflow[2], Keras[3] and sklearn. To evaluate the performance of the proposed model, standard metrics to measure machine learning models were considered – accuracy, precision, recall, F-score, Area Under Receiver Operating Characteristic curve (AUROC), Area Under Precision Recall Curve (AUPRC) and Matthew's Correlation Coefficient (MCC). We performed the evaluation of these metrics on a sample-wise basis, i.e., the predicted and actual ICD9 disease groups were compared and analyzed for each radiology report. It can be observed from the Table 6.3 that the proposed model achieved promising results: AUPRC of 0.74 and AUROC of 0.84. The accuracy of 0.77 and precision of 0.80 also indicate an effective prediction performance of the proposed approach.

We also compared the performance of the proposed approach against the current state-of-the-art ICD9 disease group prediction model (Purushotham *et al.*, 2018). As the number of records and features under consideration for both the studies are different, it is to be noted that the number of patients in both the works fall in the same range, ensuring a fair comparison. During validation experiments, it was observed that the proposed approach significantly outperformed against the state-of-the-art method by 23% considering the AUPRC metric and 9% in terms of AUROC. To encourage other comparative studies, certain additional experiments were made. We also provide the Recall & F-Score performance as well as the MCC values of the proposed model over our easily reproducible patient cohort dataset. The model showed good results in these experiments, achieving a recall of 0.77, F-score of 0.77 and MCC value of 0.50. It is to be noted that our method performed better than the state-of-the-art (Purushotham *et al.*, 2018), despite being built on a significantly larger number of patient admission data than the state-of-the-art approach (see Table 6.3). Further, we achieve this performance using only textual features and we did not make use of structured patient data or processed information from any kind of structured data to model the radiology reports of patients. Thus, there is an added advantage that the conversion from unstructured text data to a structured representation can be ignored, thereby achieving huge savings in man hours, cost and other resources.

6.2.4.1 Discussion

From our experiments, we observed a huge requirement and potential for developing prediction based CDSS using unstructured text reports rather than the usage

[2] https://www.tensorflow.org/

[3] https://keras.io/

Table 6.3: Experimental Results

Parameter	**Proposed Approach**	**Purushotham *et al.* (2018)**
Type of Data	Unstructured Text	Structured data
AUROC	0.84 ± 0.01	0.77 ± 0.01
AUPRC	0.74 ± 0.01	0.60 ± 0.02
Accuracy	0.77	*
Precision	0.80	*
Recall	0.77	*
F-Score	0.77	*
MCC	0.50	*

* Metric not reported in the study

of structured patient data and EHRs. The proposed text feature modeling was effectively able to capture the rich and latent clinical information available in unstructured radiology reports, and the neural network model used these features to effectively learn disease characteristics for prediction. The Word2Vec model generated word embedding features and the extracted terms using the Open Biomedical Annotator and SNOMED-CT ontology further enhanced the semantics of the textual features. This enabled the FFNN to generalize better and learn the feature representation, effectively resulting in prediction performance on par (in terms of AUROC) with state-of-the-art approaches built on structured data. The high values of metrics AUPRC of 0.74 and AUROC of 0.84 in comparison to the state-of-the-art AUPRC of 0.60 and AUROC of 0.77 respectively, is an indication that the unstructured text clinical notes (radiology in this case) contain abundant patient-specific information that can be used for predictive analytics applications and that the conversion process from unstructured patient text reports to structured data can be eliminated thereby saving huge man hours, cost and other resources. Moreover, the proposed approach also eliminates any dependency on structured EHRs, thus making it suitable for deployment in developing countries.

6.3 PSO-NN based Two-stage Feature Modeling for Disease Group Prediction

The approach presented in Section 6.2, i.e., the ontology driven feature modeling based disease group prediction approach, though effective, suffers from some limitations. This is due to the high dimensional nature and the sparse nature of the textual feature representation, both due to the presence of a large number of unique biomedical terms considered as features. This may adversely affect the prediction performance in certain cases. To overcome these issues, we present another disease group prediction model based on radiology notes, with a focus on improving the prediction performance by modeling the textual features in a more intuitive and context-sensitive manner.

In this section, a deep neural network model that predicts ICD9 disease groups using unstructured radiology notes based on a PSO-NN two stage feature modeling technique is presented. The workflow and processes defined as part of the proposed disease group prediction system are depicted in Figure 6.3. The same radiology reports in unstructured text format from MIMIC-III dataset used for the previous study were used for this study as well. Overall, 194,744 radiology text reports generated during 45,512 admissions of 36,447 patients were included for the study. As explained earlier, a patient may be diagnosed with multiple diseases in the same admission, and hence, it is necessary for the prediction to be a multi-label prediction task. Therefore, for each radiology report, all disease groups were considered as labels and given binary values - 0 (if the disease was not present) and 1 (if the disease was present). Table 6.1 of the previous section shows the frequency characteristics of the radiology corpus subset that is extracted from the MIMIC-III dataset.

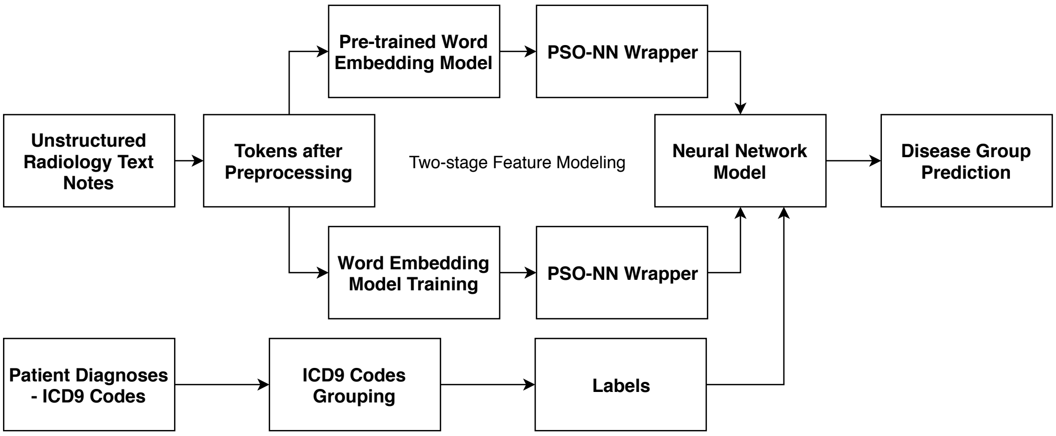

Pre-trained Word Embedding Model
PSO-NN Wrapper
Unstructured Radiology Text Notes
Tokens after Preprocessing
Two-stage Feature Modeling
Neural Network Model
Disease Group Prediction
Word Embedding Model Training
PSO-NN Wrapper
Patient Diagnoses - ICD9 Codes
ICD9 Codes Grouping
Labels

Figure 6.3: Two-stage Feature Modeling based ICD9 Disease Group Prediction

6.3.1 Data Preparation

Preprocessing. The radiology reports text corpus was first subjected to a basic NLP preprocessing pipeline consisting of tokenization and stopping as similar to previous work. The tokenization process breaks down the clinical text corpus into tokens, and the stopping process filters out unimportant words (stop words) from the corpus.

ICD9 Disease Code Grouping. The ICD9 disease grouping strategy was performed in the same way as previous study as explained in Section 6.2.2. A total of 2,593 unique ICD9 disease codes were accordingly grouped into 21 ICD9 disease groups (statistics as shown in Table 6.2 of the previous section). As a patient can suffer from multiple diseases, we consider the ICD9 group prediction task as a binary classification of multiple labels. Therefore, 21 different labels (disease groups) were considered with possible binary values: 0 (for the absence of the disease) and 1 (for the presence of the disease).

6.3.2 Two-stage Feature Modeling

To extract features and enhance the quality of the feature representation of unstructured radiology text reports, a two-stage feature modeling process is designed. The two-stage feature modeling can be considered as an ensemble pipeline of two processes - feature extraction and feature selection. In the first stage, word embeddings are extracted using a word representation model trained on the clinical text corpus along with those generated by a pre-trained word embedding model trained on biomedical terms and concepts. The two sets of word embeddings are separately fed into a PSO-NN wrapper for deriving those feature subsets that most contribute to ICD9 disease group prediction of a patient. Thus, the model not only extracts and generates relevant and discriminating features but also filters out redundant and irrelevant features, thus improving the quality of features and enhancing the performance of classification.

6.3.2.1 Word Embedding Generation

Firstly, the word embeddings for the corpus are generated using two word representation models. Initially, an openly available pre-trained word embedding model (Nédellec *et al.*, 2013) is applied to the preprocessed tokens to generate a set of textual features. This embedding model is trained on a large number of biomedical articles available in PubMed and PubMed Central (PMC), and texts extracted from English Wikipedia dump (Nédellec *et al.*, 2013). Thus, these embeddings provide rich domain knowledge, that helps generate effective feature representations for our input medical corpus. This process generates word embedding vectors of size 1x200 for each clinical index term, which are averaged to create a representation such that each preprocessed radiology report is represented as a 1x200 vector.

The preprocessed corpus is next used to train three popular neural network based word representation models – Word2Vec (Mikolov *et al.*, 2013), FastText (Joulin *et al.*, 2017) and GloVe (Pennington *et al.*, 2014) (explained in Chapter 5, Section 5.3). The generated embedding features and the ICD9 disease group labels are fed into a Random Forest classifier (explained in Section 4.2.3), and the quality of features with respect to classification performance is measured in terms of various metrics. Based on this, the best performing word embedding model was chosen and used for generating word embedding features to be fed into the PSO-NN wrapper.

6.3.2.2 PSO-NN Wrapper for Feature Selection

As the disease group prediction is a binary classification of multiple labels, performing multi-label feature selection is a challenging task, and there can be multitude of optimal feature subsets making the solution space huge. Moreover, the textual features, i.e., the word embeddings from the best word representation model and the pre-trained Word2Vec model may contain redundant and irrelevant features. With this in mind, a Particle Swarm Optimization - Feed Forward Neural Network (PSO-NN) wrapper was designed for selecting the most relevant features for effective classification of ICD9 disease groups. The process is depicted in Fig. 6.4.

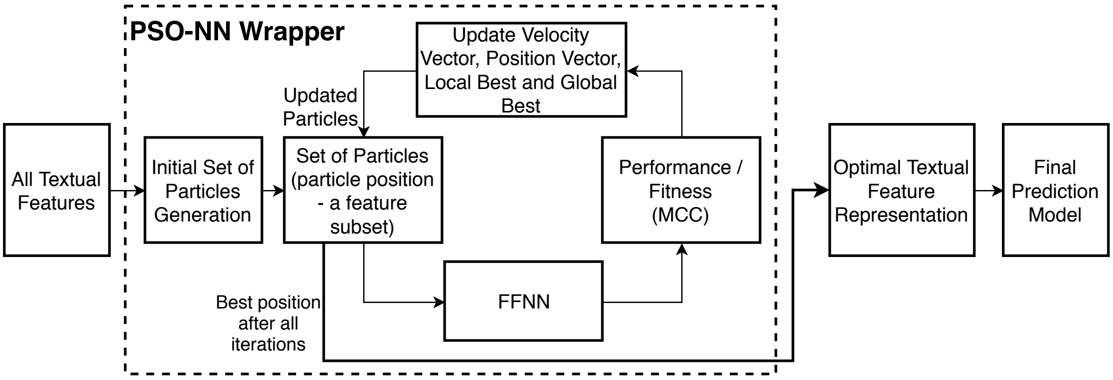

PSO-NN Wrapper
Update Velocity Vector, Position Vector, Local Best and Global Best
Updated Particles
All Textual Features
Initial Set of Particles Generation
Set of Particles (particle position - a feature subset)
Performance / Fitness (MCC)
FFNN
Best position after all iterations
Optimal Textual Feature Representation
Final Prediction Model

Figure 6.4: PSO-NN Wrapper based Two-stage Feature Modeling

Particle Swarm Optimization (Kennedy and Eberhart, 1995) is a bio-inspired, evolutionary computation algorithm modeled on the behavior of bird flocks that aims to find optimal solutions by searching a subspace of possible solutions. The position of each particle or bird in a swarm/flock represents a solution, and the updation of position based on current position and velocity of the particle is used by the algorithm to analyze possible solutions. The best position is determined by a fitness function, based on the current position of the flock or swarm. Information like *local_best* (best position of a particle/bird) and *global_best* (best position among the entire swarm/flock) is shared across all the particles/birds, which enables them to move/fly across the solution subspace converging towards the optimal solution.

For performing PSO based textual feature selection, we generate a swarm of 20 particles/birds, where each particle/bird corresponds to a subset of textual features. Each particle's position, say x_i, is initialized as a random binary vector of size n, where n is the number of features extracted, and each particle/bird travels with a velocity, say v_i, which is initialized as a zero vector of size n. Each particle's position vector, $x_i = [0, 1, 0, 1, 1....1]$ (say) of size $1 * n$, represents a feature subset, where, each element indicates if that feature is selected or not (0: not selected, 1: selected). These feature subsets, along with the respectively assigned ICD9 disease group labels, are used to train a Feed-forward Neural Network (FFNN) and testing is performed using 5-fold cross-validation. The FFNN had 4 hidden layers with 1024, 512, 256 and 128 neurons respectively, with ReLU as hidden layer activation functions and sigmoid as the output activation function. Binary cross-entropy was used as the loss function and stochastic gradient descent was used as the learning optimizer. A dropout layer with a dropout value of 20% was also used after the second layer.

The fitness function for the proposed PSO-NN wrapper is modeled on the performance of the FFNN in terms of average Matthews Correlation Coefficient (MCC). MCC is a performance measure for binary classifications introduced by Matthews (1975). Our task being development of an effective disease prediction model, the objective is not only to improve the number of true positives and true negatives (true diagnoses), but also to reduce false positives and false negatives (false alarms and wrong diagnoses). MCC takes into account not only true positives and true negatives, but also, false positives and false negatives, and returns a value between -1 and +1 (where +1 signifies perfect prediction, -1 signifies worst prediction and 0 signifies random prediction). Hence, MCC, one of the best binary classification performance measures when class imbalance is prevalent, is an apt choice as a fitness function for the proposed PSO-NN wrapper. The performance is measured sample-wise, i.e., for each patient note, the predicted and actual diseases are compared for measuring the performance, and hence, the average performance

Algorithm 4 Optimizing Textual Feature Representation using PSO-NN

Input: Word embedding feature set (n features) of the radiology text corpus & ICD9 disease code groups

Output: Optimized Textual Feature Representation (m features)

```
Generate 20 particles with their position vectors    ▷ Set to random binary vectors of size $1 \times n$
Initialize each particle's position
$x_i = [0, 1, 0, 1, 1....1]$ of size $n$
Velocity of each particle, $v_i = [0, 0, 0, 0....0]$
$local\_best_i = x_i$
$global\_best = [0, 0, 0, 0....0]$ of size $n$    ▷ 0: feature not selected, 1: feature selected
while iterations $\leq 50$ do
    for each particle do
        Calculate fitness of current particle (feature set)
        if $fitness > local\_best_i$ then
            Set $local\_best_i$ = current particle position, $x_i$
        end if
        if $fitness > global\_best$ then
            Set $global\_best$ = current particle position, $x_i$
        end if
    end for
    Update velocities and positions of particles
end while
Return $global\_best$    ▷ Optimal feature representation
```

(MCC value, in this case), is taken as the fitness value (computed as per Eq (6.1) and Eq (6.2)).

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TN + FP)(TP + FN)(TN + FN)}} \tag{6.1}$$

where, TP is the number of True Positives, TN is True Negatives, FP is False Positives and FN is False Negatives in the FFNN classification. MCC gives the performance for each patient report and hence, the fitness value, i.e., the average MCC value of the entire test set, is computed using Eq. 6.2, where, n is the number of patients in the FFNN classification testset.

$$Fitness \quad f = \sum_{j=1}^{n} \frac{MCC_j}{n} \tag{6.2}$$

Based on the fitness values, the best position for each particle at a position, say i ($local\_best_i$) and the best position among the entire swarm/flock ($global\_best$) are updated accordingly. Then, as per the theory of PSO, the new velocity, v_{i+1}, is calculated as per Eq. 6.3 and the new position of the particle, x_{i+1}, is calculated using Eq. 6.4.

$$v_{i+1} = (w * v_i) + c_1 * r_1 * (local\_best_i - x_i) + c_2 * r_2 * (global\_best - x_i) \tag{6.3}$$

$$x_{i+1} = x_i + v_{i+1} \tag{6.4}$$

The new position of each particle, x_{i+1}, determines the new set of feature subsets for which the fitness has to be found again. The entire process is repeated for 50 iterations, and the best position among the entire swarm, *global_best*, represents the feature subset that offered the best classification performance using the FFNN in the wrapper w.r.t MCC metric. The steps involved in the working of PSO-NN wrapper are depicted in Algorithm 4. The PSO-NN wrapper is applied to both the word embedding feature sets – word embeddings from the best-performing word representation model trained on the radiology text corpus and the embeddings of the same from the pre-trained Word2Vec model on PubMed and PMC articles and biomedical concepts. The optimal feature subsets from both the feature sets are determined, concatenated and then fed into the modified TextCNN based neural network architecture for the final training and prediction of ICD9 disease code groups.

6.3.3 mTextCNN: Neural Network Model for Disease Prediction

For the prediction task, a neural network architecture called the modified TextCNN (mTextCNN), an adaptation of the model TextCNN (Kim, 2014) is employed. The TextCNN model consists of an embedding layer that maps input text to a matrix representation, three convolution layers with multiple filter sizes and ReLU activation functions, a max pooling layer for each of the convolution layer, a dropout layer and finally a dense fully connected layer for the output. We adapted the TextCNN architecture for our problem by modeling it as an ensemble with the proposed PSO-NN wrapper. The embedding layer is eliminated as the data preparation and feature extraction pipeline already generates the word embeddings during the two-stage feature modeling phase (discussed in Section 6.3.2), and an enhanced representation of embeddings is generated using the PSO–NN wrapper.

The proposed mTextCNN architecture is depicted in Figure 6.5. The number of nodes in the input layer of the mTextCNN model pertains to the number of features generated by the proposed two-stage feature modeling approach. It consists of five convolution layers with ReLU activation functions, one maxpooling layer for each convolution layer, a dropout layer and finally, a dense fully-connected network with 2 hidden dense layers with 256 and 128 neurons respectively with ReLU activation functions and an output layer (21 nodes) with a sigmoid activation function. We used 1024 filters on each convolution layer with filter sizes 2, 3, 4, 5 and 6 respectively. Hyperparameters like optimizer, activation function and learning rate of the optimizer were tuned empirically over several experiments using GridsearchCV function in the python sklearn library. The output layer activation function was chosen to be sigmoid and binary cross entropy was used as the loss function, as the ICD9 disease group prediction is a binary classification across multiple labels. The major hyperparameters and configurations used in TextCNN (Kim, 2014) and the proposed mTextCNN architectures are shown in Table 6.4. The effective textual representation generated by the proposed PSO-NN wrapper from the initial sets of features are now concatenated, along with the respective ICD9 disease group labels, are fed to the mTextCNN model. The training is performed for 50 epochs and the best weights are stored and used for performance analysis. The experimental validation of the proposed model is presented in Section 6.3.4.

Table 6.4: Hyperparameters of TextCNN and mTextCNN

Parameter	**TextCNN**	**mTextCNN**
No. of convolution layers	3	5
No. of maxpooling layers	3	5
No. of filters (per filtersize)	300	1024
Filter sizes	3, 4, 5	2, 3, 4, 5, 6
Activation in Conv layers	ReLU	ReLU
Optimizer	Adam	Adam
Learning rate (lr)	0.001	0.01
No. of dense layers (fully connected layers)	1	3

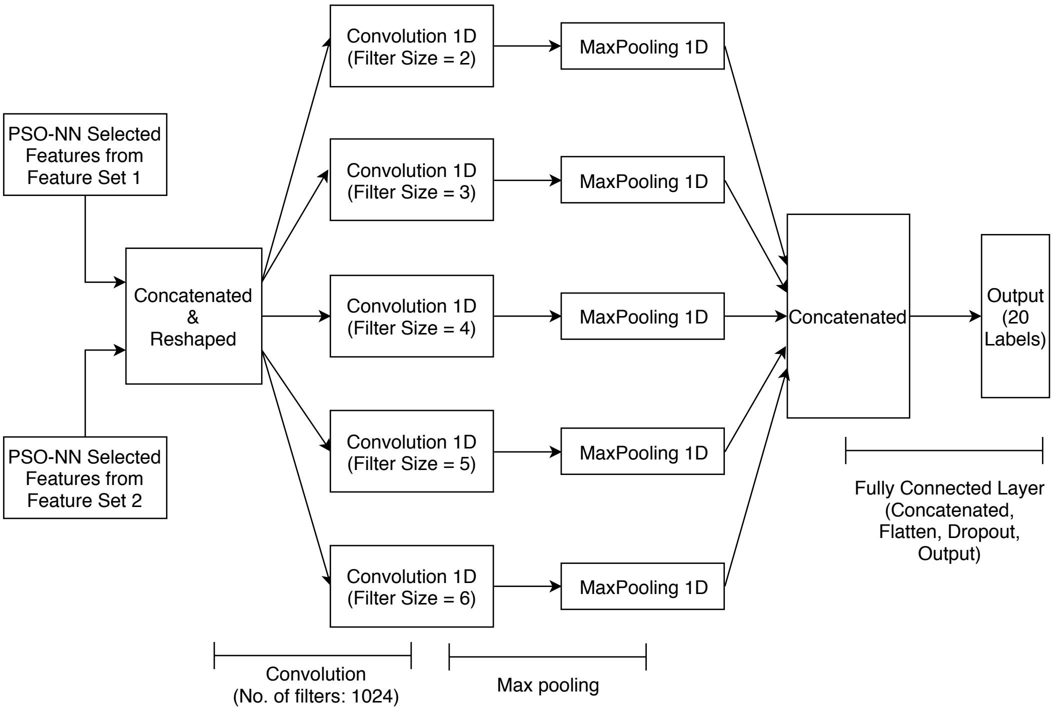

PSO-NN Selected Features from Feature Set 1
PSO-NN Selected Features from Feature Set 2
Concatenated & Reshaped
Convolution 1D (Filter Size = 2)
Convolution 1D (Filter Size = 3)
Convolution 1D (Filter Size = 4)
Convolution 1D (Filter Size = 5)
Convolution 1D (Filter Size = 6)
MaxPooling 1D
MaxPooling 1D
MaxPooling 1D
MaxPooling 1D
MaxPooling 1D
Concatenated
Output (20 Labels)
Fully Connected Layer (Concatenated, Flatten, Dropout, Output)
Convolution (No. of filters: 1024)
Max pooling

Figure 6.5: Proposed mTextCNN Model for ICD9 Code Group Prediction

6.3.4 Experimental Results and Discussion

For experimental validation of the proposed disease group prediction neural model, we performed several experiments to analyze the various aspects of the modeling pipeline and process. All such experiments were performed on an Ubuntu based High-end Server with a 56-core Intel Xeon processor, 128GB RAM, two Nvidia Tesla M40 GPUs (24GB each) and 3TB hard drive. All implementations were done in Python using packages sklearn, keras, tensorflow, gensim and matplotlib. We performed several experiments, to validate the efficacy of the proposed approach, which are discussed in detail in the following sections.

6.3.4.1 Benchmarking Word Representation Models

To understand the relative performance of the individual word representation models used in the proposed work, we conducted several experiments by applying each of them to the unstructured radiology text corpus. The word embedding vectors from three popular word representation models – Word2Vec, FastText, and GloVe were considered as features for the ICD9 disease group prediction task. For Word2Vec and FastText, both Skipgram and Continuous Bag-of-Words (CBOW) models were considered for the experiment. The features were fed into a Random

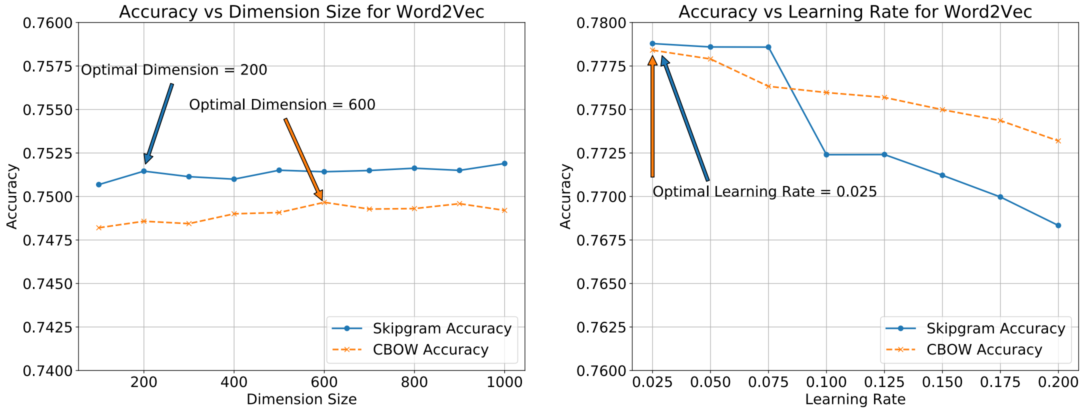
Forest classifier with 100 decision trees and 5-fold cross validation was performed for the experiments. The performance of classification based on these features was measured to determine the best-performing model in terms of accuracy w.r.t major training parameters such as dimension size and learning rate, which are described below.

Accuracy vs. Dimension. To observe the effect of vector dimension size of word embedding features on the classification performance, the size was varied while training the word representation models to generate word embedding features for each size. The models were trained with sizes – 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000 for a fixed number of 5 epochs, for each dimension size, at an initial learning rate of 0.025. After that, the embedding vectors of the preprocessed radiology text corpus were generated for each dimension size which were used as features for training a Random Forest classifier for ICD9 disease group prediction.

Accuracy vs. Learning Rate. From our observations during experiment 1, optimal dimension size was determined. Now, we varied the initial learning rate used for training each word representation model to observe the effect on classification accuracy, while keeping the dimension size constant. The initial learning rates – 0.01, 0.025, 0.05, 0.075, 0.1, 0.125, 0.15, 0.175, 0.2, were used for training the radiology text corpus on the word representation models for a fixed number of 5 epochs. Similar to previous experiments, for each learning rate, word embedding vectors of the corpus were generated to train a Random Forest classifier to predict the ICD9 disease group and the performance of each classification model was observed and analyzed.

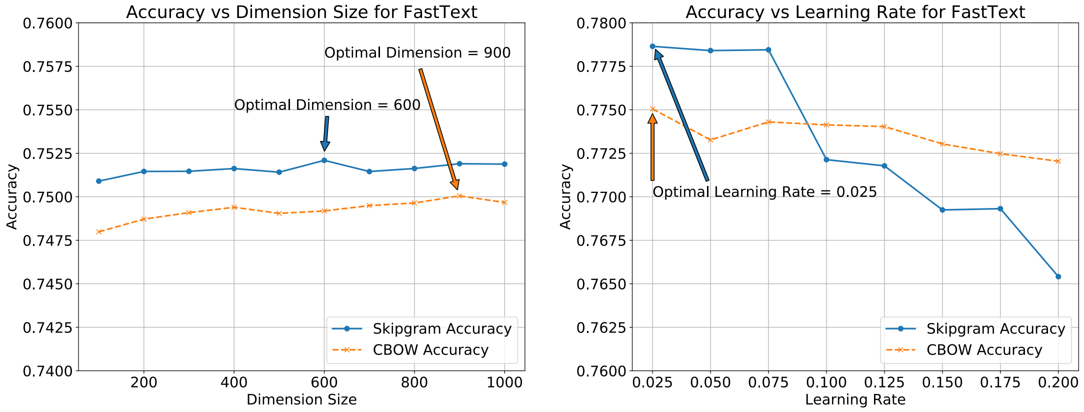
The performance of the word representation models – Word2Vec skipgram, Word2Vec CBOW, FastText skipgram, FastText CBOW and GloVe with respect to the above experiments are illustrated in Figures 6.6, 6.7 and 6.8 respectively.

From Fig. 6.6a, it can be observed that the Word2Vec skipgram model performs better at dimension sizes – 200, 500, 700, 800 and 1000. As lower dimensionality of feature representation is desirable, we chose the optimal size to be 200. Keeping size constant at 200, the initial learning rate was varied and the classification performance was observed. With this, the model performs best when the initial learning rate is 0.025 (Figure 6.6b). Hence, this configuration, i.e., dimension size of 200 and initial learning rate of 0.025, was chosen as the best model among Word2Vec skipgram models and is referred to as *W2V_skip_RF*. Similarly, dimension size/initial learning rate of 600, 0.025 were identified as the best config-



(a) Accuracy vs Dimension Size (b) Accuracy vs Learning Rate

Figure 6.6: Performance of Word2Vec Skipgram & CBOW Embedding Models

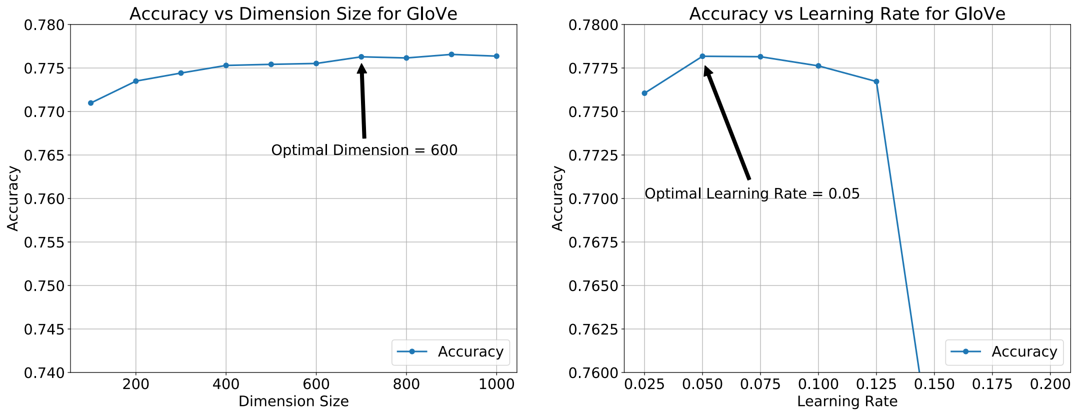


(a) Accuracy vs Dimension Size (b) Accuracy vs Learning Rate

Figure 6.7: Performance of FastText Skipgram & CBOW Embedding Models

uration for Word2Vec CBOW model ($W2V\_cbow\_RF$)(Figure 6.6) and FastText skipgram model (FT_skip_RF) (Figure 6.7); while dimension size/initial learning rate of 900, 0.025 was best for FastText CBOW model (FT_cbow_RF)(Figure 6.7). Finally, for the GloVe model, dimension size of 700 and initial learning rate of 0.05 ($Glove\_RF$) was found to be optimal (Figure 6.8). The performances of these optimal models are summarized in Table 6.5.

From Table 6.5, it is evident that all models performed consistently well with a variation of less than 0.01%. Even then, the Word2Vec skipgram model (dimension size 200 & learning rate 0.025) was judged to be the best, due to its optimal dimensionality. The hybrid combination of word embedding features generated



(a) Accuracy vs Dimension Size

(b) Accuracy vs Learning Rate

Figure 6.8: Performance of GloVe Word Embedding Model

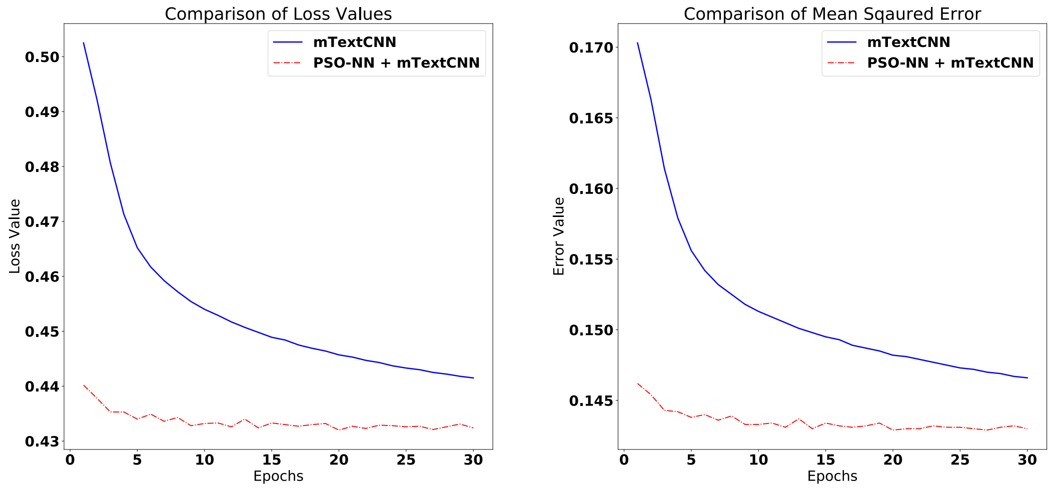
Table 6.5: Benchmarking various Word Embedding Models - Summary

Model	**Accuracy**	**Optimal Size**	**Learning Rate**
W2V_skip_RF	0.7787	200	0.025
W2V_cbow_RF	0.7784	600	0.025
FastText_skip_RF	0.7786	600	0.025
FastText_cbow_RF	0.775	900	0.025
Glove_RF	0.7781	700	0.05

by this model and the pre-trained word embedding model were then fed into the PSO-NN wrappers for improving the quality of textual feature representation and further optimize the dimensionality.

6.3.4.2 Experimental Evaluation of Two-stage Feature Modeling using PSO-NN Wrapper

From our previous experiments, the Word2Vec skipgram model (dimension size 200 & learning rate 0.025) exhibited optimal dimensionality and the features generated by this model were used to generate word embedding features of the preprocessed radiology text corpus. Along with these, word embedding features of the same corpus from the pre-trained Word2Vec model trained on PubMed and PMC articles (of dimension size 200) were also generated. The ICD9 disease group labels and both the set of word embeddings for the clinical text corpus were fed into the two PSO-NN wrappers. The output is an optimal feature representation for the problem of ICD9 disease group prediction. The details of the optimal feature sets



(a) Validation Loss vs Epochs

(b) Validation MSE vs Epochs

Figure 6.9: Performance of the proposed *PSO-NN+mTextCNN* model, when compared to *mTextCNN* model

and their performances in the PSO-NN wrappers are tabulated in Table 6.6. It is evident that the feature set derived by the proposed PSO-NN wrapper contained an optimal subset of features that effectively predicted ICD9 disease groups without performance degradation (from results reported in Table 6.5). Therefore, it can be inferred that the proposed PSO-NN wrapper was effective in deriving an effective textual feature representation that can help classifiers generalize better when used for ICD9 disease group prediction. The selected feature subsets 1 & 2 are concatenated, i.e., 209 features and respective disease group labels are fed into the proposed mTextCNN model for the final training and prediction. The performance analysis of the mTextCNN model is presented in the subsequent sub section.

6.3.4.3 Performance Analysis of mTextCNN Prediction Model

The enhanced feature representation generated by the PSO-NN wrapper along with the respective labels is fed into the mTextCNN model. To evaluate the performance, standard metrics like accuracy, precision, recall, F-score, Area under Receiver Operating Characteristic curve (AUROC), Area under Precision Recall Curve (AUPRC) and Matthew's Correlation Coefficient (MCC) were used, with 5-fold cross-validation. We measured these metrics on a sample-wise basis, i.e., for each report, the predicted and actual disease groups were compared and analyzed. Table 6.7 summarizes the results of this experiment (*PSO-NN+mTextCNN*).

Table 6.6: Performance of PSO-NN Wrapper for Feature Sets 1 & 2

Metric	**Feature set 1***	**Feature set 2**$^+$
Total number of features	200	200
Number of features selected	107	102
Global Best (MCC)	0.49	0.47
Accuracy	0.78	0.77

**Feature Set 1: Embeddings from Word2Vec trained on clinical text corpus*

$^+$*Feature Set 2: Embeddings from Word2Vec trained on PubMed and PMC articles (biomedical domain)*

To validate whether the proposed PSO-NN wrapper is effective as an ensemble to the proposed mTextCNN model, we performed an experiment using all 400 raw text embedding features (200 from both feature sets) on mTextCNN directly without the feature selection process by PSO-NN wrapper and compared it with the performance of mTextCNN used with PSO-NN (*PSO-NN+mTextCNN*). The results are tabulated in Table 6.7. The plot of validation loss function and Mean Squared Error (MSE) across epochs for both mTextCNN and *PSO-NN+mTextCNN* are given in Figure 6.9. Interestingly, from Table 6.7 and Figure 6.9, it can be observed that, even with significant dimensionality reduction, the *(PSO-NN+mTextCNN)* model still outperformed the base mTextCNN model (without PSO-NN). Therefore, it can be inferred that the proposed PSO-NN wrapper selects relevant and discriminating textual features, using which the mTextCNN model generalized effectively to classify ICD9 disease groups.

Table 6.7: Evaluation of Base mTextCNN model (on raw embedding features without feature selection) and *PSO-NN+mTextCNN* model

Metric	**mTextCNN**	**PSO-NN+mTextCNN**
Number of features	400	209
Accuracy	0.78	0.79
Precision	0.80	0.81
Recall	0.78	0.78
F-Score	0.77	0.78
MCC	0.48	0.51
AUROC	0.85	0.86
AUPRC	0.75	0.77

6.3.5 Benchmarking against State-of-the-art Structured Data Models

The proposed *PSO-NN+mTextCNN* model, on the entire corpus of radiology notes of 45,512 patient admissions, reported an AUPRC of 0.77, AUROC of 0.86 and accuracy, precision and F-score of 0.79, 0.81 and 0.78 respectively indicating good and promising performance (Table 6.7). Comparative benchmarking of the proposed approach against an existing ICD9 group prediction model (Purushotham *et al.*, 2018) (based on structured data) was considered next. We regenerated the same set of patient admissions employed by Purushotham *et al.* (2018) using the patient cohort criteria mentioned in their paper and their shared code in GitHub repository[4], which resulted in 35,849 patient admissions. The proposed *PSO-NN+mTextCNN* model was applied to the 140,710 radiology reports recorded during these patient admissions and the observed performance after performing 5-fold cross validation (as similar to experiments by Purushotham *et al.* (2018)) is tabulated in Table 6.8. The PSO-NN wrapper selected 214 textual features from the total of 400 and these features along with disease group labels were used to train the mTextCNN model.

Purushotham *et al.* (2018) benchmarked the performances of Super Learner models, Gated Recurrent Unit (GRU), Feedforward Neural Network (FFNN) and finally, proposed a Multimodal Deep Learning (MMDL) model, which was an ensemble of FFNN and GRU. The performances of these models were referred from the paper (Purushotham *et al.*, 2018) and the results of comparison of the proposed approach against these models are tabulated in Table 6.8. The comparison of AUROC and AUPRC values of the proposed *PSO-NN+mTextCNN* approach with that of MMDL (Purushotham *et al.*, 2018) is as depicted in Fig. 6.10. It can be observed that the proposed approach significantly outperformed the best performing model (MMDL) among models proposed by Purushotham *et al.* (2018) by a margin of 10% in terms of AUROC and 27% w.r.t AUPRC. To encourage other comparative studies that may use radiology notes corpus, we also benchmarked the performance in terms of other metrics like accuracy, recall, F-score performance and MCC values on this reproducible patient cohort. The model showed good results in these experiments, achieving an accuracy of 0.78, precision of 0.81, F-score of 0.77 and MCC value of 0.49. An important fact to be noted is that our model is built using only textual features from unstructured clinical notes and model by Purushotham *et al.* (2018) is based on structured patient data. As

[4] Available at https://github.com/USC-Melady/Benchmarking_DL_MIMICIII

our approach models unstructured radiology text notes for each patient, an added advantage is the elimination of the need for conversion from unstructured patient data to structured patient data, thereby achieving huge savings in person-hours, cost and other resources.

Table 6.8: Benchmarking proposed model against state-of-the-art model based on structured data Purushotham *et al.* (2018)

Parameter	**Our Approach**	**MMDL**	**Super Learner**	**GRU**	**FFNN**
Data	Unstructured Text	Structured data	Structured data	Structured data	Structured data
AUROC	**0.85**	0.77	0.75	0.72	0.71
AUPRC	**0.76**	0.60	0.54	0.51	0.50
Accuracy	0.78	-*	-*	-*	-*
Precision	0.81	-*	-*	-*	-*
Recall	0.78	-*	-*	-*	-*
F-Score	0.77	-*	-*	-*	-*
MCC	0.49	-*	-*	-*	-*

* *Metric not reported in the study*

Results referred here, are as reported by Purushotham et al. *(2018).*

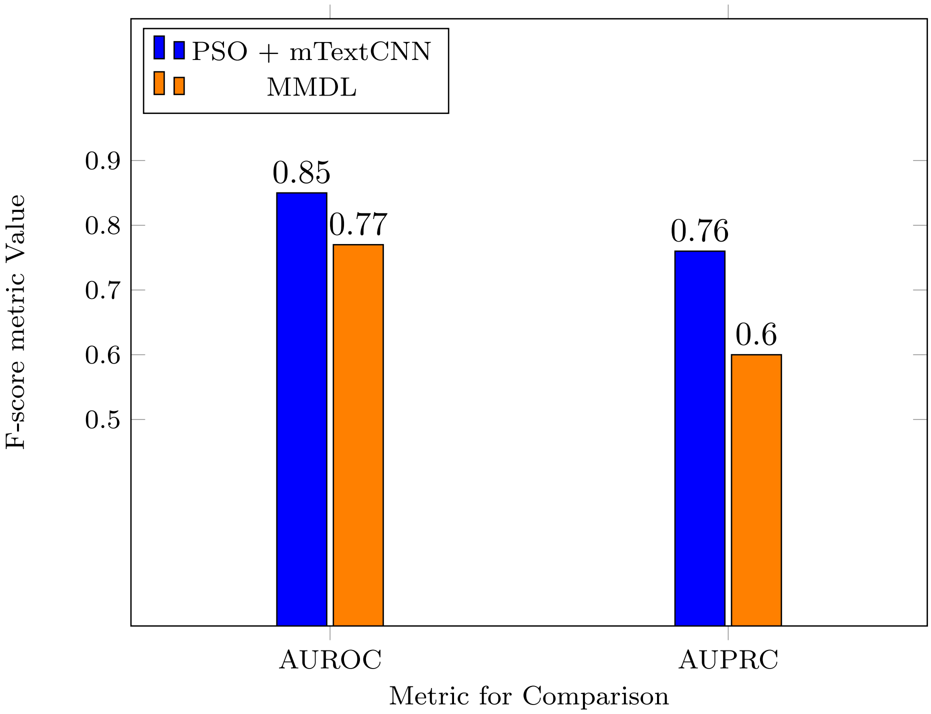


Figure 6.10: F-score comparison for All Tasks on Respective Datasets

6.3.5.1 Discussion

From our experiments, a significant potential for developing disease prediction CDSS using unstructured clinical text reports directly, rather than depend on the availability of structured patient data and EHRs, is observed. The proposed two-stage textual feature modeling approach (word embedding extraction and PSO-NN wrapper) was successful in capturing the rich, latent, patient-specific clinical information available in unstructured radiology reports, and using it to learn disease characteristics for ICD9 disease group prediction.

The Word2Vec model trained on the subjective radiology corpus with optimized parameter configuration generates word embedding features to be fed into the proposed PSO-NN wrapper. The embedding features of the corpus generated by the pre-trained model trained on PubMed and PMC articles were also fed into another instance of the proposed PSO-NN wrapper. The TextCNN architecture was modified and adapted (mTextCNN) for the problem of disease group classification and prediction. The relevant features selected from both feature sets were concatenated and fed into the mTextCNN for classification, which performed effectively, from which three factors are evident.

1. The word embeddings of corpus generated from the pre-trained Word2vec model trained on PubMed and PMC articles further enhanced and enriched the semantics of the textual features with biomedical domain knowledge.

2. The proposed PSO-NN wrapper removes irrelevant and redundant features making the textual feature representation more effective with lesser dimensions.

3. These factors have enabled the proposed mTextCNN to generalize better and learn the feature representation effectively, resulting in promising prediction performance better than state-of-the-art approaches built on structured data.

The high AUPRC value of 0.77 in comparison to the state-of-the-art (based on structured data) AUPRC of 0.60 and also the high AUROC value of 0.86 against the state-of-the-art value of 0.77 are indications that the unstructured text clinical notes (radiology reports in our case) contain abundant patient-specific information that can be used for predictive analytics applications. Moreover, the conversion process from unstructured patient text reports to structured data can be eliminated, thereby, saving huge person-hours, cost and other resources. Additionally, as our proposed approach does not depend on structured patient data, it can be deployed to hospitals in developing countries where EHR adoption is low.

6.4 Hybrid Text Feature Modeling for Disease Group Prediction

As the previously presented DPMs gave promising results for radiology notes, next, we made it our objective to observe performances of DPM based on physician's notes. This section presents such a generic disease prediction based CDSS based on unstructured physician notes using a hybrid word embedding based text feature modeling. The overall workflow is as depicted in Fig. 6.11. We used physicians' clinical notes in unstructured text form from the MIMIC-III dataset for our experiments. We extracted only the physician notes from the 'noteevents' table, which resulted in a total of 141,624 physician notes generated during 8,983 admissions. As per MIMIC documentation, physicians have reported some identified errors in notes present in the 'noteevents' table. As these notes can affect the training negatively, records with physician identified errors were removed from the cohort. Additionally, those records with less than 15 words are removed and finally, the remaining 141,209 records are considered for the study. Some characteristic features of the physician notes corpus are tabulated in Table 6.9.

We observed that there were 832 kinds of physician notes available in the MIMIC-III dataset such as Physician Resident Progress note, Intensivist note, etc. The frequency statistics of the top ten kinds of physician notes are tabulated in Table 6.10. A particular patient may suffer from multiple diagnoses during a particular admission and hence, it is necessary for the prediction to be a multi-label prediction task. Therefore, for each physician note, all the diagnosed disease groups during that particular admission were considered as labels and given binary values - 0 (if the disease was not diagnosed) and 1 (if the disease was diagnosed).

6.4.1 Preprocessing & Feature Modeling

The physician notes corpus is first preprocessed using basic NLP techniques such as tokenization and stop word removal as similar to previous works in this chapter. Using tokenization, the clinical text corpus is broken down into basic units called tokens and by the stopping process, unimportant words are filtered out. The preprocessed tokens are then fed into a pre-trained word embedding model to generate the word embedding vector representation of the corpus that can be considered

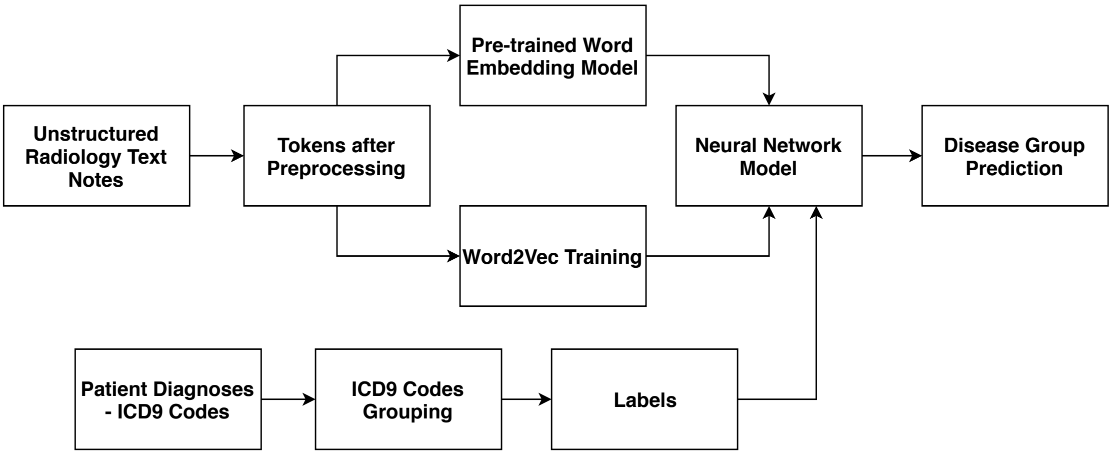

Pre-trained Word Embedding Model
Unstructured Radiology Text Notes
Tokens after Preprocessing
Neural Network Model
Disease Group Prediction
Word2Vec Training
Patient Diagnoses - ICD9 Codes
ICD9 Codes Grouping
Labels

Figure 6.11: Proposed ICD9 Disease Group Prediction Process

Table 6.9: Physician Notes Corpus Characteristics

Feature	**Total Records**
Physician Notes	141,209
Unique Words	635,531
Words in longest Note	3,443
Words in shortest Note	16
Average word length of notes	858
Unique Diseases	4,208
Disease Groups	21

Table 6.10: Top 10 Types of Physician Notes

Note Type	**Occurrences**
Physician Resident Progress	62,550
Intensivist	26,028
Physician Attending Progress	20,997
Physician Resident Admission	10,611
ICU Note - CVI	4,481
Physician Attending Admission (MICU)	3,307
Physician Resident/Attending Progress (MICU)	1,519
Physician Surgical Admission	1,102
Physician Fellow/Attending Progress (MICU)	970
Physician Attending Admission	873

as textual features. The pre-trained word embedding model used in this study is an openly available model that is trained on biomedical articles in PubMed and PMC along with texts extracted from an English Wikipedia dump (Nédellec *et al.*, 2013), hence capturing relevant terms and concepts in the biomedical domain, which helps generate quality feature representation of the underlying corpus. The pre-trained model generates word embedding vectors of size 1x200 for each word and these vectors were averaged to generate a representation such that each preprocessed physician note is represented as a 1x200 vector. The preprocessed tokens are then used to train a Word2Vec model (Mikolov *et al.*, 2013) (explained in Section 5.2.3), a neural network based word representation model that generates word embeddings based on co-occurrence of words. The Skipgram model of Word2Vec was used for training the physician notes tokens with a dimension size of 200 (same as the pre-trained model) with an initial learning rate of 0.025. The averaged word vector representation for each report tokens are extracted and then fed into the neural network model along with the vector representation extracted from the pre-trained model and the ICD9 disease group labels.

6.4.2 ICD9 Disease Code Grouping

ICD9 disease codes of patients were categorized into standard ICD9 disease groups as per standards and strategy explained in Section 6.2.2. One difference here was that we considered the V and E codes in a single category. A total of 4,208 unique ICD9 disease codes thus obtained were grouped into 20 ICD9 disease groups, i.e., potential labels. As the ICD9 group prediction task is a binary classification of multiple labels, 20 labels (disease groups) were considered with binary values: 0 (negative diagnosis of the disease) and 1 (positive diagnosis of the disease). The physician notes, modeled into two feature matrices of shape 141209×200 each, along with 20 ICD9 disease groups (labels) are now used to train the neural network model.

6.4.3 Neural Network Model

The proposed Deep Neural Network Prediction Model is illustrated in Fig. 6.12. The neural network architecture is divided into 2 parts – the first for determining the weights for the hybrid combination of features dynamically and the next for multi-label classification of ICD9 disease codes.

The process of dynamically modeling the weightage to be assigned for the combination of pre-trained word embeddings and the word embeddings generated using the physicians notes is performed as shown in Fig. 6.13. The two feature sets are fed as inputs into the neural network model, where both the input layers consist of 200 neurons, equal to the number of features generated from both models. The addition layer merges the two sets of input features in a weighted combination which is dynamically determined through backpropagation of the overall neural network architecture thereby ensuring the optimal and effective combination that offers the best classification performance possible. This architecture also ensures that the weights for the hybrid combination of features is always determined dynamically and hence can be used for any clinical text corpus. The combined set of features, i.e., the hybrid features, are then fed on to a dense Feed Forward Neural Network (FFNN) model which performs the training for multi-label classification of ICD9 disease code groups.

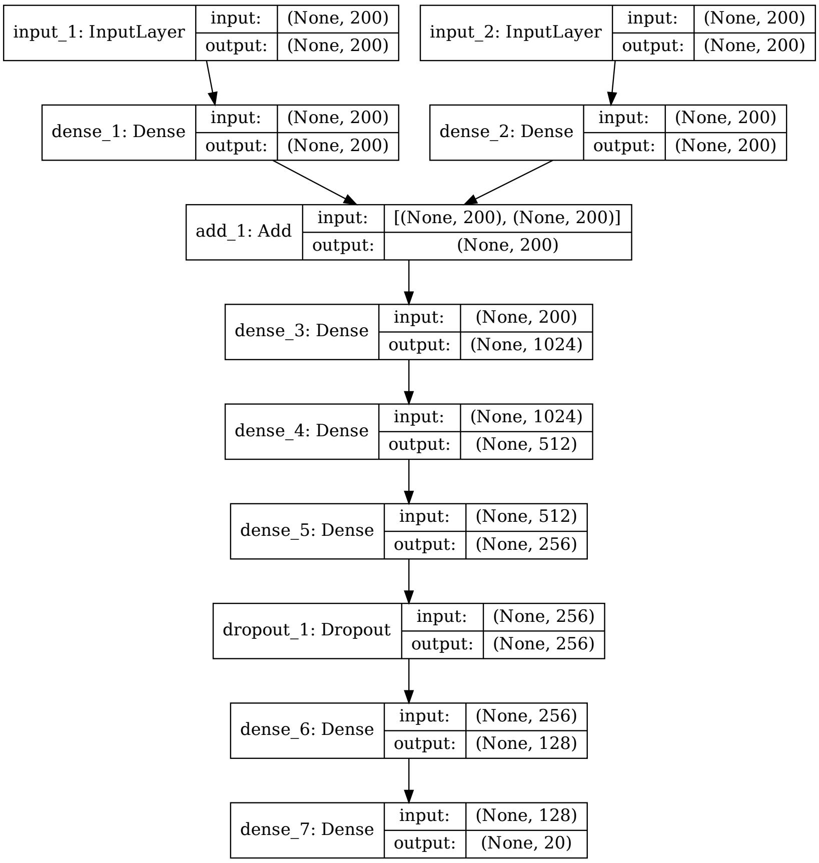


Figure 6.12: Overall Neural Network Model for ICD9 Group Prediction

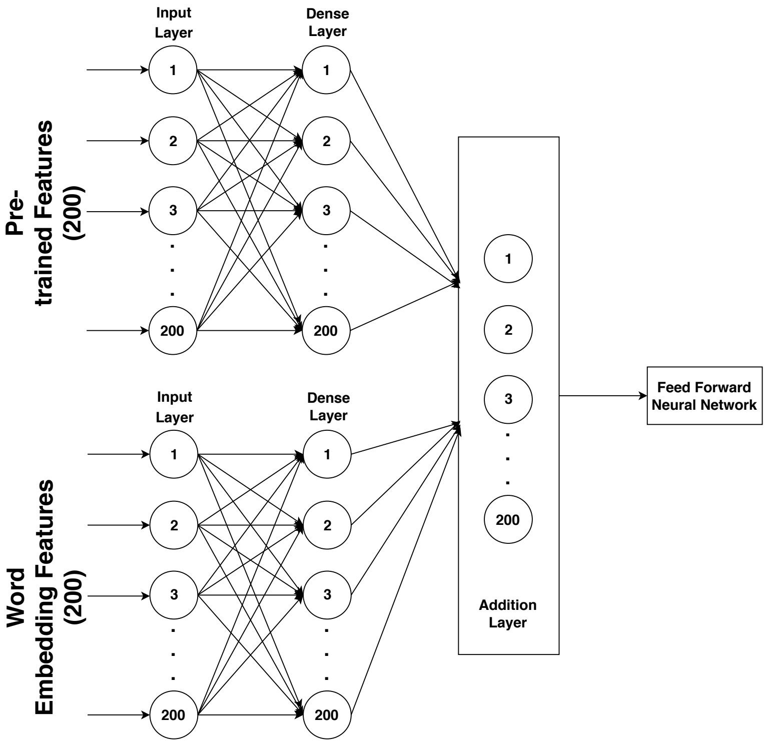

Pre-trained Features (200)
Input Layer
Dense Layer
1
2
3
200
Word Embedding Features (200)
Input Layer
Dense Layer
1
2
3
200
Addition Layer
Feed Forward Neural Network

Figure 6.13: Hybrid Features using Dynamic Weighted Addition of Feature Representations

6.4.4 Disease Prediction Model

The dynamically weighted feature matrix consisting of the hybrid word embedding features, along with ICD9 group labels are next used for training a FFNN used as the prediction model (depicted in Fig. 6.14). The input layer consists of 1024 neurons with input dimension as 200 (number of input features); followed by three hidden layers with 512, 256 and 128 neurons respectively and finally an output layer with 20 neurons, each representing an ICD9 disease group. To prevent overfitting, a dropout layer, with a dropout rate of 20% was also added to the FFNN model (see Fig. 6.12). As this is a binary classification for multiple labels, binary cross entropy was used as a loss function, while Stochastic Gradient Descent (SGD) was used as the optimizer with a learning rate of 0.01. Rectified Linear Unit ($ReLU$) activation function was used as the input and hidden layer activation functions as the feature matrix values are standardized to the range -1 and 1. The major hyperparameters for the FFNN model – the optimizer, learning rate of the optimizer and the activation function, were tuned empirically over several experiments using the GridSearchCV function in Python sklearn library. Finally, the output layer activation function is a sigmoid, again as the classification is two-class for each of the 20 labels. Training was performed for 50 epochs and then the model was applied to the validation set to predict disease groups after which the results were observed and analyzed.

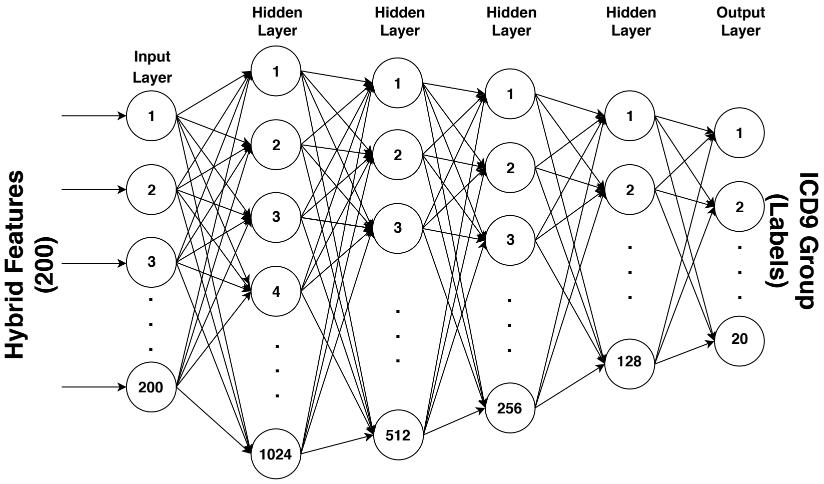


Figure 6.14: Feed-forward Neural Network - Disease Group Prediction Model

6.4.5 Experimental Results and Discussion

All experiments were performed on an Ubuntu based High-end Server with a 56-core Intel Xeon processor, 128GB RAM, two Nvidia Tesla M40 GPUs (24GB each) and 3TB hard drive. Implementations were done in Python using packages sklearn, keras, tensorflow, gensim and matplotlib. To evaluate the proposed approach, we performed several experiments using standard metrics like accuracy, precision, recall, F-score, Area under Receiver Operating Characteristic curve (AUROC), Area under Precision Recall Curve (AUPRC) and Matthew's Correlation Coefficient (MCC). We measured these metrics on a sample-wise basis, i.e., for each report, the predicted and actual disease groups were compared and analyzed. It can be observed from the Table 6.11 that the proposed model achieved promising results: AUPRC of 0.85 and AUROC of 0.89. The accuracy, precision, MCC and F-score of 0.79, 0.82, 0.57 and 0.79 respectively also indicate a good performance.

We compared the proposed model's performance against that of baseline feature modeling approaches – TF-IDF based bag-of-words approach, trained word embedding approach (only Word2Vec model) and a pre-trained word embedding approach (using word embedding model trained on PubMed, PMC and Wikipedia English articles). The term weighted bag-of-words approach uses term frequency and inverse document frequency ($tf \times idf$) scores calculated from the physician notes corpus. TF-IDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. The sklearn English stopword list was used to filter the stopwords and n-gram ($n = 1, 2, 3$) features were considered. Finally, the top 1000 features were extracted from the corpus and then fed into the neural network model for training. The other two baselines were kept the same as explained in Section 6.4.1. It is to be noted that in the neural network configuration, only one input is present, i.e., there is no hybrid weighted addition layer. The results of comparison are tabulated in Table 6.11. It can be observed that the proposed approach that involves a hybrid weighted combination of pre-trained and trained word embeddings is able to perform comparatively better in terms of all metrics.

Table 6.11: Experimental Results – Baseline Comparison

Parameter	**Proposed Approach**	**Bag-of-words (TF-IDF)**	**Only Word2Vec**	**Only Pre-trained**
AUROC	0.89	0.87	0.88	0.88
AUPRC	0.85	0.81	0.84	0.82
Accuracy	0.79	0.78	0.79	0.79
Precision	0.82	0.80	0.82	0.80
Recall	0.79	0.78	0.79	0.78
F-Score	0.79	0.78	0.79	0.79
MCC	0.58	0.53	0.57	0.56

Table 6.12: Experimental Results – Comparison with State-of-the-art

Parameter	**Our Approach**	Purushotham *et al.* (2018)
Type of Data	Unstructured text	Structured data
AUROC	0.89	0.77
AUPRC	0.85	0.60
Accuracy	0.79	-*
Precision	0.82	-*
Recall	0.79	-*
F-Score	0.79	-*
MCC	0.58	-*

* *Results not reported in the study*

Next, a comparative benchmarking of the proposed approach against the state-of-the-art ICD9 disease group prediction model developed by Purushotham *et al.*

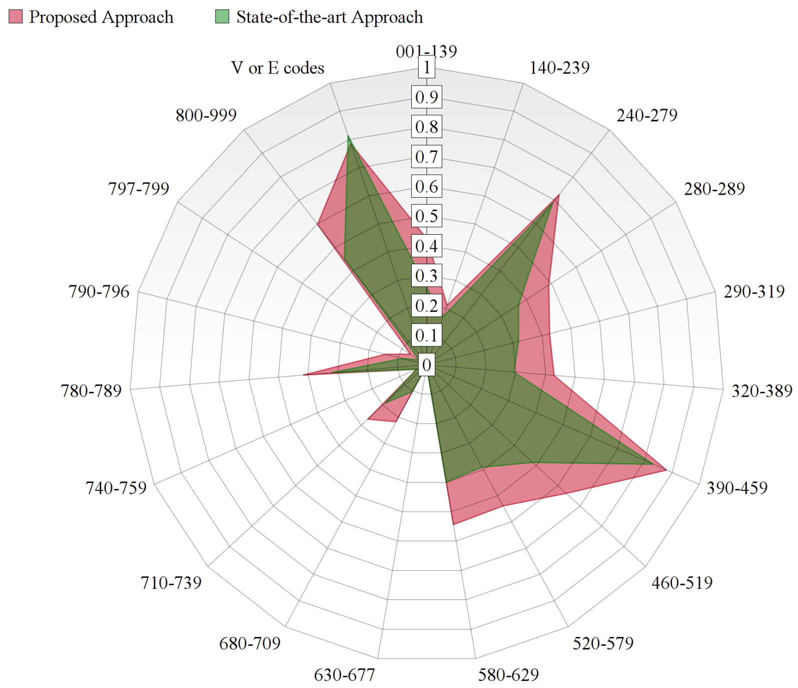


Figure 6.15: ICD9 Group Labels Statistics Comparison

(2018) was performed. Although the number of records considered for both the studies are different, it is to be noted that the labels are distributed similarly (statistics shown in Fig. 6.15) and which therefore enables a fair comparison. We consider this comparison in order to study the effect of the disease group prediction models built on structured (state-of-the-art) and unstructured patient data (physician notes in this case). The results of the benchmarking are tabulated in Table 6.12, which clearly shows the proposed approach outperformed the model by Purushotham *et al.* (2018) by 15% in terms of AUROC and 40% in terms of AUPRC. This shows that the predictive power of a model built on unstructured patient data exceeds that of those built on structured data. To encourage comparative studies that use physician notes in MIMIC-III dataset, certain additional metrics were also considered. The Recall & F-Score performance as well as the MCC values of the proposed model over our easily reproducible patient cohort data subset are also observed and provided.

6.4.5.1 Discussion

From our experiments, we observed a significant potential in developing prediction based CDSS using unstructured text reports directly, eliminating the dependency on the availability of structured patient data and EHRs. The proposed approach that involves a textual feature modeling and a neural network based prediction model was successful in capturing the rich and latent clinical information available in unstructured physician notes, and using it to effectively learn disease group characteristics for prediction. The Word2Vec model, trained on the physician notes corpus with optimized parameter configuration, generates effective word embedding features to be fed into the neural network model. The hybrid combination of these with the embedding features of the corpus generated by the pre-trained model trained on PubMed and PMC articles further enhanced and enriched the semantics of the textual features with biomedical domain knowledge. This is clear from the baseline comparison shown in Table 6.11 and it is this combination that has further enabled the FFNN to generalize better and learn the textual feature representation, effectively improving prediction performance when compared to the state-of-the-art model built on structured data.

It is interesting to note that the patient data was modeled using only textual features, without any EHRs, structured data or other processed information. The high AUPRC and AUROC values obtained in comparison to the state-of-the-art's (based on structured data) performance is an indication that the unstructured text clinical notes (physician notes in this case) contain abundant patient-specific information that is beneficial for predictive analytics applications. Moreover, the conversion process from unstructured patient text reports to structured data can be eliminated, thereby saving huge man hours, cost and other resources. The proposed approach also eliminates any dependency on structured EHRs, thus making it suitable for deployment in developing countries.

For the presented DPMs in this chapter, other insights into related challenges also came to light during our experiments. We found that the data preparation pipeline adopted for this study could be improved significantly, as it created some conflicting cases during training. This is because of the nature of the MIMIC-III dataset itself, in which the radiology reports do not have a direct link to ICD9 disease codes. To overcome this problem, we designed a strategy for extracting ICD9 codes from the DIAGNOSES_ICD table to assign them to all patients with the same patient identifier (SUBJECT_ID) and hospital admission identifier (HADM_ID) in the radiology notes corpus. An unforeseen side-effect of this strategy was that, sometimes, the ICD9 disease codes/groups assigned to radiology text reports might not be related to that particular disease. This could have reduced the model's performance because of the assignment of conflicting labels to textual features of radiology notes. Nevertheless, the presented DPMs achieved promising results and the good performance indicates that the prediction model was able to capture disease-specific features using the information present in the clinical notes during the patients' admission.

6.5 Summary

In this chapter, three disease group prediction models were presented. Firstly, a novel ontology-driven text feature modeling approach and word embedding models on radiology notes was presented, based on which effective textual feature representations were generated which were then used to train a deep neural network classifier. Next, a two-stage feature modeling approach where word embedding models and a novel PSO-NN wrapper were used on radiology notes to derive effective textual feature representation. These were then used to train a modified TextCNN classifier for disease group prediction. The third work consisted of a hybrid feature modeling approach that used dynamic weighted combination of pretrained and trained word embedding models for generating effective textual feature representation to train a deep neural network classifier. These works focused on modeling unstructured clinical notes for enabling disease group prediction, and benchmarking these against state-of-the-art disease prediction models built on structured clinical data. Experimental evaluation emphasised the superior performance of the proposed models by a significant margin. Furthermore, the proposed strategies also help eliminate dependency on availability of structured clinical records, in addition to being suitable for deployment as a real world CDSS, especially in developing countries where structured EHR adoption is low.

Publications

(based on work presented in this chapter)

1. Gokul S. Krishnan, Sowmya Kamath S., "*Ontology-driven Text Feature Modeling for Disease Prediction using Unstructured Radiological Notes*", Computación y Sistemas, ISSN 2007-9737, 23(3), 2019. (Scopus & ESCI) *(Published)*

2. Gokul S. Krishnan, Sowmya Kamath S., "*A Deep Neural Network Model for Predicting Disease Groups based on PSO-NN Two-stage Feature Model-*

ing of Unstructured Clinical Notes", ACM Transactions on Computing for Healthcare. (SCIE & Scopus) *(Under Review)*

3. Gokul S. Krishnan, Sowmya Kamath S., "*Hybrid Text Feature Modeling for Disease Group Prediction using Unstructured Physicians' Notes*", International Conference on Computational Science (ICCS) 2020. (CORE Ranked) *(Published)*

Chapter 7

Aggregation based Disease Prediction using Unstructured Clinical Data

7.1 Introduction

In the previous chapter, we presented the design of DPMs based on radiology and physician notes. In this chapter, we focus on one of the most challenging aspects of clinical data management, i.e. the streaming and incremental nature of the patient data. During every hospital visit, the patients are assessed and their conditions are diagnosed/monitored, resulting in generation of different types of notes. This essentially is temporal data, i.e., can provide insights into the evolution of patient condition since the time it was diagnosed to the current status. Clinical notes like nursing notes, radiology notes, etc. can be frequent and even redundant in a patient's single admission. Even minute variations in conditions of ICU patients are recorded and monitored regularly by trained nursing staff. Hence, nursing notes are very data-rich voluminous resources containing continuously documented subjective and objective assessments concerning a patient's state. Moreover, effective modeling of such clinical text to aid in the early identification of high-risk patients is of utmost importance, to provide prioritized care and prevent further complications. In our previous works, we considered each unstructured clinical note as a separate entity. However, in a real world hospital scenario, a real need to deal with new patient data as it is generated is a growing need and there exists a conspicuous gap in existing research towards addressing this. Thus, well-defined strategies for modeling the unstructured patient data and leveraging it for designing learnable decision-support models are essential.

7.1.1 Problem Definition

In hospitals, voluminous and varied unstructured clinical notes such as radiology notes, nursing notes, physician notes, prescriptions, etc. get recorded during the single admission of a patient. Clinical notes maintained by caregivers, record subjective assessments and crucial information concerning a patient's state, which is mostly lost when converted into structured EHRs (Dubois *et al.*, 2017). Mining and modeling such nursing notes for extracting rich patient data and utilizing this to predict clinical events and outcomes with machine learning models is a challenging process, owing to their rawness, redundancy, high-dimensionality, sparsity, complex temporal and linguistic structure, and presence of rich medical jargon and abbreviations (Dubois *et al.*, 2017; Jo *et al.*, 2017). The efficacy of using such raw clinical notes largely depends on the ability to extract and consolidate the information embedded in them effectively (Wang *et al.*, 2018). Disease coding and grouping (ICD9 code and group prediction) and risk assessment via nursing notes can aid in taking effective measures at the earliest signs of patient distress. Recognition of the onset of disease and the determination of its risk using clinical nursing notes, followed by effective communication and response by interdisciplinary care team members could be both time- and cost-efficient (Davis *et al.*, 2008), which can also lead to reduced hospital mortality rate Collins *et al.* (2013). In this context, the problem that is observed and we aim to address is defined as follows:

> *"Given multiple patient records in the form of unstructured clinical notes recorded during a single admission of an ICU patient, to design and develop effective aggregation, textual feature modeling and prediction approaches for ICD9 disease group prediction."*

7.1.2 Motivating Example

Let us again consider the earlier example, the case of *Hospital A* and *Hospital C*. For a *Patient P* consulting the physician *Dr. Alice* at *Hospital A*, *Dr. Alice* reports a tentative diagnosis, say *D1*, prescribes medicines for the same, based on the symptoms described by *Patient P*. Let us assume the patient is advised to get admitted to the hospital and undergoes the prescribed treatment, however no improvement is observed. The nurses who monitor the patient's condition frequently report back to the physician that his treatment is not being effective. *Dr. Alice* investigates further and orders some tests to be performed on *Patient P*. After consulting the test results, *Dr. Alice* discovers that *Patient P* is in fact suffering from another condition (due to some symptoms/history that the patient failed to mention), say diagnosis *D2* and refers *Patient P* to specialist treatment.

Now, let us consider the scenario if the *Patient P* had visited *Hospital C* for consulting the physician *Dr. Charlie*. During the pre-screening, the reception enters patient symptoms into the EHR system. During the actual consultation, *Dr. Charlie* sees these, and asks additional questions relating to the symptoms, which again goes into the notes section of the EHR system. The patient is admitted and then some of the first observations also go into the EHR database as nursing notes. Now, the CDSS attached to *Hospital C*, has been trained with enough historical data (that of other patients with similar symptoms) gives out a prediction for Diagnosis *D1* with a confidence level of x, and also that there is also a small probability y that the patient may be suffering from condition *D2*. This insight is provided as an alert to *Dr. Charlie*, who can immediately take informed decisions such as ordering tests or referring the patient to the specialist. The CDSS attached to the hospital has the capability to detect diseases or diagnoses in earlier stages, which reduces the risk of misdiagnosis and additional wastage of time, potentially reducing complications and deterioration of patient condition.

In this chapter, two ICD9 disease group prediction (DPM) models that are built on novel aggregation strategies for unstructured clinical notes are presented. The first DPM incorporates an aggregation strategy we refer to as '*TAGS*', an acronym for *T*erm weighting of unstructured notes *AG*gregated using fuzzy *S*imilarity. *TAGS* is a novel fuzzy similarity scoring based cleansing aggregation approach that is employed to merge or purge nursing notes so as to improve the patient data representation derived from the unstructured nursing notes. The second DPM is built on a technique called *FarSight*, an aggregation model built on *TAGS* for future lookup of disease groups for assessing early symptoms observed, to enable early prediction of onset of diseases.

7.2 *TAGS* – Fuzzy Similarity based Aggregation of Unstructured Clinical Notes

The workflow of the proposed *TAGS* aggregation approach for ICD9 disease group prediction is depicted in Fig. 7.1. In an attempt to benchmark our work on open datasets, we used the MIMIC-III database for the evaluation experiments, which contains 2,083,180 note events, of which 223,556 are nursing notes of 7,704 distinct ICU patients (subjects). The specifics of the nursing note text corpus used in the experiments described in this chapter are summarized in Table 7.1.

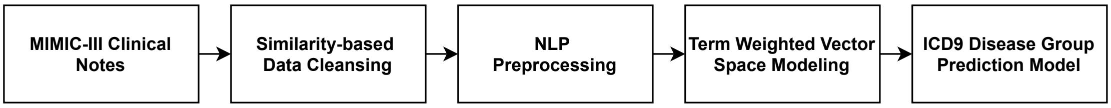

MIMIC-III Clinical Notes
Similarity-based Data Cleansing
NLP Preprocessing
Term Weighted Vector Space Modeling
ICD9 Disease Group Prediction Model

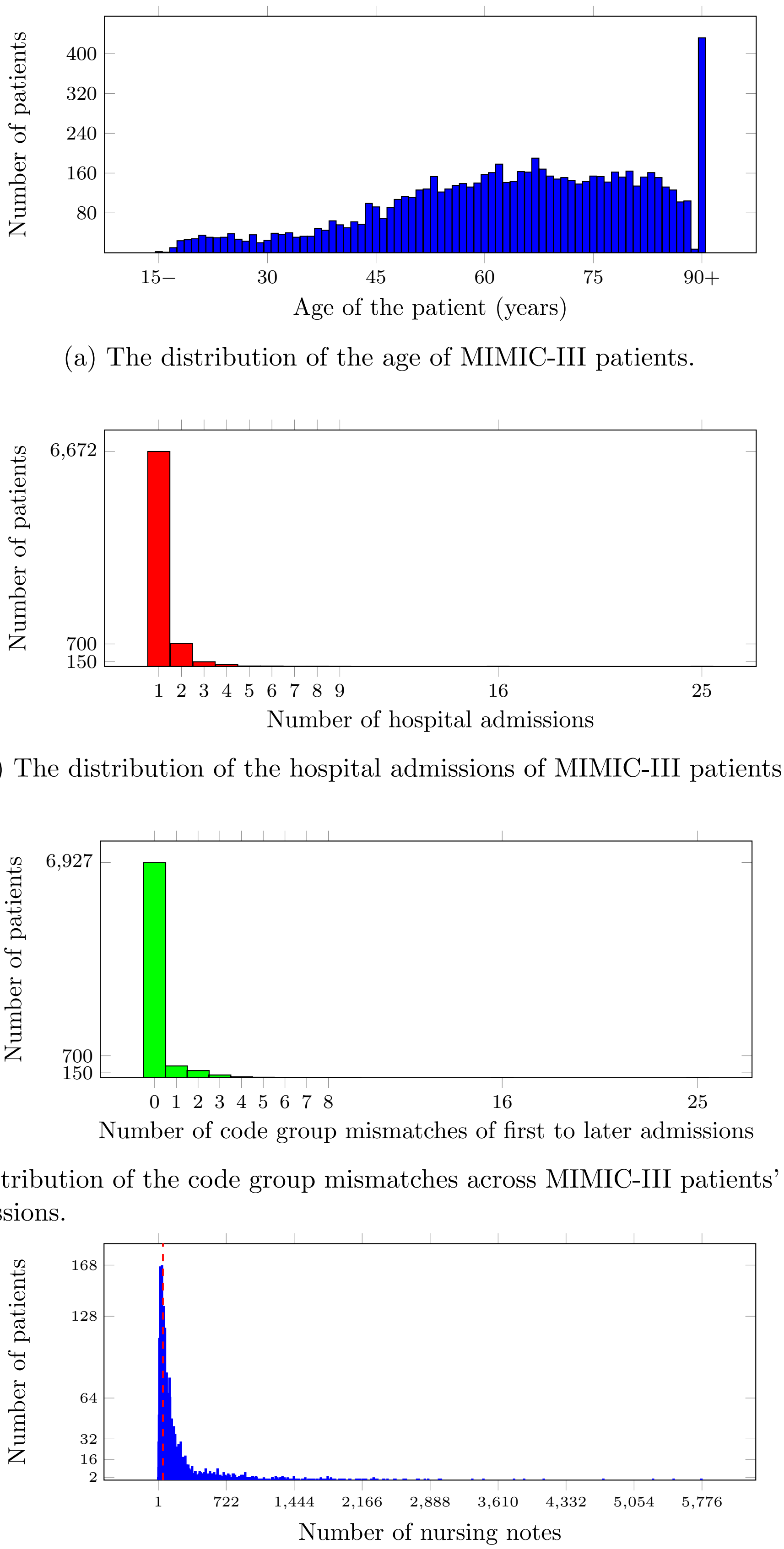
Figure 7.1: Proposed *TAGS* Model for ICD9 Disease Group Prediction based on Aggregated Clinical Notes

We considered two criteria to select the MIMIC-III subjects in the dataset preparatory phase. Firstly, the subjects with age less than 15 were identified using the age at the time of admission to the ICU. Similar to the state-of-the-art models (Johnson *et al.*, 2018; Purushotham *et al.*, 2018), only adult subjects (age 15 or above) were considered for the study. Secondly, for each MIMIC-III subject, only their first admission to the hospital was considered, and all later admissions were discarded. This was done to ensure the prediction with the earliest detected conditions, to enable faster risk prediction, avoid any information loss, and to ensure similar experimental settings as the state-of-the-art models (Johnson *et al.*, 2018; Purushotham *et al.*, 2018) considered for benchmarking. Figure 7.2c throws more light on the distribution of the number of code group mismatches across patients' first admission to their later admissions, based on which it can be observed that code groups in later admissions of over 94% of the patient nursing notes are the same as those occurring in their first hospital admission. Owing to this, we decided to consider only the first hospital admission of a MIMIC-III subject, with minimal loss of information. Based on these criteria, the selected patient cohort contained nursing notes corresponding to $7,638$ patients with a median age of 66 years. The statistics of the data extracted from the MIMIC database is shown in Figure 7.2.

Table 7.1: Statistics of the clinical nursing note text corpus.

Parameter	**Total**	**Average**
Clinical nursing notes	223,556	–
Sentences in the nursing notes	5,244,541	23.46
Words in the nursing notes	79,988,065	357.80
Unique words in the nursing notes	715,821	3.20

Due to various factors including outliers, noise, missing values, incorrect or duplicate records, and others, the data extracted from the MIMIC-III database had erroneous entries. Three major issues with the extracted data were identified



(a) The distribution of the age of MIMIC-III patients.

(b) The distribution of the hospital admissions of MIMIC-III patients.

(c) The distribution of the code group mismatches across MIMIC-III patients' first and later admissions.

(d) The distribution of nursing notes across various MIMIC-III subjects (red dashed line exhibits the distribution at 50 nursing notes).

Figure 7.2: Statistics of the data extracted from the MIMIC-III database.

and handled accordingly. Firstly, the erroneous entries in nursing notes with the ‘iserror’ attribute of the ‘noteevents’ table set to 1 were identified and removed. Secondly, some subjects that had redundant records were identified, and were deduplicated. The resulting data obtained by handling erroneous entries corresponded to $6,532$ MIMIC-III subjects. Finally, a few MIMIC-III subjects had multiple nursing notes with different ICD9 code groups, which were merged or purged using a fuzzy token-based similarity approach, referred to as *TAGS*. We discuss the methodology adopted for this process in detail in the Section 7.2.1.

7.2.1 Fuzzy Token-based Similarity Merging

The objective here to handle the multiple nursing notes of a MIMIC-III subject, by applying merging, to enable multi-label ICD9 code group classification. Figure 7.2d shows the heavy-tailed distribution of nursing notes across various patients, from which it is evident that the patient cohort has an average of 176.49 nursing notes per patient, with $4,183$ patients having more than fifty nursing notes composed of over $17,890$ words on an average. Such voluminous nursing notes often include many similar terms which could significantly affect the vector representations. To handle the voluminosity and near-duplicate nursing notes of a patient, Monge-Elkan (ME) (Monge and Elkan, 1997), a token-based fuzzy similarity scoring scheme is integrated with Jaro internal scoring scheme (Jaro, 1989) and used as a decision-making mechanism. ME similarity is used to handle clinical abbreviations, alternate names, and medical jargon in which Jaro similarity is used as an internal scoring scheme to handle typographical errors and to obtain a normalized similarity score between 0 and 1. Given two nursing notes η_i and η_j with $|\eta_i|$ and $|\eta_j|$ tokens ($\mathcal{C}_k^{(i)}$s and $\mathcal{C}_l^{(j)}$s) respectively, their ME similarity score with Jaro is,

$$\mathrm{ME_{Jaro}}(\eta_i, \eta_j) = \frac{1}{|\eta_i|} \sum_{k=1}^{|\eta_i|} \max \left\{ \mathrm{Jaro}(\mathcal{C}_k^{(i)}, \mathcal{C}_l^{(j)}) \right\}_{l=1}^{|\eta_j|} \tag{7.1}$$

where, the Jaro similarity score of two given clinical terms (tokens) $\mathcal{C}_i$ of length $|\mathcal{C}_i|$ and $\mathcal{C}_j$ of length $|\mathcal{C}_j|$ with m matching characters and t transpositions is,

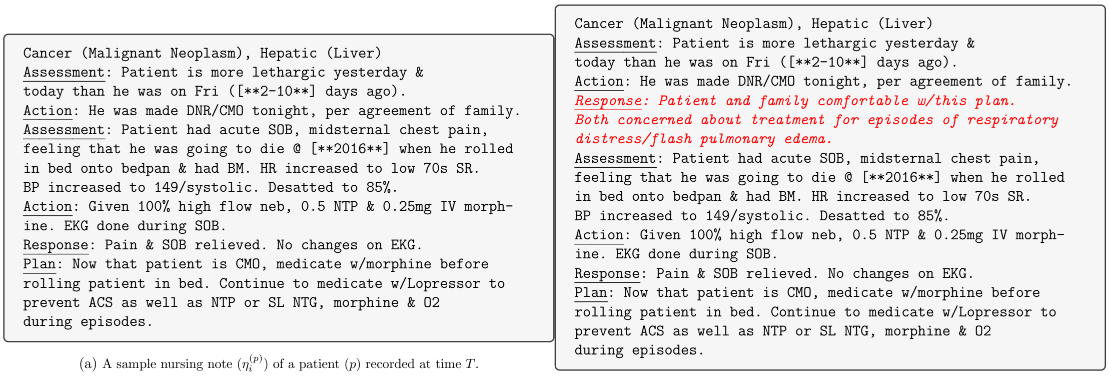
$$\mathrm{Jaro}(\mathcal{C}_i, \mathcal{C}_j) = \begin{cases} 0, & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|\mathcal{C}_i|} + \frac{m}{|\mathcal{C}_j|} + \frac{2m-t}{2m}\right), & \text{otherwise} \end{cases} \tag{7.2}$$

The nursing notes of a patient are processed in the order of oldest to the most recent. Based on the predetermined similarity threshold (θ) ranging between 0 and 1, a pair of nursing notes ($\eta_i^{(k)}$, $\eta_j^{(k)}$) corresponding to a patient ($\mathcal{P}^{(k)}$) are merged only if $\text{ME}_{\text{Jaro}}(\eta_i^{(k)}, \eta_j^{(k)})$ is less than θ, else $\eta_j^{(k)}$ is retained and $\eta_i^{(k)}$ is purged, thus maintaining only the latest of the two nursing notes. Note that, similarity merging and purging applies only to nursing notes and not to the ICD9 code groups. Corresponding ICD9 codes across various nursing notes of a patient are merged to enable multi-label classification. The resultant nursing note for a patient $\mathcal{P}^{(k)}$ after merging is hereafter referred to as the *aggregate nursing note* of that patient. For the purpose of this work, we have empirically determined the fuzzy-similarity θ to be 0.825 using grid search.

Consider two sample nursing notes ($\eta_i^{(p)}$ and $\eta_j^{(p)}$) of a patient (p) extracted from the MIMIC-III database, recorded at times T (shown in Figure 7.3a) and $T' > T$ (shown in Figure 7.3b) respectively. It can be observed that both the recorded nursing notes are quite similar—the nursing note recorded at time T' records all the details in the nursing note $\eta_i^{(p)}$, along with additional 'response' concerning the patient's state. To handle the voluminosity of the nursing notes and delete the near-duplicate nursing notes, we compute the ME similarity (with internal Jaro similarity scoring) score using Eq. (7.1). The nursing notes shown in Figure 7.3 have an ME similarity score of 0.85, which is higher than the preset threshold of 0.825. Thus, note $\eta_j^{(p)}$ is retained, and note $\eta_i^{(p)}$ is purged.

7.2.2 Preprocessing

The next phase in the NLP pipeline is to preprocess the nursing notes to achieve data (text) normalization. Transformation of text into a canonical form allows for the separation of concerns and helps maintain consistency. Preprocessing es-



Cancer (Malignant Neoplasm), Hepatic (Liver)
Assessment: Patient is more lethargic yesterday &
today than he was on Fri ([**2-10**] days ago).
Action: He was made DNR/CMO tonight, per agreement of family.
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. HR increased to low 70s SR.
BP increased to 149/systolic. Desatted to 85%.
Action: Given 100% high flow neb, 0.5 NTP & 0.25mg IV morph-
ine. EKG done during SOB.
Response: Pain & SOB relieved. No changes on EKG.
Plan: Now that patient is CMO, medicate w/morphine before
rolling patient in bed. Continue to medicate w/Lopressor to
prevent ACS as well as NTP or SL NTG, morphine & O2
during episodes.

(a) A sample nursing note ($\eta_i^{(p)}$) of a patient (p) recorded at time T.

Cancer (Malignant Neoplasm), Hepatic (Liver)
Assessment: Patient is more lethargic yesterday &
today than he was on Fri ([**2-10**] days ago).
Action: He was made DNR/CMO tonight, per agreement of family.
Response: Patient and family comfortable w/this plan.
Both concerned about treatment for episodes of respiratory
distress/flash pulmonary edema.
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. HR increased to low 70s SR.
BP increased to 149/systolic. Desatted to 85%.
Action: Given 100% high flow neb, 0.5 NTP & 0.25mg IV morph-
ine. EKG done during SOB.
Response: Pain & SOB relieved. No changes on EKG.
Plan: Now that patient is CMO, medicate w/morphine before
rolling patient in bed. Continue to medicate w/Lopressor to
prevent ACS as well as NTP or SL NTG, morphine & O2
during episodes.

(b) A sample nursing note ($\eta_j^{(p)}$) of a patient (p) recorded at time T' ($> T$).

Figure 7.3: Two sample de-identified nursing notes from the MIMIC-III database. The two nursing notes are quite similar, while the only new content is the updated response (indicated as red italicized text).

sentially includes tokenization, stopword removal, and stemming/lemmatization. First, multiple spaces, special characters, and punctuation marks are removed. During tokenization, the clinical notes' text is split into several smaller tokens (words). Stopwords from the generated tokens are removed using the NLTK English stopword corpus. Furthermore, character case folding is performed, and references to images (file names such as '*scanImage.png*') are removed. It is to be noted that, token-length based token removal was not performed to avoid the loss of important medical information (such as '*CT*' in '*CT Scan*'). Finally, stemming was performed for suffix stripping, followed by lemmatization to convert the stripped tokens to their base forms. To eliminate overfitting and lower the computational complexity, the tokens appearing in less than ten nursing notes were removed before any further processing.

7.2.3 Vector Space Modeling of Aggregated Clinical Notes

An adaptation of the Bag of Words (BoW) that weighs each token in an unsupervised way, is used as the term weighting scheme in the *TAGS* model. It is a numerical statistic that captures both the importance and specificity of a term in the given vocabulary. The weight ($W_m^{(i)}$) of a term $w_m^{(i)}$ (of total $|w^{(i)}|$ terms) in a nursing note η_i (of total N nursing notes) occurring $f_m^{(i)}$ times is given by,

$$W_m^{(i)} = \begin{cases} \left(1 + \log_2 f_m^{(i)}\right)\left(\log_2 \frac{N}{|w^{(i)}|}\right), & \text{if } f_m^{(i)} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{7.3}$$

The weight of every term in a patient's aggregate nursing note ($\mathcal{P}^{(k)}$) is computed to obtain a vector $\mathcal{V}^{(k)} \in \mathbb{R}^{|\mathbb{V}|}$. Now, the patient information is in machine processable form, which can be considered features for training machine learning classifiers.

7.2.4 ICD9 Disease Code Grouping

As discussed earlier, ICD9 codes are a taxonomy of diagnostic codes that are used by doctors, public health agencies, and health insurance companies across the world to classify diseases and a wide variety of infections, disorders, symptoms, causes of injury, and others. Owing to the high granularity of ICD9 codes, researchers suggested differentiating between category-level (group) predictions and full-code predictions (Larkey and Croft, 1995). Each ICD9 code group includes a set of similar diseases, and almost every health condition can be represented with a unique ICD9 code group. In this study, we focus on ICD9 code group predictions as a multi-label classification problem, with each patient's nursing note mapped to more than one group. All the ICD9 codes assigned to a patient's admission are grouped into ICD9 disease groups, as previously explained in Section 6.2.2. In this study, the E and V codes are classified into the same code group to lower the computational cost of training. Moreover, the ICD9 group with codes ranging from 760-779 were not considered as this group corresponds to conditions originating in the perinatal period and usually assigned to neonates (age $<$ 15). Therefore, the ICD9 groups were classified into 19 diagnosis groups.

7.2.5 ICD9 Disease Code Group Prediction Model

In this section, the proposed prediction algorithms employed to achieve the task of ICD9 code group multi-label classification are presented. We experimented with eight different prediction models conforming to various algorithmic classes including algorithm adaptation based, problem transformation based, and ensemble models. Three classifiers – KNN, Logistic Regression (LR), and Support Vector Machine (SVM) were utilized as One vs Rest (OvR) classifiers in the prediction of ICD9 diagnosis code groups. LR, as explained in Section 4.2.3, is a discriminative model that models the probabilities of possible outcomes using a logistic function.

K-Nearest Neighbors (KNN). KNN is a non-parametric instance-based learner used in regression and classification tasks. In KNN classification, the output class membership is determined by the majority vote of its K closest neighbors. In the sense of multi-label classification, KNN first identifies the K closest neighbors and then, based on the statistical inferences gained from the neighboring class label sets, maximum a posteriori principle is used to determine the class label set of an unseen instance.

Multi-Layer Perceptron (MLP). MLP is a feed-forward artificial neural network with an input layer, one or more hidden layers, and one prediction layer at the end, for classification. The first layer takes vector representations of the clinical terms as the input and uses the output of each layer as the input to the following layer with the help of a non-linear activation function such as a tanh, sigmoid, softmax or ReLU. In training, to update the weights and biases, MLP uses a supervised approach called Backpropagation (BP). BP is used to calculate the gradient of the loss function to update weights, which aids the MLP to learn the internal representations, allowing it to learn any arbitrary mappings within the network. In the case of multi-label classification, while the forward pass remains the same, the classical BP algorithm uses a global error function that addresses the dependencies between the class labels. In this study, we use vanilla neural networks with one hidden layer of 75 nodes and a ReLU activation function, which were empirically determined using grid search.

Support Vector Machines (SVM). SVM is also a discriminative approach that classifies by constructing hyperplane(s) in a high-dimensional space. For a given set of linear separable training instances, SVM finds a linear rule that maximizes (optimizes) the geometric margin (street width). In practice, most of the training sets are not usually linearly separable. Now, a trade-off between minimizing prediction error and maximizing the geometric margin must be incorporated. Kernels such as tanh, sigmoid, Radial Basis Function (RBF), and others are generally used to transform from the linearly inseparable space to a higher dimensional space where the points could be separated. The RBF kernel defines a space that is larger than linear or polynomial kernels and has properties such as being stationary, isotropic, and infinitely smooth. Thus, in this analysis, we used SVM with an RBF kernel with γ set to 1/#features.

One vs. Rest (OvR). OvR prediction strategy essentially transforms the multi-label classification problem into multiple binary relevance tasks. OvR trains a classifier such that for each class, the samples (aggregate nursing notes) of that particular class are considered as positive and the remaining samples as negative. The base classifiers produce a real-valued confidence score for the prediction decision. Then, for an unseen instance, the combined model of all such classifiers predicts all the class labels for which the corresponding base classifiers predicted a positive result.

Ensemble Approaches. Three ensemble prediction approaches including Random Forest (RF), Hard-voting Ensemble (HVE), and Stacking Ensemble (SE) were also employed in the classification of ICD9 diagnostic code groups. RF or decision tree ensembles (explained in Section 4.2.3) predict by constructing multiple Classification And Regression Trees (CARTs) during training and predict the output class as a function of the outputs of individual trees for the test data. In multi-label classification, multiple labels are present in the tree leaves, and the predictions of multiple base CARTs are combined using a simple voting scheme (such as probability distribution or majority vote). In this work, we used RF with 100 CARTs with a maximum depth of 2. HVE aggregates the predictions of multiple diverse classifiers using a majority rule. Given a set of diverse classifiers (N_is) with prediction sets $\mathcal{Y}_i s$, where each $\mathcal{Y}_i$ a subset of $\mathbb{Y}$ (set of all class labels), then the presence of a class (c) in an unseen instance ($\eta^{(m)}$) can be estimated as,

$$\mathcal{Y}^{(m)}(c) = \begin{cases} 1, & \text{if } \sum_{i=1}^{N} \mathcal{Y}_i^{(m)}(c) > \lceil \frac{N}{2} \rceil \\ 0, & \text{otherwise} \end{cases} \tag{7.4}$$

Thus, using the majority voting principle, the possible class label set for the unseen instance can be predicted. Many variations on the classifiers used in HVE were tried, starting with KNN, MLP, LR, LR as OvR, SVM as OvR, and KNN as OvR. After much experimentation, only MLP, LR as OvR, and SVM as OvR were used, due to their superior performance.

SE (Wolpert, 1992) also combines discrete learning algorithms using a meta-classifier. In the first phase, all the base classifiers (N_is) are applied to the training data which generate the predictions ($\mathcal{Y}_i$s). Then, in the second phase, a meta-level dataset is created by replacing every trained record ($\eta^{(k)}$) with the predictions for that record $(\mathcal{Y}_i^{(k)})_{i=1}^{N}$. Then, another learning algorithm (L) is used to classify the meta-level dataset. On an unseen testing instance η_m, the predicted class set is $L(\mathcal{Y}_i^{(m)})_{i=1}^{N}$. In this study, MLP, LR as OvR, and SVM as OvR are used as first-level classifiers, and MLP is used as the second-level classifier. In contrast to voting, SE learns at the meta-level, when combining multiple classifiers. The results of the comparative performance of these prediction models are discussed in Section 7.2.6.

7.2.6 Experimental Results and Discussion

To validate the proposed approach, we performed extensive experiments over the nursing notes data obtained from the MIMIC-III database. Seven standard evaluation metrics – Accuracy, Area Under the ROC Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), Matthews Correlation Coefficient (MCC), F-score, Coverage Error (CE) and Label Ranking Loss (LRL), were used to assess the performance of each prediction algorithm with reference to each data modeling approach. The implementations in the Python sklearn and gensim packages were used for the experiments and all experiments were performed in an Ubuntu based High-end Server with a 56-core Intel Xeon processor, 128GB RAM, two Nvidia Tesla M40 GPUs (24GB each) and 3TB hard drive. Exhaustive comparative study of the performance of various data and modeling approaches on the nursing notes of the MIMIC-III database was performed. For the prediction task of ICD9 code group classification, 10-fold cross-validation was performed. Furthermore, the mean and standard errors (of the mean) of the performance scores are presented. Table 7.5 shows the performance of all data modeling approaches and all prediction models using nursing notes processed using *TAGS* fuzzy token-based similarity with $\theta = 0.825$. Table 7.6 tabulates the performance of all data modeling approaches and all prediction models using nursing notes processed without any similarity modeling or aggregation. We also observed the performances of other feature modeling strategies such as embedding approach Doc2Vec and topic modeling approaches.

Doc2Vec aims at numerically representing variable length documents as fixed length low dimensional document embeddings (vectors). Doc2Vec is essentially a neural network with one shallow hidden layer that learns the distributed representations, to provide a content-related measurement. It incorporates semantic textual features obtained from the nursing notes text corpus. The Paragraph Vector (PV) Distributed Memory (PV-DM) variant of Doc2Vec was chosen over PV Distributed Bag-of-Words (PV-DBoW) due to its ability to preserve the word order in the nursing notes and its comparatively superior performance (Le and Mikolov, 2014). For an exhaustive analysis, Doc2Vec dimension sizes of 500 (trained for 25 epochs) and $1,000$ (trained for 50 epochs) were used.

A popular cluster analysis approach, Latent Dirichlet Allocation (LDA) is a generative topic model based on the Bayesian framework of a three-layer structure (documents, topics, and terms). LDA generates a soft probabilistic and flat clustering of terms into topics and documents into topics. LDA posits that each (aggregate) nursing note $\eta_i^{(k)}$ of a patient $\mathcal{P}^{(k)}$ and each term belongs to a set of d ($\ll |\mathbb{V}|$) clusters (topics) $\mathcal{T}$, with some probability ρ. Thus, each nursing note is transformed into vectors of topic probabilities, which could be considered features for training classifiers.

Similar to other clustering approaches, there is no simple way to determine the correct number of d LDA clusters. To cope with this issue, more complex models such as Hierarchical Dirichlet process (HDP) which automatically determine the number of clusters through posterior inference can be used. HDP is a hierarchical Bayesian non-parametric model that can model mixed-membership data with potentially infinite terms, in an unsupervised way. In LDA, only the mixture of topics is drawn from the Dirichlet distribution, while in HDP, a Dirichlet process is used to capture the uncertainty in the number of terms, as is typically the case with clinical documents.

Probabilistic models are commonly evaluated by measuring the log-likelihood of unseen documents. As an alternative to HDP, Topic Coherence (TC) between topics can also be used to derive the optimal number of topics. TC is a way to evaluate topic models with a much greater guarantee of human interpretability. In this work, we adopted LDA with Topic Coherence (TC) as it accounts for the semantic similarity between high scoring terms. C_v, a variant of coherence measurement is used in this study, as it accounts for high correlation with all the available human ranking data (Röder *et al.*, 2015). The higher the coherence value, the stronger is the model's human interpretability and generalization ability.

To provide exhaustive analysis, HDP with truncation level set to 150 was modeled with both BoW and term weighting. Alternatively, LDA (set to 100 topics) with TC was modeled with BoW representations. Furthermore, the number of LDA topics was determined by comparing the TC scores of several LDA models obtained by varying the number of LDA topics from 2 to 500 in the increments of 100. We observe that the Term weighting of unstructured (nursing) notes AGgregated using fuzzy Similarity ($TAGS$) model, modeled with LR as OvR, consistently outperforms more complex vector space and topic models. Furthermore, it can be observed from Figure 7.4 that the model's performance is higher when nursing notes are processed with similarity modeling, than when processed without similarity modeling.

In clinical tasks such as disease prediction, capturing True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives(FN) is of utmost importance, due to the critical nature of the task itself. As can be seen from the results in Tables 7.5 and 7.6, the AUROC metric captures True Positive Rate (TPR) and False Positive Rate (FPR), while AUPRC captures the number of true positives from positive predictions. AUPRC, unlike AUROC, varies with the change in the ratio of target classes in the data, and hence is more revealing while evaluating imbalanced data (Saito and Rehmsmeier, 2015). From Table 7.4, it can be observed that the dataset is highly class imbalanced, and hence AUPRC is more informative than AUROC. It can be seen that our approach outperforms the existing state-of-the-art method by Purushotham *et al.* (2018) in these metrics, indicating the significant decrease in the FP and FN. F-score captures both precision and recall of the prediction, while MCC score serves as a balanced measure even with class imbalance, as it takes into account TP, FP and FN. More specifically, in healthcare applications like disease or diagnosis prediction, FN (prediction miss, i.e., a disease which is present, but not diagnosed) are likely to cause more harm than false positives (false alarm) and CE captures these false negatives. LRL performs a pairwise label comparison to determine the loss of prediction. Existing works have benchmarked their performance using only AUROC and AUPRC metrics. Since all the metrics used in this research are very relevant and essential in understanding the proposed model's predictive power, we benchmark these promising results for the MIMIC-III dataset.

Furthermore, as explained earlier, the state-of-the-art work by Purushotham *et al.* (2018) is built on structured EHRs that are modeled in the form of feature sets to make clinical predictions. It is a fact that the richness and abundance of information captured by unstructured nursing notes are often lost in the structured EHRs coding process (Dubois *et al.*, 2017). Our proposed *TAGS* model combines the fuzzy similarity based data cleansing and aggregating approach with a term weighting scheme that captures the importance and rarity of clinical concepts, to model the informally written clinical nursing text into a clinically relevant and usable format effectively. From the results, it can be seen that in contrast to more complex data modeling approaches such as Doc2Vec and HDP, the *TAGS* model is able to capture all the discriminative features of the clinical nursing notes needed for the machine learning classifier to learn and generalize. We observe that using the *TAGS* model, risk stratification can be achieved well in advance, with an overall accuracy of 82.4%. Also, it can be noted that token-based similarity processing of nursing notes yields higher performance in comparison to that processed without similarity modeling. These promising results emphasize the need for reduction in redundancy and anomalous data for relieving the cognitive burden and improving the clinical decision-making process. CDSSs built on the predictive capabilities of *TAGS* could be suitable for patient-centric and evidence-based treatments, resulting in reduced mortality rates and better risk assessment.

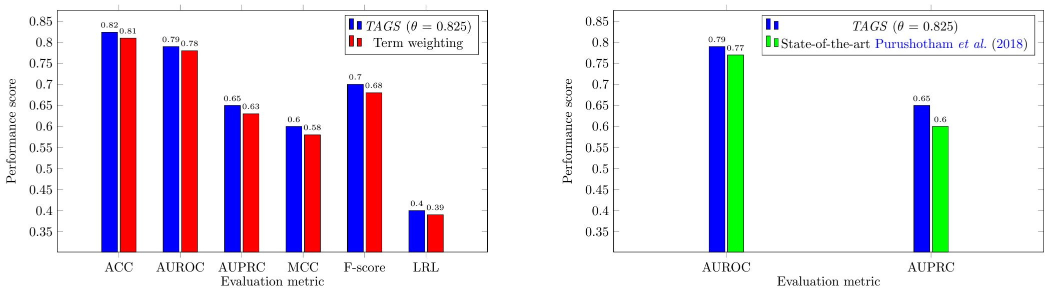


Figure 7.4: Benchmarking the best performing models (with and without fuzzy similarity modeling) against Purushotham *et al.* (2018)'s model

Table 7.2: ICD9 code group prediction using nursing notes of MIMIC-III (using fuzzy similarity with $\theta = 0.825$).

Data model	**Classifier**	**Performance scores**						
		ACC	**AUROC**	**AUPRC**	**MCC**	**F-score**	**CE**	**LRL**
TAGS $(6,532 \times 14,650)$	KNN	0.7857 ± 0.0011	0.7681 ± 0.0010	0.5904 ± 0.0016	0.5286 ± 0.0019	0.6688 ± 0.0017	18.0936 ± 0.0501	0.4181 ± 0.0018
	MLP	0.7947 ± 0.0009	0.7677 ± 0.0013	0.5987 ± 0.0018	0.5366 ± 0.0020	0.6664 ± 0.0018	18.2327 ± 0.0574	0.4226 ± 0.0024
	KNN as OvR	0.7725 ± 0.0018	0.7645 ± 0.0011	0.5738 ± 0.0021	0.5108 ± 0.0024	0.6619 ± 0.0017	**17.9385 ± 0.0791**	0.4204 ± 0.0020
	LR as OvR	**0.8239 ± 0.0011**	**0.7868 ± 0.0011**	**0.6476 ± 0.0011**	**0.5953 ± 0.0018**	**0.6981 ± 0.0016**	18.2849 ± 0.0643	**0.3978 ± 0.0021**
	SVM as OvR	0.7413 ± 0.0014	0.6801 ± 0.0011	0.5249 ± 0.0014	0.4007 ± 0.0024	0.5207 ± 0.0019	19.5542 ± 0.0206	0.5880 ± 0.0018
	RF	0.7630 ± 0.0012	0.6926 ± 0.0009	0.5486 ± 0.0014	0.4388 ± 0.0022	0.5450 ± 0.0016	19.5678 ± 0.0238	0.5728 ± 0.0014
	HVE	0.8171 ± 0.0010	0.7781 ± 0.0007	0.6367 ± 0.0007	0.5786 ± 0.0007	0.6837 ± 0.0009	18.5659 ± 0.0614	0.4132 ± 0.0014
	SE	0.7972 ± 0.0009	0.7698 ± 0.0015	0.6027 ± 0.0021	0.5421 ± 0.0016	0.6701 ± 0.0017	18.2673 ± 0.0630	0.4195 ± 0.0029
Doc2Vec 500 $(6,532 \times 500)$	KNN	0.7399 ± 0.0020	0.6628 ± 0.0027	0.5247 ± 0.0021	0.3949 ± 0.0041	0.4802 ± 0.0055	19.5644 ± 0.0278	0.6363 ± 0.0058
	MLP	0.7368 ± 0.0009	0.7102 ± 0.0012	0.5240 ± 0.0020	0.4150 ± 0.0023	0.5911 ± 0.0021	18.8039 ± 0.0450	0.5078 ± 0.0021
	KNN as OvR	0.7377 ± 0.0016	0.6674 ± 0.0024	0.5206 ± 0.0015	0.3888 ± 0.0030	0.4902 ± 0.0052	19.5144 ± 0.0269	0.6197 ± 0.0055
	LR as OvR	0.7950 ± 0.0013	0.7579 ± 0.0011	0.5970 ± 0.0018	0.5262 ± 0.0023	0.6607 ± 0.0017	**18.6491 ± 0.0375**	0.4400 ± 0.0019
	SVM as OvR	**0.8059 ± 0.0013**	**0.7666 ± 0.0010**	**0.6184 ± 0.0012**	**0.5514 ± 0.0022**	**0.6743 ± 0.0015**	18.7379 ± 0.0462	**0.4273 ± 0.0017**
	RF	0.7484 ± 0.0013	0.6787 ± 0.0010	0.5356 ± 0.0010	0.4142 ± 0.0021	0.5190 ± 0.0018	19.6208 ± 0.0225	0.5991 ± 0.0019
	HVE	0.8013 ± 0.0014	0.7636 ± 0.0011	0.6084 ± 0.0016	0.5407 ± 0.0024	0.6691 ± 0.0012	18.6652 ± 0.0149	0.4312 ± 0.0015
	SE	0.8047 ± 0.0014	0.7652 ± 0.0011	0.6164 ± 0.0008	0.5482 ± 0.0023	0.6715 ± 0.0014	18.7367 ± 0.0483	0.4296 ± 0.0017
Doc2Vec $1,000$ $(6,532 \times 1,000)$	KNN	0.7322 ± 0.0018	0.6543 ± 0.0030	0.5104 ± 0.0016	0.3741 ± 0.0036	0.4650 ± 0.0062	19.6614 ± 0.0478	0.6494 ± 0.0072
	MLP	0.7458 ± 0.0011	0.7170 ± 0.0013	0.5307 ± 0.0011	0.4291 ± 0.0025	0.5989 ± 0.0015	18.8467 ± 0.0374	0.4988 ± 0.0021
	KNN as OvR	0.7376 ± 0.0017	0.6712 ± 0.0029	0.5189 ± 0.0013	0.3883 ± 0.0035	0.5020 ± 0.0057	19.5014 ± 0.0415	0.6074 ± 0.0068
	LR as OvR	0.7735 ± 0.0015	0.7414 ± 0.0017	0.5667 ± 0.0015	0.4845 ± 0.0030	0.6374 ± 0.0019	18.7376 ± 0.0526	0.4623 ± 0.0029
	SVM as OvR	**0.8067 ± 0.0012**	**0.7693 ± 0.0013**	**0.6187 ± 0.0012**	**0.5542 ± 0.0021**	**0.6762 ± 0.0016**	**18.6286 ± 0.0472**	**0.4227 ± 0.0023**
	RF	0.7464 ± 0.0012	0.6760 ± 0.0010	0.5334 ± 0.0014	0.4102 ± 0.0020	0.5136 ± 0.0018	19.6269 ± 0.0248	0.6045 ± 0.0020
	HVE	0.7904 ± 0.0015	0.7562 ± 0.0018	0.5922 ± 0.0017	0.5201 ± 0.0033	0.6566 ± 0.0022	18.6607 ± 0.0545	0.4413 ± 0.0033
	SE	0.8052 ± 0.0015	0.7680 ± 0.0013	0.6164 ± 0.0009	0.5510 ± 0.0025	0.6738 ± 0.0016	18.6683 ± 0.0402	0.4249 ± 0.0023
HDP with BoW $(6,532 \times 150)$	KNN	0.7718 ± 0.0009	0.7422 ± 0.0009	0.5723 ± 0.0017	0.4892 ± 0.0018	0.6318 ± 0.0014	18.7632 ± 0.0514	0.4629 ± 0.0014
	MLP	**0.7912 ± 0.0011**	**0.7557 ± 0.0012**	**0.5974 ± 0.0014**	**0.5255 ± 0.0019**	**0.6502 ± 0.0019**	**18.6689 ± 0.0330**	**0.4464 ± 0.0022**
	KNN as OvR	0.7682 ± 0.0008	0.7397 ± 0.0010	0.5661 ± 0.0019	0.4822 ± 0.0018	0.6275 ± 0.0014	18.7482 ± 0.0380	0.4666 ± 0.0016
	LR as OvR	0.7815 ± 0.0010	0.7417 ± 0.0011	0.5850 ± 0.0014	0.5017 ± 0.0020	0.6251 ± 0.0016	18.9294 ± 0.0476	0.4729 ± 0.0020
	SVM as OvR	0.7511 ± 0.0011	0.6875 ± 0.0008	0.5410 ± 0.0015	0.4245 ± 0.0019	0.5284 ± 0.0017	19.4253 ± 0.0279	0.5827 ± 0.0015
	RF	0.7574 ± 0.0015	0.6915 ± 0.0014	0.5486 ± 0.0017	0.4359 ± 0.0028	0.5412 ± 0.0023	19.5291 ± 0.0314	0.5751 ± 0.0026
	HVE	0.7826 ± 0.0013	0.7404 ± 0.0015	0.5869 ± 0.0008	0.5029 ± 0.0022	0.6229 ± 0.0020	18.9688 ± 0.0626	0.4767 ± 0.0029
	SE	0.7851 ± 0.0008	0.7453 ± 0.0008	0.5874 ± 0.0014	0.5083 ± 0.0013	0.6317 ± 0.0006	18.7915 ± 0.0498	0.4660 ± 0.0014
HDP with term weighting $(6,532 \times 150)$	KNN	0.7116 ± 0.0015	0.6723 ± 0.0018	0.4887 ± 0.0023	0.3479 ± 0.0034	0.5254 ± 0.0031	19.3027 ± 0.0297	**0.5724 ± 0.0028**
	MLP	0.7409 ± 0.0016	0.6779 ± 0.0027	0.5245 ± 0.0014	0.3997 ± 0.0028	0.5158 ± 0.0056	19.5698 ± 0.0277	0.5940 ± 0.0069
	KNN as OvR	0.7076 ± 0.0014	0.6689 ± 0.0018	0.4842 ± 0.0024	0.3399 ± 0.0034	0.5213 ± 0.0031	**19.2999 ± 0.0269**	0.5764 ± 0.0028
	LR as OvR	0.7458 ± 0.0014	0.6780 ± 0.0010	0.5310 ± 0.0019	0.4082 ± 0.0027	0.5161 ± 0.0017	19.5929 ± 0.0258	0.5987 ± 0.0019
	SVM as OvR	0.7413 ± 0.0014	0.6801 ± 0.0011	0.5249 ± 0.0014	0.4007 ± 0.0024	0.5207 ± 0.0019	19.5542 ± 0.0206	0.5880 ± 0.0018
	RF	**0.7557 ± 0.0010**	**0.6880 ± 0.0008**	**0.5376 ± 0.0012**	**0.4257 ± 0.0017**	**0.5359 ± 0.0015**	19.4695 ± 0.0268	0.5800 ± 0.0015
	HVE	0.7414 ± 0.0017	0.6801 ± 0.0013	0.5249 ± 0.0015	0.4007 ± 0.0029	0.5207 ± 0.0023	19.5542 ± 0.0083	0.5880 ± 0.0022
	SE	0.7414 ± 0.0017	0.6801 ± 0.0013	0.5249 ± 0.0015	0.4007 ± 0.0029	0.5207 ± 0.0023	19.5542 ± 0.0083	0.5880 ± 0.0022
LDA with TC $(6,532 \times 100)$	KNN	0.7883 ± 0.0016	0.7512 ± 0.0015	0.5939 ± 0.0014	0.5201 ± 0.0029	0.6440 ± 0.0021	18.7220 ± 0.0465	0.4554 ± 0.0025
	MLP	**0.8037 ± 0.0010**	**0.7657 ± 0.0014**	**0.6181 ± 0.0016**	**0.5544 ± 0.0021**	**0.6663 ± 0.0020**	**18.5933 ± 0.0463**	**0.4341 ± 0.0026**
	KNN as OvR	0.7838 ± 0.0012	0.7479 ± 0.0010	0.5859 ± 0.0008	0.5098 ± 0.0017	0.6389 ± 0.0015	18.7532 ± 0.0544	0.4593 ± 0.0019
	LR as OvR	0.8018 ± 0.0010	0.7644 ± 0.0012	0.6157 ± 0.0014	0.5505 ± 0.0019	0.6624 ± 0.0017	18.6514 ± 0.0467	0.4361 ± 0.0023
	SVM as OvR	0.7773 ± 0.0014	0.7272 ± 0.0013	0.5852 ± 0.0016	0.4949 ± 0.0026	0.5999 ± 0.0021	19.1559 ± 0.0464	0.5087 ± 0.0025
	RF	0.7569 ± 0.0014	0.6945 ± 0.0011	0.5531 ± 0.0013	0.4415 ± 0.0023	0.5462 ± 0.0019	19.4421 ± 0.0404	0.5694 ± 0.0022
	HVE	0.8018 ± 0.0011	0.7633 ± 0.0011	0.6160 ± 0.0011	0.5498 ± 0.0017	0.6607 ± 0.0012	18.6970 ± 0.0587	0.4384 ± 0.0020
	SE	0.7983 ± 0.0012	0.7570 ± 0.0010	0.6096 ± 0.0012	0.5408 ± 0.0017	0.6504 ± 0.0013	18.7473 ± 0.0621	0.4513 ± 0.0017

Table 7.3: ICD9 code group prediction using nursing notes of MIMIC-III (without similarity modeling).

Data model	**Classifier**	**Performance scores**						
		ACC	**AUROC**	**AUPRC**	**MCC**	**F-score**	**CE**	**LRL**
Term weighting $(6,532 \times 14,665)$	KNN	0.7866 ± 0.0012	0.7689 ± 0.0016	0.5920 ± 0.0025	0.5306 ± 0.0032	0.6697 ± 0.0021	18.0463 ± 0.0691	0.4168 ± 0.0027
	MLP	0.7962 ± 0.0011	0.7694 ± 0.0015	0.6009 ± 0.0026	0.5400 ± 0.0029	0.6685 ± 0.0024	18.2134 ± 0.0530	0.4199 ± 0.0026
	KNN as OvR	0.7741 ± 0.0017	0.7662 ± 0.0014	0.5764 ± 0.0027	0.5144 ± 0.0032	0.6639 ± 0.0020	**18.1744 ± 0.0644**	0.4179 ± 0.0023
	LR as OvR	**0.8143 ± 0.0014**	**0.7804 ± 0.0017**	**0.6378 ± 0.0032**	**0.5845 ± 0.0035**	**0.6874 ± 0.0030**	18.2934 ± 0.0389	**0.3985 ± 0.0030**
	SVM as OvR	0.7414 ± 0.0015	0.6801 ± 0.0015	0.5249 ± 0.0026	0.4007 ± 0.0036	0.5207 ± 0.0028	19.5542 ± 0.0368	0.5880 ± 0.0024
	RF	0.7653 ± 0.0011	0.6951 ± 0.0013	0.5517 ± 0.0024	0.4449 ± 0.0031	0.5484 ± 0.0023	19.5449 ± 0.0387	0.5695 ± 0.0022
	HVE	0.8064 ± 0.0014	0.7782 ± 0.0014	0.6369 ± 0.0031	0.5788 ± 0.0032	0.6832 ± 0.0026	18.5193 ± 0.0489	0.4132 ± 0.0023
	SE	0.7971 ± 0.0013	0.7693 ± 0.0018	0.6017 ± 0.0032	0.5412 ± 0.0034	0.6682 ± 0.0029	18.2290 ± 0.0363	0.4207 ± 0.0030
Doc2Vec 500 $(6,532 \times 500)$	KNN	0.7134 ± 0.0013	0.5986 ± 0.0021	0.4719 ± 0.0024	0.3111 ± 0.0040	0.3323 ± 0.0059	19.9011 ± 0.0208	0.7824 ± 0.0048
	MLP	0.7370 ± 0.0011	0.7081 ± 0.0017	0.5217 ± 0.0022	0.4113 ± 0.0029	0.5885 ± 0.0026	18.8870 ± 0.0421	0.5113 ± 0.0028
	KNN as OvR	0.7177 ± 0.0013	0.6091 ± 0.0020	0.4783 ± 0.0020	0.3167 ± 0.0035	0.3627 ± 0.0054	19.8782 ± 0.0171	0.7533 ± 0.0048
	LR as OvR	0.7970 ± 0.0007	0.7586 ± 0.0009	0.5999 ± 0.0020	0.5291 ± 0.0016	0.6659 ± 0.0016	**18.6661 ± 0.0346**	0.4382 ± 0.0017
	SVM as OvR	**0.8068 ± 0.0010**	**0.7678 ± 0.0012**	**0.6206 ± 0.0024**	**0.5527 ± 0.0025**	**0.6774 ± 0.0018**	18.7267 ± 0.0269	**0.4245 ± 0.0021**
	RF	0.7490 ± 0.0014	0.6801 ± 0.0016	0.5351 ± 0.0027	0.4142 ± 0.0037	0.5232 ± 0.0029	19.6314 ± 0.0357	0.5942 ± 0.0027
	HVE	0.8011 ± 0.0006	0.7627 ± 0.0008	0.6083 ± 0.0024	0.5387 ± 0.0013	0.6701 ± 0.0011	18.6705 ± 0.0216	0.4318 ± 0.0014
	SE	0.8054 ± 0.0009	0.7659 ± 0.0010	0.6179 ± 0.0028	0.5489 ± 0.0022	0.6740 ± 0.0018	18.7635 ± 0.0400	0.4279 ± 0.0018
Doc2Vec 1,000 $(6,532 \times 1,000)$	KNN	0.7141 ± 0.0016	0.6058 ± 0.0026	0.4754 ± 0.0028	0.3192 ± 0.0045	0.3520 ± 0.0069	19.8945 ± 0.0179	0.7643 ± 0.0058
	MLP	0.7442 ± 0.0011	0.7159 ± 0.0017	0.5312 ± 0.0024	0.4270 ± 0.0030	0.5995 ± 0.0027	18.8172 ± 0.0321	0.4992 ± 0.0028
	KNN as OvR	0.7162 ± 0.0018	0.6112 ± 0.0034	0.4781 ± 0.0037	0.3219 ± 0.0058	0.3671 ± 0.0091	19.8661 ± 0.0200	0.7493 ± 0.0076
	LR as OvR	0.7749 ± 0.0005	0.7425 ± 0.0007	0.5698 ± 0.0018	0.4864 ± 0.0017	0.6418 ± 0.0015	18.7278 ± 0.0397	0.4592 ± 0.0010
	SVM as OvR	**0.8071 ± 0.0009**	**0.7684 ± 0.0012**	**0.6194 ± 0.0027**	**0.5528 ± 0.0026**	**0.6768 ± 0.0022**	18.6731 ± 0.0429	**0.4239 ± 0.0020**
	RF	0.7455 ± 0.0014	0.6760 ± 0.0014	0.5313 ± 0.0023	0.4077 ± 0.0032	0.5138 ± 0.0025	19.6283 ± 0.0375	0.6034 ± 0.0025
	HVE	0.7915 ± 0.0009	0.7559 ± 0.0014	0.5943 ± 0.0037	0.5200 ± 0.0035	0.6588 ± 0.0029	**18.6419 ± 0.0225**	0.4410 ± 0.0022
	SE	0.8061 ± 0.0011	0.7674 ± 0.0013	0.6179 ± 0.0035	0.5508 ± 0.0032	0.6750 ± 0.0025	18.6649 ± 0.0241	0.4256 ± 0.0022
HDP with BoW $(6,532 \times 150)$	KNN	0.7778 ± 0.0011	0.7505 ± 0.0014	0.5792 ± 0.0024	0.5033 ± 0.0027	0.6407 ± 0.0019	18.5832 ± 0.0558	0.4502 ± 0.0024
	MLP	**0.7946 ± 0.0013**	**0.7574 ± 0.0016**	**0.6026 ± 0.0031**	**0.5336 ± 0.0036**	**0.6518 ± 0.0028**	18.6202 ± 0.0417	**0.4467 ± 0.0028**
	KNN as OvR	0.7733 ± 0.0013	0.7476 ± 0.0017	0.5726 ± 0.0030	0.4949 ± 0.0037	0.6367 ± 0.0026	**18.5783 ± 0.0456**	0.4536 ± 0.0027
	LR as OvR	0.7878 ± 0.0016	0.7453 ± 0.0020	0.5932 ± 0.0030	0.5183 ± 0.0042	0.6307 ± 0.0033	18.7679 ± 0.0444	0.4723 ± 0.0033
	SVM as OvR	0.7623 ± 0.0014	0.6926 ± 0.0017	0.5510 ± 0.0029	0.4450 ± 0.0038	0.5411 ± 0.0032	19.5415 ± 0.0398	0.5776 ± 0.0029
	RF	0.7619 ± 0.0015	0.6982 ± 0.0017	0.5535 ± 0.0029	0.4468 ± 0.0039	0.5563 ± 0.0030	19.5531 ± 0.0314	0.5606 ± 0.0030
	HVE	0.7886 ± 0.0011	0.7438 ± 0.0016	0.5941 ± 0.0027	0.5183 ± 0.0029	0.6286 ± 0.0024	18.8647 ± 0.0482	0.4759 ± 0.0031
	SE	0.7886 ± 0.0006	0.7431 ± 0.0011	0.5935 ± 0.0023	0.5172 ± 0.0017	0.6288 ± 0.0018	18.8853 ± 0.0417	0.4766 ± 0.0022
HDP with term weighting $(6,532 \times 150)$	KNN	0.7108 ± 0.0010	0.6718 ± 0.0018	0.4885 ± 0.0025	0.3476 ± 0.0030	0.5262 ± 0.0026	**19.3230 ± 0.0378**	**0.5728 ± 0.0027**
	MLP	0.7413 ± 0.0014	0.6783 ± 0.0016	0.5253 ± 0.0029	0.4009 ± 0.0037	0.5167 ± 0.0033	19.5623 ± 0.0396	0.5934 ± 0.0046
	KNN as OvR	0.7067 ± 0.0012	0.6685 ± 0.0020	0.4837 ± 0.0028	0.3393 ± 0.0036	0.5221 ± 0.0029	19.3410 ± 0.0392	0.5767 ± 0.0030
	LR as OvR	0.7455 ± 0.0016	0.6779 ± 0.0016	0.5301 ± 0.0030	0.4072 ± 0.0041	0.5161 ± 0.0030	19.5868 ± 0.0369	0.5984 ± 0.0026
	SVM as OvR	0.7414 ± 0.0015	0.6801 ± 0.0015	0.5249 ± 0.0026	0.4007 ± 0.0036	0.5207 ± 0.0028	19.5542 ± 0.0368	0.5880 ± 0.0024
	RF	**0.7559 ± 0.0012**	**0.6862 ± 0.0018**	**0.5386 ± 0.0030**	**0.4259 ± 0.0039**	**0.5313 ± 0.0033**	19.4848 ± 0.0370	0.5854 ± 0.0030
	HVE	0.7444 ± 0.0023	0.6789 ± 0.0012	0.5286 ± 0.0038	0.4058 ± 0.0049	0.5179 ± 0.0023	19.5742 ± 0.0588	0.5948 ± 0.0031
	SE	0.7413 ± 0.0016	0.6800 ± 0.0010	0.5248 ± 0.0025	0.4007 ± 0.0031	0.5206 ± 0.0024	19.5566 ± 0.0507	0.5882 ± 0.0015
LDA with TC $(6,532 \times 100)$	KNN	0.7872 ± 0.0011	0.7517 ± 0.0012	0.5937 ± 0.0023	0.5197 ± 0.0027	0.6449 ± 0.0024	18.7065 ± 0.0454	0.4539 ± 0.0020
	MLP	**0.8039 ± 0.0011**	**0.7669 ± 0.0014**	**0.6182 ± 0.0025**	**0.5547 ± 0.0028**	**0.6681 ± 0.0023**	**18.5665 ± 0.0489**	**0.4311 ± 0.0025**
	KNN as OvR	0.7824 ± 0.0008	0.7482 ± 0.0013	0.5851 ± 0.0022	0.5087 ± 0.0026	0.6392 ± 0.0021	18.7217 ± 0.0364	0.4581 ± 0.0021
	LR as OvR	0.8018 ± 0.0013	0.7639 ± 0.0014	0.6152 ± 0.0027	0.5497 ± 0.0033	0.6626 ± 0.0025	18.6916 ± 0.0466	0.4367 ± 0.0024
	SVM as OvR	0.7778 ± 0.0016	0.7297 ± 0.0015	0.5858 ± 0.0028	0.4961 ± 0.0036	0.6050 ± 0.0027	19.1415 ± 0.0275	0.5024 ± 0.0025
	RF	0.7587 ± 0.0015	0.6962 ± 0.0013	0.5527 ± 0.0027	0.4424 ± 0.0032	0.5487 ± 0.0024	19.4452 ± 0.0393	0.5655 ± 0.0022
	HVE	0.8009 ± 0.0009	0.7613 ± 0.0009	0.6141 ± 0.0022	0.5469 ± 0.0020	0.6584 ± 0.0018	18.7753 ± 0.0523	0.4423 ± 0.0019
	SE	0.7975 ± 0.0011	0.7566 ± 0.0013	0.6078 ± 0.0027	0.5388 ± 0.0023	0.6509 ± 0.0025	18.7774 ± 0.0599	0.4510 ± 0.0029

7.3 FarSight: Long-Term Aggregation by Future Lookup

Our next work is an attempt to model the rich patient information embedded in multiple clinical notes of a patient for disease prediction, using vector space (Doc2Vec) and topic modeling (Nonnegative Matrix Factorization (NMF)), for deriving optimal patient-specific data representations. *FarSight*, a variant of patient record aggregation mechanism intended to detect the onset of the disease with the earliest recorded symptoms, infections, and disorders, along with a deep learning prediction model, forms the core of this work. The proposed *FarSight*-aggregated unstructured modeling was evaluated against naïve note aggregation strategy and structured EHR based state-of-the-art model using standard evaluation metrics. The overall workflow of the proposed disease group prediction system using the *FarSight* approach is as illustrated in Figure 7.5.

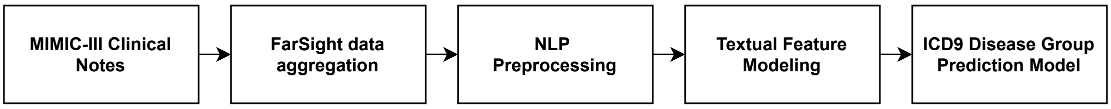


Figure 7.5: *FarSight* Model for ICD9 Disease Group Prediction

The patient cohort considered for this work was the same one used in the previous work, extracted from the MIMIC-III dataset, as explained in Section 7.2. First admissions of patients aged 15 and above were considered for the study, and finally erroneous and duplicated records were filtered out. The resultant patient cohort obtained after handling the erroneous entries contained nursing notes corresponding to 6,532 patients (140,792 clinical notes)—the data in these nursing notes were aggregated using the *FarSight* approach to detect the onset of the disease with the earliest recorded symptoms.

7.3.1 Modeling Multiple Clinical Records

As the need for critical care facilities like ICUs grows, the limited availability of resources including specialized monitoring equipment and trained clinical staff, is often a major bottleneck. In addition, a lack of precise knowledge concerning the etiology of ICU complications can lead to delayed and imprecise recognition of patients at high-risk, thus hindering preemptive treatment options. As a result, the requisite care is often delivered only after the development of a particular complication. Therefore, detection of disease onset when the earliest recorded infections or symptoms are observed is of utmost importance, as it can significantly reduce mortality and morbidity rates. Towards this objective, we present *FarSight*, a long-term aggregation mechanism, which facilitates the aggregation of the patient data using a future lookup on all the later detected symptoms and diseases[1].

Let $\mathcal{P}$ be the set of all patients, indexed by p. For each patient, we have a sequence of clinical nursing notes, $\Phi^{(p)} = \{(\eta_n^{(p)}, \mathcal{I}_n^{(p)})\}_{n=1}^{N(p)}$, with each nursing note $\eta_n^{(p)}$ and the corresponding ICD9 diagnostic code (group) $\mathcal{I}_n^{(p)}$ indexed by n, and with $N(p)$ number of notes (of total N nursing notes) for a patient p. Furthermore, the nursing notes of a patient are ordered from oldest to the most recent. Now, the aggregation of the ICD9 code groups across the nursing notes of a patient is performed using *FarSight*, through a future lookup of the diseases in the long run (dependent on the number of nursing notes recorded for that specific patient (p), concerning several episodes during single hospital admission), resulting in $\Phi^{(p)} = \{(\eta_n^{(p)}, \mathcal{I}^{(p)})\}_{n=1}^{N(p)}$, where $\mathcal{I}^{(p)} = \{\mathcal{I}_n^{(p)}\}_{n=1}^{N(p)}$. Note that, while the aggregation of diagnostic code groups seems to be incremental in nature, the objective is to predict the diseases and complications that are most likely to be observed in the subsequent episodes of a patient's current hospital admission—*FarSight* facilitates such prediction through aggregation of diagnostic code groups across all the episodes recorded for a patient, thus performing long-term aggregation through future lookup. Ultimately, our goal is to learn a generalizable function ($\mathcal{G}$) that estimates the probability of classifying a given clinical nursing note $\eta_n^{(p)}$ into a set of ICD9 diagnostic codes:

$$\mathcal{G}(\Phi^{(p)}) \approx Pr(\mathcal{I}^{(p)} \mid \eta_n^{(p)}) \tag{7.5}$$

It is to be noted that the proposed *FarSight* mechanism facilitates multi-label classification by aggregating the diagnostic code groups across a patient's multiple medical records, rather than aggregating the raw medical text in the nursing notes. Such an aggregation facilitates risk assessment at the initial stages of the disease, with the earliest detected infections and symptoms.

Consider the nursing notes ($\{\eta_n^{(p)}\}_{n=1}^{N(p)}$) of a patient ($p$) ordered chronologically; assuming $N(p)$ to be three, we have three medical records of the patient p corresponding to three (distinct) diagnostic code groups ($\{\mathcal{I}_n^{(p)}\}_{n=1}^{N(p)=3}$). By employing the *FarSight* aggregation mechanism, we map each of the three medical records to all the ICD9 code groups observed in the patient p's nursing notes,

[1] In this context, 'later detected diseases' at time T are the diseases recorded in the medical records after time T.

i.e., $\{\mathcal{I}_n^{(p)}\}_{n=1}^{N(p)=3}$. Simply put, each $\eta_n^{(p)}$ corresponds to $\{\mathcal{I}_1^{(p)}, \mathcal{I}_2^{(p)}, \ldots, \mathcal{I}_{N(p)}^{(p)}\}$. It is important to stress that, *FarSight* aggregation is only effective when the disease symptoms are progressive and related (e.g., *sore throat* $\longrightarrow$ *cold* $\longrightarrow$ *fever* vs. *sore throat* $\longrightarrow$ *leukaemia*). Since this study specifically considers the first ICU admission of a MIMIC-III subject, *FarSight* can be employed to stratify risk using the earliest detected symptoms. However, through naïve aggregation of nursing notes using patient identification numbers, we have $\eta_1^{(p)} \oplus \eta_2^{(p)} \oplus \ldots \oplus \eta_{N(p)}^{(p)}$ mapping to $\{\mathcal{I}_1^{(p)}, \mathcal{I}_2^{(p)}, \ldots, \mathcal{I}_{N(p)}^{(p)}\}$ ($\oplus$ denotes concatenation). Thus, *FarSight*-aggregated data can help train the underlying classifier in identifying all the possible diagnostic groups, by capturing the episode-specific characteristics, i.e., training at the clinical note level. In contrast, patient-based aggregated data aids in training the predictor at the patient level. Furthermore, the diagnostic code groups of $\eta_i^{(p)}$ $(1 < i < N(p))$ predicted using a model trained on *FarSight*-aggregated data $(\{(\eta_n^{(p)}, \mathcal{I}_n^{(p)})\}_{n=1}^{N(p)=3})$ would include (with high probability) $\mathcal{I}_i^{(p)}$, owing to the training at nursing note granularity. However, employing a classifier trained on naïvely aggregated data to predict the diagnostic code groups of $\eta_i^{(p)}$ $(1 < i < N(p))$ might not include $\mathcal{I}_i^{(p)}$, as episode-specific characteristics are lost.

7.3.2 Preprocessing

Despite the inherent content-rich nature of the patient-specific information available in the clinical nursing notes, they are raw, sparse, informally-written, complexly structured, and voluminous. Thus, any transformation of raw medical text into a canonical form extends the learnability and generalizability of the underlying deep neural architectures. Such normalization not only allows for the separation of concerns but also helps maintain consistency. To achieve this, we subject all the notes to NLP processing, which included tokenization, stopword removal, and stemming/lemmatization. First, we removed multiple spaces and special characters. Next, we experimented with multiple tokenizers including MedPost[2], Penn bio tagger[3], NLTK, Stanford log-linear part-of-speech tagger[4], and GENIA tagger[5], to segment the medical text in the nursing notes into several primary building blocks (tokens). MedPost tokenizer splits the input text at hyphens, slashes, internal periods, and punctuation within numbers (e.g., *IL-20 i.e. 1,000 U/ml* is split as *IL␣-␣20␣i␣.␣e␣.␣1␣,␣000␣U␣/␣ml*), while Penn bio tagger splits the words

[2] ftp://ftp.ncbi.nlm.nih.gov/pub/lsmith/MedPost/medpost.tar.gz
[3] https://www.seas.upenn.edu/~strctlrn/BioTagger/BioTagger
[4] https://nlp.stanford.edu/software/tagger.shtml
[5] http://www.nactem.ac.uk/tsujii/GENIA/tagger/

at slashes (e.g., *0.05 U/ml* is split as *0.05␣U␣/␣ml*), and hence are not employed in this study.

We observed that the NLTK tokenizer was similar (in the splitting scheme) to Stanford log-linear part-of-speech and GENIA taggers, with respect to DNA sequences (e.g., *CCAAAGCGTAAAGG*), words with numbers and letters (e.g., *15th*), and hyphenated compound words (e.g., *x-ray*). Thus, we employed the NLTK tokenizer to facilitate the tokenization of nursing text. Utilizing the NLTK English stopword corpus, we removed stopwords from the generated tokens. Furthermore, punctuation marks (except hyphens and slashes) were also removed. References to images (e.g., *MRI_Scan.jpeg*) were removed, and character case folding was performed. Note that, word-length based token removal was not performed to eliminate the loss of important medical information (e.g., *CT*, *DEXA*, *MRI*, and *PET*). Before any further processing, medical concept normalization through disambiguation of abbreviations (into their respective long forms) was facilitated using CARD, an open-source framework for clinical abbreviation recognition and disambiguation (Wu *et al.*, 2016). Lastly, suffix stripping was performed through stemming, followed by lemmatization for the conversion of the stripped tokens into their respective base forms. Additionally, we eliminated the tokens appearing in less than ten nursing notes (e.g., *spot*, *cope*, and *inch*) in order to lower the computational complexity of training (the total number of tokens pre- and post-elimination were $188,742$ and $32,687$ respectively) and mitigate problems arising due to overfitting.

7.3.3 Topic Modeling of Clinical Notes

Topic modeling aims at finding a set of topics (collection of terms) from a collection of documents that best represents the underlying corpus. Latent semantic analysis and other traditional methods of information retrieval compute the Singular Value Decomposition (SVD) of the BoW or TW matrix to generate a lower approximation of the matrix—such methods often deal with matrix computations of high complexity. NMF is a popular multivariate analysis approach that aims at factoring a data matrix ($M \in \mathbb{R}^{|\mathbb{V}| \times N}$) by minimizing the reconstruction error, with nonnegativity constraints. This can be viewed as learning an unnormalized probability distribution over the topics (Stevens *et al.*, 2012). Formally, NMF seeks a factorization model for a given data matrix M and a target rank $\mathcal{T}$ (number of topics) to explain the data matrix (M), where $W \geq 0$, $H \geq 0$, and $\mathcal{T} \leq \min\{|\mathbb{V}|, N\}$ (as shown in Eq. (7.6)). The unnormalized probabilities are learned by randomly initializing each set of probabilities and then updating them according to a set of iterative rules defined in Eq. (7.7).

$$M \approx WH^{\intercal},\ W \in \mathbb{R}^{|\mathbb{V}| \times \mathcal{T}},\ H \in \mathbb{R}^{N \times \mathcal{T}} \tag{7.6}$$

$$H \longleftarrow H \cdot \frac{W^{\intercal}M}{W^{\intercal}WH^{\intercal}} \qquad W \longleftarrow W \cdot \frac{MH}{WH^{\intercal}H} \tag{7.7}$$

At first glance, NMF is an alternative factoring model similar to SVD that considers different constraints (orthogonality) on the latent factors. However, the effectiveness of NMF when modeling real-life nonnegative data (e.g., text, images, and audio spectra) has sparked widespread interest in the fields of signal processing and data analytics (Fu *et al.*, 2018). Representing real-life data into nonnegative matrices and factoring them into latent factors yields intriguing results, and thus, NMF is popularly recognized as a workhorse in data analytics.

As is the case with other clustering approaches, determining the optimal number of NMF clusters is a challenging problem. Moreover, learning topics from a multinomial distribution of words from sparse and noisy textual data can often be hard to interpret. Semantic Coherence (SC) is a way of evaluating models with a higher guarantee of human interpretability. In our work, we adopt NMF with SC, as it accounts for the semantic similarity between high scoring clinical terms. We employ the C_v variant of the coherence measurement with the Normalized Pointwise Mutual Information Score (NPMI) as the confirmation measure, owing to its more significant correlation with the available human-judged data (Röder *et al.*, 2015).

Let $\mathcal{T}_i = \{t_1, t_2, \ldots, t_n\}$ be a topic generated from a topic model, represented by its top$-n$ most probable terms (t_ks). A topic depicts greater coherence when the average pairwise similarity among the terms of that topic is high. Given a predefined similarity score $(\mathrm{Sim}(t_k, t_l))$[6], we compute the SC score using:

$$\mathrm{SC}(\mathrm{Sim}, \mathcal{T}_i) = \frac{\sum_{\substack{1 \le k \le n-1 \\ k+1 \le l \le n}} \mathrm{Sim}(t_k, t_l)}{\binom{n}{2}} \tag{7.8}$$

where $t_k, t_l \in \mathcal{T}_i$. The NPMI similarity score is used in finding collocations and associations between the words and is computed as per (7.9) and (7.10). To obtain the final conformation score, we average the individual confirmation scores

[6] In our work, we use NPMI as the similarity measure.

obtained for all the topics ($\mathcal{T}_i$s).

$$\mathrm{NPMI}(t_k, t_l) = \frac{\mathrm{PMI}(t_k, t_l)}{-\log_2(\Pr(t_k, t_l))} \tag{7.9}$$

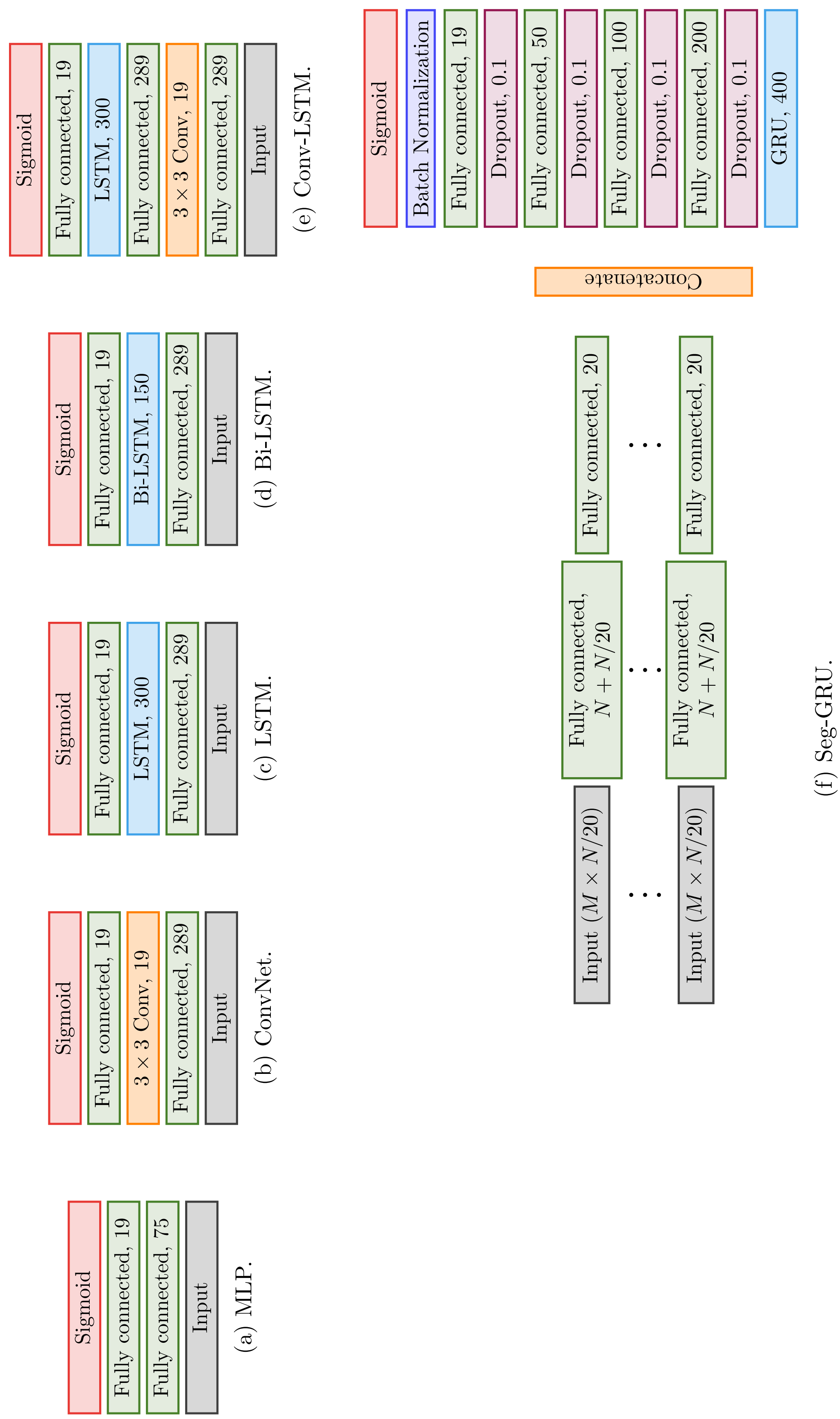
$$\mathrm{PMI}(t_k, t_l) = \log_2\left(\frac{\Pr(t_k, t_l)}{\Pr(t_k)\Pr(t_l)}\right) \tag{7.10}$$

The optimal number of topics in NMF was determined to be 100, by comparing the coherence scores of several NMF models obtained by heuristically varying the number of topics from 2 to 500. In this study, we built the NMF matrices on both BoW and TW matrices, to enable an exhaustive comparison. Moreover, we model the BoW and TW matrices using NMF without coherence scoring (set to 150 topics, which was determined empirically using grid-search). The implementations available in the Python Gensim package were employed to implement the NMF models.

7.3.4 ICD9 Code Group Prediction

In our work, we focus on ICD9 code group prediction as a multi-label classification problem, where each nursing note of every patient is mapped to multiple diagnostic code groups. The ICD9 codes of a given admission from MIMIC-III are mapped into 19 distinct diagnostic groups, similar to the previous study as explained in Section 7.2.4. Here also our dataset does not contain any records in the ICD9 code range of $760 - 779$ and all E-codes and V-codes are classified into the same code group to lower the computational complexity of training.

The prediction model is a Convolutional Long Short-Term Memory (Conv-LSTM) based architecture, where the convolution layer effectively extracts the high-level features from a given precomputed embedding of a clinical nursing note. However, to capture the long-term dependencies in the nursing notes, we need a substantial number of convolutional layers—such dependencies are easily captured and retained by an LSTM network. Thus, a hybrid Conv-LSTM architecture is designed in which both captures the high-level features and retains the long-term dependencies over time. We used a hybrid Conv-LSTM network with one fully connected layer of 289 ReLU processing units, one convolution or ConvNet layer with 3×3 convolution window and a feature map size of 19, followed by another fully connected layer of 289 ReLU processing units, and an LSTM layer with 300 hidden nodes. ICD9 code group prediction is facilitated by a sigmoid activation of the final LSTM output (see Fig. 7.6e).



(a) MLP.

(b) ConvNet.

(c) LSTM.

(d) Bi-LSTM.

(e) Conv-LSTM.

(f) Seg-GRU.

Figure 7.6: Schematic overview of the deep neural architectures employed in this study.

For comparison, we implemented several other models. A Multi-Layer Perceptron (MLP) (explained in Section 7.2.5) is a fully connected feed-forward artificial neural network with multiple layers of processing elements (neurons) interacting through weighted connections. Typically, MLP consists of an input layer, one or more hidden layers, and one classification layer at the top to solve the prediction task. The input to the first layer is comprised of a $d-$dimensional embedding (topics) of $\eta_n^{(p)}$, and the output of each layer serves as the input to the subsequent layer. In our study, we use an MLP network with one hidden layer of 75 ReLU processing units and a prediction layer of 19 sigmoid processing units (see Fig. 7.6a).

ConvNets are a regularized variation of the deep MLP architecture which are aimed at minimal processing. They utilize layers with convolving filters, which are applied to local features. Due to their transition invariance characteristics and shared-weights architecture, ConvNets are space invariant. They are shown to be effective in a variety of NLP tasks, including semantic parsing, sentence modeling, and search query retrieval (Kim, 2014). Consider that a clinical nursing note $\eta_n^{(p)}$ is modeled to produce an $n-$dimensional embedding. A convolution operation involving a filter is applied to a window of say h terms to produce a new feature. To produce a feature map, we now apply this filter to every possible window of terms in the embedding. Here, we extract one feature from one filter, and this process can be extended to obtain multiple features from multiple filters (of varying sizes). The features from the penultimate layer are passed to a fully connected layer using a nonlinear activation function. We employed one fully connected layer of 289 ReLU processing units, and one ConvNet layer with 3×3 convolution window and a feature map size of 19. Finally, the code group prediction is facilitated by a fully connected layer of 19 sigmoid processing units (see Fig. 7.6b).

LSTM is a special type of Recurrent Neural Network (RNN) that effectively overcomes the gradient vanishing problem and captures long-term dependencies, which is crucial to predict the code groups using nursing notes accurately. To determine the extent to which LSTM memory units must memorize the current state and retain the previous state, LSTMs employ an adaptive gating mechanism. More specifically, an LSTM memory unit is composed of four gates: the input gate, the forget gate, the output gate, and the candidate value for the cell state. In nursing notes, the semantic meaning of a term is often influenced by the terms before and after it. Thus, the predictability of diagnostic codes can be enhanced by accessing both past and future input features for a given time—a Bi-LSTM network can facilitate such functionality. In doing so, we can effectively utilize future features (via backward states) and past features (via forward states) for a particular time frame. We set the dimensions of the embedding and LSTM hidden state to 289 (17 time steps with 17 features each) and 300 respectively. A sigmoid activation of the final LSTM output facilitates the multi-label classification (see Fig. 7.6c). The Bi-LSTM architecture is similar to that of LSTM, except that Bi-LSTM employs two LSTM layers of 150 hidden states each (see Fig. 7.6d).

Gated Recurrent Units (GRUs) are a gating mechanism in RNNs, similar to LSTM networks (with output gate) but with fewer parameters, as it lacks an output gate. Each recurrent unit in a GRU network adaptively captures dependencies of different time scales. A GRU memory unit is composed of two gates: the reset gate and the update gate. The GRU computes a candidate hidden state and then smoothly extrapolates it (gated by the update gate). We employ a segment-level GRU, where the input embedding of a nursing note (of size N) is split column-wise into 20 segments, followed by a fully connected layer of $N + N/20$ ReLU processing units, and another fully connected layer of 20 ReLU processing units. The outputs from various segments are then concatenated channel-wise and are flattened. Regularization prevents co-adaptation of the hidden units and is hence necessary. The flattened output is passed through a series of 10% dropouts and fully connected ReLU processing units for regularization. Finally, the obtained output is subject to batch normalization to stabilize the network and reduce the covariance shift. A sigmoid activation of the normalized output facilitates the prediction (see Fig. 7.6f). Next, all the models discussed are evaluated as per a set of standard metrics on the standard dataset, for qualitative assessment of their relative merits and drawbacks, and suitability for use as CDSSs in practice.

7.3.5 Experimental Results and Discussion

To validate our approach, we performed exhaustive benchmarking experiments on the clinical nursing notes obtained from the MIMIC-III database as per the defined cohort. All experiments were performed on an Ubuntu based High-end Server with a 56-core Intel Xeon processor, 128GB RAM, two Nvidia Tesla M40 GPUs (24GB each) and 3TB hard drive. A significant challenge was the multi-label prediction, where a set of probable ICD9 code groups were to be predicted for a given clinical nursing note. To assess the predictability of the proposed approaches, we employ a pair-wise comparison of the actual and the predicted diagnostic code group sets, performed via 5-fold cross-validation (the means and the standard errors of the mean are presented). Furthermore, to accurately assess the performance of

Table 7.4: Code group prediction using the data in the nursing notes aggregated using *FarSight*.

Data model	**Classifier**	**Performance scores**				
		ACC	**MCC**	**F-score**	**AUPRC**	**AUROC**
Doc2Vec $(140,792 \times 500)$	MLP	0.7873 ± 0.0006	0.5587 ± 0.0011	0.7103 ± 0.0027	0.6577 ± 0.0012	0.7795 ± 0.0014
	ConvNet	0.8053 ± 0.0005	0.5938 ± 0.0011	0.7332 ± 0.0006	0.6810 ± 0.0011	$\mathbf{0.7967 \pm 0.0004}$
	LSTM	0.7986 ± 0.0016	0.5804 ± 0.0025	0.7250 ± 0.0026	0.6705 ± 0.0027	0.7885 ± 0.0016
	Bi-LSTM	0.8018 ± 0.0008	0.5861 ± 0.0018	0.7265 ± 0.0043	0.6758 ± 0.0023	0.7906 ± 0.0026
	Conv-LSTM	$\mathbf{0.8069 \pm 0.0023}$	$\mathbf{0.5961 \pm 0.0040}$	$\mathbf{0.7338 \pm 0.0022}$	$\mathbf{0.6824 \pm 0.0039}$	0.7962 ± 0.0019
	Seg-GRU	0.7779 ± 0.0020	0.5332 ± 0.0051	0.6794 ± 0.0054	0.6505 ± 0.0030	0.7605 ± 0.0035
NMF−BoW $(140,792 \times 150)$	MLP	0.7829 ± 0.0006	0.5498 ± 0.0009	0.7029 ± 0.0016	0.6530 ± 0.0017	0.7744 ± 0.0007
	ConvNet	0.7965 ± 0.0007	0.5750 ± 0.0013	0.7187 ± 0.0041	0.6688 ± 0.0018	0.7860 ± 0.0020
	LSTM	0.7921 ± 0.0005	0.5652 ± 0.0016	0.7093 ± 0.0030	0.6638 ± 0.0018	0.7794 ± 0.0017
	Bi-LSTM	0.7894 ± 0.0007	0.5596 ± 0.0015	0.7042 ± 0.0020	0.6619 ± 0.0017	0.7758 ± 0.0012
	Conv-LSTM	$\mathbf{0.8048 \pm 0.0021}$	$\mathbf{0.5897 \pm 0.0042}$	$\mathbf{0.7240 \pm 0.0034}$	$\mathbf{0.6806 \pm 0.0031}$	$\mathbf{0.7911 \pm 0.0024}$
	Seg-GRU	0.7945 ± 0.0063	0.5666 ± 0.0137	0.7039 ± 0.0114	0.6698 ± 0.0057	0.7772 ± 0.0080
NMF−TW $(140,792 \times 150)$	MLP	0.7953 ± 0.0004	0.5740 ± 0.0010	0.7167 ± 0.0010	0.6696 ± 0.0015	0.7850 ± 0.0005
	ConvNet	0.8174 ± 0.0006	0.6181 ± 0.0008	0.7489 ± 0.0016	0.6948 ± 0.0014	0.8091 ± 0.0014
	LSTM	0.8129 ± 0.0015	0.6062 ± 0.0028	0.7347 ± 0.0020	0.6908 ± 0.0024	0.7992 ± 0.0012
	Bi-LSTM	0.8076 ± 0.0016	0.5952 ± 0.0034	0.7280 ± 0.0037	0.6839 ± 0.0024	0.7936 ± 0.0024
	Conv-LSTM	$\mathbf{0.8282 \pm 0.0023}$	$\mathbf{0.6368 \pm 0.0042}$	$\mathbf{0.7562 \pm 0.0021}$	$\mathbf{0.7089 \pm 0.0046}$	$\mathbf{0.8157 \pm 0.0019}$
	Seg-GRU	0.8249 ± 0.0021	0.6273 ± 0.0050	0.7434 ± 0.0057	0.7089 ± 0.0019	0.8073 ± 0.0040
NMF−BoW with SC $(140,792 \times 100)$	MLP	0.7820 ± 0.0004	0.5476 ± 0.0007	0.7011 ± 0.0008	0.6517 ± 0.0013	0.7735 ± 0.0006
	ConvNet	0.7956 ± 0.0002	0.5731 ± 0.0007	0.7174 ± 0.0021	0.6672 ± 0.0017	0.7852 ± 0.0011
	LSTM	0.7905 ± 0.0004	0.5619 ± 0.0003	0.7066 ± 0.0035	0.6623 ± 0.0020	0.7777 ± 0.0019
	Bi-LSTM	0.7889 ± 0.0009	0.5598 ± 0.0005	0.7076 ± 0.0040	0.6598 ± 0.0024	0.7774 ± 0.0023
	Conv-LSTM	$\mathbf{0.8003 \pm 0.0015}$	$\mathbf{0.5817 \pm 0.0033}$	$\mathbf{0.7218 \pm 0.0038}$	$\mathbf{0.6735 \pm 0.0024}$	$\mathbf{0.7885 \pm 0.0027}$
	Seg-GRU	0.7918 ± 0.0041	0.5622 ± 0.0087	0.7034 ± 0.0112	0.6659 ± 0.0022	0.7763 ± 0.0070
NMF−TW with SC $(140,792 \times 100)$	MLP	0.7961 ± 0.0003	0.5753 ± 0.0009	0.7175 ± 0.0010	0.6703 ± 0.0016	0.7856 ± 0.0006
	ConvNet	0.8192 ± 0.0006	0.6199 ± 0.0025	0.7466 ± 0.0043	0.6983 ± 0.0013	0.8077 ± 0.0023
	LSTM	0.8142 ± 0.0014	0.6087 ± 0.0034	0.7367 ± 0.0050	0.6918 ± 0.0016	0.8003 ± 0.0030
	Bi-LSTM	0.8096 ± 0.0006	0.5998 ± 0.0012	0.7318 ± 0.0030	0.6860 ± 0.0011	0.7961 ± 0.0019
	Conv-LSTM	$\mathbf{0.8343 \pm 0.0031}\bullet$	$\mathbf{0.6459 \pm 0.0073} \bullet \circ$	$\mathbf{0.7602 \pm 0.0068} \bullet \circ$	$\mathbf{0.7170 \pm 0.0045} \bullet \circ$	$\mathbf{0.8192 \pm 0.0046} \bullet \circ$
	Seg-GRU	0.8285 ± 0.0028	0.6350 ± 0.0064	0.7502 ± 0.0060	0.7131 ± 0.0034	0.8120 ± 0.0048

the proposed methods, we employed five standard evaluation metrics including Accuracy (ACC), MCC score, F-score, Area Under the Precision-Recall Curve (AUPRC), and Area Under the ROC Curve (AUROC). In this study, a pairwise comparison of the predicted and actual diagnostic code groups is presented.

For the clinical task of multi-label ICD-9 code group prediction, we compared the Conv-LSTM model's performance against five other deep neural architectures: MLP, ConvNet, LSTM, Bi-LSTM, and Seg-GRU (depicted in Fig. 7.6). We used the implementations available in the Python Keras package with the Tensorflow backend. Grid-search was used to determine the optimal values of the hyperparameters employed in the underlying deep neural models. The deep neural models were trained to minimize a cross-entropy loss (mean squared error prediction loss) function using Adam optimizer.

The results of our experiments and the related studies are tabulated in Ta-

Table 7.5: Code group prediction (with deep learners) using nursing notes aggregated naïvely by patient identification numbers.

Data model	Classifier	Performance scores				
		ACC	MCC	F-score	AUPRC	AUROC
Doc2Vec $(6,532 \times 500)$	MLP	0.7898 ± 0.0031	0.5195 ± 0.0088	0.6542 ± 0.0069	0.5914 ± 0.0089	0.7556 ± 0.0030
	ConvNet	0.7729 ± 0.0028	0.4841 ± 0.0049	0.6341 ± 0.0056	0.5679 ± 0.0056	0.7399 ± 0.0036
	LSTM	**0.8018 ± 0.0030**	**0.5427 ± 0.0062**	**0.6731 ± 0.0110**	**0.6098 ± 0.0054**	**0.7634 ± 0.0067**
	Bi-LSTM	0.7964 ± 0.0033	0.5308 ± 0.0083	0.6673 ± 0.0081	0.6003 ± 0.0094	0.7594 ± 0.0055
	Conv-LSTM	0.7989 ± 0.0027	0.5321 ± 0.0044	0.6604 ± 0.0035	0.6050 ± 0.0039	0.7554 ± 0.0030
	Seg-GRU	0.7673 ± 0.0046	0.4558 ± 0.0121	0.5991 ± 0.0109	0.5533 ± 0.0092	0.7179 ± 0.0064
NMF−BoW $(6,532 \times 150)$	MLP	0.7810 ± 0.0026	0.4995 ± 0.0066	0.6179 ± 0.0070	0.5838 ± 0.0067	0.7354 ± 0.0030
	ConvNet	**0.7972 ± 0.0029**	**0.5392 ± 0.0085**	**0.6555 ± 0.0083**	**0.6068 ± 0.0072**	**0.7596 ± 0.0053**
	LSTM	0.7801 ± 0.0042	0.4960 ± 0.0074	0.6225 ± 0.0066	0.5776 ± 0.0085	0.7366 ± 0.0041
	Bi-LSTM	0.7776 ± 0.0042	0.4904 ± 0.0094	0.6189 ± 0.0089	0.5735 ± 0.0070	0.7333 ± 0.0061
	Conv-LSTM	0.7870 ± 0.0033	0.5141 ± 0.0075	0.6363 ± 0.0068	0.5920 ± 0.0083	0.7449 ± 0.0038
	Seg-GRU	0.7893 ± 0.0074	0.5218 ± 0.0117	0.6436 ± 0.0026	0.5973 ± 0.0108	0.7495 ± 0.0025
NMF−TW $(6,532 \times 150)$	MLP	0.7878 ± 0.0042	0.5153 ± 0.0108	0.6321 ± 0.0103	0.5939 ± 0.0093	0.7445 ± 0.0058
	ConvNet	**0.8065 ± 0.0033**	**0.5616 ± 0.0083**	**0.6739 ± 0.0075**	**0.6231 ± 0.0073**	**0.7707 ± 0.0050**
	LSTM	0.7858 ± 0.0026	0.5083 ± 0.0076	0.6327 ± 0.0068	0.5881 ± 0.0091	0.7410 ± 0.0041
	Bi-LSTM	0.7800 ± 0.0044	0.4950 ± 0.0106	0.6249 ± 0.0054	0.5796 ± 0.0107	0.7349 ± 0.0041
	Conv-LSTM	0.7876 ± 0.0014	0.5167 ± 0.0074	0.6440 ± 0.0123	0.5936 ± 0.0064	0.7482 ± 0.0074
	Seg-GRU	0.7946 ± 0.0034	0.5304 ± 0.0103	0.6432 ± 0.0128	0.6044 ± 0.0101	0.7497 ± 0.0077
NMF−BoW with SC $(6,532 \times 100)$	MLP	0.7787 ± 0.0039	0.4910 ± 0.0091	0.6075 ± 0.0087	0.5790 ± 0.0073	0.7295 ± 0.0054
	ConvNet	**0.7956 ± 0.0028**	**0.5358 ± 0.0075**	**0.6550 ± 0.0067**	**0.6046 ± 0.0081**	**0.7599 ± 0.0039**
	LSTM	0.7770 ± 0.0012	0.4885 ± 0.0058	0.6176 ± 0.0114	0.5722 ± 0.0034	0.7336 ± 0.0069
	Bi-LSTM	0.7757 ± 0.0039	0.4832 ± 0.0104	0.6118 ± 0.0116	0.5698 ± 0.0081	0.7292 ± 0.0055
	Conv-LSTM	0.7823 ± 0.0042	0.5041 ± 0.0090	0.6331 ± 0.0076	0.5828 ± 0.0092	0.7436 ± 0.0047
	Seg-GRU	0.7800 ± 0.0042	0.4969 ± 0.0164	0.6267 ± 0.0259	0.5796 ± 0.0095	0.7411 ± 0.0160
NMF−TW with SC $(6,532 \times 100)$	MLP	0.7884 ± 0.0037	0.5162 ± 0.0105	0.6316 ± 0.0090	0.5943 ± 0.0093	0.7441 ± 0.0052
	ConvNet	**0.8062 ± 0.0029**	**0.5608 ± 0.0085**	**0.6718 ± 0.0087**	**0.6229 ± 0.0087**	**0.7695 ± 0.0046**
	LSTM	0.7847 ± 0.0033	0.5049 ± 0.0101	0.6307 ± 0.0131	0.5848 ± 0.0079	0.7404 ± 0.0073
	Bi-LSTM	0.7804 ± 0.0033	0.4949 ± 0.0097	0.6211 ± 0.0121	0.5784 ± 0.0066	0.7335 ± 0.0085
	Conv-LSTM	0.7919 ± 0.0036	0.5246 ± 0.0096	0.6436 ± 0.0095	0.5997 ± 0.0069	0.7489 ± 0.0054
	Seg-GRU	0.7928 ± 0.0049	0.5268 ± 0.0156	0.6446 ± 0.0158	0.6009 ± 0.0120	0.7511 ± 0.0093

Table 7.6: Code group prediction (with machine learners) using nursing notes aggregated naïvely by patient identification numbers.

Data model	Classifier	Performance scores				
		ACC	MCC	F-score	AUPRC	AUROC
BoW $(6,532 \times 14,665)$	KNN	0.7741 ± 0.0023	0.4912 ± 0.0025	0.6320 ± 0.0019	0.5454 ± 0.0022	0.7405 ± 0.0019
	LR as OvR	**0.8056 ± 0.0019**	**0.5418 ± 0.0026**	**0.6668 ± 0.0012**	**0.6094 ± 0.0026**	**0.7348 ± 0.0012**
	SVM as OvR	0.7549 ± 0.0015	0.5064 ± 0.0016	0.6148 ± 0.0021	0.5789 ± 0.0018	0.6452 ± 0.0007
	RF ensemble	0.7255 ± 0.0027	0.4067 ± 0.0012	0.5182 ± 0.0025	0.5133 ± 0.0023	0.6670 ± 0.0014
TW $(6,532 \times 14,665)$	KNN	0.7866 ± 0.0012	0.5306 ± 0.0032	0.6697 ± 0.0021	0.5920 ± 0.0025	0.7689 ± 0.0016
	LR as OvR	**0.8143 ± 0.0014**	**0.5845 ± 0.0035**	**0.6874 ± 0.0030**	**0.6378 ± 0.0032**	**0.7804 ± 0.0017**
	SVM as OvR	0.7414 ± 0.0015	0.4007 ± 0.0036	0.5207 ± 0.0028	0.5249 ± 0.0026	0.6801 ± 0.0015
	RF ensemble	0.7653 ± 0.0011	0.4449 ± 0.0031	0.5484 ± 0.0023	0.5517 ± 0.0024	0.6951 ± 0.0013

bles 7.4, 7.5, and 7.6. In Table 7.4, the performance of the proposed modeling approaches that are built on *FarSight*-aggregated clinical nursing data is summarized. Table 7.5 shows the performance of all the proposed modeling approaches built on data obtained by naïvely aggregating the patients' nursing notes using their identification numbers. We observe that the NMF$-$TW with SC approach built on *FarSight*-aggregated data and modeled using Conv-LSTM consistently outperforms other data modeling and classification approaches with respect to all the metrics. Also, the performance of the proposed models drastically increased by 2.47% in Accuracy, 16.07% in MCC, 13.43% in F-score, 16.13% in AUPRC, and 6.50% in AUROC when the data was aggregated using the *FarSight* long-term aggregation mechanism.

We compared the actual and predicted (using Conv-LSTM trained on NMF$-$TW with SC representations) number of clinical notes that received a particular diagnostic code group. It was observed that the diagnostic code ranges including $001-139$, $280-289$, $320-389$, $460-519$, $630-677$, and $780-789$ had less than 100 mismatches ($< 0.007\%$); $520-579$, $580-629$, and $800-999$ had less than 500 mismatches ($< 0.35\%$) between the actual and predicted ICD9 code groups across $140,792$ nursing notes. We also remarked that the maximum number of mismatches (over $3,500$) corresponded to Ref and V-codes ($4,078$ – 2.90%), and the code range of $710-739$ ($4,366$ – 3.10%). Note that the statistics presented above were measured as the maximum mismatches across all the cross-validation folds. Table 7.6 illustrates the ICD9 code group prediction performance of conventional machine learning models including K-Nearest Neighbors (KNN) with $K = 15$, LR as One-vs-Rest (OvR) with stochastic average gradient solver, Support Vector Machines (SVM) as OvR with radial basis function kernel, and Random Forest (RF) ensemble with 100 trees (maximum depth of 2), using nursing notes aggregated naïvely by patient identification numbers. This study does not include BoW or TW modeling on *FarSight*-aggregated data, owing to the high-dimensionality and sparsity of such statistical transformations of the underlying corpus ($140,792 \times 32,687$). From Fig. 7.7, it is evident that the proposed deep learners trained on *FarSight*-aggregated data outperform the conventional machine learners and deep learners trained on naïvely aggregated data.

The ability of a model to effectively capture the TP, TN, FP and FN in risk assessment is of paramount importance, owing to the critical nature of the task itself. As explained in the previous work, AUPRC measures the number of true positives from the set of positive predictions, while AUROC captures the TPR and FPR. AUPRC varies with a change in the ratio of the target classes in the data

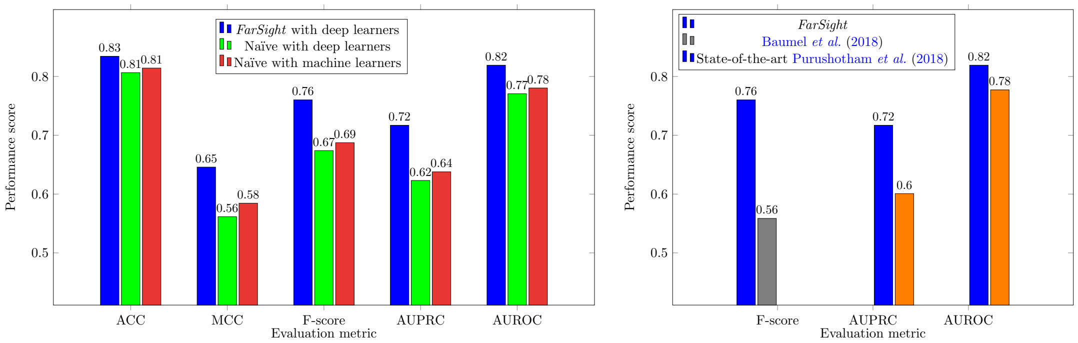


Figure 7.7: Comparison of the best performing models (aggregated with and without *FarSight*) and benchmarking works in ICD9 code group prediction.

and hence, is more revealing than AUROC in this context (Saito and Rehmsmeier, 2015). Precision captures the proportion of the patient records that the proposed model predicted to have a risk that actually had a risk, while recall expresses the ability to find all the patients at risk. These are captured using the F-score, while the MCC score accounts for TP, FP and FN, thus serving as a balanced measure even with class imbalance.

To facilitate the prediction of clinical outcomes, most existing works, including the state-of-the-art model (considered for benchmarking) (Purushotham *et al.*, 2018), rely on the structured nature of the EHRs, modeled as feature sets. From Fig. 7.7, it can be observed that our model built on the unstructured nursing text and modeled using *FarSight*-aggregated data, significantly outperforms the state-of-the-art model by 19.34% in AUPRC and 5.41% in AUROC, and the hierarchical attention GRU model (Baumel *et al.*, 2018) by 35.71% in F-score. Moreover, most of the existing works benchmarked their performance only on the AUPRC and AUROC metrics, while neglecting to assess the performance of their models with metrics most suited in the cases of imbalanced data, as is the case with most of the real-world data. We argue that the reliability and other critical aspects of the underlying CDSS can be accurately and explicitly captured by assessing the performance of a model using targeted metrics like Accuracy, F-score, and MCC scores, which are incorporated in our work. The NMF$-$TW with SC model was able to capture discriminative features of the nursing notes needed for the classifier to learn and generalize. Also, the *FarSight* aggregation strategy effectively facilitates accurate risk assessment, well in advance, with an overall accuracy of 83.3%. Thus, a CDSS equipped with the predictive capabilities of *FarSight*-aggregation and NMF$-$TW with SC modeling could demonstrate evidence-based and patient-centric risk assessments. Furthermore, these observations corroborate the suitability of *FarSight* for clinical decision support in real-world hospital scenarios, especially in developing countries with limited resources and low structured EHR adoption rates.

7.4 Summary

In this chapter, two strategies for performing aggregation of unstructured clinical notes namely – *TAGS* and *FarSight* were presented, based on which disease group prediction models were built. *TAGS* strategy involved a fuzzy similarity scoring based decision mechanism to merge/purge nursing notes and a term weighting based vector space model for effective textual feature representation, which were then used to train ML classifiers. *FarSight* involved a future look-ahead mechanism to map admission records to aggregated diagnosis groups across the admission records. The model also involved a SC based NMF topic modeling approach to effectively represent the textual features, which were then used to train deep learning architectures. Both the approaches facilitate early detection of onset of disease groups and were found to outperform state-of-the-art approaches for disease group prediction.

Publications

(based on work presented in this chapter)

1. Tushaar Gangavarapu, Aditya Jayasimha, Gokul S Krishnan, Sowmya Kamath S, "*Predicting ICD-9 Code Groups with Fuzzy Similarity based Supervised Multi-Label Classification of Unstructured Clinical Nursing Notes*", Knowledge-Based Systems, Elsevier, Volume 190, 2020. (SCI & Scopus, IF:5.921) *(Published)*

2. Tushaar Gangavarapu, Gokul S Krishnan, Sowmya Kamath S, Jayakumar Jeganathan, "*FarSight: Long-Term Disease Prediction Using Unstructured Clinical Nursing Notes*", IEEE Transactions on Emerging Topics in Computing. (SCI & Scopus, IF: 6.043) *(Published)*

3. Gangavarapu, Tushaar, Gokul S. Krishnan, and Sowmya Kamath S. "*Coherence-Based Modeling of Clinical Concepts Inferred from Heterogeneous Clinical Notes for ICU Patient Risk Stratification*", In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL) (pp. 1012-1022), Hong Kong, China, November 2019. (CORE Ranked) *(Published)*

PART V

Towards Population Analytics based CDSSs

Chapter 8

Population-centric Predictive Analytics

8.1 Introduction

Population health is an important contributor to the financial, socio-economic, defense and behavioral aspects of a country (Holt *et al.*, 2016; Organization, 2000). Comprehensive investigations and vigilant surveillance of its population's health is a task of utmost importance for a country's government agencies so as to prevent epidemic outbreak scenarios, medicine shortage during epidemics, vaccine outrage and even for detecting signs of bio-warfare and devising strategies for circumventing them. Automating population health surveillance based on the confluence of technologies like Big Data Analytics, Data Mining and ML has thus created huge scope in gaining potentially latent insights to help govern the health of a population and also drive public health policies (Darcy *et al.*, 2016; Krumholz, 2014). Patient level data available in hospitals can provide government agencies with systematic data on instances of disease or virus outbreak, which can help put effective prevention and quarantine procedures in place. At the same time, it also provides long-term data for future prediction of similar outbreaks even before the symptoms manifest themselves. Statuses and posts on Online Social Network (OSN) sites such as Twitter, Facebook, etc. have proven to be an abundant source of useful information for such population based analytics. Several research works that use Big Data Analytics and ML have been proposed over the years that prove the same.

Public health policies have made vaccination mandatory for several endemic diseases, which has turned out to be an important cornerstone of a country's fundamental public health framework. Governments frame vaccination policies with reference to the possibility of disease outbreak risks or biowar factors, and have been hugely successful in completely eradicating contagious diseases like smallpox and polio, for instance, India's commendable efforts towards implementing long-term vaccination policies for all citizens. This underscores the need for effective framing of vaccination policies which in turn requires meticulous research and data collection on vaccine-related behaviours, side effects and people's beliefs (Zell *et al.*, 2000). Understanding public opinion towards vaccination and outcome planning is critical for the development of public health policies and effective implementation of such policies for maximising impact (Downs *et al.*, 2008). Public opinion on vaccinations can be diverse - for example, some people voluntarily get vaccination shots for Influenza, while many others may choose not to, despite the recommendations of medical practitioners, due to their own inherent biases. Several factors affect effective policy formulation - vaccination behaviour defines if people received the shot or not, and vaccination hesitancy indicates an inherent disinclination towards vaccines or people's negative opinions towards vaccination (Joshi *et al.*, 2018). It is crucial for the government health agencies to discern the vaccination coverage or the statistics on the number of people who received the vaccination shot, thus making vaccination behaviour (or shot) collection a task of critical importance. Normally, such information is collected from the citizens through surveys and targeted interactions, which are not only difficult and time-consuming procedures but also under-represent a lot of categories of citizens due to sampling interval and framing errors (Huang *et al.*, 2017; Keeter *et al.*, 2006; Parker *et al.*, 2013). This requirement has led to extensive research towards automated vaccine behavior modeling and mining of public opinion on vaccination, through potential alternate means, like mining OSN data for inferring people's views, experiences, biases, and potential hindrances to effective implementation of vaccination policies (Dredze *et al.*, 2016).

Depression, a debilitating health condition that impairs sufferers' quality of life, is a leading illness worldwide and is a major contributor towards the overall burden of global diseases (Shen *et al.*, 2017). It is estimated by the WHO that around 300 million people of various ages suffer from depression and also that more than 800,000 people suffering from chronic depression commit suicide each year[1]. Depression is a long lasting illness that can be diagnosed by differentiating behaviours and can have fatal side effects like suicidal tendencies. Statistics show that over 70% of people in early stages of depression would not consult psychological health personnel as they are unaware or even ashamed of their condition (Shen *et al.*, 2017). However, it is observed that people actually open up in the relative anonymity and potential outlet offered by social media platforms

[1] https://www.who.int/news-room/fact-sheets/detail/depression

like Facebook and Twitter to share or vent out their emotions/moods through messages and status updates. Such data opens up significant possibilities towards understanding disease physiology and how it affects individuals, so that effective early intervention measures and automated surveillance systems can be designed.

8.1.1 Problem Definition

Automated population health surveillance systems are seen as a vital solution to the problems defined above and mining OSN data can provide essential insights for government health agencies for informed decision making. Computational techniques like NLP and ML aid in performing predictive analytics based on social media data. Most existing OSN based prediction or analysis models are designed to perform a specific task, say vaccine sentiment or depression detection. However, in the real world scenario, this means that for each prediction task to be performed, different learning architectures have to be designed and deployed, which makes it challenging for governing bodies to manage and maintain these models. Hence, there is a need for prediction architectures that are dynamic and effective for multiple prediction tasks. Therefore, the problem to be solved can be stated as follows:

> *"Given openly available OSN data, build a generic prediction model architecture that can perform multiple population health based prediction tasks enabling insights into population health."*

8.1.2 Motivating Example

The Ministry of Health and Family Welfare, Government of India, actively monitors population health of Indian citizens. Consider that an associated organization, *PopHealth* collects health data from various state governments, who in turn collect data from government hospitals, private hospitals and even from the numerous primary health care centres introduced under the AB-PMJAY scheme. Now let us consider one part of the data is the counts of say, flu affected people, across the country. This data collection typically takes a lot of time to be generated because of operational reasons and takes even more time to analyze, to enable actionable insights.

Because of the huge numbers of social network users and availability of OSN data providers, *PopHealth* decides to add a technical module that monitors OSN users in India and analyze these data also for signs of flu outbreak. The module with NLP and ML capabilities can analyze numerous posts such as Tweets and Facebook posts, which can provide clearer insights of the health conditions of citizens and might also give an idea of spread of the flu outbreak (as is evidenced by popular services like Google FluTrends$^{\text{TM}}$). This technical module, along with the incoming data from the hospitals and health organizations across India enables *PopHealth* to make informed decisions on several factors like availability of medicines, treatment measures and medical personnel, using military measures or even quarantine measures (if necessary). In addition to general disease outbreaks, *PopHealth* may also be able to assess sentiment and intent of people towards health policies like vaccine policies, new healthcare initiatives, etc. as they can analyze the OSN data of their citizens.

8.2 MDSHA: Multi-task Deep Social Health Analyzer using Particle Swarm Optimization aided Topic Modeling

The overall workflow of the proposed MDSHA approach is as depicted in Figure 8.1. The tweets dataset corpus (required for the task) was subjected to a basic preprocessing strategy where, initially, special characters except white spaces were removed. Tokenization was performed so as to strip down the corpus into tokens. Stemming and Lemmatization were performed to reduce the words to its root form. Finally, stopping was also performed to remove unimportant words from the corpus.

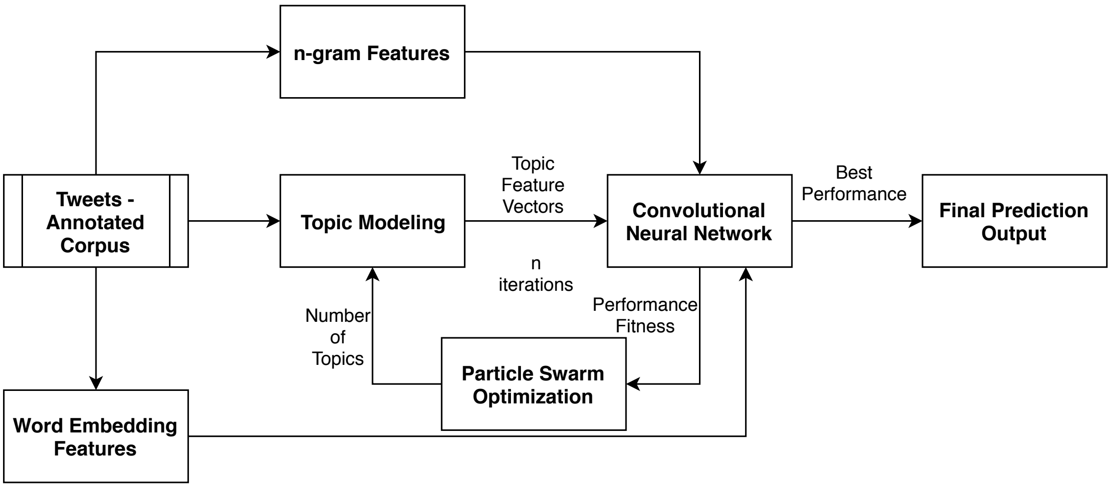


Figure 8.1: Workflow of Proposed MDSHA

8.2.1 Term Weighted n-gram Features

Next, the preprocessed tokens obtained from the corpus are then vectorized using a TF-IDF (Term Frequency - Inverse Document Frequency) vectorizer to create their TF-IDF weighted n-gram features. TF-IDF is a statistical measure that signifies the importance or relevance of a word within a document and is often considered as weighing factor in text mining approaches. In this approach, the top 2000 TF-IDF weights for n-grams (i.e., unigrams, bigrams and trigrams) were considered and were used as Text Feature Set 1 (hereafter referred to as FS1).

8.2.2 Word Embedding Features

Word Embeddings are dense real number vector representations of words or groups of words that can be used as textual features for tasks like classification, clustering, etc. They are considered effective as they are generated taking concepts such as context and co-occurrence of words into consideration. In our work, we adopted the Skipgram model of Word2Vec Mikolov *et al.* (2013) (explained in Section 5.2.3), a neural network based word representation model that generates word embedding vectors based on prediction of context given an input word. We used a dimension size of 500 per word and averaged the vector representation such that each tweet is represented by a 1x500 vector(hereafter referred to as Text Feature Set 2 , i.e., FS2).

8.2.3 Latent Dirichlet Allocation

Topic modeling is an unsupervised method that determines a set of topics based on the documents in a corpus, that most represent the documents. The objective of topic modeling is to derive a representation such that - "each document is best described by a set of topics and each topic is best described by a set of words". Latent Dirichlet Allocation (LDA) (explained in Section 7.2.5), a generative probabilistic model, is a popular topic modeling approach that assigns documents in a corpus to a set of topic clusters. LDA hypothesizes that each document, given that it consists of a set of particular words, belongs to a set of topics with a certain probability value. In this paper, the preprocessed tweets corpus is subjected to LDA and the topic vectors that contain the probabilities of a document (that it belongs to each of the topics) are extracted and used as features for supervised classification, i.e., Text Feature Set 3 is hereafter referred to as FS3.

Similar to any unsupervised clustering technique, automatically determining the number of topics during LDA is a challenging task and remains an open research problem. In this paper, we utilize Particle Swarm Optimization (PSO), an evolutionary computation approach for dynamically determining the optimal number of topics for various prediction tasks. The application of PSO for the same is explained in detail in subsequent sections.

8.2.4 PSO-LDA Topic Modeling

The challenging task of determining the optimal number of LDA topics for topic modeling also includes deriving the optimal number of features in the topic feature vector generated by the topic model. For a particular prediction task, dynamically determining the number of topics is critical and this is decided by modeling the optimal feature set. To derive the optimal feature set, the solution subspace to be searched is quite large and for this reason, evolutionary computation is an apt choice to go with. Towards this objective, a wrapper module based on PSO and Convolutional Neural Networks (CNN), referred to as PSO-CNN, is proposed, which dynamically determines the number of topic clusters based on classification performance of CNN using the combined set of all the textual features - FS1, FS2 and FS3.

The algorithm for the PSO aided topic modeling using the PSO-CNN wrapper is as shown in Algorithm 5. The combination of the three feature sets – FS1, FS2 and FS3, is fed into a PSO-CNN wrapper to determine the number of topics based on the CNN's classification performance. The neural network model, i.e., the CNN part of the wrapper was adopted from the TextCNN model (Kim, 2014) for effectively classifying text using textual features. The fitness performance of the PSO-CNN wrapper was measured in terms of F-score (Eq. 8.3) as it considers not only True Positives (TP), but also False Positives (FP) and False Negatives (FN).

$$Precision = \frac{TP}{TP + FP} \tag{8.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{8.2}$$

$$F\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{8.3}$$

At the beginning, a swarm or flock of particles, along with particle positions were initialized as a list of number of topics candidates for the LDA model. For

Algorithm 5 Topic Modeling using PSO-CNN Wrapper

Input: Text Feature Sets – FS1, FS2 and FS3 and labels for the given prediction task

Output: Optimal number of topics for Topic Modeling and Performance for the optimized Feature Sets (m features)

```
Initialize 8 particles with their positions and velocities
Position vector, $x = [250, 500, 750, 1000, 1250, 1500, 1750, 2000]$
Velocity of each particle, $v = [0, 0, 0, 0, 0, 0, 0, 0]$
$local\_best = [-\infty, -\infty, -\infty, -\infty, -\infty, -\infty, -\infty, -\infty]$
$global\_best = 0$
while iterations $\leq 50$ do
    for each particle do
        Use particle position value for LDA topic modeling and generate FS3
        Concatenate FS3 with FS1 and FS2
        Feed the combined feature sets and labels to PSO-CNN wrapper for training
        Calculate fitness of feature set (F-score performance of PSO-CNN
        if $fitness > local\_best_i$ then
            Set $local\_best_i$ = current particle position, $x_i$
        end if
        if $fitness > global\_best$ then
            Set $global\_best$ = current particle position, $x_i$
        end if
    end for
    Update velocities and positions of particles
end while
Return $global\_best$                              ▷ Optimal number of topics
```

each of the position i, the best classification performance of the PSO-CNN in terms of F-score, $local\_best_i$, and that of the entire swarm or flock, the $global\_best$ is calculated and the positions and velocities of the particles are updated using standard PSO equations (Eq. 8.4 and 8.5). In our work, a set of eight particles were used, and the PSO-CNN wrapper was set to 50 iterations. The best position of the swarm or flock after all the iterations was considered as the optimal number of topics for the LDA topic modeling for a given prediction task. The performance at this position, $global\_best$, was considered the result metric.

$$v_{i+1} = w * v_i + c_1 * r_1 * (local\_best_i - x_i) + c_2 * r_2 * (global\_best - x_i) \tag{8.4}$$

$$x_{i+)} = x_i + v_{i+1} \tag{8.5}$$

where, $c1$ and $c2$ are empirically determined constants 0.5 and 0.2 respectively and $r1$ and $r2$ are randomly generated real numbers.

The CNN model in the PSO-CNN wrapper consisted of three one dimensional convolution layers consisting of 512 filters with kernel (filter) sizes - 5,6 and 7 respectively. The input layer consisted of the number of units pertaining to the number of topics determined by the PSO aided Topic modeling module and the total number of features in the combined feature set. Each convolution layer was provided with a Rectified Linear Unit (ReLU) activation function. Respective maxpooling layers (one dimensional) were added to each of the convolution layers, after which a concatenation and flatten layers were provided and finally, a dropout of 50% was also added to avoid overfitting. The output layer included a sigmoid activation function and rmsprop optimizer was used for training with binary cross-entropy as the loss function. The overall architecture of the CNN model used in the proposed approach is illustrated in the Figure 8.2.

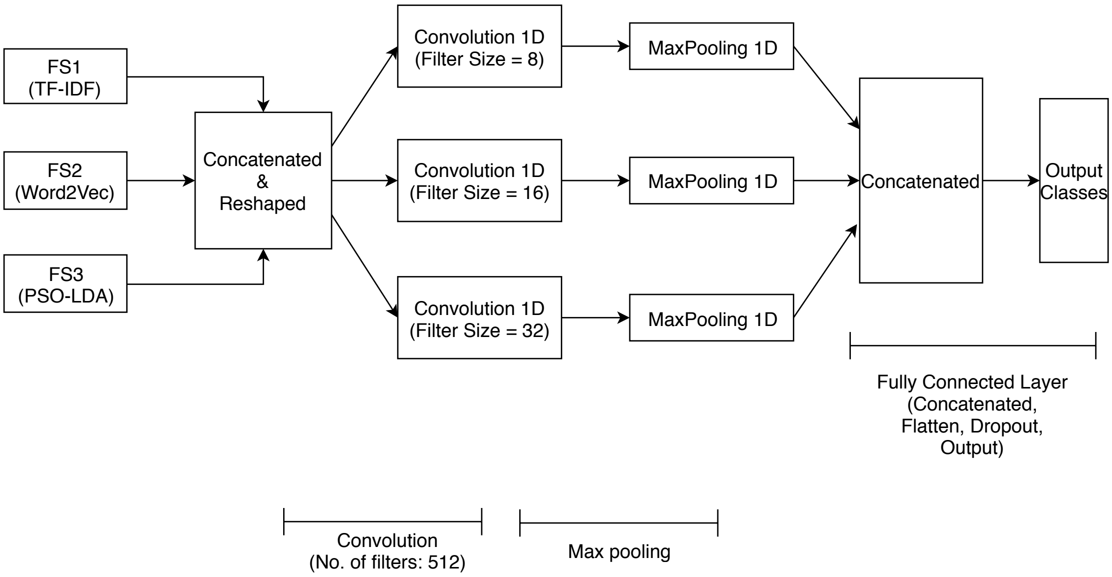


Figure 8.2: CNN model used in PSO-CNN Wrapper

8.2.5 Experimental Results & Discussion

The proposed MDSHA approach was benchmarked for three different population analytics based prediction tasks – Flu Vaccine Hesitancy (Flu Vaccine Intent), Flu Vaccine Behaviour (Flu Vaccine Shot Received or Not) and finally, Depression Detection on respective datasets. The performances of the prediction models were measured and compared in terms of standard ML metrics – precision, recall and

F-score. The proposed MDSHA approach is also benchmarked against the state-of-the-art approaches for the respective tasks. The experimental results and analysis are presented in detail in subsequent sections.

8.2.5.1 Task 1: Flu Vaccine Hesitancy (Intent)

The Intent prediction task uses tweets to determine the intent or hesitancy of a user towards an influenza vaccine, i.e., whether the user intends to receive the vaccine or not. We used the DS1 dataset collated and provided by Huang *et al.* (2017) for this task. The dataset consists of 10,000 tweets that are based on influenza vaccine and its characteristics are as tabulated in Table 8.1. It is to be noted that only the relevant tweets were used (irrelevant tweets are labelled in the dataset) and that rows with missing labels were dropped, after which we were left with 9513 rows. MDSHA was applied to the dataset and the PSO-CNN wrapper based Topic modeling module was used to determine the optimal number of topics as 634, along with 500 Word2Vec embeddings and 2000 top n-gram features, thereby making the total number of textual features as 3134. The performance of the proposed MDSHA approach is compared to that of the state-of-the-art approach by Huang *et al.* (2017), where TF-IDF weighted n-gram features and Logistic Regression classifier were used for the task. Similar to their approach, the performance of MDHSA was measured after a 5-fold cross validation process. The comparison of performance is as shown in Table 8.2.

Feature	**Frequency**
Unique tweets	9,865
Users	9,334
Words	1,54204
Intend/Receive (Positive)	3,148
Negative Intend	6,365
Only Intends	2,354
Shot Received	743

Table 8.1: Dataset Statistics of DS1

From Table 8.2, it can be observed that the proposed MDSHA approach outperformed Huang *et al.* (2017)'s approach in terms of Recall and F-score by 5% and 2% respectively. Higher values of recall and F-score indicate that MDSHA was able to reduce the number of False Negatives (FN), thus making the classification

Approach	Precision	Recall	F-Score
Huang *et al.* (2017)	0.84	0.80	0.82
MDSHA	0.84	0.84	0.84

Table 8.2: Flu Vaccine Intent: Comparison on DS1

much more effective.

8.2.5.2 Task 2: Flu Vaccine Behaviour (Shot Received or Not)

In this prediction task, the Vaccine Behaviour of a user is to be determined, i.e., whether a twitter user received an influenza vaccine shot or not, depending on the users' tweets. We used two datasets for this task - Huang *et al.* (2017)'s dataset (DS1) and Joshi *et al.* (2018)'s dataset (DS2), used by these authors, in their respective studies. For DS1, we applied the proposed MDSHA approach for the 3097 tweets (2354 positive and 743 negative). It is to be noted that only the relevant tweets were chosen in which the labels were either intends to receive (negative) or shot received (positive) as similar to the tasks performed by Huang *et al.* (2017). The PSO-CNN wrapper based topic modeling module was used to determine the optimal number of topics to be 2000 and the total number of textual features added to 4500. The performance was compared against the Logistic Regression based approach on n-gram features by Huang *et al.* (2017). Similar to their approach, the performance was measured after 5-fold cross validation and the results are as shown in Table 8.3.

Table 8.3: Flu Vaccine Shot Detection: Comparison on DS1

Approach	Precision	Recall	F-Score
Huang *et al.* (2017)	0.90	0.95	0.93
MDSHA	0.92	0.91	0.91

Dataset DS2 is actually the training dataset used in SM4HH 2018 workshop (shared task #4), for which Joshi *et al.* (2018) presented the cross-validation results on the training dataset, making it easier for us to compare with, as we do not have access to the test dataset of the task. This dataset also consisted of tweets with respective labels (positive or negative) and the respective characteristics of the dataset are provided in Table 8.4. The proposed MDSHA approach was applied on the dataset and 10-fold cross-validation was performed (as similar to that of Joshi *et al.* (2018)). The PSO based topic modeling module determined the optimal number of LDA topic clusters to be 289, making the total of textual features to be 2789. The MDSHA approach was compared against the performance of the approach by Joshi *et al.* (2018) and the results are tabulated in Table 8.5.

Table 8.4: Dataset Statistics of DS2

Feature	**Frequency**
Tweets	5,391
Words	84,826
Shot Received	1,558
Shot not Received	3,883

Table 8.5: Flu Vaccine Shot Detection: Comparison on DS2

Approach	**Precision**	**Recall**	**F-Score**
Joshi *et al.* (2018)	*	*	0.81
MDSHA	0.86	0.86	0.86

* indicates metric not reported in study

From Table 8.3 and Table 8.5, it can be observed that the proposed MDSHA approach outperforms the existing approaches in terms of precision metric and F-score metric respectively for DS1 and DS2. For DS1, the MDSHA approach slightly underperforms in case of F-score when compared to Joshi *et al.* (2018)'s approach. High value of precision indicates that MDSHA achieved better number of TP while reducing FP, thus making the classification much more effective in determining whether the user took the vaccine shot or not. Moreover, the proposed MDSHA approach significantly outperformed the existing method on DS2 by 6% in terms of F-score.

8.2.5.3 Task 3: Depression Detection

This prediction task aims to determine whether a Twitter user is depressed or prone to depression, based on his or her tweet attributes and content. For the task of depression detection, we used the dataset created and benchmarked by Shen *et al.* (2017) (DS3). DS3 consists of users and their tweets with demographic details, including three subsets - D1 (6493 tweets labeled as depressed), D2 (5384 tweets labeled as non-depressed) and D3 (a large unlabeled set). D1 and D2 were used for classification performance benchmarking and D3 was used by Shen *et al.* (2017) for behavior analysis of users. The overall statistics of DS3 are as tabulated in Table 8.6. We applied the proposed MDSHA approach on the combined set of D1 and D2 and performed 5-fold cross validation. The optimal number of LDA topic clusters determined by the PSO aided topic modeling module was 250, making the total number of textual features to be 2750. The results were compared with the performance of Shen *et al.* (2017)'s model and are tabulated in Table 8.7.

Table 8.6: Dataset Statistics of DS3

Feature	**Frequency**
Tweets	11,877
Users	11,059 (5,899 depressed and 5,160 non-depressed)
Words	194,082
Depressed (Tweets)	6,493
Non-depressed (Tweets)	5,384

Table 8.7: Depression Detection: Comparison on DS3

Approach	**Precision**	**Recall**	**F-Score**
Shen *et al.* (2017)	*	*	0.85
MDSHA	0.97	0.97	0.97

* indicates metric not reported in study

From Table 8.7, it can be observed that the proposed MDSHA approach significantly outperforms Shen *et al.* (2017)'s approach by 14% in terms of F-score. This indicates that the proposed approach is able to detect depressed users based on tweets by users in a much more effective way than the state-of-the-art approach. A visualization of the F-score metrics of the proposed approach against existing state-of-the-art approaches is provided in Figure 8.3 for illustrating the improvement in the performance.

8.2.6 Discussion

From Tables 8.2, 8.3, 8.5, 8.7 and also Figure 8.3, it can be observed that the proposed MDSHA approach outperforms the existing approaches for the tasks of flu vaccine intend (DS1), vaccine shot detection (DS2), depression detection (DS3)

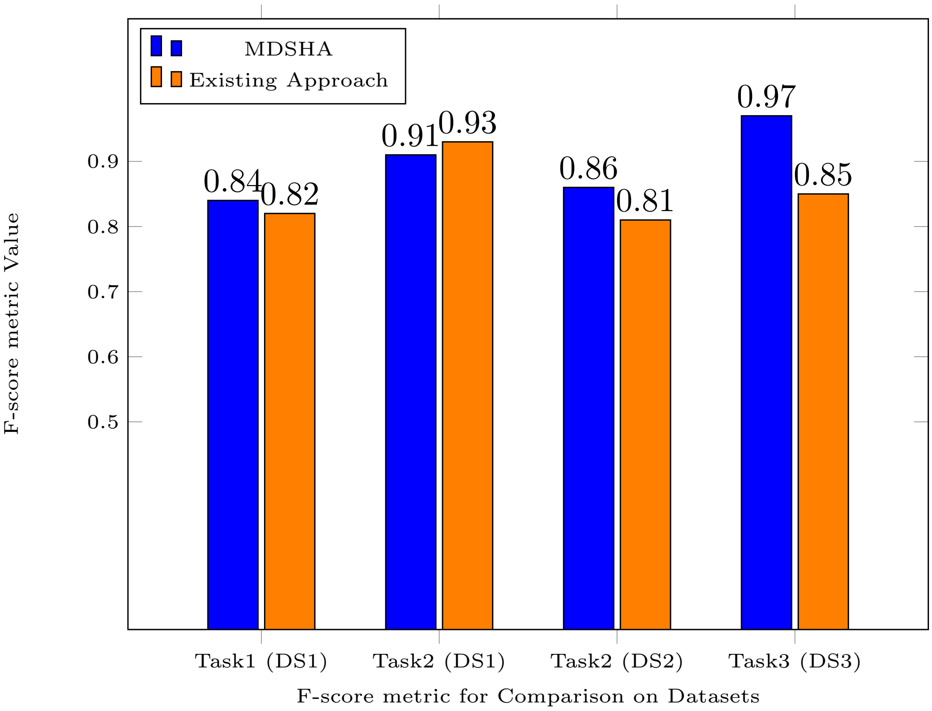


Figure 8.3: F-score comparison for All Tasks on Respective Datasets

in terms of most metrics. While it falls below the existing approach's F-score for flu vaccine behaviour (shot detection) on DS1 by 2%, the precision has improved by 2%, ensuring the reduced number of FP. The FP, in this case, is a critical measure as it detects a user to have received the shot even though he has not and hence should be reduced. Moreover, the proposed MDSHA approach significantly improved over the state-of-the-art approach by 6% in terms of F-score on DS2, ensuring its efficacy for the task of flu vaccine behaviour (shot detection). It is also to be noted that the proposed MDSHA approach is based on purely tweet text and does not include any other kind of features. The topic modeling probabilities, along with other textual features such as word embeddings and n-grams, the PSO and neural network based MDSHA approach is able to generalize the context and make the classification tasks more effective. Moreover, the ability of PSO based topic modeling approach that determines the optimal number of topic clusters for LDA topic modeling, makes MDSHA well-suited for any task due to its dynamic nature. Having proven to be effective in multiple tasks of population based analytics on OSN data, MDSHA can be considered to be a generic approach for real-world applications for such tasks.

8.3 Summary

In this chapter, a generic model for population health analytics based on OSN data called MDSHA was presented. The proposed approach involved a novel PSO based topic modeling approach for effective textual feature representation and a CNN based classifier for prediction. The proposed MDSHA approach was trained and tested for three population analytics based prediction tasks – flu vaccine hesitancy (intent recognition), flu vaccine behaviour (shot received or not) and depression detection. For each task, the proposed approach outperformed the respective state-of-the-art approaches.

Publications

(based on work presented in this chapter)

1. Gokul S. Krishnan, Sowmya Kamath S., "*Multi-task Deep Social Health Analytics with Particle Swarm Optimization based Topic Modeling*", Evolutionary Intelligence, Springer. (Scopus & ESCI Indexed) *(Revision Submitted)*

Chapter 9

Conclusion & Future Work

9.1 Conclusion

Big Data Analytics in Healthcare is an emerging field that has the potential to revolutionize evidence based medicine, genomic analytics, patient profile analytics, healthcare delivery, clinical operations, research & development and public health policies. Predictive Analytics based CDSSs that demonstrate superior performance over traditional rule based systems have helped caregivers in superior diagnosis and intervention decisions. In this thesis, several strategies for building intelligent individual-centric and population-centric predictive analytics CDSSs were designed and evaluated. The lack of effective feature representation approaches observed from extensive literature review was taken into account. The defined objectives included designing feature modeling and patient representation generation techniques for both structured and unstructured clinical data, for enabling large-scale analytics applications like mortality prediction, disease prediction, population analytics etc.

In Chapter 4, approaches for building CDSSs for patient-centric predictive analytics using structured data were presented. First, an empirical study was performed to observe the effect of feature selection on mortality prediction performance using the combined feature set of traditional severity scoring based MPMs. The approach involved a Logistic regression based RFE wrapper feature selection process and a Random Forest classifier which outperformed the traditional severity scoring based MPMs by a margin of 12-16% in terms of accuracy. From this work, it was determined that feature selection can prove to be effective in improving mortality prediction performance. Next, towards the same, a GA-ELM model was proposed in this research to determine the relevant lab events that helps in effective patient-specific mortality prediction. The proposed approach involved a

GAWFS model that determined the best contributing labevents to mortality prediction and an ELM classifier to effectively predict the patient-specific mortality risk. The proposed approach was found to outperform traditional severity scoring based MPMs by 11-29% and state-of-the-art ML based MPMs by upto 14% in terms of AUROC metric performance.

In Chapter 5, approaches for modeling unstructured clinical textual data to build mortality risk prediction models were discussed. A patient-specific mortality prediction using ECG text reports for cardiac patients was proposed. This work included an unsupervised data cleansing approach to filter out anomalous and special cases and an ELM classifier to perform the prediction. The proposed ELM based approach outperformed the best performing traditional severity scoring SAPS-II based MPM, by nearly 19% in terms of AUROC performance. Next, a benchmarking study on performance of word representation models for patient specific mortality prediction using unstructured clinical notes was also performed. It was determined that the Word2Vec skipgram model generated the better feature representation and the random forest classifier trained on these features outperformed traditional severity scoring based MPMs by a significant margin of 26-38% in terms of AUROC metric.

Chapter 6 presented in detail, the various approaches for building effective disease prediction models based on unstructured clinical notes, in which three ICD9 disease group prediction approaches were proposed. First, an ontology-driven feature modeling approach was proposed to effectively improve the disease group prediction performance. The proposed approach assisted by pretrained embeddings and deep neural networks, outperformed the state-of-the-art disease group prediction model (built on structured data) by 9% in terms of AUROC metric. Second, a two-stage feature modeling approach based on word embeddings and a PSO-NN wrapper was proposed to further enhance the performance of disease group prediction model using unstructured patient records. The performance of the proposed approach, in comparison against the state-of-the-art disease group prediction model was found to be 10% better in terms of AUROC metric. Thirdly, a hybrid feature modeling approach was proposed to effectively model unstructured patient records for disease group prediction. The proposed approach involved dynamically determined weighted combination of word embedding features and deep neural networks, the AUROC performance of which was better than that of state-of-the-art disease group prediction approach by 15%.

Approaches for aggregation of unstructured patient records using *TAGS* and *FarSight* were explored for effective disease group prediction indicating possible capability for early disease diagnosis, details of which are presented in Chapter 7. *TAGS* approach involved a fuzzy similarity scoring based decision-making mechanism to merge/purge clinical notes and a vector space model that generated effective feature representation to train machine learning classifiers. *FarSight* approach involved an aggregation mechanism that enabled future diagnosis lookup, mapping clinical notes to collectively aggregated diagnosis across medical records during an admission. The approach also involved topic modeling approach for effective textual feature representation which was trained using deep learning models. Both the approaches – *TAGS* and *FarSight* based DPMs outperformed state-of-the-art disease group prediction models built on structured data by a margin of 3% and 20% in terms of AUPRC metric respectively.

Chapter 8 presented the techniques that paved the way towards population-centric predictive health analytics through analysis of OSN data. Towards population-centric predictive health analysis tasks such as flu vaccination hesitancy (intent recognition), flu vaccination behaviour (shot detection) and depression detection, a generic model MDSHA was proposed which uses a novel PSO based topic modeling approach for effective feature representation and predictive modeling. The performance of the proposed MDSHA approach was compared to that of existing state-of-the-art approaches for respective prediction tasks – vaccination hesitancy, vaccination behaviour and depression detection using according datasets and the former outperformed them in all tasks by 2%, 6% and 14% respectively in terms of F-score metric.

All CDSS approaches developed as part of this thesis are intended to assist healthcare personnel with insights based on the prediction outputs of the respective CDSS models. The outputs in terms of probabilities can also tell the health personnel the certainty with which the CDSS model makes a particular prediction, with which the health personnel can make informed decisions.

E.g., Currently, the mortality risk of the ICU patients are estimated using traditional scoring based systems, for which a multitude of (sometimes unnecessary) lab events may be required for calculating the mortality risk. With the availability of our proposed approach (discussed in Section 4.3), the hospitals will be able to determine the optimal subset of labevents that are required to be prescribed with respect to their infrastructure and patients. The proposed mortality risk estimation model can determine the mortality risk of a patient based on the optimal subset lab event values at the earliest possible time instant and provide the doctors with a probability value between zero and one, based on which further treatment decisions can be taken. Therefore, the model ensures that the hospital will be able to determine the mortality risk effectively, with reduced number of lab events, which can result in savings in terms of cost, time and other resources.

While the previously provided example showcases a use case of the proposed mortality risk estimation model based on structured patient data, we are providing one more example with respect to disease group (generic diagnosis) prediction models using unstructured data. Let us consider a patient admitted into the ICU due to his/her condition. The continuously monitoring nursing staff enter their narrations or notes into the integrated CDSS of the hospital, where a ML based text classification model (such as the ones presented in Chapter 6), can generate diagnosis predictions based on the input notes. The aggregation strategies presented in Chapter 7 can ensure that the textual feature representations are intelligently generated for which the ML based model can predict possible diagnoses. The predicted diagnoses are also in terms of probabilities which range between 0 and 1 for each of the disease groups, which may enable the doctors to make decisions accordingly.

The contributions in this thesis has explored various avenues of developing effective decision support systems using structured and unstructured data sources that are available in the domain of healthcare at patient level, hospital level and even at population level. We believe that the technical capabilities of decision support systems proposed in this thesis, such as NLP, AI and other data modeling strategies, has an excellent potential for deployment as real-world CDSS applications. We also believe that in doing so, the works proposed in this research can assist healthcare personnel significantly in a positive manner which can ultimately improve the overall quality of healthcare.

9.2 Future Directions

This thesis put forth several approaches towards design and development of CDSSs for various prediction tasks in the domain of Healthcare Informatics. A complete and integrated CDSS for a hospital will consist of one or more such models or systems, which will be designed to work based on task-specific inputs and outputs, as required by end-user applications.

E.g., The scenarios that described our contributions through the example consisting of Hospitals B and C discussed in earlier chapters will benefit from integrated CDSSs integrated with their existing hospital systems. The doctors' notes in Hospital C may be processed by one of the modules in the integrated CDSS and the prediction output (say, diagnosis) may be recommended to the doctor with good level of accuracy (which will improve as the system is fed a wide variety of patient records) who can then make an “informed decision”. Assuming a patient is currently admitted in the ICU for continuous care and monitoring, the monitored lab test values and the recorded nursing notes can be fed into the CDSS models. Once this data is processed the doctor is provided with available progressive and additional insights which may reflect the patient’s current condition in terms of mortality risk or even a change in the diagnosis.

Such integrated systems require dedicated physical/cloud infrastructure for database related tasks for storing EHRs and also for training and deployment of ML based CDSS models. It is also to be noted that such integrated EHR based CDSSs require extensive desktop/mobile/web application development that includes a rich and well-designed user interface for all the stakeholders (health personnel, patients, hospital staff, admin, etc.) to work on. Designing frameworks for effectively inputting patient data into the EHR database is a challenging task as well. A large number of healthcare personnel do not prefer to use such systems due to lack of good user interfaces and due to the time consuming nature of currently existing frameworks’ input mechanisms. An intuitive user interface is a mandatory requirement for such use cases and a lot of effort and brainstorming needs to be put into this before implementing such frameworks in a hospital. Training hospital staff and health personnel to use the systems required for the CDSSs to function will also be a challenge to be overcome. Another key issue to be handled will be the hybrid nature of the patient data streaming into the integrated CDSS application. The new data streaming into the system needs to be stored efficiently and also has to be handled based on the kind of data, i.e., respective data representation strategies need to be incorporated which can generate effective feature representations for the ML models to consume. Yet another issue will be the efficiency of such models to perform tasks in almost real time, which means that there has to be high performance computing based models in place to ensure parallel execution wherever possible. Finally, patient data is meant to be private between the doctors and the patients and has to be stored securely. This means that there has to be effective security on the servers and an authentication system that ensures that only the authenticated users with specific roles can access the data.

The above paragraphs discussed the requirements and challenges while deploying the research presented in this thesis as real world CDSS applications. However, there is abundant scope for research as well in the domain of Healthcare Analytics and Informatics. Some directions to explore for further research are as follows:

- Currently, the DPMs predict the ICD9 disease groups only, i.e. they predict at a higher level of granularity than actual disease codes. Effective models to accurately predict ICD9 codes based on the predicted ICD9 groups using deep learning and evolutionary computation shall be explored in future.

- Currently, the works presented in this thesis takes either structured or unstructured data as input. Observing the performances of CDSS models which hybridize and model structured and unstructured data is a research we wish to pursue.

- Unstructured clinical reports are now an input to the research works presented in this thesis. Clinical report generation using information extracted from structured and unstructured patient data is an avenue intended to be explored.

- The population based prediction model presented in this thesis considers only OSN data. Using real population health data (such as government data on epidemic case counts for various regions) in association with OSN data for effective population-centric predictive analytics is also an interesting avenue that we wish to explore in the future.

- Finally, a major volume of healthcare data is the image format. In this thesis, medical images are not considered for modeling. Multi-modal prediction based CDSS models based on unstructured text and images will be explored, which show significant potential and promise.

Publications based on Research Work

Journal Publications

1. Gokul S. Krishnan, Sowmya Kamath S., "*A novel GA-ELM model for patient-specific mortality prediction over large-scale lab event data*", Applied Soft Computing, Elsevier, Volume 80, 2019, Pages 525-533, ISSN 1568-4946, (SCIE & Scopus, IF: 5.472) *(Published)*

2. Gokul S. Krishnan, Sowmya Kamath S., "*Ontology-driven Text Feature Modeling for Disease Prediction using Unstructured Radiological Notes*", Computación y Sistemas, ISSN 2007-9737, 23(3), 2019. (Scopus & ESCI) *(Published)*

3. Tushaar Gangavarapu, Aditya Jayasimha, Gokul S Krishnan, Sowmya Kamath S, "*Predicting ICD-9 Code Groups with Fuzzy Similarity based Supervised Multi-Label Classification of Unstructured Clinical Nursing Notes*", Knowledge-Based Systems, Elsevier, Volume 190, 2020. (SCI & Scopus, IF:5.921) *(Published)*

4. Tushaar Gangavarapu, Gokul S Krishnan, Sowmya Kamath S, Jayakumar Jeganathan, "*FarSight: Long-Term Disease Prediction Using Unstructured Clinical Nursing Notes*", IEEE Transactions on Emerging Topics in Computing. (SCI & Scopus, IF: 6.043) *(Published)*

5. Gokul S. Krishnan, Sowmya Kamath S., "*A Deep Neural Network Model for Predicting Disease Groups based on PSO-NN Two-stage Feature Modeling of Unstructured Clinical Notes*", ACM Transactions on Computing for Healthcare. (SCIE & Scopus) *(Under Review)*

6. Gokul S. Krishnan, Sowmya Kamath S., "*Multi-task Deep Social Health Analytics with Particle Swarm Optimization based Topic Modeling*", Evolutionary Intelligence, Springer. (Scopus & ESCI Indexed) *(Revision Submitted)*

Conference Publications

1. Gokul S. Krishnan and Sowmya Kamath S. "*Evaluating the quality of word representation models for unstructured clinical Text based ICU mortality prediction*". In Proceedings of the 20th International Conference on Distributed Computing and Networking (ICDCN '19), ACM, IISc Bangalore, January 2019. (CORE Ranked) *(Published)*

2. Krishnan, Gokul. S., & Kamath, Sowmya. S., "*A Supervised Learning Approach for ICU Mortality Prediction Based on Unstructured Electrocardiogram Text Reports*", In the proceedings of the 23rd International Conference on Applications of Natural Language to Information Systems (NLDB 2018), Springer, Paris, France, June 2018. (CORE Ranked) *(Published)*

3. Gokul S Krishnan and Sowmya Kamath S, "*A Supervised Approach for Patient-specific ICU Mortality Prediction using Feature Modeling*", 7th International Conference on Frontier Computing (FC 2018), Springer, Kuala Lumpur, Malaysia, July 2018. *(Published)*

4. Gangavarapu, Tushaar, Gokul S. Krishnan, and Sowmya Kamath S. "*Coherence-Based Modeling of Clinical Concepts Inferred from Heterogeneous Clinical Notes for ICU Patient Risk Stratification*", In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL) (pp. 1012-1022), Hong Kong, China, November 2019. (CORE Ranked) *(Published)*

5. Gokul S. Krishnan, Sowmya Kamath S., "*Hybrid Text Feature Modeling for Disease Group Prediction using Unstructured Physicians' Notes*", International Conference on Computational Science (ICCS) 2020. (CORE Ranked) *(Published)*

References

Achrekar, H., A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu, Predicting flu trends using twitter data. *In Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*. IEEE, 2011.

Alshammari, S. M. and R. D. Nielsen, Less is More: With a 280-character limit, Twitter Provides a Valuable Source for Detecting Self-reported Flu Cases. *In Proceedings of the 2018 International Conference on Computing and Big Data*. ACM, 2018.

Aramaki, E., S. Maskawa, and M. Morita, Twitter catches the flu: detecting influenza epidemics using Twitter. *In Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2011.

Awad *et al.* (2017). Patient length of stay and mortality prediction: A survey. *Health Services Management Research*, 0951484817696212.

Baron, R. J. (2010). What's keeping us so busy in primary care? a snapshot from one practice.

Basak, D., S. Pal, and D. C. Patranabis (2007). Support vector regression. *Neural Information Processing-Letters and Reviews*, 11(10), 203–224.

Baud, R. H., C. Lovis, P. Ruch, and A.-M. Rassinoux, A light knowledge model for linguistic applications. *In Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001.

Baumel, T., J. Nassour-Kassis, R. Cohen, M. Elhadad, and N. Elhadad, Multilabel classification of patient notes: case study on icd code assignment. *In Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

Bayegan, E. (2002). Knowledge representation for relevance ranking of patient-record contents in primary-care situations. *Dr. ingeniøravhandling, 0809-103X*, 144.

Bayegan, E. and S. Tu, The helpful patient record system: problem oriented and knowledge based. *In Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2002.

Belle, A., R. Thiagarajan, S. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian (2015). Big data analytics in healthcare. *BioMed research international*, 2015.

Bennett, C. and T. Doub, Data mining and electronic health records: Selecting optimal clinical treatments in practice. *In Proceedings of the 6th International Conference on Data Mining*. 2010.

Berner, E. S., *Clinical decision support systems*volume233. Springer, 2007.

Black, A. D., J. Car, C. Pagliari, C. Anandan, K. Cresswell, T. Bokun, B. McKinstry, R. Procter, A. Majeed, and A. Sheikh (2011). The impact of ehealth on the quality and safety of health care: a systematic overview. *PLoS medicine*, 8(1), e1000387.

Boughorbel, Sabri, *et al.* (2017). Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PloS one*, 12(6), e0177678.

Braa, J., E. Monteiro, and S. Sahay (2004). Networks of action: sustainable health information systems across developing countries. *MIS quarterly*, 337–362.

Brown, P. J. and P. Sönksen (2000). Evaluation of the quality of information retrieval of clinical findings from a computerized patient database using a semantic terminological model. *Journal of the American Medical Informatics Association*, 7(4), 392–403.

Byrd, K., A. Mansurov, and O. Baysal, Mining Twitter data for influenza detection and surveillance. *In Proceedings of the International Workshop on Software Engineering in Healthcare Systems*. ACM, 2016.

Calvert *et al.* (2016*a*). Using ehr collected clinical variables to predict medical intensive care unit mortality. *Annals of Medicine and Surgery*, 11, 52–57.

Calvert, J., Q. Mao, A. J. Rogers, C. Barton, M. Jay, T. Desautels, H. Mohamadlou, J. Jan, and R. Das (2016*b*). A computational approach to mortality prediction of alcohol use disorder inpatients. *Computers in biology and medicine*, 75, 74–79.

Campbell, A. J., J. A. Cook, G. Adey, and B. H. Cuthbertson (2008). Predicting death and readmission after intensive care discharge. *British journal of anaesthesia*, 100(5), 656–662.

Celi, L. A., S. Galvin, G. Davidzon, J. Lee, D. Scott, and R. Mark (2012). A database-driven decision support system: customized mortality prediction. *Journal of personalized medicine*, 2(4), 138–148.

Chaudhry, B., J. Wang, S. Wu, M. Maglione, W. Mojica, E. Roth, S. C. Morton, and P. G. Shekelle (2006). Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. *Annals of internal medicine*, 144(10), 742–752.

Che, Z., S. Purushotham, K. Cho, D. Sontag, and Y. Liu (2018). Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1), 6085.

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.

Cheng, Y., F. Wang, P. Zhang, and J. Hu, Risk prediction with electronic health records: A deep learning approach. *In Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 2016.

Choi, E., M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, Doctor ai: Predicting clinical events via recurrent neural networks. *In Machine Learning for Healthcare Conference*. 2016.

Clermont, G., D. C. Angus, S. M. DiRusso, M. Griffin, and W. T. Linde-Zwirble (2001). Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. *Critical care medicine*, 29(2), 291–296.

Cocos, A., A. G. Fiks, and A. J. Masino (2017). Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. *Journal of the American Medical Informatics Association*, 24(4), 813–821.

Collins, S. A., K. Cato, D. Albers, K. Scott, *et al.* (2013). Relationship between nursing documentation and patients' mortality. *American Journal of Critical Care*, 22(4), 306–313.

Cooper, G. F. and E. Herskovits (1992). A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4), 309–347.

Coorevits, P., M. Sundgren, G. O. Klein, A. Bahr, B. Claerhout, C. Daniel, M. Dugas, D. Dupont, A. Schmidt, P. Singleton, *et al.* (2013). Electronic health records: new opportunities for clinical research. *Journal of internal medicine*, 274(6), 547–560.

Darcy, A. M., A. K. Louie, and L. W. Roberts (2016). Machine learning and the profession of medicine. *Jama*, 315(6), 551–552.

Davis, D. A., N. V. Chawla, N. Blumm, N. Christakis, and A.-L. Barabasi, Predicting individual disease risk based on medical history. *In Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 2008.

Demner-Fushman, D., W. W. Chapman, and C. J. McDonald (2009). What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5), 760–772.

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.

Domingos, P. and G. Hulten, Mining high-speed data streams. *In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2000.

Donnelly, K. (2006). Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121, 279.

Dorr, D., L. M. Bonner, A. N. Cohen, R. S. Shoai, R. Perrin, E. Chaney, and A. S. Young (2007). Informatics systems to promote improved care for chronic illness: a literature review. *Journal of the American Medical Informatics Association*, 14(2), 156–163.

Downs, J. S., W. B. de Bruin, and B. Fischhoff (2008). Parents' vaccination comprehension and decisions. *Vaccine*, 26(12), 1595–1607.

Dredze, M., D. A. Broniatowski, M. C. Smith, and K. M. Hilyard (2016). Understanding vaccine refusal: why we need social media now. *American journal of preventive medicine*, 50(4), 550–552.

Dubois, S., N. Romano, D. C. Kale, N. Shah, and K. Jung (2017). Learning effective representations from clinical notes. *arXiv preprint arXiv:1705.07025*.

Dybowski, R., V. Gant, P. Weller, and R. Chang (1996). Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. *The Lancet*, 347(9009), 1146–1150.

Efron, B. (1969). Student's t-test under symmetry conditions. *Journal of the American Statistical Association*, 64(328), 1278–1302.

Evans, R. S., S. L. Pestotnik, D. C. Classen, T. P. Clemmer, L. K. Weaver, J. F. Orme Jr, J. F. Lloyd, and J. P. Burke (1998). A computer-assisted management program for antibiotics and other antiinfective agents. *New England journal of medicine*, 338(4), 232–238.

Farkas, R. and G. Szarvas, Automatic construction of rule-based icd-9-cm coding systems. *In BMC bioinformatics*volume9. BioMed Central, 2008.

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 2013 IEEE International Conference on Healthcare Informatics*. IEEE, 2013.

Fialho, A. S., F. Cismondi, S. M. Vieira, S. R. Reti, J. M. Sousa, and S. N. Finkelstein (2012). Data mining using clinical physiology at discharge to predict icu readmissions. *Expert Systems with Applications*, 39(18), 13158–13165.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Fiszman, M., W. W. Chapman, D. Aronsky, R. S. Evans, and P. J. Haug (2000). Automatic detection of acute bacterial pneumonia from chest x-ray reports. *Journal of the American Medical Informatics Association*, 7(6), 593–604.

Fontelo, P., F. Liu, and M. Ackerman (2005). ask medline: a free-text, natural language query tool for medline/pubmed. *BMC Medical Informatics and Decision Making*, 5(1), 5.

Friedman, N., D. Geiger, and M. Goldszmidt (1997). Bayesian network classifiers. *Machine learning*, 29(2-3), 131–163.

Fu, X., K. Huang, N. D. Sidiropoulos, and W.-K. Ma (2018). Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications. *arXiv preprint arXiv:1803.01257*.

Gall, L. *et al.* (1984). A simplified acute physiology score for icu patients. *Critical care medicine*, 12(11), 975–977.

Gall, L. *et al.* (1993). A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, 270(24), 2957–2963.

Gentimis, T., A. Ala'J, A. Durante, K. Cook, and R. Steele, Predicting hospital length of stay using neural networks on mimic iii data. *In 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*. IEEE, 2017.

Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232), 1012–1014.

Ginsburg, G. S. and H. F. Willard (2009). Genomic and personalized medicine: foundations and applications. *Translational research*, 154(6), 277–287.

Göbel, G., S. Andreatta, J. Masser, and K. P. Pfeiffer (2001). A mesh based intelligent search intermediary for consumer health information systems. *International journal of medical informatics*, 64(2), 241–251.

Greenhalgh, T. (1997). How to read a paper. the medline database. *BMJ: British Medical Journal*, 315(7101), 180.

Grnarova, P., F. Schmidt, S. L. Hyland, and C. Eickhoff (2016). Neural document embeddings for intensive care patient mortality prediction. *arXiv preprint arXiv:1612.00467*.

Gunter, T. D. and N. P. Terry (2005). The emergence of national electronic health record architectures in the united states and australia: models, costs, and questions. *Journal of medical Internet research*, 7(1), e3.

Guyon *et al.* (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1), 389–422.

Harutyunyan, H., H. Khachatrian, D. C. Kale, and A. Galstyan (2017). Multitask learning and benchmarking with clinical time series data. *arXiv preprint arXiv:1703.07771*.

Haug, P. J., R. M. Gardner, R. S. Evans, B. H. Rocha, and R. A. Rocha, Clinical decision support at intermountain healthcare. *In Clinical decision support systems*. Springer, 2007, 159–189.

Herland, M., T. M. Khoshgoftaar, and R. Wald (2014). A review of data mining using big data in health informatics. *Journal of Big data*, 1(1), 2.

Himes, B. E., Y. Dai, I. S. Kohane, S. T. Weiss, and M. F. Ramoni (2009). Prediction of chronic obstructive pulmonary disease (copd) in asthma patients using electronic medical records. *Journal of the American Medical Informatics Association*, 16(3), 371–379.

Hinton, G. E., S. Osindero, and Y.-W. Teh (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7), 1527–1554.

Hira, Z. M. and D. F. Gillies (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics*, 2015.

Holt, D., F. Bouder, C. Elemuwa, G. Gaedicke, A. Khamesipour, B. Kisler, S. Kochhar, R. Kutalek, W. Maurer, P. Obermeier, and Others (2016). The importance of the patient voice in vaccination and vaccine safety—are we listening? *Clinical Microbiology and Infection*, 22, S146—-S153.

Huang, G., G.-B. Huang, S. Song, and K. You (2015). Trends in extreme learning machines: A review. *Neural Networks*, 61, 32–48.

Huang, G.-B., Q.-Y. Zhu, and C.-K. Siew, Extreme learning machine: a new learning scheme of feedforward neural networks. *In Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*volume2. IEEE, 2004.

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.

Huang, X., M. C. Smith, M. J. Paul, D. Ryzhkov, S. C. Quinn, D. A. Broniatowski, and M. Dredze, Examining patterns of influenza vaccination in social media. *In Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

Jain, H. and C. G. Huimin Zhao, David P. Klemer, A scheme for symptom based retrieval of electronic medical records. *In Proceedings of the Fourth Workshop on e-Business (WeB 2005)*. Las Vegas, NE, 2005.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406), 414–420.

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.

Jo, Y., L. Lee, and S. Palaskar (2017). Combining lstm and latent topic modeling for mortality prediction. *arXiv preprint arXiv:1709.02842*.

Johnson *et al.* (2013). A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy. *Critical care medicine*, 41(7), 1711–1718.

Johnson, A. E., T. J. Pollard, L. Shen, L.-w. H. Lehman, 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.

Johnson, A. E., D. J. Stone, L. A. Celi, and T. J. Pollard (2018). The mimic code repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25(1), 32–39.

Jonquet, C., N. H. Shah, and M. A. Musen (2009). The open biomedical annotator. *Summit on translational bioinformatics*, 2009, 56.

Joshi, A., X. Dai, S. Karimi, R. Sparks, C. Paris, and C. R. MacIntyre, Shot or not: Comparison of nlp approaches for vaccination behaviour detection. *In Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*. 2018.

Joulin, A., E. Grave, P. Bojanowski, and T. Mikolov (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Joulin, A., E. Grave, P. Bojanowski, and T. Mikolov, Bag of tricks for efficient text classification. *In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*volume2. 2017.

Kawamoto, K., C. A. Houlihan, E. A. Balas, and D. F. Lobach (2005). Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *Bmj*, 330(7494), 765.

Keeter, S., C. Kennedy, M. Dimock, J. Best, and P. Craighill (2006). Gauging the impact of growing nonresponse on estimates from a national RDD telephone survey. *International Journal of Public Opinion Quarterly*, 70(5), 759–779.

Kennebeck, S. S., N. Timm, M. K. Farrell, and S. A. Spooner (2012). Impact of electronic health record implementation on patient flow metrics in a pediatric emergency department. *Journal of the American Medical Informatics Association*, 19(3), 443–447.

Kennedy, J. and R. Eberhart, Particle swarm optimization (pso). *In Proc. IEEE International Conference on Neural Networks, Perth, Australia*. 1995.

Kim, S., W. Kim, and R. W. Park (2011). A comparison of intensive care unit mortality prediction models through the use of data mining techniques. *Healthcare informatics research*, 17(4), 232–243.

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.

Knaus, A. William, Draper, *et al.* (1985). Apache ii: a severity of disease classification system. *Critical care medicine*, 13(10), 818–829.

Knaus *et al.* (1981). Apache-a physiologically based classification system. *Critical care medicine*, 9(8), 591–597.

Knaus *et al.* (1991). The apache iii prognostic system: risk prediction of hospital mortality for critically iii hospitalized adults. *Chest*, 100(6), 1619–1636.

Koh, H. C., G. Tan, *et al.* (2011). Data mining applications in healthcare. *Journal of healthcare information management*, 19(2), 65.

Kohavi, R. and G. H. John (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2), 273–324.

Krumholz, H. M. (2014). Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Affairs*, 33(7), 1163–1170.

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

Larkey, L. S. and W. B. Croft (1995). Automatic assignment of icd9 codes to discharge summaries. Technical report, Technical report, University of Massachusetts at Amherst, Amherst, MA.

Le, Q. and T. Mikolov, Distributed representations of sentences and documents. *In International Conference on Machine Learning*. 2014.

Leape, L. L. (1994). Error in medicine. *Jama*, 272(23), 1851–1857.

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.

Leroy, G. and H. Chen (2001). Meeting medical terminology needs-the ontology-enhanced medical concept mapper. *IEEE Transactions on Information Technology in Biomedicine*, 5(4), 261–270.

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

Li, C., Y. Zhang, and X. Li, Ocvfdt: one-class very fast decision tree for one-class classification of data streams. *In Proceedings of the Third International Workshop on Knowledge Discovery from Sensor Data*. ACM, 2009.

Li, M., Z. Fei, M. Zeng, F. Wu, Y. Li, Y. Pan, and J. WanMulti-label classification of patient notes: case study on ICD code assignmentg (2018). Automated icd-9 coding via a deep learning approach. *IEEE/ACM transactions on computational biology and bioinformatics*.

Liang, Z., G. Zhang, J. X. Huang, and Q. V. Hu, Deep learning for healthcare decision making with emrs. *In Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*. IEEE, 2014.

Lindberg, D. A., B. L. Humphreys, and A. T. McCray (1993). The unified medical language system. *Yearbook of Medical Informatics*, 2(01), 41–51.

Lipton, Z. C., D. C. Kale, C. Elkan, and R. Wetzell (2015). Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*.

Liu, Z. and W. W. Chu, Knowledge-based query expansion to support scenario-specific retrieval of medical free text. *In Proceedings of the 2005 ACM symposium on Applied computing*. ACM, 2005.

Lovis, C. (2011). Clinical information systems: cornerstone for an efficient hospital management. *Studies in health technology and informatics*, 169, 992–5.

Lowe, H. J. and G. O. Barnett (1994). Understanding and using the medical subject headings (mesh) vocabulary to perform literature searches. *Jama*, 271(14), 1103–1108.

Malet, G., F. Munoz, R. Appleyard, and W. Hersh (1999). A model for enhancing internet medical document retrieval with "medical core metadata". *Journal of the American Medical Informatics Association*, 6(2), 163–172.

Mamlin, B. W., J. M. Overhage, W. Tierney, P. Dexter, and C. J. McDonald, Clinical decision support within the regenstrief medical record system. *In Clinical Decision Support Systems*. Springer, 2007, 190–214.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2), 442–451.

McDonald, C. J., J. M. Overhage, W. M. Tierney, P. R. Dexter, D. K. Martin, J. G. Suico, A. Zafar, G. Schadow, L. Blevins, T. Glazener, *et al.* (1999). The regenstrief medical record system: a quarter century experience. *International journal of medical informatics*, 54(3), 225–253.

McDonald, C. J. and W. M. Tierney (1988). Computer-stored medical records: their future role in medical practice. *Jama*, 259(23), 3433–3440.

McManus, K., E. K. Mallory, R. L. Goldfeder, W. A. Haynes, and J. D. Tatum (2015). Mining Twitter data to improve detection of schizophrenia. *AMIA Summits on Translational Science Proceedings*, 2015, 122.

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

Michelson, J. D., J. S. Pariseau, and W. C. Paganelli (2014). Assessing surgical site infection risk factors using electronic medical records and text mining. *American journal of infection control*, 42(3), 333–336.

Mikolov, Chen, *et al.* (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miotto, R., L. Li, B. A. Kidd, and J. T. Dudley (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6.

Moja, L., K. H. Kwag, T. Lytras, L. Bertizzolo, L. Brandt, V. Pecoraro, G. Rigon, A. Vaona, F. Ruggiero, M. Mangia, *et al.* (2014). Effectiveness of computerized decision support systems linked to electronic health records: a systematic review and meta-analysis. *American journal of public health*, 104(12), e12–e22.

Monge, A. and C. Elkan (1997). An efficient domain-independent algorithm for detecting approximately duplicate database records.

Moreno *et al.* (2005). Saps 3 - from evaluation of the patient to evaluation of the intensive care unit. *Intensive care medicine*, 31(10), 1345–1355.

Mullenbach, J., S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein (2018). Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*.

Musen, M. A., B. Middleton, and R. A. Greenes, Clinical decision-support systems. *In Biomedical informatics*. Springer, 2014, 643–674.

Nachtigall, I., S. Tafelski, M. Deja, E. Halle, M. Grebe, A. Tamarkin, A. Rothbart, A. Uhrig, E. Meyer, L. Musial-Bright, *et al.* (2014). Long-term effect of computer-assisted decision support for antibiotic treatment in critically ill patients: a prospective 'before/after' cohort study. *BMJ open*, 4(12), e005370.

Nédellec, C., R. Bossy, J.-D. Kim, J.-J. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum, Overview of bionlp shared task 2013. *In Proceedings of the BioNLP Shared Task 2013 Workshop*. 2013.

Nguyen, P., T. Tran, N. Wickramasinghe, and S. Venkatesh (2017). Deepr: A convolutional net for medical records. *IEEE journal of biomedical and health informatics*, 21(1), 22–30.

Nikfarjam, A., A. Sarker, K. O'Connor, R. Ginn, and G. Gonzalez (2015). Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3), 671–681.

Nimgaonkar, A., D. R. Karnad, S. Sudarshan, L. Ohno-Machado, and I. Kohane (2004). Prediction of mortality in an indian intensive care unit. *Intensive care medicine*, 30(2), 248–253.

Orabi, A. H., P. Buddhitha, M. H. Orabi, and D. Inkpen, Deep learning for depression detection of twitter users. *In Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. 2018.

Organization, W. H., *The world health report 2000: health systems: improving performance*. World Health Organization, 2000.

Pakhomov, S. V., J. D. Buntrock, and C. G. Chute (2006). 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.

Parker, A. M., R. Vardavas, C. S. Marcum, and C. A. Gidengil (2013). Conscious consideration of herd immunity in influenza vaccination decisions. *American journal of preventive medicine*, 45(1), 118–121.

Patel, P. and B. Grant (1999). Application of mortality prediction systems to individual intensive care units. *Intensive care medicine*, 25(9), 977–982.

Pennington, J., R. Socher, and C. Manning, Glove: Global vectors for word representation. *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.

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.

Pirracchio *et al.* (2015). Mortality prediction in intensive care units with the super icu learner algorithm (sicula): a population-based study. *The Lancet Respiratory Medicine*, 3(1), 42–52.

Plovnick, R. M. and Q. T. Zeng (2004). Reformulation of consumer health queries with professional terminology: a pilot study. *Journal of medical Internet research*, 6(3), e27.

Pollak, V. E. (1983). Computerization of the medical record: Use in care of patients with endstage renal disease. *Kidney international*, 24(4), 464–473.

Purushotham, S., C. Meng, Z. Che, and Y. Liu (2018). Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics*.

Ravì, D., C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang (2017). Deep learning for health informatics. *IEEE journal of biomedical and health informatics*, 21(1), 4–21.

Reddy, K. S., V. Patel, P. Jha, V. K. Paul, A. S. Kumar, L. Dandona, L. I. G. for Universal Healthcare, *et al.* (2011). Towards achievement of universal health care in india by 2020: a call to action. *The Lancet*, 377(9767), 760–768.

Röder, M., A. Both, and A. Hinneburg, Exploring the space of topic coherence measures. *In Proceedings of the eighth ACM international conference on Web search and data mining*. ACM, 2015.

Rodriguez, J. J., L. I. Kuncheva, and C. J. Alonso (2006). Rotation forest: A new classifier ensemble method. *IEEE transactions on pattern analysis and machine intelligence*, 28(10), 1619–1630.

Saeed, M., C. Lieu, G. Raber, and R. G. Mark, Mimic ii: a massive temporal icu patient database to support research in intelligent patient monitoring. *In Computers in Cardiology, 2002*. IEEE, 2002.

Saeed, M., M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark (2011). Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5), 952.

Saito, T. and M. Rehmsmeier (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3), e0118432.

Salakhutdinov, R. and G. Hinton, Deep boltzmann machines. *In Artificial Intelligence and Statistics*. 2009.

Salton, G. and C. Buckley (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5), 513–523.

Sánchez-Maroño, N., A. Alonso-Betanzos, and M. Tombilla-Sanromán (2007). Filter methods for feature selection–a comparative study. *Intelligent Data Engineering and Automated Learning-IDEAL 2007*, 178–187.

Santillana, M., A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie, and J. S. Brownstein (2015). Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS computational biology*, 11(10), e1004513.

Sarker, A., R. Ginn, A. Nikfarjam, K. O'Connor, K. Smith, S. Jayaraman, T. Upadhaya, and G. Gonzalez (2015). Utilizing social media data for pharmacovigilance: a review. *Journal of biomedical informatics*, 54, 202–212.

Sarker, A. and G. Gonzalez (2015). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics*, 53, 196–207.

Shen, G., J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, and W. Zhu, Depression detection via harvesting social media: A multimodal dictionary learning solution. *In IJCAI*. 2017.

Signorini, A., A. M. Segre, and P. M. Polgreen (2011). The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, 6(5), e19467.

Silva, I., G. Moody, D. J. Scott, L. A. Celi, and R. G. Mark, Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. *In Computing in Cardiology (CinC), 2012*. IEEE, 2012.

Simpao, A. F., L. M. Ahumada, J. A. Gálvez, and M. A. Rehman (2014). A review of analytics and clinical informatics in health care. *Journal of medical systems*, 38(4), 45.

Snomed, C. (2011). Systematized nomenclature of medicine-clinical terms. *International Health Terminology Standards Development Organisation.*

Sood, S. P., S. N. Nwabueze, V. W. Mbarika, N. Prakash, S. Chatterjee, P. Ray, and S. Mishra, Electronic medical records: A review comparing the challenges in developed and developing countries. *In Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*. IEEE, 2008.

Sood, S. P. and M. Tech (2004). Implementing telemedicine technology: lessons from india. *World Hospitals and Health Services*, 40(3), 29–32.

Stead, W. W. and W. E. Hammond (1983). Computerized medical records. *Journal of medical systems*, 7(3), 213–220.

Stevens, K., P. Kegelmeyer, D. Andrzejewski, and D. Buttler, Exploring topic coherence over many models and many topics. *In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012.

Suarez, H., X. Hao, and I. Chang, Searching for information on the internet using the umls and medical world search. *In Proceedings of the AMIA Annual Fall Symposium*. American Medical Informatics Association, 1997.

Sun, W., A. Rumshisky, and O. Uzuner (2013). Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5), 806–813.

Suykens, J. A. and J. Vandewalle (1999). Least squares support vector machine classifiers. *Neural processing letters*, 9(3), 293–300.

Tang, P. C., M. A. Jaworski, C. A. Fellencer, N. Kreider, M. LaRosa, and W. Marquardt, Clinician information activities in diverse ambulatory care practices. *In Proceedings of the AMIA Annual Fall Symposium*. American Medical Informatics Association, 1996.

Trivedi, M. H., J. Kern, A. Marcee, B. Grannemann, B. Kleiber, T. Bettinger, K. Altshuler, and A. McClelland (2002). Development and implementation of computerized clinical guidelines: barriers and solutions. *Methods of information in medicine*, 41(05), 435–442.

Vincent *et al.* (1996). The sofa score to describe organ dysfunction/failure. *Intensive care medicine*, 22(7), 707–710.

Wakamiya, S., Y. Kawai, and E. Aramaki (2018). Twitter-based influenza detection after flu peak via tweets with indirect information: text mining study. *JMIR public health and surveillance*, 4(3), e65.

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. J., B. Middleton, L. A. Prosser, C. G. Bardon, C. D. Spurr, P. J. Carchidi, A. F. Kittler, R. C. Goldszer, D. G. Fairchild, A. J. Sussman, *et al.* (2003). A cost-benefit analysis of electronic medical records in primary care. *The American journal of medicine*, 114(5), 397–403.

Wang, Y., N. Afzal, S. Fu, L. Wang, F. Shen, M. Rastegar-Mojarad, and H. Liu (2018). Medsts: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 1–16.

Weibel, S., J. Kunze, C. Lagoze, and M. Wolf (1998). Dublin core metadata for resource discovery. Technical report.

Weiskopf, N. G. and C. Weng (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1), 144–151.

White, C. (2007). Health Care Spending Growth: How Different Is The United States From The Rest Of The OECD? *Health Affairs*, 26(1), 154–161. PMID: 17211024.

Williams, R. J. and D. Zipser (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2), 270–280.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2), 241–259.

Wong, L. and J. Young (1998). A comparison of icu mortality prediction using the apache ii scoring system and artificial neural network.

Wu, J., J. Roy, and W. F. Stewart (2010). Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches. *Medical care*, 48(6), S106–S113.

Wu, Y., J. C. Denny, S. Trent Rosenbloom, R. A. Miller, *et al.* (2016). A long journey to short abbreviations: developing an open-source framework for clinical abbreviation recognition and disambiguation (card). *Journal of the American Medical Informatics Association*, 24, 79–86.

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.

Xu, H., S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, and J. C. Denny (2010). Medex: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1), 19–24.

Yang, J. and V. Honavar, Feature subset selection using a genetic algorithm. *In Feature extraction, construction and selection*. Springer, 1998, 117–136.

Yuan, Q., E. O. Nsoesie, B. Lv, G. Peng, R. Chunara, and J. S. Brownstein (2013). Monitoring influenza epidemics in china with search query from baidu. *PloS one*, 8(5), e64323.

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)*. IEEE, 2019.

Zell, E. R., T. M. Ezzati-Rice, M. P. Battaglia, and R. A. Wright (2000). National immunization survey: the methodology of a vaccination surveillance system. *Public health reports*, 115(1), 65.

Zeng, M., M. Li, Z. Fei, Y. Yu, Y. Pan, and J. Wang (2019). Automatic icd-9 coding via deep transfer learning. *Neurocomputing*, 324, 43–50.

Zhang, Y., S. Fong, J. Fiaidhi, and S. Mohammed (2012). Real-time clinical decision support system with data stream mining. *BioMed Research International*, 2012.

Zimmerman *et al.* (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.

Bio-data

Name:	Gokul S Krishnan
Current Address:	Research Scholar, Department of Information Technology, NITK Surathkal Mangaluru, Karnataka India - 575025.
Permanent Address:	Manakkattu Madom, Manipuzha, Podiyadi PO, Thiruvalla, Kerala India - 689110.
Email:	gsk1692@gmail.com
Mobile No:	+91 99448 46833
Qualification:	Ph.D. in Information Technology Department of Information Technology National Institute of Technology Karnataka, Surathkal Mangalore, India. M.Tech in Computer Science & Engineering VIT University, Vellore, Tamil Nadu, India. B.Tech in Computer Science & Engineering Cochin University of Science and Technology Kerala, India.
Research Area:	Healthcare Analytics, Natural Language Processing, Machine Learning, Web Semantics