Predicting Post Procedural Complications using MIMIC-III

Master of Science in Computer Science

Faisal Maqbool

MSDS17027

Session: 2017 – 2019

DEPARTMENT OF COMPUTER SCIENCE

INFORMATION TECHNOLOGY UNIVERSITY

LAHORE, PAKISTAN

Predicting Post Procedural Complications Using MIMIC-III

A thesis submitted in partial fulfillment of the requirements for the
Degree of Master of Science in
Data Science

Faisal Maqbool

Dr. Saeed Ul Hassan

Committee Member Name
Committee Member Name

Declaration

This thesis is a presentation of my original research work. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature, and acknowledgement of collaborative research and discussions. I also declare that this work is the result of my own investigations, except where identified by references and free from plagiarism of the work of others.

Signature: ………….

Faisal Maqbool

Date: ………………..

The undersigned hereby certify that they have read and recommend the thesis entitled "…………………………..” by ………………….. For the degree of Master of Science in Data Science.

Supervisor Name (ITU), Thesis Advisor

Committee Member Name (ITU), Thesis Committee Member

Committee Member Name, Thesis Committee Member

Chairperson Name, Chairperson of the Department

Acknowledgment

I thank my advisor Dr. Saeed Ul Hassan for being the supportive advisor. His precious insights, guidance and support throughout the research journey has not only helped me become a better student but also examine data for research and contribute to previously done research. Though he is very busy, he always made me feel like a priority. He took time to review my findings, answer questions, and give thoughtful feedback for which I am extremely grateful. This thesis would not have been possible without them.

I thank Dr. Mohsin Ali for his help, advice and support throughout my work and for agreeing to serve on my thesis committee.

Moreover, I thank my family especially my parents and my friends for me letting me follow my dreams and for their encouragement which has been the vital part of the research.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

MIMIC: Medical Information Mart for Intensive Care III

PHI: Protected Health Information

ICU: Intensive Care Unit

ETL: Extraction, Transformation, Load

ICD: International Classification of Diseases

HIPPA: Health Insurance Portability and Accountability Act

BMI: Body Mass Index

i2b2: Information for Integrating Biology and the Bedside

ACDF: Anterior cervical discectomy and fusion

SMOTE: Synthetic Minority Over-Sampling Technique

ADASYN: Adaptive Synthetic Sampling

PFS: Progression Free Survival

RSF: Random Survival Forests

PD-1: Anti-Programmed Death-1

ABSTRACT

Analyzing the health of patients using Electronic Health Records (EHR) and take precise decisions is an important problem in healthcare research. To predict outcomes and take such decisions patients admitted in ***ICU*** *requires specific features of medical records: worth, capacity, access and dimensionality. A substantial amount of data needs to be stored in proper infrastructure and comprehensive analysis which can aid doctors, clinicians, medical experts and families is required. In this Thesis, an investigation of medical data is carried out. We propose an ETL approach to extract features and train machine learning models to predict the post-procedural complications seeking if those complications can lead to further complications or mortality. To derive insights for that, we used well-known clinical dataset named Medical Information Mart for Intensive Care III (****MIMIC-III****) with two types of data sampling techniques:* ***SMOTE*** *and* ***ADASYN****. For both techniques our results showed that* ***Random Forrest*** *outperforms other linear models with an accuracy of* ***>80%*** *and* ***AUROC*** *of* ***0.83.***

CHAPTER: 1

1.1 INTRODUCTION

With advancements in digital technologies in medical sciences, different techniques making it possible to explore big data precisely and predict tasks of clinicians, labs and doctors using patient records of diagnoses or procedures. Intensive care unit **(ICU)** is where severely ill patients are admitted and requires accurate predictors that can help doctors and take respective measures.

Diagnostic and medical technologies have evolved rapidly and doctors have to take complex decisions. In the book **[23]** it is being said that unfortunately, majority of clinical decisions are not based on evidence. According to **2012** Institute of Medical Committee Report, a small percentage of decision were evidence based. This problem ranges to not have proper clinical guidelines.

In this thesis we investigate the different methodologies for extracting, transforming and **ETL** techniques **[15], [16]**, which obtain data from original source to perform informative analysis and features extraction to aid model to predict post-procedural (diagnoses & procedure) complications of critically ill patients and investigating those complications if those can lead to mortality of patient or not. The methods use demographics, data from different hospital system, lab events, diagnoses, notes and other engineered information regarding each patient. The database used for the study is Medical Information Mart for Intensive Care **MIMIC-III [1]** which is de-identified data and abide by protecting health information **(PHI).**

Other researches about MIMIC-III data is also presented to motivate our problem, establish understanding of dataset, key findings and recommendations for future investigations. The question of predicting post-procedural complications from data science perspective and critical health perspective is not only important for doctors, administrators but also for the patient as well. For administrators this would help managing patients and required resources. Avoiding predicted complications can further be avoided if such information is known during the stay of patient at ICU.

On MIMIC-III researchers widely contributed to support the cause. For any severely sick patient obesity is not so important feature to consider. **[9]** Presented an investigation of **BMI** against obesity. They conjecture that illness scores would give results differently if **BMI** was considered. Their setting included documented weight and height. The assessment was based on **184402** from **184** different ICU's across United States and assessment showed that **4%** were classified as under-weight, **30%** as normal and same for over-weight. Apart from these **28%** were mentioned as obese and **8%** as morbidly obese. To further explore the obesity paradox **[10]** presented a study recently where they evaluated the relation between **BMI** with survival and ensured the gender interactions. The design of their study was retrospective with total cohort samples **N = 139** from June 2014 to September 2016**.** The main results were Survival **(OS)** and **(PFS)**. Analysis was performed using **Random Forests**. Their findings showed that the paradox of obesity exists under overweight in the real world. The result showed that sarcopenia body mass composition can be used as predictors.

A study from *Journal of intensive care medicine* **[11]** having the design of retrospective study and setting of single tertiary academic medical center examine the impact of overstays of patients and discharge delays on in-hospital morbidity and mortality. For the interventions, for all patients, from the bed request in ward to discharge time was calculated. Created bins for greater and less than **24** hours discharge delays. To find out the relationship between delays and ICU outcome they used multivariate linear regression and logistic regression. There was no association between long delays and subsequent mortality.

Many other researcher have contributed in prediction of other critical factors related to medicine and the health of patient. A study from 2018 **[12]** with a design of retrospective study examined that severity scores may also lead to misclassification of critically ill obese patients using the weight of patient and the obesity score having documented results of their diagnoses prior to hospital admissions. They further compared the laboratory results between the normal patients having less obesity scores and high obesity score. They showed that an unconventionality is found within the scores of creatinine, white blood cells and urea nitrogen in blood and compared between them. These results could also be improved and used for the mortality and also in predicting other important clinical tasks.

1.2 ICD, HIPPA AND COMPLICATIONS

The International Classification of Diseases **(ICD)** **[17]** is the foundation which identifies and evaluates the statistics of diseases globally. Creates standards for all the diagnoses, diseases which further can be used for research purposes. Under revision of ICD9 codes, the code **996** defines complications particular to certain specified procedures and diagnoses. Most complications are caused due to cardiac, vascular or other used devices and some of them relates to reaction caused due to a procedure performed. In our work, we are focusing on such complications and investigating if those can lead to the mortality of patient. **HIPAA** (Health Insurance Portability and Accountability Act of 1996) **[37]** provides privacy and protection of health records for safeguarding and protecting (PHI) medical information. To protect health information MIMIC-III provided anonymized data, still we need to make sure we are following HIPPA compliance rules so that our research does not conflict with any of the standards defined.

1.3OBJECTIVE

A predictor that can evaluate complications using the **EHR** can help hospital administrators, experts, families and doctors to take decisions that are required for the safety of the patient. For any hospital if the management already knows the precise expected stay of patient they can plan better to use their resources to provide the best possible care. For doctors predictors can evidence, statistics and labels which can be used for taking valuable decisions, for families who will pay for the expenses against the patient can better plan their billing using such predictors. Moreover, extended diagnoses and patients stay at the hospital due to complications caused during the stay at hospital or after is associated with not only the health of patient, cost, increased number of deaths but also increased number of readmissions. Each of these parameters defines the hospital performance. So, our objective is to produce insights that can complement these parameters ranging from cost to patient health. To be specific, our objective is to extract features against not so rare diagnoses and procedures with defined **ETL** process, balancing of our dataset using distributions and sampling techniques and predict **996** (complication or no complication) (In general). Furthermore, to derive insights that can determine whether those complications can lead to death or not.

1.4 APPROACH FOLLOWED

As open sourced MIMIC-III data is available for research purposes for free. Our main focus was to use tools and technologies in a way that can complement already done research. Hence, we used **Python**, **R** and **PostgreSQL** for ETL process, features selection and modeling our features to extract results which I will explain in respective chapters in details.

We have used the basic Data Science (mining) model as full cycle **[19]** of development in our research.

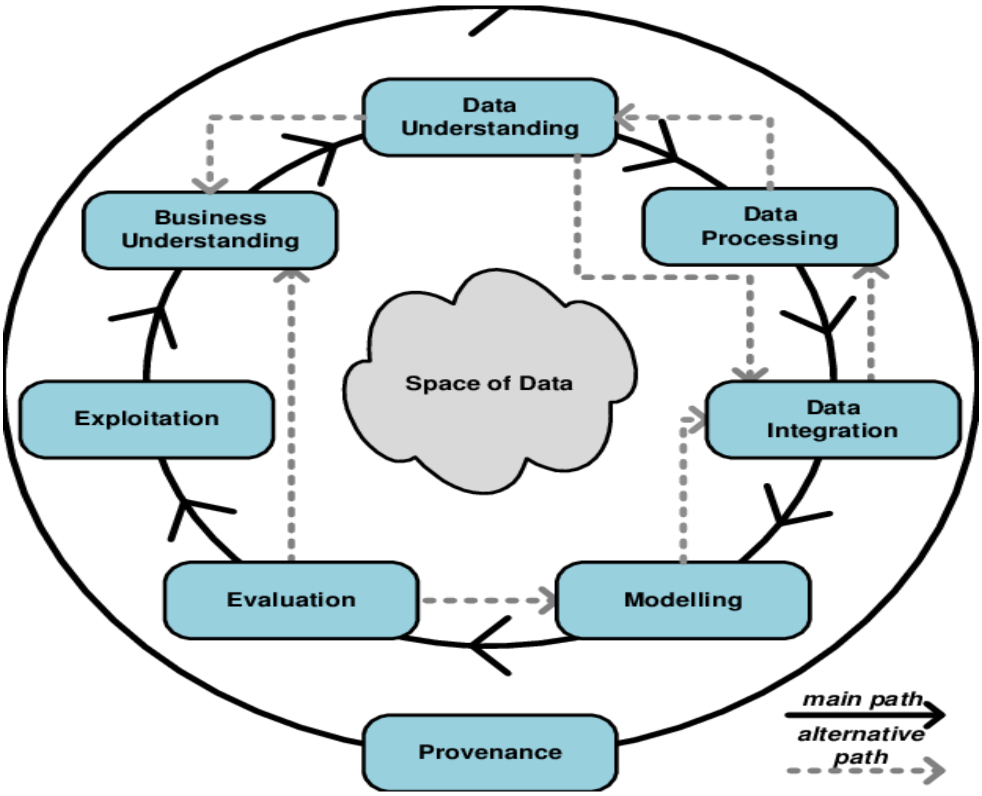

Figure 1: Methodology

The model divides several phases which are executed in iteration to get to goal which we can see in **Figure 1.** In the first phase data extraction and understanding is built, then comes the phase of data processing, handling the information as per required business transformations. The next step is the integration of all the transformations into one source to further use by next step of feature space extraction and model building. The later part of process includes models evaluations and optimization.

1.4 CONTRIBUTIONS

The prediction of complications is subjective to type of complication, data availability, resources used by management to record EHR. Against each type of diagnoses, subclasses of complications are already defined and regularly updated by ICD. Domain experts have contributed to individual complications defined by ICD, but we generalized all the complications and implemented models to predict a general class of whether a patient will have complication or not. Moreover, one must consider the existence of unstructured data available in MIMIC-III which lead to extensive search for variables without access to actual system deployed at hospitals. Because of that a considerable part was dedicated to the **ETL** and a novel approach was implemented to extract meaningful features which can lead to reliable prediction models. We used **PostgreSQL** for the SQL implementation and **Python** as it provide large number of libraries and models to implement. To be specific, we implemented an ETL approach, generalized all the complications to one class, implemented models to predict that class. (If 1, patient will have complication else no complication)

1.5 DESCRIPTION OF THE CONTENT

The thesis is organized as follows: In chapter **1** we introduce the problem, explains purpose of our work, methodologies followed and detail of contents. In Chapter **2**, we reviewed the research and studies already done by researchers and contributors. In Chapter **3,** we explained data source MIMIC-III in detail. In Chapter **4**, we explained the ETL process to obtain data and process it effectively for posterior use. In Chapter **5**, we explain the preprocessing and analysis of data which is used of feature engineering and model selection process. In Chapter **6**, results, Chapter **7,** conclusion and future work is been discussed.

CHAPTER: 2 RELATED WORK

Health and medicine are one of the key sectors that requires use of new technologies to produce new possibilities and cause a greater impact in the society. Some of the recent research studies mentioned above have been conducted on **MIMIC-III** dataset. Numerous researches have also been conducted on the subject matter and overall on the usage of MIMIC-III for creating new possibilities of research and scientific areas.
Following are some of the researches that helped us motivate our problem and contribute by applying information retrieval and data science techniques.

Before we take the recent contributions related to predicting complications and other clinical tasks into account, let us first explain the fact that structured data may not always contain accurate and required information. As we are working with a huge medical data mart, the volume, diversity, dimensions and ETL matters. To identify the patient cohort from whole mart by searching the structured tables for diagnoses, procedures, chart events, demographics and other features is a burdensome task keeping the scoring systems **[24]** and authenticate the results of predictors with cost and time complexity **[25]**. The studies with **EHR** databases that have both structural and unstructured data have not only help researchers in identifying the new possibilities but also helped them to recognize patients having higher risks **[26].**

Several studies **[27, 28, 29, 30, 31, 32]** upon structured data related to **ICD** (International Classification of Diseases) have helped us carry our research as they extracted structured information related to patient, diagnoses and procedure and shown to have good precision and **AUROC**. Working with large data sources, retrieval of information is time taking and difficult when multiple data sources are involved in the ETL process **[36]**. From IEEE *International conference of Healthcare Informatics* a study **[42]** explored the use of **SVM** on assigning the ICD-9 code using the clinical notes.

Specific to **MIMIC-III,** from learning about the relationship between healthcare processes **[35]**, the study **[33]** which made us realize that laboratory research is based on information retrieval irrespective of the design of study whether retrospective or prospective. All the management, doctors and clinicians are not involved in utilities of labs and how to proceed with **MIMIC** data and its structures. Another study **[34]** combined description of notes from the EHR and clinical trials using the state of the art NLP models to evaluate find relationships within trials and notes. Which, is our future plan to incorporate the textual features combining with complications to increase the performance of predicting complications accurately.

LOS prediction for the patient can provide valuable information to the management of the hospital but also for patient's health. **[2]** Explored the NN usage for predicting the LOS prediction within time range of **(<5)** days or **(>5)** days after the patient left the Intensive care unit. They used a subset of **MIMIC-III** and written all their models in **R** and **PostgreSQL** using a supercomputer provided by the Florida Polytechnic University. Their model predictions achieved **80%** accuracy and outperformed any other linear models previously used of predicting the length of stay. They used a neural net of **2** layers with **5** and **3** nodes at each level. In their final dataset, they had total **31018** data points, from which, they randomly chose subsets of **15000** points and trained the NN on the **Florida Polytechnic University** supercomputer.

Another study in **2019 [3]** mapped ICD-9 codes using the clinical notes from physicians, doctors and other staff against each patient automatically using deep learning. Their main focus was to evaluate the performance of deep learning for mapping the ICD-9 codes. Their pipeline had extraction of data from **Notes and Diagnoses**, removing special characters and stop words from each note. They extracted non-sequential features and applied **tfidf and word2vec**. They baseline models **LR, RF, FRNN** on non-sequential features and on sequential features from word sequences + embedding matrix they applied **Conv1D**, **LSTM** and **GRU**. Overall they applied multiple experiments and showed that deep learning outperforms linear models to map ICD-9 codes with **0.6957** $\mathbf{F_1}$ score and accuracy of **0.8967**.

Research published in *BMC Medical Informatics and Decision Making* **[4]** proposed a new approach by combining rule-based features and DL model with prior knowledge for effective disease classification. They evaluated their method on **i2b2** obesity challenge and demonstrated that their model outperform other method used for disease classification. They based their method on **Solt's Systems [43]** (The Perl's[1] Implementation of Solt provided by author) for first identifying the major triggers terms and phrases from notes, then predicting the classes with rare terms and phrases and then trained convolutional neural network. Another mortality prediction case study published in *Machine Learning for HealthCare Conference* **[5]** demonstrated large variability in studies

[1] https://github.com/yao8839836/obesity/tree/master/perl_classifier

of mortality prediction. They reviewed the performance of the related studies based on MIMIC-III and compared it to **Gradient Boosting** and LR ran on extracted data.

Apart from above mentioned researches a lot of other researchers contributed in predicting the complications. A study **[6]** explored the model performance in predicting after operation complications following anterior cervical discectomy and fusion **(ACDF)**. They applied Logistic Regression, Random Forrest and ANN to achieve their results. **20,879** patients and **ANN** outperformed **LR (p < 0.05).** Similarly, **[7]** used machine learning to derive and validate the hospital readmission. Their model reached (area under the receiver operating curve, **0.76**). **[8]** A study with design of retrospective study, on predicting the real time complications used **RNN** against mortality, renal failure, and postoperative bleeding and operation revision. 47 559 samples of ICU were included. Their deep learning models produced **PPV 0·90** and **sensitivity 0·85** for mortality, **0·87** and **0·94** for renal failure, and **0·84** and **0·74** for bleeding.

Another study using MIMIC-III **[13]** implemented a deep rule based fuzzy systems for predicting the accurate in-hospital mortality. Their main contribution to the system was proposing a system which can handle multiple categorical data.

CHAPTER: 3 MIMIC-III CRITICAL CARE DATABASE

This chapter explains the structure, context and development researchers have done on data source MIMIC-III.

3.1 ABOUT MIMIC-III

Over the past decade, much have been written about the field of data science regarding the explosion of big data. In health care, every decision made for a critical patient requires precision by clinicians and doctors. To carry out research to aid clinicians and doctors to make better, reliable and quick decision using research and applications. This demands privacy existence of data and wide-ranging analysis.

Researcher and developer community is contributing in the **MIMIC-III (Medical Information Mart for Intensive Care III)** database, openly accessible at https://mimic.physionet.org/ This database created by the Lab of Computational Physiology of The Massachusetts Institute for Technology (MIT) to fulfil the purpose of providing different techniques of data analytics and sciences. MIMIC-III is a large, freely-available relational database comprising de-identified **[38]** health related data associated with over forty thousand patients who stayed in Intensive Care Units at Beth Israel Deaconess Medical Center (Boston, Massachusetts). The data spans June 2001 October2012.

MIMIC-III is a comprehensive collection of de-identified data from **53,423** distinct critical care hospital admissions from **38,597** distinct adult patients. The data has been compiled into 26 tables which contain, for example, an average of **4579** noted values and 380 lab charted measurements for each hospital admission as well as a total of **3.8** gigabytes of unstructured textual data from various healthcare provider notes and analyses. In addition to de-identifying patient data, MIT requires training in the protection of patient data for anyone requesting access to the MIMIC dataset. After completing the prescribed training, data can be downloaded as 26 comma separated files (csv) files representing the **26** tables in the MIMIC-III database. Sample SQL code can be acquired from GitHub (https://github.com/MIT-LCP/mimic-code) for establishing relationships between the tables. Additionally, there is a published data dictionary which can be found at https://mimic.physionet.org/mimictables/admissions/. There is inconsistency in the usage of “unique” attributes and definition of primary keys between the sample SQL code and the published data dictionary. For example, every table has an attribute called “ROW_ID”, and the sample SQL code consistently declares this attribute as “unique” and/or as a “primary key” for every table despite the fact that tables like the “PATIENTS” table have a unique identifier (SUBJECT_ID) that is intended to be the primary key and serve as foreign key in child relations that refer to the “PATIENTS” table.

After downloading and analyzing the MIMIC source tables, implementation occurs in 5 additional steps:

- Create tables with attribute rules (data types) and identify the primary key for each table. 1. Load records from csv files into each table.
- Declare the indexes for each table.
- Define foreign keys in each table and establish table relationships.
- Implement user interface (with appropriately granted permissions) for the database.

This database includes patients hospital level demographic data, ICU level laboratory test results, Labs and resources measurements, procedures, diagnoses, clinical notes, staff, notes and services. The class distribution of data shown in table **1** (Taken from https://www.nature.com/articles/sdata201635/tables/3)

Table 1: Class Distribution of data for MIMIC-III Dataset

Class of data	Description
Billing	Coded data recorded primarily for billing and administrative purposes. Includes Current Procedural Terminology (CPT) codes, Diagnosis-Related Group (DRG) codes, and International Classification of Diseases (ICD) codes.
Descriptive	Demographic detail, admission and discharge times, and dates of death.
Dictionary	Look-up tables for cross referencing concept identifiers (for example, International Classification of Diseases (ICD) codes) with associated labels.
Interventions	Procedures such as dialysis, imaging studies, and placement of lines.
Laboratory	Blood chemistry, hematology, urine analysis, and microbiology test results.
Medications	Administration records of intravenous medications and medication orders.
Notes	Free text notes such as provider progress notes and hospital discharge summaries.
Physiologic	Nurse-verified vital signs, approximately hourly (e.g., heart rate, blood pressure, respiratory rate).
Reports	Free text reports of electrocardiogram and imaging studies.

3.2 MIMIC-III TABLES

MIMIC-III consists of 26 files with relational mapping from which we created following tables in PostgreSQL. The tables, dimensions and their summary is explained below:

Table 2: MIMIC-III Tables Summary (Taken from: https://mimic.physionet.org/gettingstarted/access/)

File Name	Dimension	Summary
ADMISSIONS	(58976, 19)	The ADMISSIONS table gives information regarding a patient's admission to the hospital.
CALLOUT	(34499, 24)	The CALLOUT table provides information about ICU discharge planning
CAREGIVERS	(7567, 4)	This table provides information regarding care givers. For example, it would define if a caregiver is a research nurse (RN), medical doctor (MD), and so on.
CHARTEVENTS	(330712483, 15)	CHARTEVENTS contains all the charted data available for a patient.
CPTEVENTS	(573146, 12)	The CPTEVENTS table contains a list of which current procedural terminology codes were billed for which patients. This can be useful for determining if certain procedures have been performed (e.g. ventilation).
D_CPT	(134, 9)	This table gives some high level information regarding current procedural terminology (CPT) codes. Unfortunately, detailed information for individual codes is unavailable.
D_ICD_DIAGNOSES	(14567, 4)	This table defines International Classification of Diseases Version 9 (ICD-9) codes for diagnoses. These codes are assigned at the end of the patient's stay and are used by the hospital to bill for care provided.
D_ICD_PROCEDURES	(3882, 4)	This table defines International Classification of Diseases Version 9 (ICD-9) codes for procedures. These codes are assigned at the end of the patient's stay and are used by the hospital to bill for care provided.
D_ITEMS	(12487, 10)	The D_ITEMS table defines ITEMID, which represents measurements in the database.
D_LABITEMS	(753, 6)	D_LABITEMS contains definitions for all ITEMID associated with lab measurements in the MIMIC database.

DATETIMEEVENTS	(4485937, 14)	DATETIMEEVENTS contains all date measurements about a patient in the ICU.
DIAGNOSES_ICD	(651047, 5)	This table defines ICD-9 codes for diagnoses. The ICD codes are generated for billing purposes at the end of the hospital stay.
DRGCODES	(125557, 8)	This table defines HCFA-DRG and APR-DRG codes which provide information regarding Diagnosis-Related Group recorded primarily for billing and administrative purposes.
ICUSTAYS	(61532, 12)	This table gives information regarding ICU hospital stays.
INPUTEVENTS_CV	(17527935, 22)	This table contains data of fluid input events (serums, intravenous medication, insulin, etc.) regarding Carevue database source associated to ICU episodes.
INPUTEVENTS_MV	(3618991, 31)	This table contains input data for patients.
LABEVENTS	(27854055, 9)	Contains all laboratory measurements for a given patient, including outpatient data.
MICROBIOLOGYEVENTS	(631726, 16)	Contains microbiology information, including tests performed and sensitivities.

NOTEEVENTS	(2083180, 9)	This table contains all notes for patients took in a manual way by their caregivers.
OUTPUTEVENTS	(4349218, 13)	This table contains output data for patients.
PATIENTS	(46520, 8)	This table contains hospitalization-independent data for all patients such as, gender, date of birth, etc.
PRESCRIPTIONS	(4156450, 19)	This table contains medication related order entries, i.e. prescriptions.
PROCEDUREEVENTS_MV	(258066, 25)	This table contains procedures for patients
PROCEDURES_ICD	(17527935, 22)	Contains ICD procedures for patients, most notably ICD-9 procedures. The ICD codes are generated for billing purposes at the end of the hospital stay and are recorded for all patient hospitalizations.
SERVICES	(73343, 6)	The SERVICES table describes the service that a patient was admitted under. This service admission can be elective or caused due to a number of reasons, including bed shortage.

3.3 MIMIC-III CONSTRUCTION

MIMIC-III was constructed based upon hospital level, patient level, ICU level & used systems level. Furthermore it includes billing, notes and reports as shown in below figure

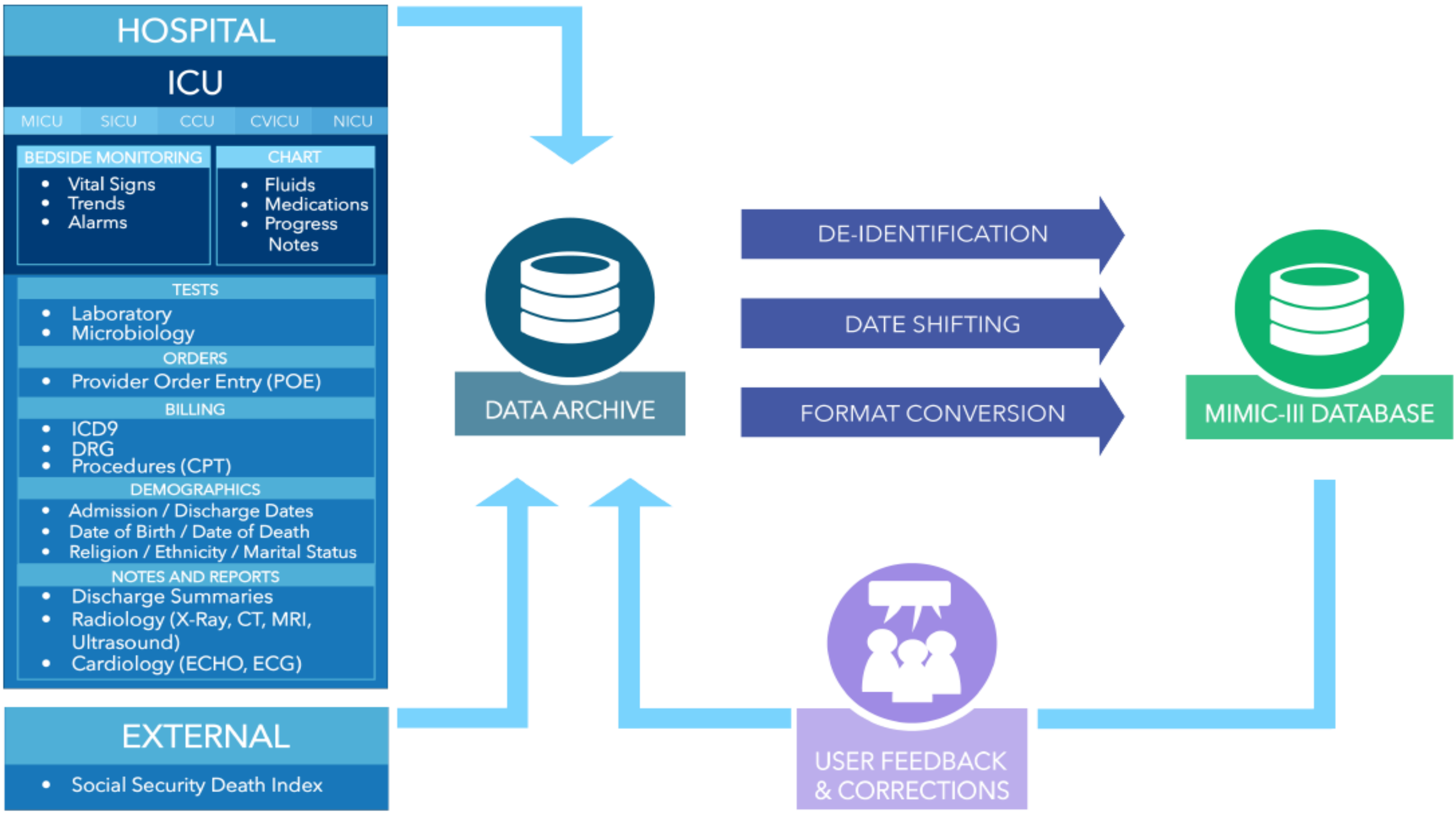

Figure 2: MIMIC-III Construction Model. Reprinted from (http://pi.cs.oswego.edu/~jmiles3/mimic/Miles-MIMIC-Project_report.pdf)

3.4 MIMIC-III DERIVED CONCEPTS

The active researchers have contributed to already given data with additional scripts to generate new concepts and insights at MIMIC code repository which includes views and tables as well. They also encourage other researchers to contribute to derived insights which helps to distinct between the original data and derived data and one can use as per the problem they are solving and contribute as well.

Following are the major concepts that are being used frequently by researchers.

Table 3 : Derived Concepts

Class of Data	Summary
Comorbidity	These scripts derive binary flags indicating the presence of various comorbidities using billing codes (ICD-9) assigned to the patient at hospital discharge.
First day	The first day subfolder contains scripts used to calculate various clinical concepts on the first day of a patient's admission to the ICU, such as the highest blood pressure, lowest temperature, etc. This folder contains many useful scripts which can be adapted to capture data outside the first day.
Sepsis	Definitions of sepsis, a common cause of mortality for intensive care unit patients.
Severity Scores	Severity of illness scores which summarize the acuity of a patient's illness on admission to the intensive care unit (usually in the first 24 hours).
Durations	Start and stop times for administration of various treatments or durations of various phenomena, including: medical agents which have a vasoactive effect on a patient's circulatory system, continuous renal replacement therapy (CRRT), and mechanical ventilation.
Organ Failure	This script derives binary flags for major organ failures

[2]The tables are linked by identifiers which usually have the suffix "**ID**". For example **HADM_ID** refers to a unique hospital admission and **SUBJECT_ID** refers to a unique patient. One exception is **ROW_ID**, which is simply a row identifier unique to that table.

Tables pre-fixed with "**D_**" are dictionaries and provide definitions for identifiers. For example, every row of **OUTPUTEVENTS** is associated with a single **ITEMID** which represents the concept measured, but it does not contain the actual name of the drug. By joining **OUTPUTEVENTS** and **D_ITEMS** on **ITEMID**, it is possible to identify what concept a given **ITEMID** represents

CHAPTER: 4 ETL AND DATASET BUILDING

In this section, we introduce the ETL followed by us to derive certain insights which lead us to conclusion of stated problem. The section is divided into extraction, transformation and loading sections to reach to our features.

4.1 TECHNICAL PROCESS

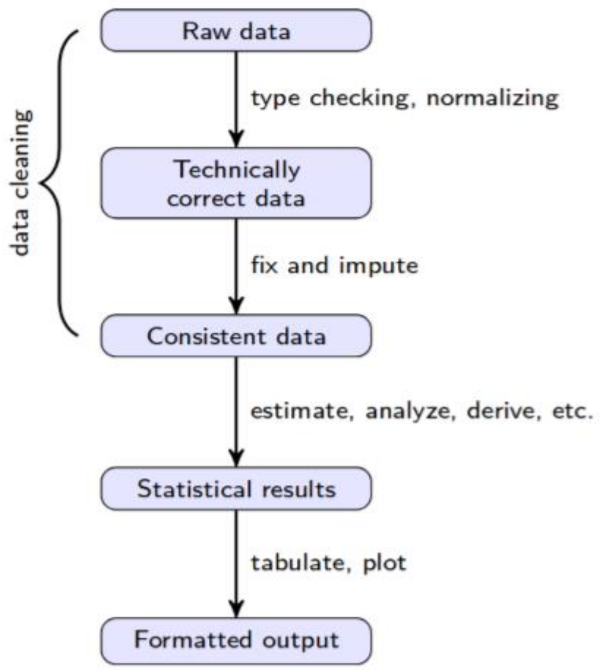

Figure 3: Technical Chain of Steps

[2] https://mimic.physionet.org/gettingstarted/overview/

Given the size of data mart and the volume of raw data, we devoted most of time to extraction and transformation of data.

In the first step prior to requesting access to **MIMIC**, you will need to complete the **CITI** "Data or Specimens Only Research" course by registering yourself on CITI program. After getting data access we are provided links to the **26** comma separated file containing patient, hospital and ICU related data[3].

Following are key steps covered in technical process to engineer features:

- Tables Creation
- Relationship Mapping (Indexes and Keys)
- Materialized views from already given tables
- Trim down values for ICD-9 Codes
- Filter rows with subject id lookup and pass it to items lookup for certain diagnoses and procedure
- ICD-9 Codes for class complications which is 996
- Making it to binary classes with 1 and all other classes to 0
- Extracting derived features from chart events and lab events with batch processing
- Consolidate all other features with derived concept
- Format all features, fill out invalid fields and normalize features for model training

Code available at: https://github.com/faisalmaqbool94/Thesis-Bioinformatics-MIMICIII-

4.2 RELATIONAL MAPPING, BATCH PROCESSING AND PREPROCESSING

All 26 files are relationally mapped with each other[4]. After getting these files we created a database of all those file and created respective tables. To improve the performance indexes and constraints were added. Some of them are very huge and requires pre and post processing. The smaller files were dealt with PSQL but on the other hand, the big files caused a problem for not only creating table but also of processing those files in RAM. To handle such problems with huge files, we implemented Python script for asyn batch processing using Pandas[5] which is an open sourced library to manipulate structured data and very highly efficient because of its reliable data frame objects along with transformation tools available with it.

[3] https://mimic.physionet.org/gettingstarted/access/

[4] https://mit-lcp.github.io/mimic-schema-spy/

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

As the data sources and research work is now publicly available. Researchers have contributed in the form of code, new concepts, and optimization of previously written script and in many other ways. Similarly we have used and created features set containing top diagnoses and procedures performed on ICU patients.

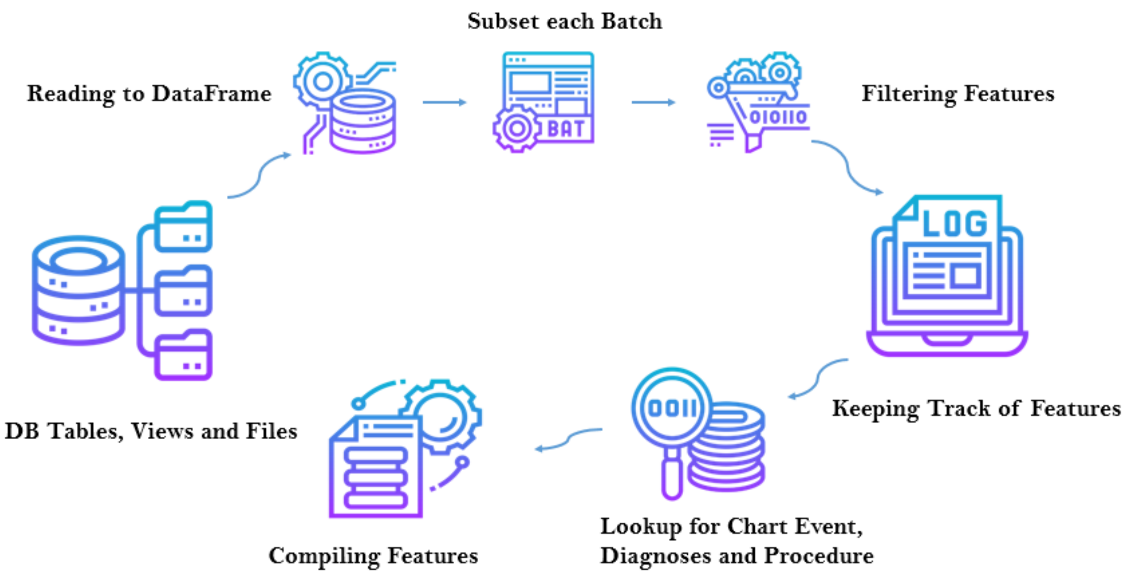

Figure 4: Batch Processing for Huge Files Using Python

To complete ETL process, PostgreSQL and Python played important role. Multiple SQL scripts were written for creation of tables, indexes, materialized views and derived tables. All of which are presented on a public repository **[19]**.
https://github.com/faisalmaqbool94/Thesis-Bioinformatics-MIMICIII-
Extraction of major chart events and lab events against each patient involved filtering of specific patients, lookup against particular diagnoses and procedures.

4.3 DIAGNOSES

From **14328** unique diagnoses that were available in MIMIC-III, we selected those diagnoses which were more common with a threshold of happening more than **30.** The number is randomly selected and experimented with. We can change number with other experiments to see if the models can outperform earlier built models. The reason behind choosing this particular number was that using these number of groups, if we use any aggregated function to fill out missing information, there will be a low probability of creating features containing near to zero variance.

4.4 PROCEDURES

From **3882** unique procedures registered in MIMIC-III, we selected more common ones. Same as we did for diagnoses. For procedures there were only few fields that required any preprocessing.

In the **Figure 4** the subset creations and filtering involved lookups where we created separate files for segregating the subjects which are only appearing in above criteria of procedures and diagnoses.

Item	Counts
Diagnoses	14328
Procedure	3882

Table 4: Diagnoses and Procedures Count

4.5 FEATURE SPACE

Now that we have explored and discussed about the dataset. Now we discuss the features that are valuable to models. Once these features are identified we have to define the processes of our models and reach our goals.

From potential features which can impact the predictor are included, physicians from **PIMS** hospital Islamabad and **Islamabad Diagnostic Centre**, selected and helped us engineer our features they know are important from their medical knowledge, experience or intuition. Following are our selected variables:

a) **General**: Insurance, Martial status, Hospital Expire Flag, Length of Stay, Calculated Bicarbonate, TotalCo2, Chloride, Free Calcium, Glucose, Hematocrit, Hemoglobin, Lactate, Oxygen, Oxygen Saturation, PCO2, PH, Potassium, Sodium, Temperature, Calcium Total, Centromere, Creatinine, Globulin, Blood Glucose, Blood Lipase, Blood Magnesium, Blood Potassium, Blood Sodium, Platelets Counts, Red Blood Cells, White Blood Cells, Lymphocytes

b) **Engineered Concepts (Derived from Table 3)**: Hypothyroidism, Renal Failure, Liver Disease, Peptic Ulcer, Aids, Lymphoma, Metastatic Cancer, Solid Tumor, Rheumatoid Arthritis, Coagulopathy, Obesity, Weight Loss, Fluid Electrolyte, Blood Loss Anemia, Alcohol Abuse, Drug Abuse, Psychoses, Depression, Congestive Heart Failure, Cardiac Arrhythmias, Valvular Disease, Pulmonary Circulation, Peripheral Vascular, Hypertension, Paralysis, Other Neurological, Chronic Pulmonary, Diabetes Uncomplicated, Diabetes Complicated

4.6 FEATURES ENGINEERING

Once all the featured got extracted for certain subjects and against hospital admissions. We had to distinguish between categorical and numerical variables.
For categorical features One Hot Encoding technique was performed.

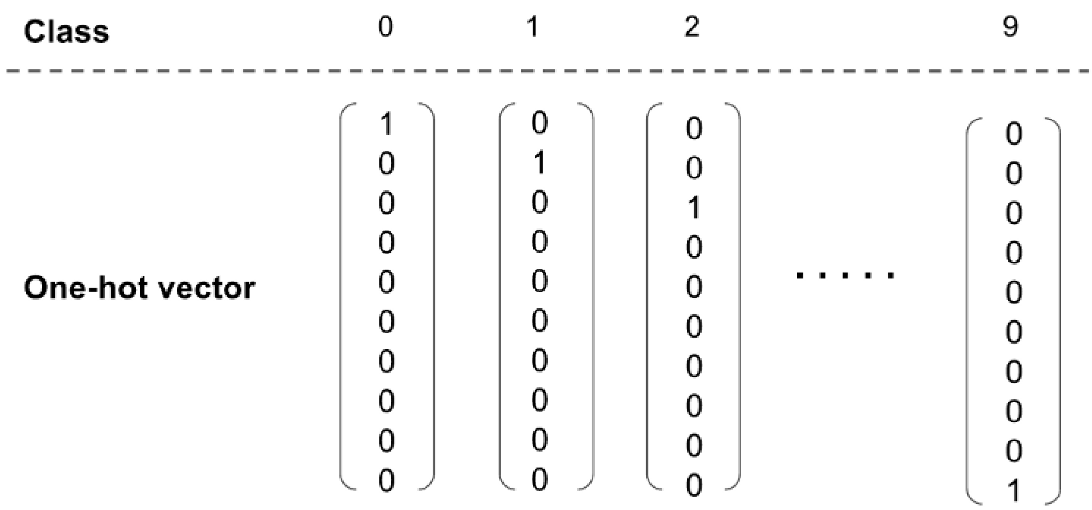

Figure 5 : One Hot Encoding for Categorical Features

Missing values got treated by the average of all the diagnoses and same goes for procedures. Average value is taken because we have extracted diagnoses and procedures which are commonly occurring and average of each group was taken.

Our target variable **'Icd9_Code'** got converted into binary variable and mapped to (0, 1) where **0** indicates the non-complication and **1** indicates the occurrence of complication. All the complications are further subdivided into thousands of categories but we were just interested in the main class of complication which is indicated by code **996.**

Class	Count
Complication	2754
No Complication	30491

Table 5: Class Count

4.7 SAMPLING

As we have considered a very sensitive topic which requires a lot of domain knowledge and predicting a complication requires precision. Although we collected and engineered our dataset for targeting our goals but as the **table 5** shows that we clearly have class imbalance problem. To tackle this problem we applied over sampling and down sampling techniques which are explained below.

4.7.1 SMOTE: Synthetic Minority Over-Sampling Technique

SMOTE **[14]** proposed by Chawla et al., is an oversampling method. This method interpolate the minority class neighbors to construct new minority class samples randomly. The method can be described as follows. First, for each minority class sample $\boldsymbol{x}$ one gets it's k-nearest neighbors from other minority class samples. Second, one chooses one minority class sample $\widetilde{\boldsymbol{x}}$ among $\boldsymbol{k}$ neighbors. Finally, one generates the synthetic sample $\boldsymbol{x_{new}}$ by interpolating between $\boldsymbol{x}$ and $\widetilde{\boldsymbol{x}}$ as follows:

$$\boldsymbol{x_{new} = x + rand(0,1) * (\widetilde{x} - x)}$$

Where (0, 1) refers to random number between (0,1]. In view of geometry, SMOTE can be regarded as interpolating between two minority class samples. The decision space for the minority class is expanded that allows the classifier to have a higher prediction on unknown minority class samples.

The SMOTE algorithm is simple and effective while generating synthetic samples, and the overfitting problem is avoided. It expands the decision space for the minority class.

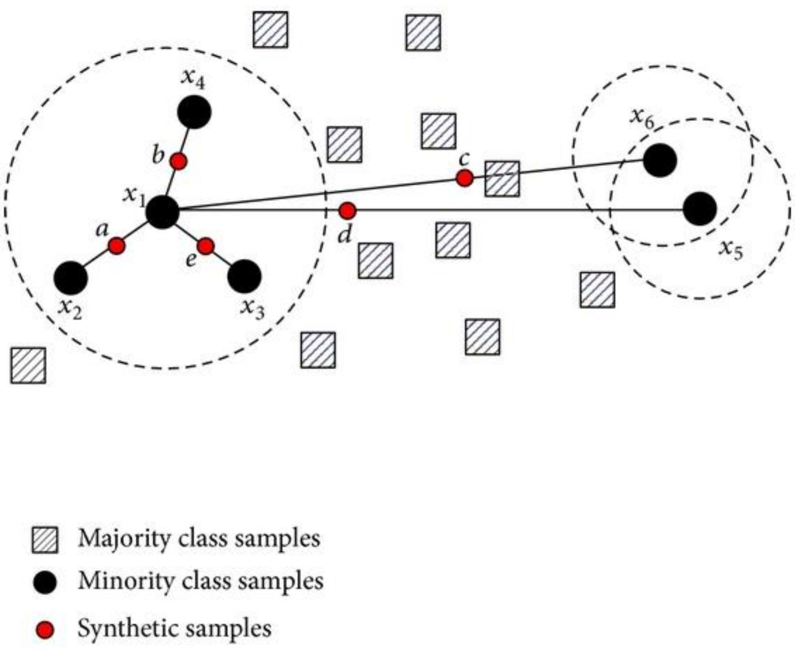

Figure 6: SMOTE (Sampling) (Reprinted From: https://www.hindawi.com/journals/mpe/2013/694809/)

4.7.2 ADASYN: Adaptive Synthetic Sampling

ADASYN [44] is another oversampling method which interpolates new minority class samples by first calculating the number of synthetic samples of minority class and for each minority sample find the k nearest neighbors by calculating the Euclidean distance.

Input:

D_r with $\boldsymbol{m}$ samples with$\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$, $\boldsymbol{i = 1}$ *to* $\boldsymbol{m}$, where $\mathbf{x_i}$ is an n-dimensional vector in feature space and $\mathbf{y_i}$ is the corresponding class. Let $\boldsymbol{m}_r$ and $\boldsymbol{m}_x$ be the number of minority and majority class samples respectively, such that $\boldsymbol{m}_r \leq \boldsymbol{m}_x$ and $\boldsymbol{m}_r + \boldsymbol{m}_x = \boldsymbol{m}$

Procedure [44]:

I. Calculate the *Degree of Imbalance*, $\mathbf{d} = \mathbf{m}_r / \mathbf{m}_x$

II. If $d < d_x$ (where d_x is the preset threshold for maximum tolerated imbalance) then:

a. Calculate the number of synthetic samples to be generated from the minority class: $\boldsymbol{G} = (\boldsymbol{m}_{\mathrm{x}} - \boldsymbol{m}_{\mathrm{r}}) \times \boldsymbol{\beta},$ β is the balance level of the synthetic samples generated. β = 1 means there is a total balance between two classes.

b. For each $x_{\mathrm{i}} \in$ minority samples, find the k-nearest neighbors based on Euclidean distance and calculate the ratio r_{i}, $\boldsymbol{r}_{\mathrm{i}} = \frac{\boldsymbol{\Delta}_{\mathrm{i}}}{\boldsymbol{K}}$

c. Normalize $\boldsymbol{r}_{\mathrm{x}} \leftarrow \boldsymbol{r}_{\mathrm{i}} / \sum \boldsymbol{r}_{\mathrm{i}},$ such that r_x is now a density distribution.

d. Calculation of synthetic sample generated for each minority data point $g_{\mathrm{i}} = r_{\mathrm{x}} \times G,$ where G is the total number of synthetic data examples that need to be generated for the minority class as defined in aforementioned Equation.

e. For each minority class data example $\mathbf{x}_\mathbf{i}$, generate $\mathbf{g}_\mathbf{i}$ synthetic data examples according to the following steps:

- Do the Loop from **1 to $\mathbf{g}_\mathbf{i}$**

  (a) Randomly choose one minority data example, $\mathbf{x}_\mathbf{u}$, from the K nearest neighbors for data $\mathbf{x}_\mathbf{i}$.

  (b) Generate the synthetic data example: $\boldsymbol{s}_{\mathrm{i}} = \boldsymbol{x}_{\mathrm{i}} + (\boldsymbol{x}_{\mathrm{u}} - \boldsymbol{x}_{\mathrm{i}}) \times \boldsymbol{\lambda}$ where $(\boldsymbol{x}_{\mathrm{u}} - \boldsymbol{x}_{\mathrm{i}})$ is the difference vector in n-dimensional spaces, and λ is a random number: $\lambda \in [0, 1]$. **END Loop**

The major difference between SMOTE and ADASYN is the *difference in the generation of synthetic sample points* for minority data points. In ADASYN, we consider a **density distribution $\mathbf{r}_\mathbf{x}$** which thereby decides the number of synthetic samples to be generated for a particular point, whereas in SMOTE, there is a uniform weight for all minority points.

CHAPTER 5: METHODOLOGY AND EVALUATIONS

5.1 ABOUT BINARY CLASSIFICATION

As we have defined our problem as binary class, either the patient would have complication or not, and defined a target feature which will be used to identify that. Depending upon the number of samples in each class, we dealt with balanced and unbalanced labeled dataset. For unbalanced dataset we applied above mentioned two over-sampling techniques. To notion of metric performance is called accuracy which we used to validate our models defined in below section. Accuracy is defined as:

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y_i})$$

Here $\hat{y_i}$ is the predicted class label for the iTh iteration using f(a defined function), the total number of samples are defined by n and the i index represents the each sample. $I(y_i = \hat{y_i})$ Is the indicator variable that equals one if classified correctly and zero if the result is negative. One can also define the accuracy metric as follow:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP and TN are positive and negative instances correctly classified by the model and FP are negative instances classified positive by model, similarly FN are positive instances classified negative by the model.

5.2 SELECTED MODELS

For our prediction task, we applied following models and are explained below:

- Logistic Regression (LR)
- Linear SVN (Linear Support Vector Classification)
- Random Forrest (Decision Tree)
- ANN (Artificial Neural Network)

To apply all these models, **Python** sklearn, imblearn, matlplotlib, Pandas and Numpy libraries were used.

5.2.1 Logistic Regression

Logistic regression is a predictor which provide the output by mapping the dependent variables and uses a logistic function. The range of the model is as follow $[0 - 1]$ and being continuous and differentiable.
Property the **sigmoid** function, which we denote $P(x)$ as follows:

$$P(x) = \frac{1}{1 + e^{-x}}$$

Figure below illustrate the sigmoid function:

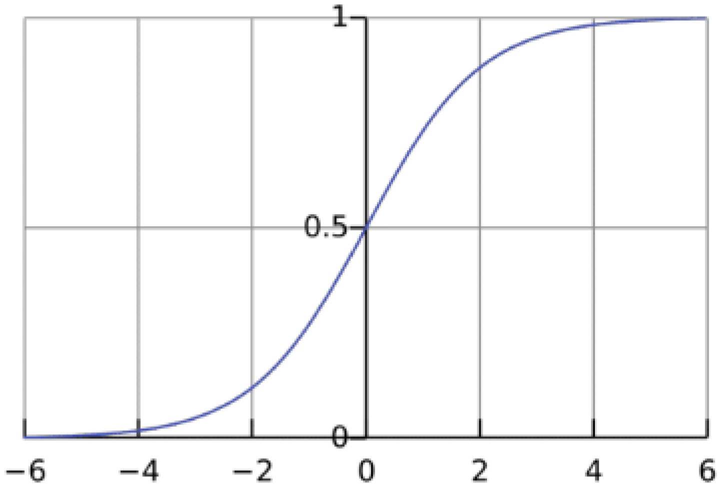

Figure 7: Sigmoid Function: Taken from [45]

5.2.2 Random Forrest

Random forests **[39]** are predictors which on given labeled data create random forests for reaching towards the target of each instances and map all the features to those trees. The actual output is then generated by averaging the output of all the forests.

Tin Kam Ho initially presented the concept of Random Forrests in the year of **1995**. In 2001 **[40]** where Breimen extend the concept by extending already built algorithm and work of Amit **[41]**.

The random forest creates many decision trees from which it calculates the actual and optimal path to find the expected label. It average the performance from all the forests, the major concepts due to which model was named Random Forrest are as follow:

1. While building new forests, sample the data in random fashion
2. While splitting the parent and leaf nodes, creation of subsets of multiple features

5.2.2 Linear SVC

A linear SVC support vector classifier (SVM) a supervised predictor which given a labeled dataset finds the "best fit" hyperplane which divides and segregate the data into different classes.

It has the cost function like logistic regression defined below:

$$h_\theta(x) = \begin{cases} 1 & \text{if } \theta^T x >= 0 \\ 0 & \text{otherwise} \end{cases}$$

A comparison is given with the logistic cost function below:

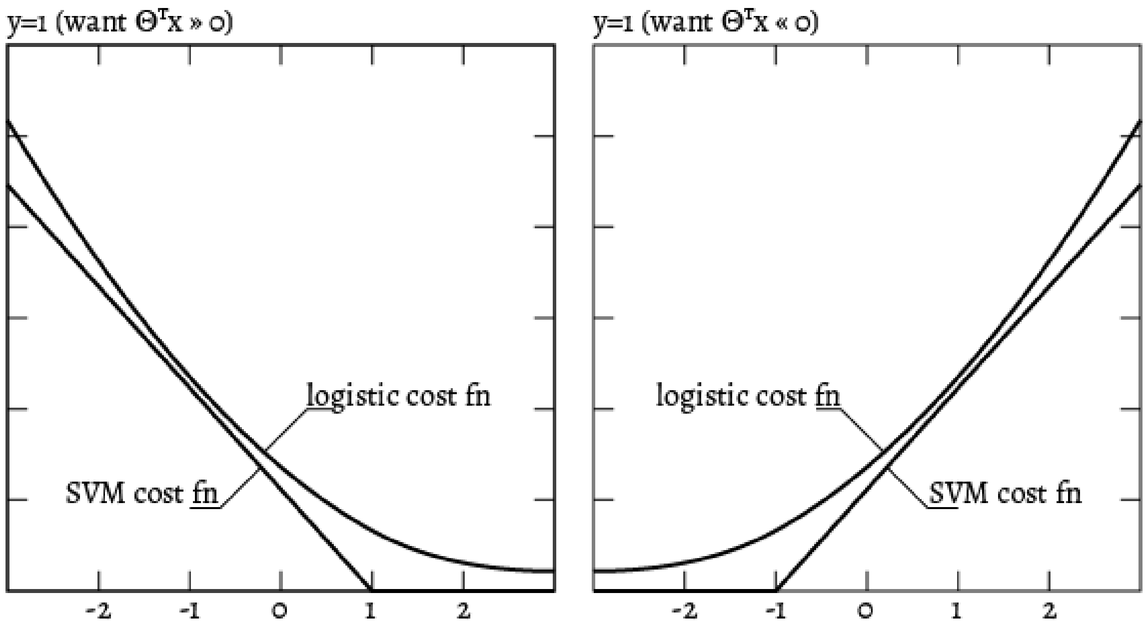

Figure 8: Logistic Vs Linear SVM

5.2.3 Artificial Neural Network (ANN)

Artificial Neural Networks (ANN) or just Neural Networks (NN) are objectively the main tool in machine learning appropriate for handling large data sets. Neural Networks are a combination of “neurons” and “synapses” consisting of three main layers: input, hidden layers and an output layer. These three parts create what is called an n-layer Neural Network. Each layer is connected with a set of weights and a bias value to the next one. Also, in each hidden layer a choice of activation function must be defined, but if that is fixed in the beginning of the analysis, only the weights and bias values will affect the output, thus training a Neural Network is a process of fine tuning the weights and bias values to get a better accuracy through a complicated Stochastic Gradient Descent method.

Every iteration in training the neural network contains two main steps: **Backpropagation** and **Feedforward**. Feedforward is the process of calculating the predicted output and Backpropagation is the process of updating weights and biases after a specified number of iterations.

Below are the major components of a perceptron:

1) Inputs: All the features available in the training dataset become the input for a perceptron. Also, an extra value called a bias value is fed as one of the inputs.

2) Weights: The value of weights are initiated randomly (most of the times zero for all) and these values are updated accordingly by reviewing the training error

3) Weighted sum: This is the summation of all the values obtained after multiplying each weight with its associated input value and adding the bias at the end.

4) Activation function: These functions convert an input signal of a node to an output signal. Some of the commonly used activation functions are **tanh**, **sigmoid**, **relu [21]**, **softmax**, exponential and linear. The flexibility of these activation functions is one of the reasons neural networks perform better than traditional multilinear models.

5) Output: The weighted sum is passed into the activation function and becomes the input 13 value of the next layer. As a first step, the weight vector is initialized. All the features available in the training dataset are fed as input to the perceptron. These input features are then multiplied with the corresponding weights and the values are summed up including the bias value. The new computed value is fed to the activation function in order to get the predicted output. If the predicted value doesn’t match with the actual value, the error is calculated and the weights are updated in order to reduce the error for the next iteration.

This process is repeated until the error is reduced to a prescribed level, or if a certain number of steps is achieved.

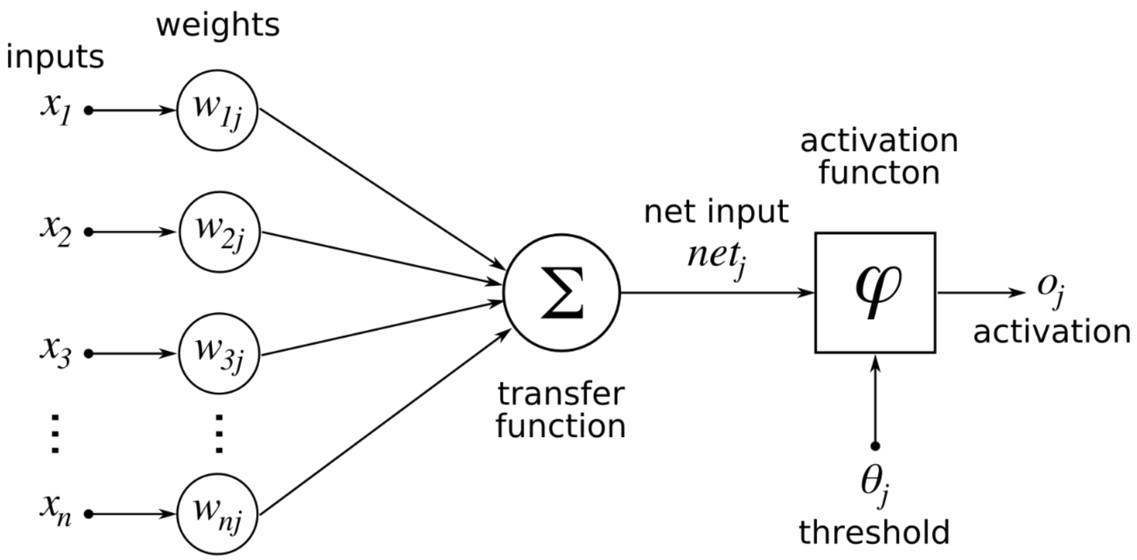

Figure 9: Artificial Neural Network (Perceptron Basic Model)

Our ANN model in compiled form given below shows number of nodes and hidden layers.

```
Model: "sequential_8"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_6 (Dense)              (None, 2048)              135168
_________________________________________________________________
dropout_5 (Dropout)          (None, 2048)              0
_________________________________________________________________
dense_7 (Dense)              (None, 2048)              4196352
_________________________________________________________________
dropout_6 (Dropout)          (None, 2048)              0
_________________________________________________________________
dense_8 (Dense)              (None, 2048)              4196352
_________________________________________________________________
dropout_7 (Dropout)          (None, 2048)              0
_________________________________________________________________
dense_9 (Dense)              (None, 2048)              4196352
_________________________________________________________________
dropout_8 (Dropout)          (None, 2048)              0
_________________________________________________________________
dense_10 (Dense)             (None, 1)                 2049
=================================================================
Total params: 12,726,273
Trainable params: 12,726,273
Non-trainable params: 0
```

Training of our ANN model are shown in below images:

Figure 10: Training and Validation Error

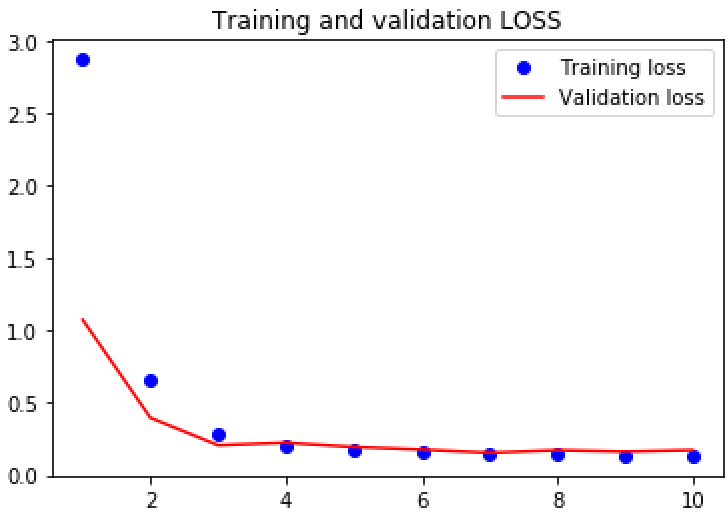

Figure 11: Training and Validation Loss

CHAPTER 6: EXPERIMENTAL RESULTS

6.1 Statistics

In this section, we present some of the basic numerical descriptors of our dataset and the results of our initial analysis. Following are some statistics from our data.

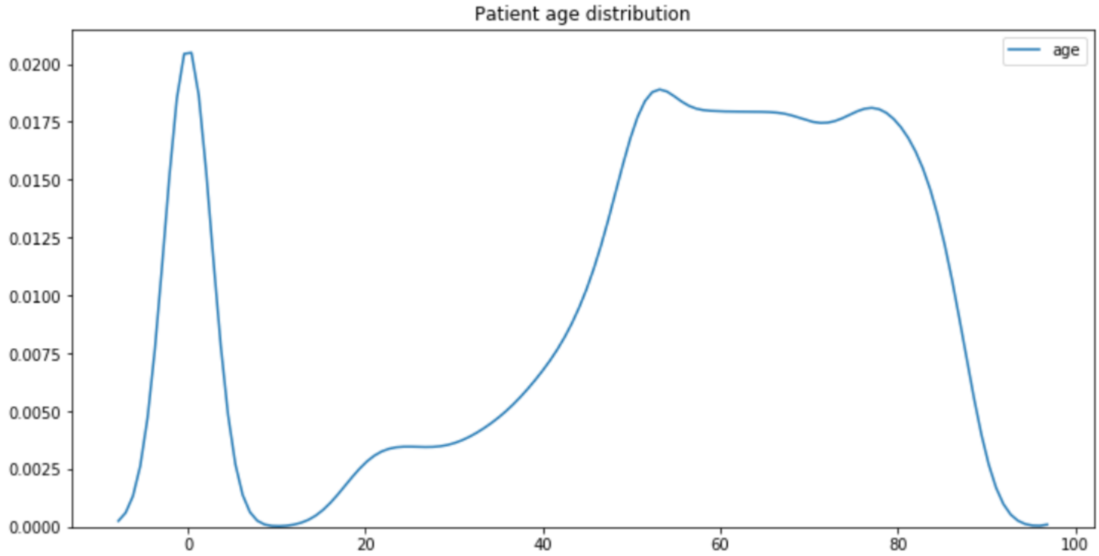

Figure 12: Patient Age Distribution

The above figure shows the patient age distribution which clearly impacted every feature of patient.
Similarly, we plotted the length of stay distribution for patients. The figure below shows that:

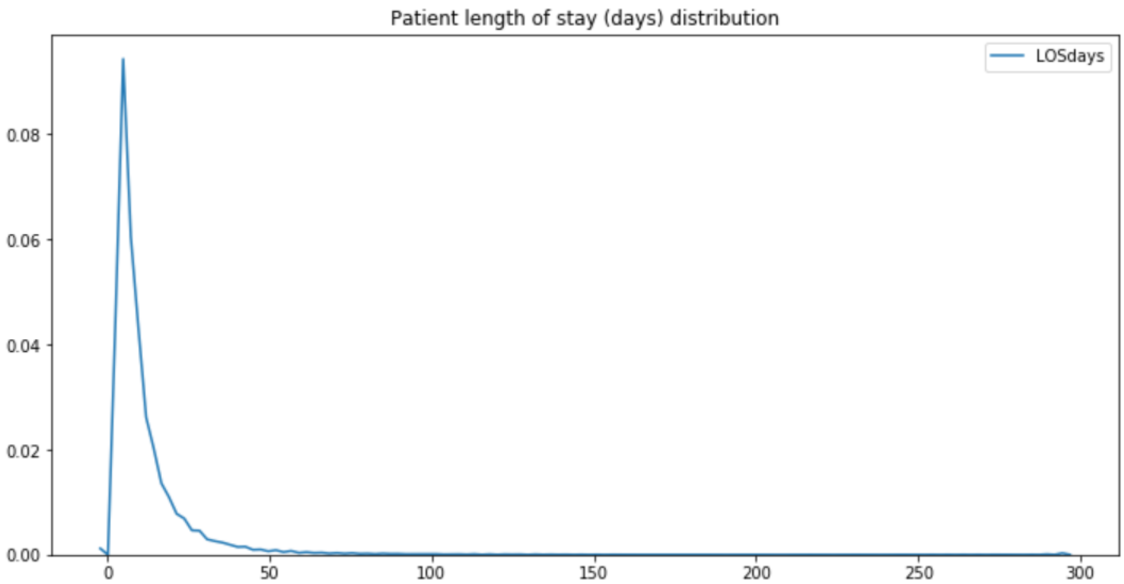

Figure 13: Patient Length of Stay Distribution

As we have considered the insurance types as well. Insurance type do impact in cases of complications because the complications are not only based on certain medication only, they also are based upon certain tools, systems or an items used on patients due to which a particular complication occurred. For example, complications related to stents. The insurance covers such type of things that a patients will get what kind of services hence we did not neglect this feature. Following is the distribution on types of insurances.

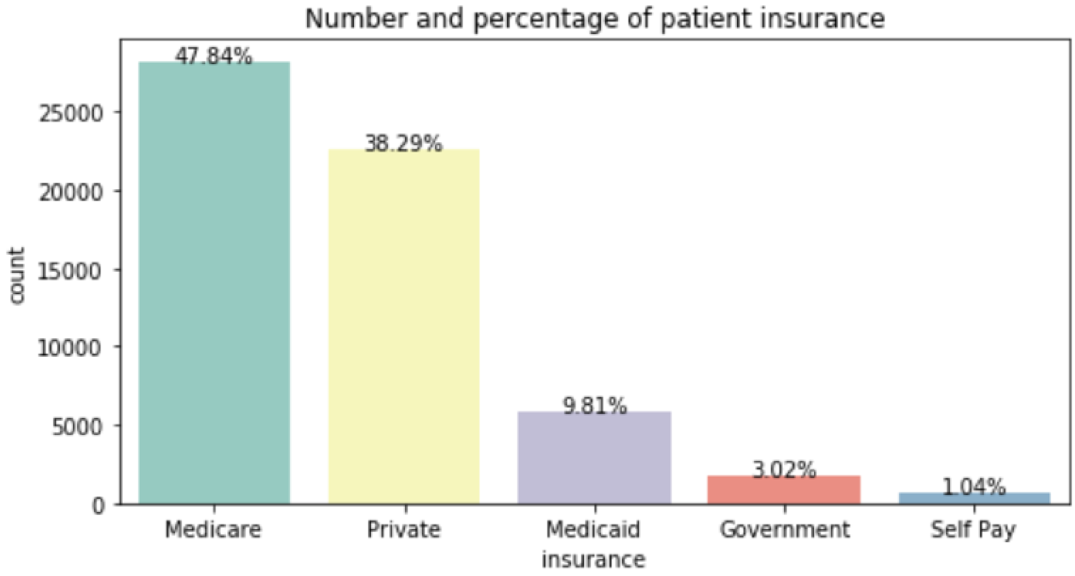

Figure 14: Insurance Types Distribution

6.2 Results

After the data preparation, problem statement and the decision of performance metric there were several model candidates to run. The candidate models that we selected are Logistic Regression, Linear SVC (SVM), Random Forrest and ANN.

First we standardized our data and normalized our data frames to be passed to models to predict the complications. We used standard min-max Scaler for normalization of our data. For NN we used a multilayer perceptron for Complications predictions which used **ReLU** as activation function.

Before each model, two data sampling processes were executed. SMOTE and ADASYN as we had class imbalance problem. After getting the interpolated data we passed that dataset to each of the model to predict the complications.

The different result metric resulted from each of the model against each data sampling technique is shown in below table:

Model	Accuracy	AUROC
Logistic Regression	65 %	0.72
Linear SVC	66%	0.72
Random Forrest	86%	0.83
ANN	81%	0.82

Table 6: Results with ADASYN Data Sampling

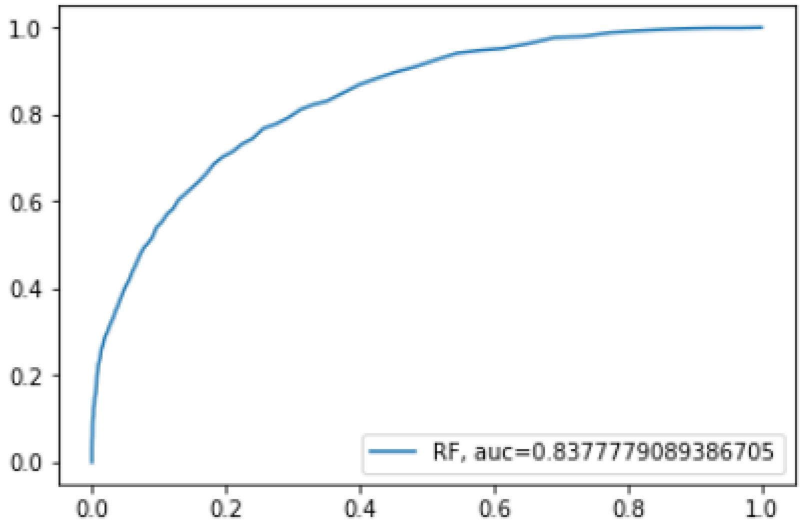

Figure 15: AUC for RF

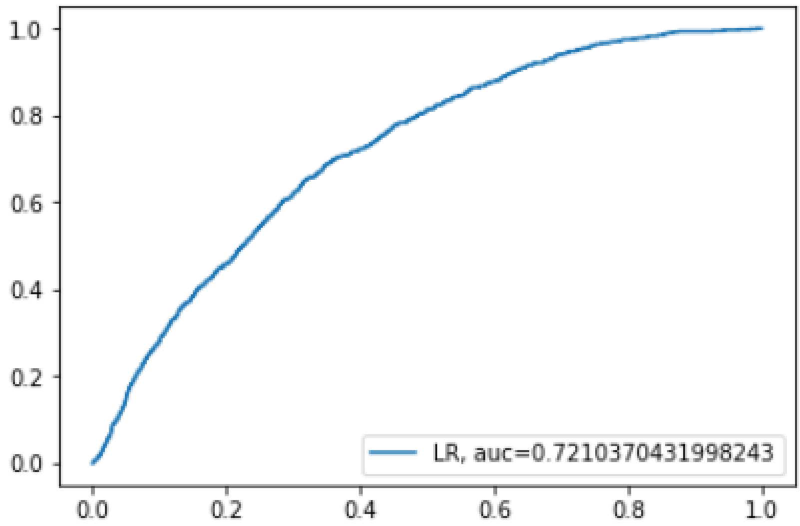

Figure 16: AUC for LR

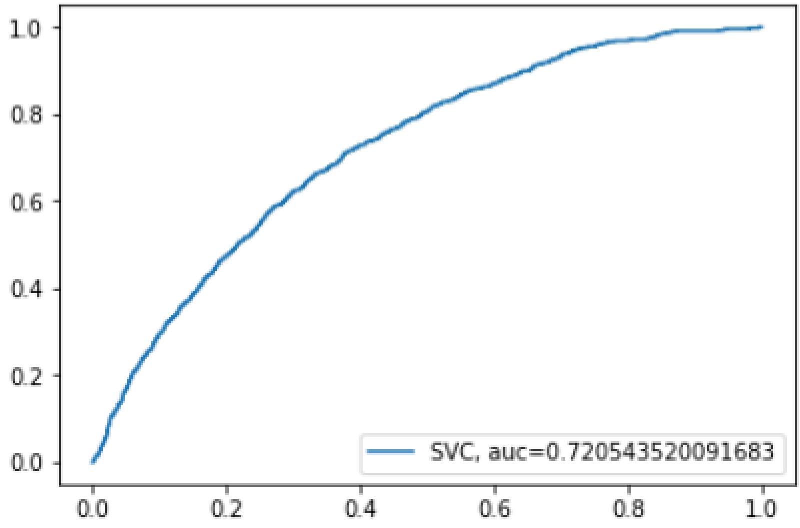

Figure 17: AUC for SVC

The above table showed result after applying the ADASYN data sampling technique. After that we implement SMOTE as well to see if the models vary with the type of static sampling instead of interpolating data based on the distribution as in ADAYSN

Model	Accuracy	AUROC
Logistic Regression	67 %	0.72
Linear SVC	67%	0.73
Random Forrest	85%	0.86
ANN	84%	0.83

Table 7: Results with SMOTE Data Sampling

The later table shows result with SMOTE data sampling technique. It clearly shows that RF still outperforms other models.

Patients with Complications	Patients Died at Hospital
2754	576

Table 8: Complications vs Expiry

We cannot evidently say that, complications lead to death but there can certainly be some parameters which leads to mortality based on complications. In our future development we will try to map those features with deep domain knowledge and expand our study to contribute more.

CHAPTER 7: CONCLUSION

Extended diagnoses and patients stay at the hospital is associated with not only the health of patient, cost, increased number of deaths but also increased number of readmissions. Each of these parameters defines the hospital performance. So, our focus was to produce insights that can complement these parameters ranging from cost to patient health. In order to propose a new method or optimize models that are already built. We can improve the data quality enhance methods of feature engineering and tune our models to outperform earlier methods.

To do that we applied following steps:

- For high quality: feature engineering with **ETL**, batch processing and data sampling
- Apply different sampling techniques and implement different models

These are presented in the order in above sections. The data is important for each predicting task, if the data is not available, there is no point of using complex model and optimize with resources and conversely if the data is available but the models are not complex enough to utilize that data can create performance drawbacks **[22]**.

Firstly, the use of MIMIC-III database to analyze electronic health records is probably the best to develop studies and researches that can contribute to the society. As it is being shared with the researchers, educational community and scientists, people are aggressively contributing to the cause and creating a huge impact. But, one must consider the existence of unstructured data available in MIMIC-III which lead to extensive search for variables without access to actual system deployed at hospitals. Because of that a considerable part was dedicated to the ETL to extract meaningful features which can lead to reliable prediction models. We used **Python** as it provide large number of libraries and models.

Finally, in our study, **Random Forrest Classifier** which gives the best prediction after both the sampling of data **ADASYN** and **SMOTE.** The accuracy also indicated that by including the derived concept combined with chart and lab events against each patient.

By extraction of derived features of organ failure expanded our problem and features space which lead to the analysis of the importance of the variables that have relationship between each other and to complications, We have achieved models with great prognostic capacity using demographic, concepts, lab events, chart events features and interpolating the minority class with different techniques which are not only intuitive for management’s view, for patient’s health and for doctors as well.

7.1 Future Investigation

The future studies regarding the complications and mortality of patients due to those complications are evident. As we can do multiclass classification which drills down the complications related to specific types instead of binary class of having complications or not. It was more adequate to start off with linear models and then further moved towards complex models, so, to improve our models we can adapt complex models in our future work.

Another line of research would be to engineer more features, create new concepts, and combine NLP techniques for textual features and applying complex models to contribute more the already done research.

REFERENCES

[1] Johnson, Alistair EW, Tom J. Pollard, Lu Shen, H. Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. "MIMIC-III, a freely accessible critical care database." *Scientific data* 3 (2016): 160035. Available at: https://mimic.physionet.org/

[2] Gentimis, Thanos, Alnaser Ala'J, Alex Durante, Kyle Cook, and Robert 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)*, pp. 1194-1201. IEEE, 2017.

[3] Huang, Jinmiao, Cesar Osorio, and Luke Wicent Sy. "An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes." *Computer Methods and Programs in Biomedicine* 177 (2019): 141-153.

[4] Yao, Liang, Chengsheng Mao, and Yuan Luo. "Clinical text classification with rule-based features and knowledge-guided convolutional neural networks." *BMC medical informatics and decision making* 19, no. 3 (2019): 71.

[5] Johnson, Alistair EW, Tom J. Pollard, and Roger G. Mark. "Reproducibility in critical care: a mortality prediction case study." In *Machine Learning for Healthcare Conference*, pp. 361-376. 2017.

[6] Arvind, Varun, Jun S. Kim, Eric K. Oermann, Deepak Kaji, and Samuel K. Cho. "Predicting Surgical Complications in Adult Patients Undergoing Anterior Cervical Discectomy and Fusion Using Machine Learning." *Neurospine* 15, no. 4 (2018): 329.

[7] Rojas, Juan C., Kyle A. Carey, Dana P. Edelson, Laura R. Venable, Michael D. Howell, and Matthew M. Churpek. "Predicting intensive care unit readmission with

machine learning using electronic health record data." *Annals of the American Thoracic Society* 15, no. 7 (2018): 846-853.

[8] Meyer, A., Zverinski, D., Pfahringer, B., Kempfert, J., Kuehne, T., Sündermann, S.H., Stamm, C., Hofmann, T., Falk, V. and Eickhoff, C., 2018. Machine learning for real-time prediction of complications in critical care: a retrospective study. *The Lancet Respiratory Medicine*, *6*(12), pp.905-914.

[9] Deliberato, Rodrigo Octavio, Ary Serpa Neto, Matthieu Komorowski, David J. Stone, Stephanie Q. Ko, Lucas Bulgarelli, Carolina Rodrigues Ponzoni, Renato Carneiro de Freitas Chaves, Leo Anthony Celi, and Alistair EW Johnson. "An evaluation of the influence of body mass index on severity scoring." *Critical care medicine* 47, no. 2 (2019): 247-253.

[10] Naik, Girish S., Sushrut S. Waikar, Alistair EW Johnson, Elizabeth I. Buchbinder, Rizwan Haq, F. Stephen Hodi, Jonathan D. Schoenfeld, and Patrick A. Ott. "Complex inter-relationship of body mass index, gender and serum creatinine on survival: exploring the obesity paradox in melanoma patients treated with checkpoint inhibition." *Journal for immunotherapy of cancer* 7, no. 1 (2019): 89.

[11] Bose, Somnath, Alistair EW Johnson, Ari Moskowitz, Leo Anthony Celi, and Jesse D. Raffa. "Impact of intensive care unit discharge delays on patient outcomes: a retrospective cohort study." *Journal of intensive care medicine* 34, no. 11-12 (2019): 924-929.

[12] Deliberato, Rodrigo Octávio, Stephanie Ko, Matthieu Komorowski, M. A. Armengol de La Hoz, Maria P. Frushicheva, Jesse D. Raffa, Alistair EW Johnson, Leo Anthony Celi, and David J. Stone. "Severity of illness scores may misclassify critically ill obese patients." *Critical care medicine* 46, no. 3 (2018): 394-400.

[13] Davoodi, Raheleh, and Mohammad Hassan Moradi. "Mortality prediction in intensive care units (ICUs) using a deep rule-based fuzzy classifier." *Journal of biomedical informatics* 79 (2018): 48-59.

[14] Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16 (2002): 321-357.

[15] Bergamaschi, Sonia, Francesco Guerra, Mirko Orsini, Claudio Sartori, and Maurizio Vincini. "A semantic approach to ETL technologies." *Data & Knowledge Engineering* 70, no. 8 (2011): 717-731.

[16] Vassiliadis, Panos, Anastasios Karagiannis, Vasiliki Tziovara, Alkis Simitsis, and Ioannina Hellas. "Towards a benchmark for etl workflows." (2007).

[17] *International Classification of Diseases*. Available at https://www.cdc.gov/nchs/icd/icd9.htm

[18] Hofmann, Markus, and Brendan Tierney. "An enhanced data mining life cycle." In *2009 IEEE Symposium on Computational Intelligence and Data Mining*, pp. 109-117. IEEE, 2009.

[19] *Our Code Repository:* https://github.com/faisalmaqbool94/Thesis-Bioinformatics-MIMICIII-

[20] Repository of code shared by the research community provided by MIT Laboratory for Computational Physiology available at https://github.com/MIT-LCP/mimic-code

[21] Eckle, Konstantin, and Johannes Schmidt-Hieber. "A comparison of deep networks with ReLU activation function and linear spline-type methods." *Neural Networks* 110 (2019): 232-242.

[22] Alon Halevy, Peter Norvig, and Fernando Pereira. The Unreasonable Effectiveness of Data. IEEE Computer Society (2009). Available at: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/35179.pdf

[23] Data, MIT Critical. *Secondary Analysis of Electronic Health Records*. Springer International Publishing, 2016.

[24] Rapsang, Amy Grace, and Devajit C. Shyam. "Scoring systems in the intensive care unit: a compendium." *Indian journal of critical care medicine: peer-reviewed, official publication of Indian Society of Critical Care Medicine* 18, no. 4 (2014): 220.

[25] Kury, Fabrício SP, Vojtech Huser, and James J. Cimino. "Reproducing a prospective clinical study as a computational retrospective study in MIMIC-II." In *AMIA Annual Symposium Proceedings*, vol. 2015, p. 804. American Medical Informatics Association, 2015.

[26] Bates, David W., Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. "Big data in health care: using analytics to identify and manage high-risk and high-cost patients." Health Affairs 33, no. 7 (2014): 1123-1131.

[27] Segal, Jodi B., and Neil R. Powe. "Accuracy of identification of patients with immune thrombocytopenic purpura through administrative records: a data validation study." American journal of hematology 75, no. 1 (2004): 12-17.

[28] Eichler, April F., and Elizabeth B. Lamont. "Utility of administrative claims data for the study of brain metastases: a validation study." Journal of neuro-oncology 95, no. 3 (2009): 427-431.

[28] Kern, Elizabeth FO, Miriam Maney, Donald R. Miller, Chin-Lin Tseng, Anjali Tiwari, Mangala Rajan, David Aron, and Leonard Pogach. "Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes." Health services research 41, no. 2 (2006): 564-580.

[29] Perotte, Adler, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. "Diagnosis code assignment: models and evaluation metrics." Journal of the American Medical Informatics Association 21, no. 2 (2013): 231-237.

[30] Mullen, Michael T., Charles J. Moomaw, Kathleen Alwell, Jane C. Khoury, Brett M. Kissela, Daniel Woo, Matthew L. Flaherty et al. "ICD9 codes cannot reliably identify

hemorrhagic transformation of ischemic stroke." Circulation. Cardiovascular quality and outcomes 6, no. 4 (2013): 505.

[31] Lita, Lucian Vlad, Shipeng Yu, Stefan Niculescu, and Jinbo Bi. "Large scale diagnostic code classification for medical patient records." In Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II. 2008.

[32] Baumel, Tal, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noemie Elhadad. "Multi-label classification of patient notes: case study on ICD code assignment." In Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[33] Huang, Yuan-Lan, Tony Badrick, and Zhi-De Hu. "Using freely accessible databases for laboratory medicine research: experience with MIMIC database." Journal of Laboratory and Precision Medicine 2, no. 6 (2017).

[34] Shao, Jianyin, Ram Gouripeddi, and Julio C. Facelli. "2166: Semantic characterization of clinical trial descriptions from ClincalTrials. gov and patient notes from MIMIC-III." *Journal of Clinical and Translational Science* 1, no. S1 (2017): 12-12.

[35] Mandalapu, Varun, Benjamin Ghaemmaghami, Renee Mitchell, and Jiaqi Gong. "Understanding the relationship between healthcare processes and in-hospital weekend mortality using MIMIC III." Smart Health 14 (2019): 100084.

[36] Bache, Richard, Simon Miles, and Adel Taweel. "An adaptable architecture for patient cohort identification from diverse data sources." Journal of the American Medical Informatics Association 20, no. e2 (2013): e327-e333.

[37] *Health Insurance Portability and Accountability Act (HIPAA)* https://www.hhs.gov/hipaa/for-professionals/index.html

[38] Neamatullah, Ishna, Margaret M. Douglass, H. Lehman Li-wei, Andrew Reisner, Mauricio Villarroel, William J. Long, Peter Szolovits, George B. Moody, Roger G. Mark, and Gari D. Clifford. "Automated de-identification of free-text medical records." *BMC medical informatics and decision making* 8, no. 1 (2008): 32.

[39] Ho, Tin Kam. "Random decision forests." In Proceedings of 3rd international conference on document analysis and recognition, vol. 1, pp. 278-282. IEEE, 1995.

[40] Breiman, Leo. "Random forests." Machine learning 45, no. 1 (2001): 5-32.

[41] Amit, Yali, and Donald Geman. "Shape quantization and recognition with randomized trees." Neural computation 9, no. 7 (1997): 1545-1588.

[42] Ferrao, Jose C., Filipe Janela, Mónica D. Oliveira, and Henrique MG Martins. "Using structured EHR data and SVM to support ICD-9-CM coding." In 2013 ieee international conference on healthcare informatics, pp. 511-516. IEEE, 2013.

[43] Solt, Illés, Domonkos Tikk, Viktor Gál, and Zsolt T. Kardkovács. "Semantic classification of diseases in discharge summaries using a context-aware rule-based classifier." Journal of the American Medical Informatics Association 16, no. 4 (2009): 580-584.

[44] He, Haibo, Yang Bai, Edwardo A. Garcia, and Shutao Li. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning." In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322-1328. IEEE, 2008.

[45] Nwankpa, Chigozie, Winifred Ijomah, Anthony Gachagan, and Stephen Marshall. "Activation functions: Comparison of trends in practice and research for deep learning." arXiv preprint arXiv: 1811.03378 (2018).

Appendix A

Resources

Available code repository:

- https://github.com/faisalmaqbool94/Thesis-Bioinformatics-MIMICIII-

Repository of code shared development community provided by MIT Laboratory for Computational Physiology

- https://github.com/MIT-LCP/mimic-code