Using Natural Language Processing to Detect Breast Cancer Recurrence in Clinical Notes: A Hierarchical Machine Learning Approach

by

Sujan Subendran

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Masters of Applied Science
in
System Design Engineering

Waterloo, Ontario, Canada, 2021

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The vast amount of data amassed in the electronic health records (EHRs) creates needs and opportunities for automated extraction of information from EHRs using machine learning techniques. Natural language processing (NLP) has the potential to substantially reduce the burden of manual chart reviewing to extract risk factors, adverse events, or outcomes, that are documented in unstructured clinical reports and progress notes. In this thesis, an NLP pipeline was built using open-source software to process a corpus of electronic clinical notes extracted from an integrated health care system in Cancer Care Manitoba (CCMB) which contains a cohort of women with early-stage incident breast cancers. The goal is to identify whether and when recurrences were diagnosed. We developed and evaluated the system using 117,365 clinical notes from 892 patients receiving EHR-documented care at CCMB between 2004 to 2007. We used a hierarchical architecture, where a model is built to provide the patient-level recurrence status, then the NLP pipeline is used to detect notes which contains information about recurrence and the date of recurrence. Class imbalance was a significant issue as the proportion of positive to negative notes was at approximately 1:22 ratio. Various techniques including undersampling and cost-based classification were used to mitigate this issue. The XGBoost classifier was the best performing model which achieved a balanced accuracy of 0.924, with sensitivity of 0.867, specificity of 0.981, precision of 0.886 and ROC of 0.924. In addition, more data was collected from the years 2008 to 2012 in a similar cohort. This dataset was used to validate the performance of the models, which include 615 patients with 78,460 notes. The model performed well with a balanced accuracy of 0.909, sensitivity of 0.843, specificity of 0.974, precision of 0.575 and Area Under the ROC Curve (AUC) value of 0.909.

The study has demonstrated the ability to use natural language processing and machine learning techniques to assist in chart review by 1) excluding a large amount of notes which contain no relevant information, 2) identifying notes that most likely contain relevant recurrence information, in order to accurately identify the timing of recurrence.

Acknowledgements

I would like to thank my supervisor Professor Helen Chen for providing guidance and offering continuous support throughout my MASc program.

I would also like to thank CancerCare Mantioba for providing funding, access to the dataset and clinical knowledge to support me throughout this research project. Special thanks to Dr. Marshall Pitz, Dr. Harminder Singh, Pascal Lambert, Elizabeth Harland and Kathleen Decker.

I would also like to thank my research group members George Michalopoulos, Hussam Kaka and Rena Yang for their assistance in data collection and preprocessing, and advice on modelling approaches. In addition, special thanks to Kam Sharma and Therese Tisseverasinghe for extensive support in literature review of the clinical background.

Dedication

This is dedicated to all my loved ones.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Breast Cancer Recurrence

Cancer is defined as a group of diseases that are a result of some sort or abnormal change of the cells in the body resulting in abnormal growth of those cells into lumps or masses [1]. Breast cancer is the abnormal growth of cells in the breast tissue with most breast cancers starting out in the lobules or milk glands of the breast [1].

Breast cancer is typically considered a very dangerous cancer as there are few to no signs and symptoms in the beginning stages of tumor development [2], [3]. For this reason, it is extremely important that efficient early detection methods are set in place. Regular practices include checking for hard lumps around the breast or swelling in the underarm area near the lymph nodes [4]. Some less common signs and symptoms include pain in the breast, swelling or discharge from the breast [5].

Typically, most breast cancer is diagnosed at a physical examination at a physicians screening or is noticed by the patient once a lumps has formed. A mammogram which is a machine that compresses the breast and detects the presence of lumps is also used [6]. Most tumors detected during a mammogram tend to be benign (not cancerous). If either of these detection methods show a positive result, usually, a needle biopsy is completed in order to confirm the result and suspected positive test.

The most common type of breast cancer is the invasive or infiltrating kind. Invasive breast cancer is a type of cancer that is characterized when cells have emerged through the glands in the breast and have now started to invade and spread to the breast tissue around the gland [7].

Breast cancer is the umbrella term for a group of cancer related diseases that originate in the breast tissue. There are 4 different molecular subtypes and a minimum of 21 histological subtypes [8]. All of these subtypes of beast cancer have a very large range of risk factors, recurrence rates, and outcomes [8]. Histological subtypes of breast cancer are defined by the size, shape and cancer cell arrangement. Molecular subtypes of breast carcinoma are defined based on their gene expression and can be easily recognized by determining their biological markers like ER, PR, HER2 and more [8].

Treatment for breast cancer varies largely based on the subtype, stage, the patients preferences, the patients age and the stage the cancer has progressed to. Once diagnosed with cancer, the goal for a physicians treatment plan is to essentially remove all the cancer from the patients body. Surgical treatments includes performing a mastectomy which is surgical removal of the entire breast. However, there are situations in which a partial mastectomy can be performed in which only the tumor with a reasonable margin of healthy tissue is also removed. Usually, after the tumor has been removed surgically, multiple rounds of radiation take place in order to ensure no cancerous cells remain [9]. Radiation is also beneficial to reduce the chances of a recurrence. Systemic therapies like drugs that travel through the entire body via the blood stream are also typically administered however they are likely to cause damage to other parts of the body as well [9]. Chemotherapy is one of the types of systemic drug therapies that are commonly used to treat cancer usually for metastatic cancer in younger women [9]. Many more treatments exist such as hormonal therapy, targeted therapy and immunotherapy to name a few [10].

It was reported by the American Cancer Society in 2019 that 268,600 new cases of breast cancer are diagnosed with 41,760 death from the disease in the United States itself [11]. Unfortunately, these numbers only seem to increases as the years go on. This is undoubtedly a very large amount of deaths making breast cancer related deaths one of the most common cancer related deaths in women in the United States [11].

Breast cancer has a recurrence rate of 20% to 30% which is a relatively high rate compared to other cancers, therefore, in order to determine a better treatment plan for breast cancer and to lower the recurrence rate, it is essential that further research be completed to completely understand previous patient charts and medical treatments [12], [13].

A countless number of people have developed methods to detect and understand medical information from medical documents [14]–[17]. cTAKES is and application that does exactly this by analyzing medical and clinical notes and through this method, this application can be very helpful to help flag patients that may show signs of breast cancer recurrence by identifying the keywords that are important in identifying signs of possible breast cancer recurrence.

Recurrence is defined as the occurrence of a second cancer diagnosis of breast cancer after the primary one was completely treated and considered to be cured. Recurrence of breast cancer is predominantly due to the presence of residual diseased cells that were not destroyed during the first treatment of the disease [18]. The prevalence and detection of recurrence with cancer is a very important topic to study and research because with research into effective treatments, the measure of recurrence-free survival and cancer treatment control, a better treatment plan can be determined to reduce the rate of recurrence significantly.

Currently, cancer registries do not have any means to identify cancer recurrences and for this reason, prior studies conducted their research using chart reviews. However, in order to do this, charts must be read individually which is a very time consuming and simply a very expensive process. In order to lessen the time required for this task, administrative health care data can be used as a way to identify cancer recurrences.

1.2 Natural Language Processing in Breast Cancer Recurrence Detection

A vast majority of healthcare data is unstructured text. It is extremely challenging for health researchers to extract insights from this data at scale. Finding the critical information in the EMR and extract them into high-quality, research-grade datasets is important for conducting clinical research, and making clinical and financial decisions for healthcare providers. Extracting correct information is time-consuming, error-prone and expensive. Manual chart review by train clinical professionals are the gold standard to extract information from unstructured text. Coding errors may rise due to several reasons. First, in clinical descriptions, abbreviations and synonyms are commonly used, which may cause ambiguity for interpretation. Second, in many cases, several diagnosis descriptions are closely related and should be combined into codes, or multiple codes need to be assigned for a particular procedure. Third, because the codes are organized in a hierarchical structure, an overly generic code can be assigned instead of a more specific code [19]. Natural language processing (NLP) is a method of analyzing text in order to achieve language processing at a human-like level for different applications. A completely functional NLP System would be able to paraphrase input text, translate test from one language to another, accurately answer questions about the input text, and lastly draw inferences from the information in the text.

A study conducted on contralateral breast cancer detection using NLP can be very effectively applied to the detection of breast cancer recurrence as this study specifically looked at the recurrence of breast cancer in the opposite breast after the the primary cancer was already detected and treated [20].

It has been determined that women have a 2-6 times higher risk of developing contralateral breast cancer after a primary one has already been treated or detected [20]. Understanding the set of causes that result in the recurrence of breast cancer can help devise a better understanding of how breast cancer actually develops and a more effective treatment plan as well in order to prevent recurrence all together. Studies have been completed in order to determine the relationship between the primary and secondary diagnosis of breast cancer by analyzing information such as family history, exposure from the environment, and possible genetic mutations. However, in order to get accurate results in these studies, it is important that medical records and notes be accurately and efficiently analyzed. Electronic Health Records (EHRs) have allowed for very large groups of medical information to be accessed and used for studies including studies conducted for breast cancer recurrence [20]. However, even with the presence of the EHRs, the amount of reports and clinical notes a researcher would have to sort through and analyze is astronomical. For this reason, applications of NLP are very helpful as a method to help sort through large cohorts of medical documentation effectively.

A problem does occur when trying to use NLP as a way to analyze medical records as it is required that the medical reports like pathology reports be well completed but in a standard format. However, machine learning and NLP models are a good alternative to solve this problem as this algorithm learns patterns from free text that have been labeled and then later applies those algorithms to other further text that is not labeled [20].

1.3 Challenges of Breast Cancer Recurrence Identification

Based on the importance of studying long-term outcomes for cancer patients, the inefficiencies of chart reviews, and the limitations of previous recurrence algorithm development, our goal is to develop cancer-site specific algorithms that use high quality, complete data. We have access to administrative health data (i.e., data that is generated through the routine administration of health care programs) and electronic medical record (EMR) data (i.e., the computer-based CancerCare Manitoba (CCMB) cancer patient chart). The EMR data includes structured data and unstructured data found in the health care provider's notes.

We hypothesize that recurrence algorithms which utilize unstructured data from the EMR health care provider's notes will have higher sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) compared to algorithms that include only administrative health data and structured EMR data. Our objective is to develop algorithms using administrative health data and EMR data that can accurately capture recurrence in breast cancer and colorectal cancer cohorts.

This study will access multiple sources of population-based, complete, high-quality administrative health data and information from an EMR. The developed algorithms have the potential to increase the efficiency and reduce the costs of epidemiological and health services research on cancer treatment effectiveness and outcomes. The recurrence algorithms will be specific to Manitoba and will be useable by future research teams enhancing CCMB research productivity and quality of care assessments. One limitation is that recurrences that are not followed-up with contact with medical, surgical, or radiation oncology at CancerCare Manitoba may not be captured. However, we expect that this impact will be minimal as the cancers selected for this study have high survival rates and the usual practice pattern of referring almost all patients with recurrent disease to oncologists.

1.4 Motivation and Contribution

The motivation for this research is to improve automatic identification of breast cancer recurrence. By using a tool to identify positive-recurrence patient notes, many hours of manual charts review can be saved and more time can be dedicated to other research endeavours using these identified notes. The current process is completely manual, if the tool can assist in charts review by screening cases it has the potential to be very useful to researchers. However, the data that is available to train models for the recurrence identification poses some challenges such as class imbalance and unstructured/semi-structured information. Intelligent techniques are required to handles these challenges and provide a useful model with minimal misclassifications.

1.5 Thesis Outline

In Chapter 2 the relevant literature is reviewed include the technology and software used for NLP and ML. Chapter 3 discusses the methods which includes data collection, the NLP software used to analyze clinical notes and the hierarchical machine learning approach. In this section, the dataset characteristics and patient cohorts are presented. The architecture and steps for the hierarchical machine learning approach are explained. In Chapter 4 an in depth analysis of the NLP is explored. The components of the NLP engine are discussed as well as examples of clinical notes processed by the NLP engine. Chapter 5 explains the machine learning algorithms that are used and the internal mechanisms of these models. In addition, modelling techniques that are used to improve classification on imbalanced datasets are discussed. The experimental results are shown in Chapter 6. Finally, Chapter 7 and 8 contain the discussions and conclusions.

Chapter 2

Literature Review

2.1 cTAKES Package and Its Application in Processing Clinical Text

cTAKES stands for Clinical Text Analysis and Knowledge Extraction System. This system was created by the Mayo Clinic [21]. cTAKES is a natural language processing system created as a way to highlight important information from electronic medical and pathology reports. cTAKES was derived from and build upon already existing technology called the Unstructured Information Management Architecture framework and OpenNLP toolkit [21].

cTAKES combines both rule-based and machine learning techniques to perform information extraction of clinical text. cTAKES was developed with gold standard datasets that consist of linguistic labels and clinical concepts from the Mayo Clinic's electronic health records (EMR). Standard evaluation metrics were used to measure and improve performance of cTAKES, this includes sensitivity, positive predictive value (PPV), F-score and accuracy.

EMR is the main source of medical and clinical information, with the majority of data captured as unstructured free text, such as discharge summaries, radiology notes and progress notes. This information is essential for clinical decision making, yet it is hard to go through the huge amount of free text for clinical care, and for extracting data for research or regulatory reporting. The gold standard of data extraction from these free text is through chart review, where researchers have to manually read every report. Chart review is an extremely time consuming process and is also prone to errors. Using a natural language processing tool specifically adjusted to handle medical text, such as the cTAKES package, we are capable of identifying facts or entities of interest. This is very helpful for the clinicians and researchers to navigate through a large amount of text. It helps to improve quality of care and support clinicians to make more effective treatment decisions [22]–[24].

One notable research paper from Harvard Medical School examined the use of natural language processing and machine learning approaches to classify clinical notes into medical subdomains (i.e. Cardiology, Gastroenterology, Neurology, Psychiatry, Pulmonary and Nephrology) [25]. The developed pipeline was tested on two datasets, a public clinical dataset called iDASH (integrating data for analysis, anonymization and sharing) and a hospital dataset from MGH (Massachusetts General Hospital). The paper achieved promising results, AUC of 0.957 and 0.964, and F1 scores of 0.932 and 0.934 for the iDASH and MGH datasets, respectively. The research group had also published the vocabulary dictionary that was used with Apache cTAKES for medical term extraction from clinical notes. Since this was a curated dictionary specifically for clinical notes it proved to be a very useful starting point for extracting terminology from clinical notes in the breast cancer recurrence subdomain.

cTAKES was compared to other NLP application such as MetaMAP [26], with a relatively comparable performances. It is noteworthy that both algorithms were shown to be more effective compared to manual searches and extraction of information [27].

2.2 Classification using Machine Learning

Machine Learning (ML) methods have attracted significant attention in health research over the past decade. There have been some successful applications using ML methods in disease detection [24], precision medicine [28] and risk of readmission prediction [29]. Benefits of using ML include minimal domain assumptions, flexibility of model selection and automation of the model construction and evaluation. Significant improvements in performance of ML classifiers have encouraged wide-spread exploration of ML for breast cancer recurrence predication.

2.2.1 Decision Tree

Breiman first introduced the decision tree algorithm [30]. Decision tree uses a simple tree structure to classify data points into a predefined set of categories(classes). Each tree node represents a feature and each branch represents a possible split of the feature value. For example, for a tree node with a numerical feature the branches may split the feature by some finite number which is learned during training. Classification of a sample is a simple process once the decision tree structure is learned. The decision tree is traversed starting from the root node and following the branches corresponding to the provided sample. The tree is traversed until a leaf node is reached, at which point a class is predicted for that sample.

A decision tree classifier is built in two stages, namely tree building and tree pruning. At the first stage, the training dataset is recursively divided based on a locally criterion until each of the branches contains the same , or almost the same class label. Generally, the smaller each partition is, the more likely all data points in this partition has the same label. However, too many branches in a decision tree could lead to overfitting, thus pruning is performed to reduce the size of the decision tree. Careful pruning can improve the model's generalizability. Starting from the bottom of the tree, we exam each non-leaf subtree. If we replace a subtree with a leaf, or with its most frequently used branch, a lower predicted error rate is achieve, then we will prune the tree [31]. One of the commonly used decision tree algorithms is C4.5 [32], this algorithm is first performs tree building to construct the decision tree then follows up with pruning to remove branches and replace them with leaf nodes.

2.2.2 AdaBoost

AdaBoost is one of the ensemble learning methods that aims to improve classification performance by variance and bias reduction. The base learner for AdaBoost are decision stumps. These are decision trees with one step from the root node to the terminal nodes. The decision stumps have high bias but low variance. In order to improve classification performance once such technique that can be applied is called boosting. In this approach, a weak learner is sequentially trained across the dataset while using weights for individual observations. These weights are adjusted (increased for misclassification and decreased for correct classification) as the training progresses, the idea is that subsequent trained models iteratively improve upon the errors of previous models. The training algorithm chooses a cut-off at which it accepts a learner in the ensemble. The AdaBoost algorithm [33] is one such method that implements the boosting technique with decision trees most commonly used as the base learner. When we have an highly imbalanced dataset, standard learning methods often fail to achieve an acceptable performance on the rare class due to the bias in the dataset. AdaBoost is capable in reducing learning bias, thus often performs better when dealing with imbalanced datasets.

2.2.3 Random Forest

A random forest [34] algorithm is an ensemble of decision trees, where each tree is built with an independent set of random vectors of a dataset.

There are four steps in the random forest algorithm:

1. Generate a random bootstrap, i.e. select a subset of n data points
2. Grow a decision tree from this bootstrap by a) randomly select k features without replacement and b) partition the node using the features that provides the best split according to the optimal function
3. Repeat the steps 1-2 m times
4. Assign the class label by majority vote

The size of the bootstrap n can impact the correlation of the trees in the random forests. The smaller n tends to make trees less correlated. Therefore the algorithm is superior when training with a dataset that contains a very large number of input variables. This algorithm also requires less run time since only a subset of the features needs to be examined at each node. In addition, a individual tree may suffer from high variance, but averaging multiple trees will result in a more robust model that performs better, and is less susceptible to overfitting.

2.2.4 Gradient Boosting

Gradient boosting is a machine learning technique that produces prediction models in the form of decision trees for regression and classification problems. The models are built stagewise similar to other boosting methods. It does this by optimizing arbitrary differentiable loss functions.

Jerome H. Friedman developed explicit regression gradient boosting algorithms which can be explained to be iterative functional gradient descent algorithms [35]. Iterative functional gradient descent algorithms include algorithms that optimize a cost function over function space. This is accomplished by repetitively choosing a function (weak hypothesis) that points in the negative gradient direction. The development of boosting algorithms via the functional gradient concept has allowed for enhancements in areas of machine learning and statistic in not just regression and classification but much more.

2.3 Modeling with Highly Imbalanced Datasets and Performance Evaluation of Clinical Models

The class imbalance problem is an issue present in many datasets in the healthcare field. There are a few approaches to dealing with class imbalance such as maniuplating the data using sampling techniques to improve the class balance ratio while maintaining integrity of the dataset. Alternatively, machine learning algorithms can introduce bias to a particular class in order to mitigate issues from class imbalance. This bias can be used to help adjust the algorithm such that it performs better on the minority class.

To deal with the imbalance of datapoints a number of sampling methods were utilized as given below:

1. Random Undersampling
2. Oversampling SMOTE

Random Undersampling includes sampling the majority class such that the number of datapoints are in comparison with the minority class. Near miss Undersampling is about sampling the datapoints from the majority class which aid in discriminating between classes. Oversampling SMOTE (Synthetic Minority Over-sampling Technique) synthetically generates samples by interpolating between samples and the original dataset.

Chapter 3

Methods

3.1 Study Design

The study includes two phases: an algorithm development phase and a validation phase. During the algorithm development phase (I), there are two steps as the following:
Step 1: Chart Review. The data collected to develop the algorithm includes individuals diagnosed with breast cancers from 2004 to 2007 at the Cancer Care Manitoba since the information contained in the EMR has remained stable from 2004 onward. The cohort of breast cancer includes stage I to III cases that are estrogen receptor (ER) negative, progesterone receptor (PR) negative, or human epidermal growth factor receptor 2 (HER2) positive. These factors are related to poorer outcomes including recurrence. Charts were reviewed by trained research assistants. Recurrences were recorded as any recurrence as well as loco-regional (reappearance of cancer in the same region or in the lymph nodes) or distant (reappearance of cancer in another part of the body). Recurrence label and dates from the chart review are considered the "gold standard".
Step 2. Model development. The recurrent prediction models are developed using machine learning algorithms (details of the model development are presented in Chapter 5). In addition, any mis-classification of recurrences predicted from the selected model, including false positive and false negative, are investigated further to verify that chart information was not missing or misinterpretated. During the validation phase, we determine if the algorithms developed are generalizable to new cancer cases that are not included in the dataset for training the model. The validation dataset is extracted via a chart review of cases diagnosed from 2008-2012 with stage I, II, and III cancer.

To ensure the quality of the chart review process, a second experienced Cancer Registrar was included to independently review 10% of the cancer cases randomly selected (in both phases) to provide inter-rater reliability.

3.2 Dataset

The training dataset (Phase I) was collected from breast cancer patients in Manitoba, from the years 2004 to 2007. The validation set (Phase II) was collected from breast cancer patients from 2008 to 2012, also from Cancer Care Manitoba [36]. The dataset consists of clinical notes throughout the patients' medical journey and several administrative variables collected for each patient. In the training dataset there are 916 patients analyzed with 233,715 notes. The validation dataset consists of 615 patients and 159,107 notes. Using several preprocessing steps the number of total notes were reduced by combining them per patient, per day. This allowed for easier identification of the first date of recurrence without confusion of multiple notes for the same patient on the same day. The resulting number of notes for the training set was 38,372 combined notes and the validation set had 78,460 combined notes.

The following administrative health data sources are included: the Manitoba Cancer Registry (MCR), Manitoba Health Medical Claims, Hospital Abstracts, and Drug Program Information Network (DPIN), Diagnostic Services Manitoba (DSM) data, and the Manitoba Health Population Registry (MHPR). The MCR will be used to identify the cohorts and will include diagnosis dates, cancer sites, treatment (date and type of every surgery, the date of the first systemic anti-cancer therapy within the calendar year, and the date of the first radiation treatment per line of radiation treatment). Medical Claims and Hospital Abstracts will be used to identify diagnostic and treatment procedures and provider type. DPIN data will be used to identify systemic anti-cancer therapies (eg. Capecitabine, Tamoxifen, and aromatase inhibitors). DSM data will be used to identify carcinoembryonic antigen (CEA) and Ca15-3 blood test results. The source of EMR data is the CCMB cancer patient chart. The CCMB cancer patient chart includes the following structured data: dates and types of first-line treatment, systemic anti-cancer therapy administration dates, drug identification information, systemic anti-cancer therapy cycles, dose for each systemic anti-cancer therapy cycle, radiotherapy administration dates, radiation dose, radiation fraction number, radiation site treated, CEA and Ca15-3 blood test results. Since the CCMB EMR does not have a mandatory structured field to document recurrences, unstructured data from the health care provider's notes are reviewed to identify the presence of recurrence and recurrence details (whether additional treatment was provided, type of treatment, the diagnostic procedures that were involved and dates, clinical breast examination findings).

Variables	Positive Patients, n(%)	Negative Patients, n(%)	All Patients, n(%)
Recurrence	53 (5.91)	844 (94.09)	897 (100)
Radiation Therapy			
Yes	23 (2.56)	84 (9.36)	107 (11.93)
No	30 (3.34)	760 (84.73)	790 (88.07)
Chemotherapy			
Yes	28 (3.12)	70 (7.80)	98 (10.93)
No	25 (2.79)	774 (86.29)	799 (89.07)
Breast Surgery			
Yes	25 (2.79)	25 (2.79)	50 (5.57)
No	28 (3.12)	819 (91.30)	847 (94.43)
Other Surgery			
Yes	6 (0.67)	22 (2.45)	28 (3.12)
No	47 (5.24)	822 (91.64)	869 (96.88)
Palliation			
Yes	7 (0.78)	36 (4.01)	43 (4.79)
No	46 (5.13)	808 (90.08)	854 (95.21)
CEA Test			
Yes	7 (0.78)	59 (6.58)	66 (7.36)
No	46 (5.13)	785 (87.51)	831 (92.64)
CA 15-3 Test			
Yes	17 (1.90)	77 (8.58)	94 (10.48)
No	36 (4.01)	767 (85.51)	803 (89.52)

Table 3.1: Characteristics of the patients in the training dataset stratified by administrative variables and recurrence.

The administrative variables and the primary phrases used in the NLP are shown in Table 3.2. The administrative variables were collected as part of the data collection initiative. The primary phrases used for NLP were identified with domain experts (oncologists and clinical researchers from CancerCare Manitoba). These primary phrases were recorded with the corresponding Concept Unique Identifier (CUI) from the UMLS metathesaurus. The CUIs are presented for the primary phrases in the table as an alphanumeric identifier beginning with the letter "C" followed by the numeric ID. The complete dictionary of CUIs and phrases that were used in this study can be found in the Appendix.

Administrative Variables (7)	Radiation Therapy
	Chemotherapy
	Breast Surgery
	Other Surgery
	Palliation
	CEA Test
	CA 15-3 Test
Primary Phrases for NLP (24)	C0438204 Axillary Recurrence
	C0006142 Breast Cancer
	C0438105 Chest Wall Recurrence
	C0015726 Concern about Recurrence
	C0069515 HER-2
	C0007099 Intraductal Carcinoma
	C1134719 Infiltrating Breast Cancer
	C0205281 Invasive
	C0441989 Ipsilateral Recurrence
	C0343834 Lesion
	C1268990 Lobular
	C0851238 Lumpectomy
	C0024881 Masectomy
	C0027627 Metastatic
	C0278488 Metastatic Breast Cancer
	C0034897 Recurrence
	C0278489 Spread Ductal
	C0027628 Spread to Lung
	C0441771 Stage III
	C0441772 Stage IV
	C2348819 Triple Negative
	C0438105 Tumor Recurrent
	C0278488 Tumor Spread

Table 3.2: Table of administrative variables and primary phrases used in NLP.

3.3 Hierarchical Machine Learning (HML) Approach

The HML approach allows for the classification model to train on a smaller subset of patients that are identified as positive for having recurrence. This approach helps deal with the class imbalance by reducing the number of negative recurrence notes by removing them at the patient level.

Other approaches that were considered and tested include undersampling techniques and balanced bagging techniques. In both these cases we noticed significantly less performance compared to using the HML approach to reduce the class imbalance and focus the model for identifying the positive recurrence notes.

The HML architecture is seperated into two levels, the patient level and the note level. The purpose of this division is to separate the tasks of predicting if a patient is positive and the task of identifying if a note mentions that the patient has recurrence.

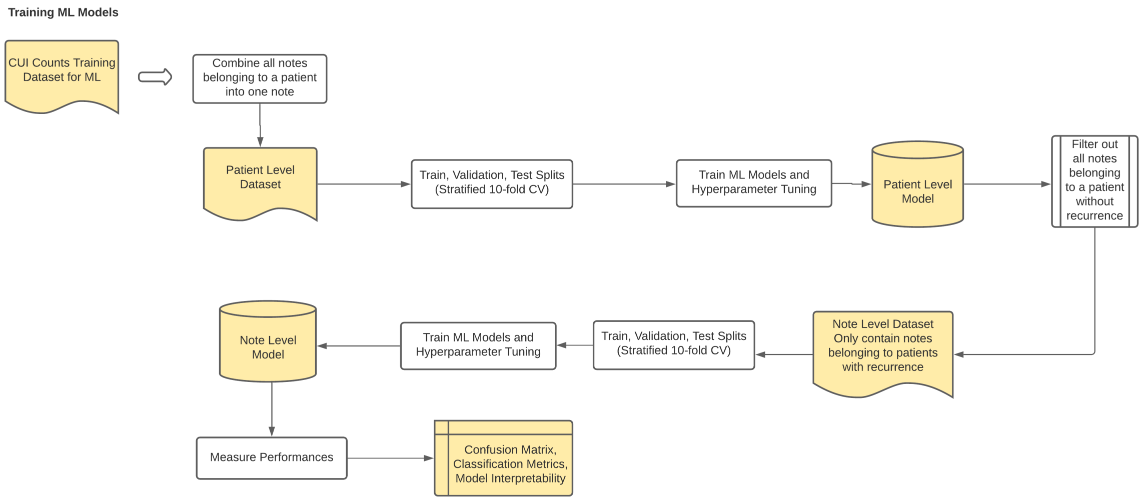
Training ML Models
CUI Counts Training Dataset for ML
Combine all notes belonging to a patient into one note
Patient Level Dataset
Train, Validation, Test Splits (Stratified 10-fold CV)
Train ML Models and Hyperparameter Tuning
Patient Level Model
Filter out all notes belonging to a patient without recurrence
Note Level Dataset Only contain notes belonging to patients with recurrence
Train, Validation, Test Splits (Stratified 10-fold CV)
Train ML Models and Hyperparameter Tuning
Note Level Model
Measure Performances
Confusion Matrix, Classification Metrics, Model Interpretability

Figure 3.1: The architecture of the hierarchical machine learning pipeline.

3.4 Natural Language Processing (NLP) for Analyzing Clinical Notes

cTAKES is the base for the NLP pipeline used in this thesis. cTAKES accepts plain text input. In our pipleline, we used the following cTAKES components/annotators:

- Sentence boundary detector
- Tokenizer
- Normalizer
- Part-of-speech (POS) tagger
- Named entity recognition (NER) annotator (including negation annotators)

The cTAKES tokenizer is made up of two components. The first divides the internal text stream of a sentence based on space and punctuation. The context-dependent tokenizer, on the other hand, merges tokens to produce date, fraction, measurement, person title, range, roman numeral, and time tokens by applying rules (implemented as finite state machines) to each of these forms.

Normalization is a process of mapping several instances of the same word in the input data that don't have the same string representations. The cTAKES normalizer is a wrapper around the SPECIALIST Lexical Tools component "norm," which generates a representation for each word in the input text that is normalised with respect to a variety of lexical properties, such as "alphabetic case," "inflection," "spelling variants," "punctuation," "genitive markers," "stop words," "diacritics," "symbols," and "ligatures." This off-the-shelf normalizer is used to boost the recall of the NER annotator. Every word in the text is normalized, and the dictionary look-up mentioned below uses both normalised and non-normalized forms [21].

Within a noun-look-up window, the cTAKES NER component implements a terminology-agnostic dictionary look-up algorithm. Each named entity is mapped to a terminology definition using the dictionary look-up. We use a dictionary that is a subset of UMLS, version 2008AB, with SNOMED CT and a controlled vocabulary unique for breast cancer recurrence that is used by BR oncologists and clinicians after detailed consultations with CCMB researchers and practitioners. As described in the dictionary, each word is a member to one of the semantic types: disorders/diseases with a separate group for signs/symptoms, procedures, anatomy, the latter includes terms from the Orange Book that have an RxNORM code [21].

Synonyms from UMLS and a list of customized words were added to this dictionary. The algorithm seeks all noun phrases based on the shallow parser's output, which becomes the look-up window. To allow for non-lexical differences, the dictionary is searched for permutations of variations of the head and modifiers within the noun sentences. When defining several words in the same text span, the NER component does not overcome ambiguities. The NegEx algorithm, which is a pattern-based method for identifying terms and phrases indicating negation close to named entities mentioned is implemented by the negation annotator. For identifying appropriate terms and phrases that signify the state of a named entity, the status annotator takes a similar approach.

Each named entity discovered belongs to one of the dictionary semantic categories and has attributes for (1) the text span associated with the named entity ("span" attribute), (2) the terminology/ontology code the named entity maps to ("concept" attribute), (3) if the named entity is negated ("negation" attribute), and (4) the text span associated with the named entity ("concept" attribute) (4) the status associated with the named entity with a value of current, history of, family history of, possible ("status" attribute) [21].

These semantic types and their characteristics were chosen in conjunction with clinical researchers and physicians who are investigating a set of clinical questions and retrieval queries, which may include diseases, clinical drugs, signs and symptoms, and procedures which were the most frequently used UMLS types and classes. Since every potential occurrence is regarded as speculative, the status value will be set to "possible." Allergies of a certain drug are treated by setting the medication's negation attribute to "is negated."

3.5 Patient Level Modelling

The purpose of the patient level modelling was to reduce the scope for downstream classification tasks to only the patients who had cancer recurrence. This tiered modelling approach helps to reduce some of the class imbalance by removing non-cancer recurrent patients from the cohort. One of the main concerns with using patient level modelling is that misclassifications of a positive patient as negative is detrimental since that patient will not be seen in downstream classification tasks and removes vital information. Therefore, the main consideration for the patient level modelling is to ensure that false negatives are avoided as much as possible. This can be enforced using multiple methods such as imposing high costs for false negatives or using data sampling techniques.

3.6 Note Level Modelling

The purpose of the note level modelling was to predict if a clinical note mentions the patient was positive for cancer recurrence. This model was used to find the data of cancer recurrence for a particular patient by analyzing all of their notes and finding the first note mentioning positive cancer recurrence. In this modelling task, the performances for both positives and negatives must be adequate in order to identify the date of cancer recurrence accurately.

3.7 Model Evaluation Procedure

The model evaluation is separated into two categories. Firstly, the model is trained and evaluated on the training dataset using stratified 10-fold cross validation. In this method we divide the dataset into 10 subsets with equal class balance in relation to the overall training dataset. Then, the training is performed on nine of these subsets and the model is tested on the final subset. This procedure is conducted ten times such that each subset has been used as a testing subset once.

The second evaluation method employed is to test on the external validation dataset which is a dataset collected at another time period from the training dataset from the same site. This dataset has some nuances that make it challenging for prediction when only the training dataset is used to build the model. In this evaluation method, the best model and hyperparameters are selected using the stratified 10-fold cross validation from the training dataset. This model is then trained on all of the training dataset and tested on the external validation dataset. The approach reduces bias in the model performance and improves our confidence in the reproducibility of the results [37].

Chapter 4

Natural Language Processing (NLP) Architecture for Breast Cancer Recurrence Detection

4.1 NLP Pipeline

The NLP system consists of three main modules: **prepossessing** module; **NLP** module and **Rule** module. The process pipeline is depicted in Fig 4.1.

The function of the **prepossessing** module is to remove the information that was not tagged as correct due to the semi-structured format of the notes. We tracked the number of repetitions of the sentences between the patients and only removed the sentences that were not connected with a filled box and can be observed for multiple patients. The logic behind this policy is that, in structured data, the sentences need to be repeated between the patient forms and cannot change over multiple patients. Therefore, unique sentences are intentionally added by a physician for a specific patient.

For the extraction of information from the unstructured notes, we developed an **NLP** module with the open-source Apache clinical Text Analysis and Knowledge Extraction System (cTAKES) [21]. cTAKES is a natural language processing tool, whose components are specifically trained for the clinical domain. Our NLP pipeline was composed of (i) A *concept coding* module where we tried to identify terms of interest for which a concept unique identifier (CUI) exists in UMLS [38]. In this module, the sentences were split into tokens and the tokens were normalized to their based form. A POS tagging sub-module was

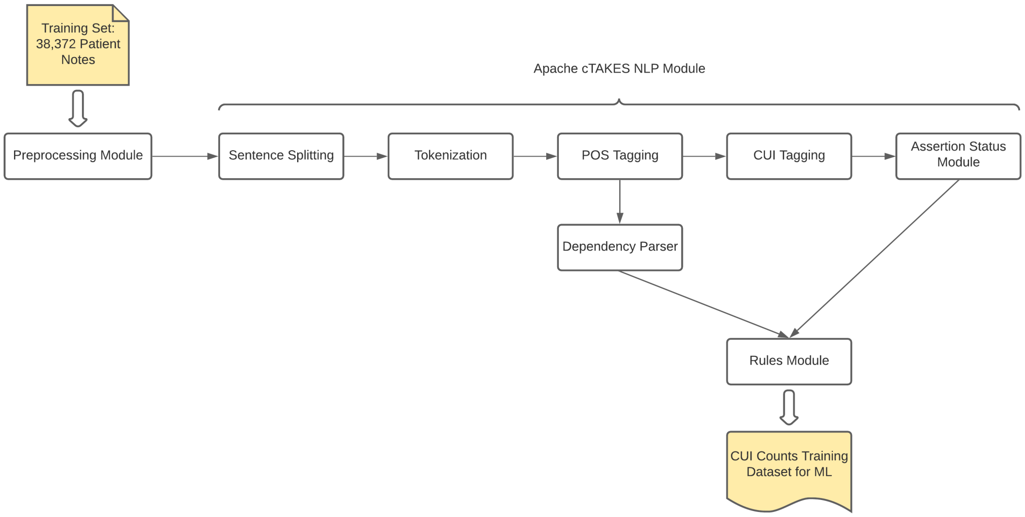

Figure 4.1: Pipeline of the NLP system

then used to tag the normalized tokens to their part-of-speech and finally a named-entity recognition algorithm used this information to identify terms of interest that exist in the UMLS library. (ii) *A*n assertion status annotation module where additional information for each CUI were discovered in the texts, such as whether a term is negated or whether it expresses uncertainty (iii) A *coreference annotation* [39] module which provides information about whether the person of interest for a specific CUI is the patient or a family member or someone irrelevant. Finally, after reviewing the initial results with only the terms from the UMLS dictionary, our expert examined the notes of patients that were flagged as false-negative and he provided new terms that are indicators of a specific risk factor. We created a new custom dictionary with the additional CUI that were determined to be relevant to the identification of different risk factors.

In the **Rule** module, different rules were created in order to determine the existence of a connection between a patient and a specific risk factor. Our first rule was that a patient is diagnosed with a specific risk factor only if all the conditions below were true: (i) A term/s that was/were connected to this risk factor was found in his notes (ii) The term/s was/were not negated (iii) The term/s did not express uncertainty (iv) The subject of the term/s was/were the patient. Finally, we investigated if a term was connected to the following sentences by using the dependency parser, which was provided by cTAKES: (i) can cause, (ii) risk of, (iii) voicing concerns about, (iv) raising concerns about, (v) no prior history of, and (vi) free. We chose these sentences as they express uncertainty or negation but cTAKES failed to identify them. By analyzing the structure of the dependency tree, our system could understand the terms that were connected to the above sentences and if they were indeed connected to one of them, then we did not consider these terms as proof that a patient is connected to a specific risk factor. Python 3.7 was used for the creation of this module and the prepossessing module.

The NLP architecture uses the Apache cTAKES as the core processing engine to generate counts of tags present in the clinical notes. In order to use the Apache cTAKES engine there are several preprocessing steps used to arrange the data for processing. Afterwards, the pre-processed data is sent to the Apache cTAKES engine in a sequential order. Each note is processed and the output of the NLP engine is stored in an xml file that contains the tags identified, uncertainty information for each tag and polarity information for each tag (i.e., negation of a tag).

The Apache cTAKES engine comprises modular components that can be included or excluded depending on the requirements. Some of these components include identification of subject, polarity and negation. In addition, included phrases from a custom dictionary is possible to extend the tagging capabilities and provide unique tags that may benefit downstream classification tasks. All components that are added to the Apache cTAKES engine are processed individually for each note and the output of that component is added to the xml output for each processed note.

Once all notes have been processed the xml files are analyzed to generate a datas t that can be used to build a classification model. The post-processing procedure consists of parsing the xml files and generating a tabular dataset with one patient per row and the counts for each tag for the columns. This procedure has several steps to ensure the integrity of the tags including identifying negation of individual tags, filtering phrases that have high uncertainty (e.g., “risk of” or “concerns about”) and verifying the tag is relevant to the patient (i.e., differentiate family history compared to patients’ current medical condition).

4.2 Custom Dictionary of Phrases for Breast Cancer Recurrence

Custom dictionaries were used to extend the vocabulary for breast cancer recurrence. These phrases in the custom dictionaries were collected from medical experts that analyze notes on a regular basis. The appearance of these phrases were tracked using Concept Unique Identifiers (CUIs). Multiple phrases can be assigned to a single CUI such that various forms of a phrase can be linked to a single CUI. For example, the CUI breast_cancer_recurrence can be used to track multiple phrases such as “breast cancer recurrence”, “breast cancer recurrent” or “br ca recur.”

Table 4.1 shows some of the custom dictionary terms that were used. The first column indicates the CUI and the second column indicates the phrases identified for that particular CUI.

Concept Unique Identifiers (CUIs)	**Custom Phrases**
C0007097: cancer	adenoca
	adenocarc
	adenocarcinoma
	ca
	cancer
	carc
	carcinoma
C0006142: breast cancer	adenoca br
	adenoca breast
	adenoca dctl
	adenoca ductal
	adenoca lobular
	adenocarc br
	adenocarcinoma br
	br ca
	br cancer
	ductal adenoca
	ductal adenocarc
	ductal adenocarcinoma
	lobular adenoca
C0027627: metastatic	met
	metastatic
	metastases
	metastasis
	metastasize
C0034897: recurrence	recur
	recuring
	recurrence
	recurrent
	recurring
	recured
	recurred
	recurrnet
	re currnet

Table 4.1: Examples of CUIs and custom phrases used in the NLP tagging, the complete dictionary can be found in the Appendix.

4.3 Identifying Negation, Uncertainty and Relation

One of the concerns in the NLP tagging process is identifying certain status modifiers that may change the interpretation of the sentence. For example, some of these status modifiers are negation of phrases, uncertainty (probabilistic keywords such as "likely" or "possibly") and relation of phrases (if the phrase pertains to family, patient or other). These status modifiers can completely change the understanding of a sentence, therefore it is critical to record and handle them appropriately. To indicate a negated tag, "∼" is prepended to the CUI. For cases with probabilistic keywords or relationship of phrases to members other than the patient, the CUI is not recorded.

> *COMMUNITY CANCER PROGRAM NOTE Oct 13, 2006 Patient of Dr. X with a T3 N0, ER/PR positive, **HER-2** positive left **breast carcinoma** completing FEC/Taxotere in May 2006 and then radiation because of the T3 tumor size in August 2006.*
>
> **"HER-2"** generic for patient
> Receptor
> C0069515 HER-2_C0069515
> **"breast carcinoma"** for patient
> Disorder
> C0006142 breast cancer_C0006142
> **"breast"** generic for patient
> Body Location
> C1268990 breast_C1268990
> **"carcinoma"** for patient
> Disorder
> C0007097 cancer_C0007097

Chapter 5

Classification of Breast Cancer Recurrence

5.1 Classification Algorithms

5.1.1 Algorithm Level Techniques

There are several approaches to dealing with the class imbalance problem at the algorithm level. One commonly used strategy is to apply bias to improve the classification performance on the minority class [40]. Each algorithms has different hyper-parameters that can be adjusted to improve the performances for the class imbalance scenarios. For example, decision trees can apply different technique for pruning to improve performance on the minority class. Similarly, methods based off the decision tree such as AdaBoost, XGBoost and Random Forest can benefit from these techniques.

5.1.2 Data Level Techniques

The data-level techniques to improve performance in the class imbalance scenario, a number of sampling methods were utilized as the following:

1. Random Undersampling [41]
2. Synthetic Minority Over-sampling Technique (SMOTE) [42]

Undersampling includes sampling the majority class such that the number of datapoints are in comparison with the minority class. Near miss Undersampling is about sampling the datapoints from the majority class which aid in discriminating between classes [41].

Generating synthetic data points of the minority class is another approach to improving the class balance. Oversampling SMOTE (Synthetic Minority Over-sampling Technique)-synthetically generates samples by interpolating between samples and original dataset[42].

An important issue in this kind of resampling techniques is what is or how to decide the optimal class distribution in a given dataset. Weiss and Provost studied the effect of a training set's class ratio on a classifier's performance [43]. According to their study, With respect to the classification performance evaluated by AUC, a balanced class distribution with a 1:1 ratio could achieve a good performance, yet not necessarily optimal. Optimal class distributions differ from data to data, and in some cases, depends on the perceived cost of mis-classification of a certain class.

Another important issue is how to effectively re-sample the training data. Random sampling is straightforward to implement, yet may not be sufficient in many cases. In our case of a bi-class classification (i.e.recurrence or no recurrence), the recurrence class is the minority class, and the vast number of "no recurrence" class is the majority class. The characteristics of recurrence is far more important in the modelling, thus a randomly under sampling of the majority class is deemed more favorable solution. However, such an informative resampling process increases the cost for data analysis. With a very large number, and variation of notes that contains no recurrence information (no recurrence class), it is difficult to set the criterion in selecting samples.

We experimented techniques including random oversampling the minority class, and random undersampling the majority class and using a combination of both these methods. The undersampling technique produced the best results, and the model performance using this technique is included in Chapter 6.

5.2 Design and Training of Patient-Level and Note-Level Models

The modelling started with data preprocessing to understand the class distribution of the clinical notes. The positive recurrence patients are of higher significance since the objective is to identify breast cancer recurrence. The dataset was filtered to examine positive recurrence patients, followed by stratification on note level recurrence (i.e., positive and negative notes). The stratification is show in Table 5.1. This table shows that 20.62% and 15.93% of patients were identified as positive for recurrence in the training and validation datasets, respectively. Among these groups the positive notes make up a very small fraction of 4.33% and 3.80% of the total number of notes for the training and validation datasets, respectively.

Dataset	Number of Patients, n(%)	Number of Notes, n(%)
Training Dataset	897 (100)	117365 (100)
Positive Patients	185 (20.62)	38372 (32.69)
Positive Notes	-	5082 (4.33)
Negative Notes	-	33290 (28.36)
Validation Dataset	615 (100)	78460 (100)
Positive Patients	98 (15.93)	21071 (26.86)
Positive Notes	-	2985 (3.80)
Negative Notes	-	18086 (23.05)

Table 5.1: The training and validation dataset statistics stratified by recurrence status.

Due to the low number of positive notes relative to the total number of notes, data level and algorithm level class imbalance techniques were utilized to improve performance of prediction on the minority class. In addition, appropriate trade-offs were made to minimize the number off false negatives at the cost of false positives. One such example of implementing the trade-offs is the use of specific performance metrics such as the F-2 score for model selection. The F-2 score places higher importance on the sensitivity over precision compared to the F-1 score. In the case of breast cancer recurrence identification, the cost of failing to identify a positive case (false negative) outweighs the cost of misidentifying a negative case as positive (false positive). These concerns were taken into consideration and the results are described in Chapter 6.

Chapter 6

Results

6.1 Model Training and Selection

The testing method used was stratified holdout testing with 80% for training and 20% for testing. The distribution and class balance for the training and testing sets are shown in Figure 6.1. The orange represents the positive recurrence notes and the blue represents the negative recurrence notes. Since the sets are stratified they both have the same class balance ratio of 1:6.5 (positive to negative ratio).

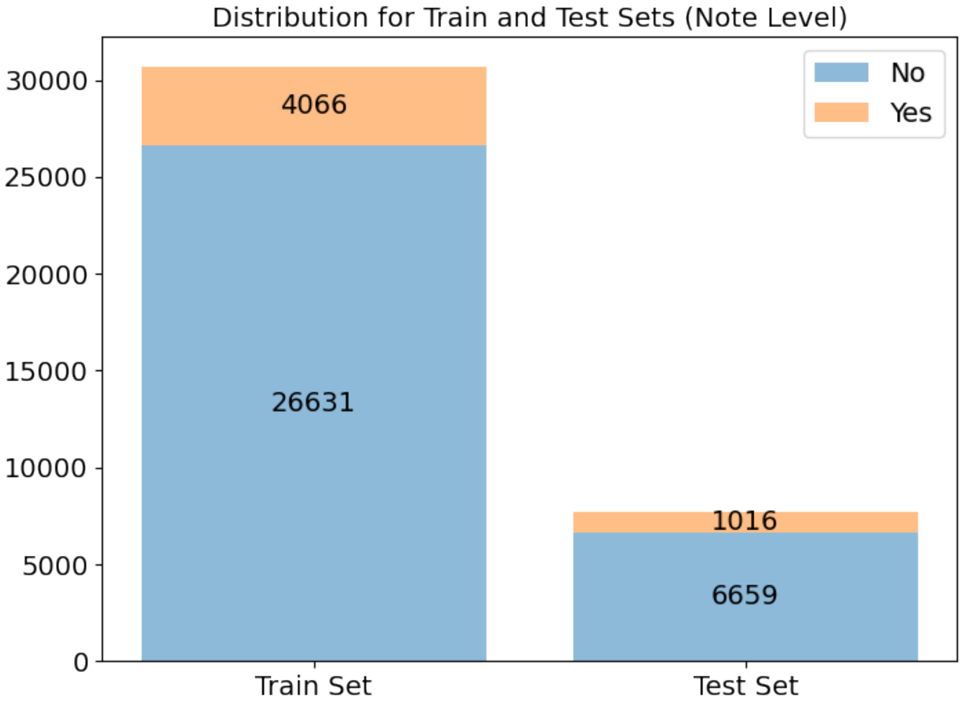

Figure 6.1: The train test distribution with 80% of the dataset used in training and 20% holdout for testing.

The performances of the machine learning models were assessed using a set of standard classification metrics. These metrics include balanced accuracy, sensitivity, specificity, AUC, PPV and NPV. The classification models that were evaluated are Decision Tree, Random Forest, AdaBoost, XGBoost and Logistic Regression. These classification algorithms have some hyperparameters that need to be tuned based on the dataset to achieve good performances. In order to perform this task, the stratified 10-fold cross validation technique is used on the training set. This strategy splits the training set into 10 splits that have equally proportional class balance. One split is used as the test set while the other nine sets are used for training. The performances are computed on the test set then repeated again 10 times, in each iteration a different set is used for testing until each set has been used at least once for testing. The benefit of this method is that it minimizes overfitting on the training dataset by using multiple different subsets of the training dataset to find appropriate hyperparameters.

The stratified 10-fold cross validation is repeated with different sets of hyperparameters for each of the classification algorithms. To select the best hyperparameters a selected classification metric is used to compare performances between the different sets of hyperparameters. In this case, the f-beta measure is used with beta=2. The f-beta metric takes into consideration the sensitivity and precision which are two important metrics in breast cancer recurrence classification. The beta value is used to control the weight of sensitivity over precision or vice versa. Beta values between 0 and 1 put greater weight for precision whereas beta values higher than 1 place higher weight on sensitivity. Once the best hyperparameters are selected from the training set the model is evaluated on the test set (unseen data).

The classification performances of the models using the training dataset with holdout of 20% is shown in Table 6.1. Finally, the best model is chosen using the f-beta measure with beta=2. This model performs well across the other classification metrics as well.

Model	Bal. Acc.	Sens.	Spec.	PPV	NPV	AUC	F-1	F-2
DT	0.924	0.889	0.959	0.785	0.981	0.924	0.833	0.866
RF	0.915	0.861	0.970	0.825	0.977	0.915	0.842	0.853
XGB	0.924	0.867	**0.981**	**0.886**	0.978	0.924	**0.876**	0.871
LR	0.932	0.902	0.961	0.795	0.983	0.932	0.845	0.878
AdaBoost	0.931	0.890	0.972	0.840	0.981	0.931	0.864	0.879
Undersampling XGB	**0.940**	**0.969**	0.911	0.644	**0.994**	**0.940**	0.774	**0.880**
ROSE XGB	0.932	0.954	0.909	0.636	0.992	0.932	0.764	0.868

Table 6.1: Balanced Accuracy, Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), ROC AUC, F-1 and F-2 classification measures of the models tested on the 20% holdout of the **training dataset**. Bolded are the best performances for each measure.

		actual outcome		
		P	**N**	**total**
prediction outcome	**P**	965	152	1117
	N	150	6543	6693
	total	1115	6695	

Figure 6.2: The confusion matrix for the XGBoost model on the 20% holdout of the **training dataset**.

The XGB classifier was further analyzed with the use of ROC AUC curves. This plot is shown in Figure 6.3. This plot was generated using 10-fold stratified cross-validation, the result from each iteration of the train-test splits is shown in the light colours and the mean AUC is represented by the thicker blue line.

The calibration curve is one method to indirectly measure the performance of the model with respect to the population. Figure 6.4 shows the calibration curve for the XGBoost classifier. There were two calibration techniques applied to the XGBoost classifer to improve the prediction probability outputs from the model, these techniques are Isotonic Regression and Platt scaling with Sigmoid function.

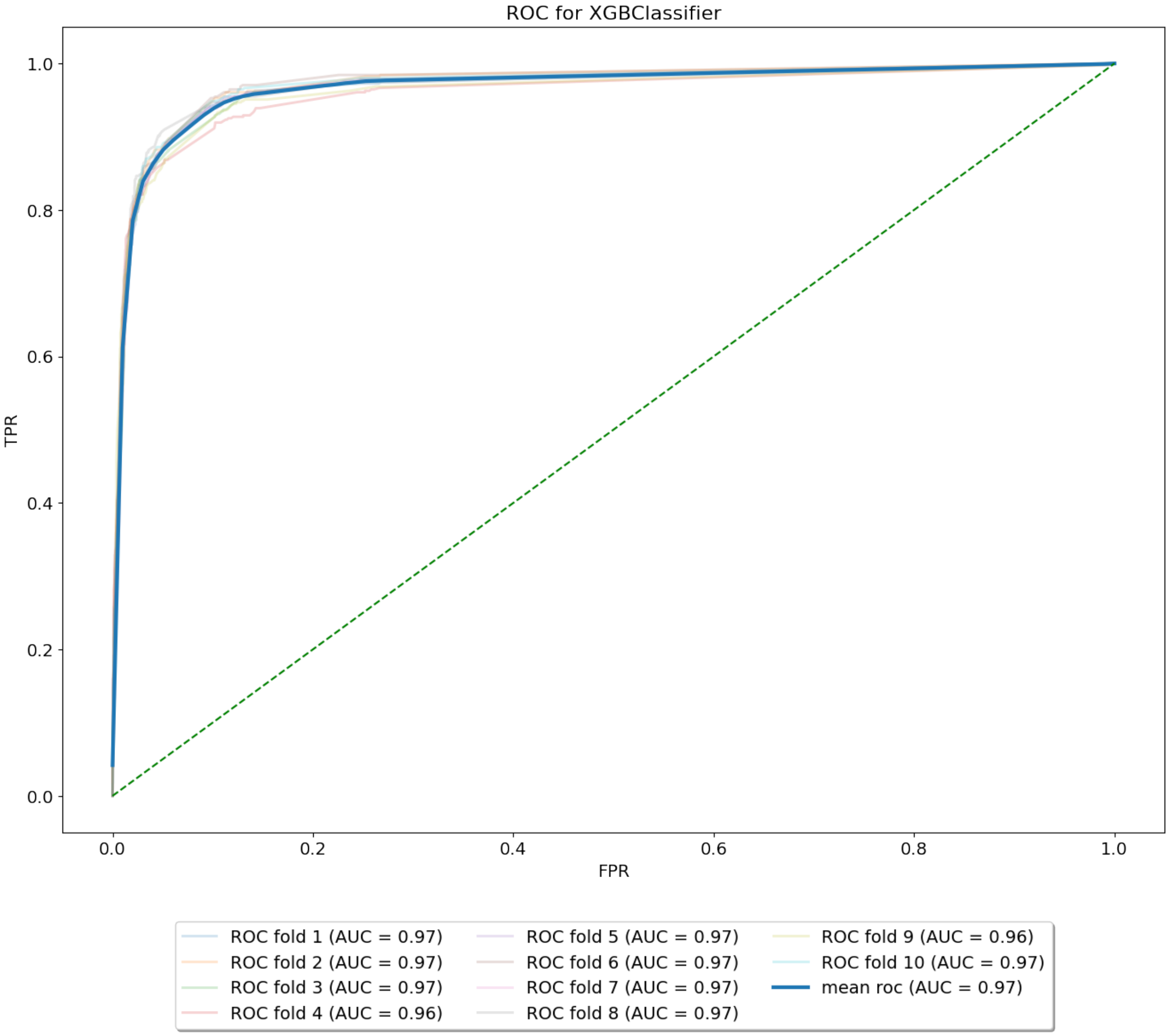

Figure 6.3: The AUC plot for the XGBoost classifier.

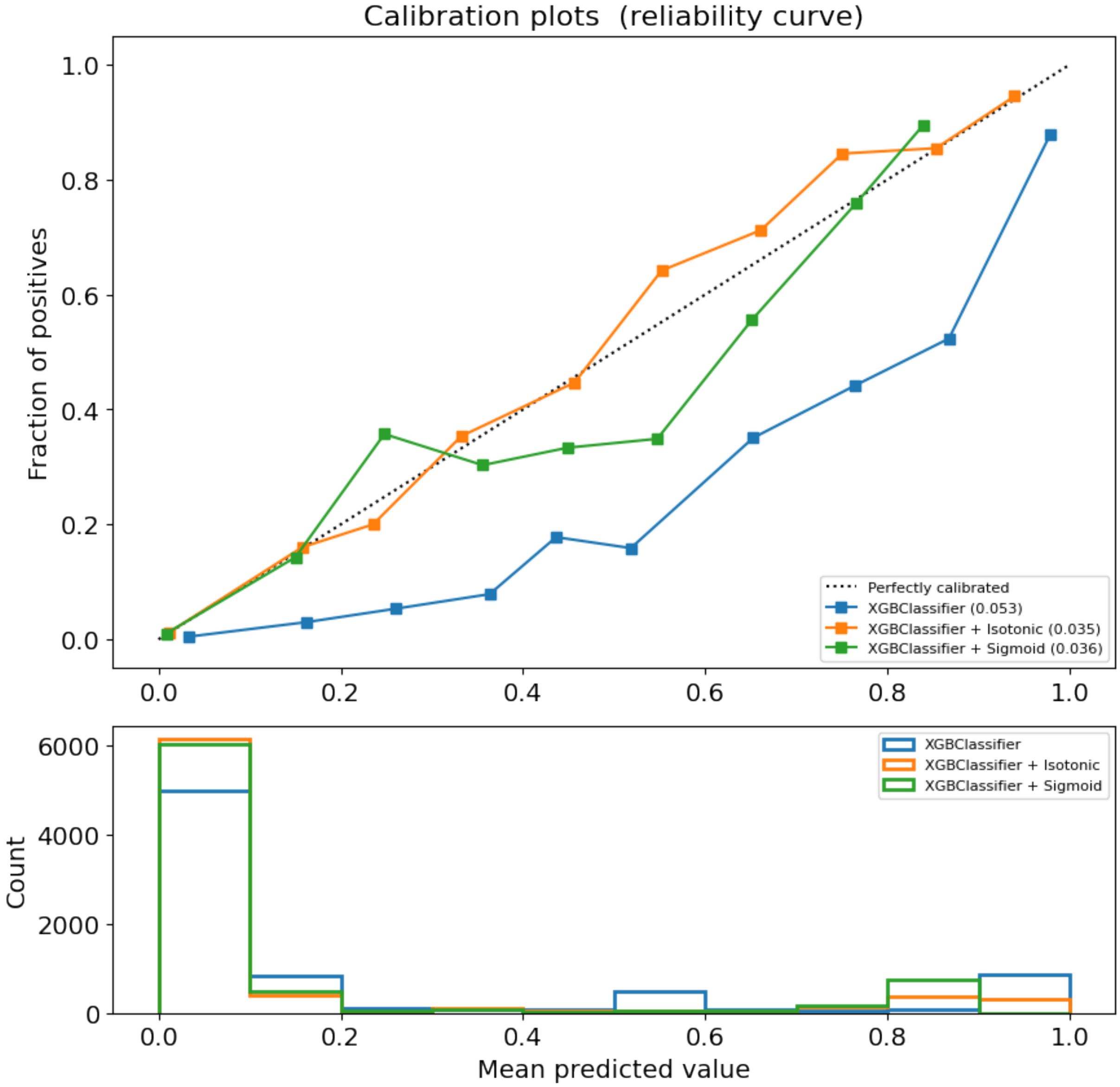

Figure 6.4: The calibration plot for the XGBoost classifier.

6.2 Model Interpretability and Explainability

In order to better understand the models, several model interpretability techniques were used. These techniques aim to provide clarity in which features are most significant for prediction and to understand the relative contributions of individual features for model predictions [44]–[47].

The first model interpretability technique used is a model specific method where the XGBoost classifier is used to generate feature importances. In this method, the feature importances are based on the internal model structure of the XGBoost classifier. The XGBoost classifier is an ensemble technique that utilizes groups of decision trees. Decision trees are considered inherently interpretable in the literature, whereas ensemble methods such as the XGBoost model is not. However, since the XGBoost model is a combination of decision trees, there are methods to compute the average contribution of each feature to the prediction. One such metric is called gain, this is computed by comparing the improvement in accuracy when a new split is created on a feature in each decision tree. The feature importance based on the XGBoost model is reported in Figure 6.5.

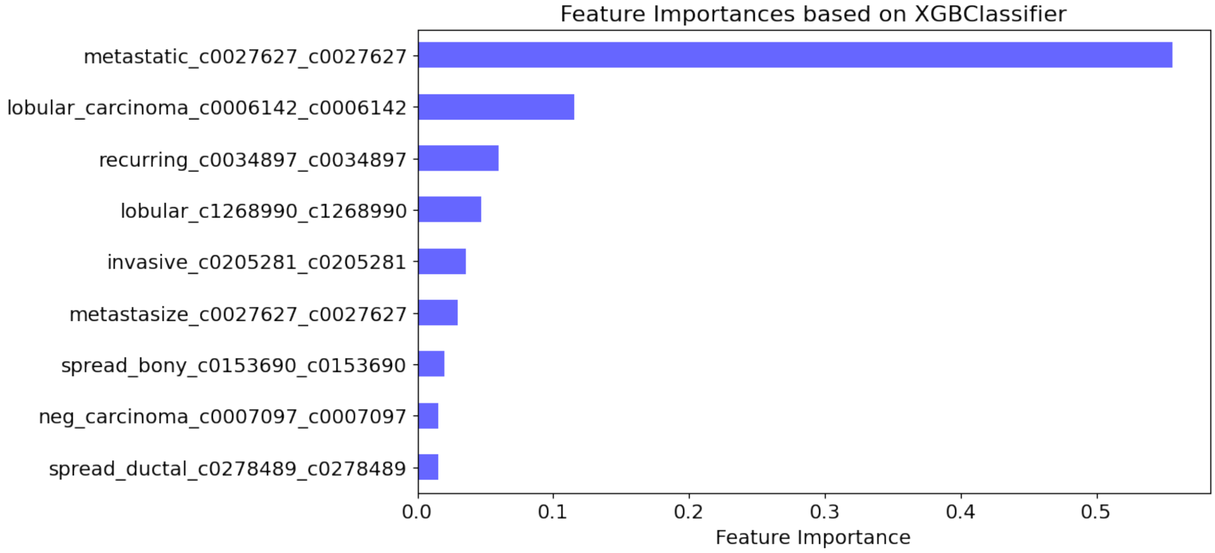

Figure 6.5: The feature importance plot created using the XGBoost classifier.

One of the model agnostic methods for interpretability is called Shapley Additive Explanations (SHAP) [48]. The SHAP framework calculates the contribution of each feature to the prediction of individual samples (patient notes). Shapely values are used to indicate the contribution of a feature to the model output. Negative values indicate a feature that on average "pushes" the model output towards negative and positive values indicate the opposite. These Shapely values are calculated based on the average marginal contribution of a feature value over all possible coalitions. The SHAP feature effects plot is shown in Figure 6.6. This plot is generated by plotting the Shapely values for each sample by feature (along the horizontal). The colours along the horizontal indicate presence of the feature (red) and absence of the feature (blue). The features are ranked based on mean absolute value of SHAP values which is an indicator of overall feature importance.

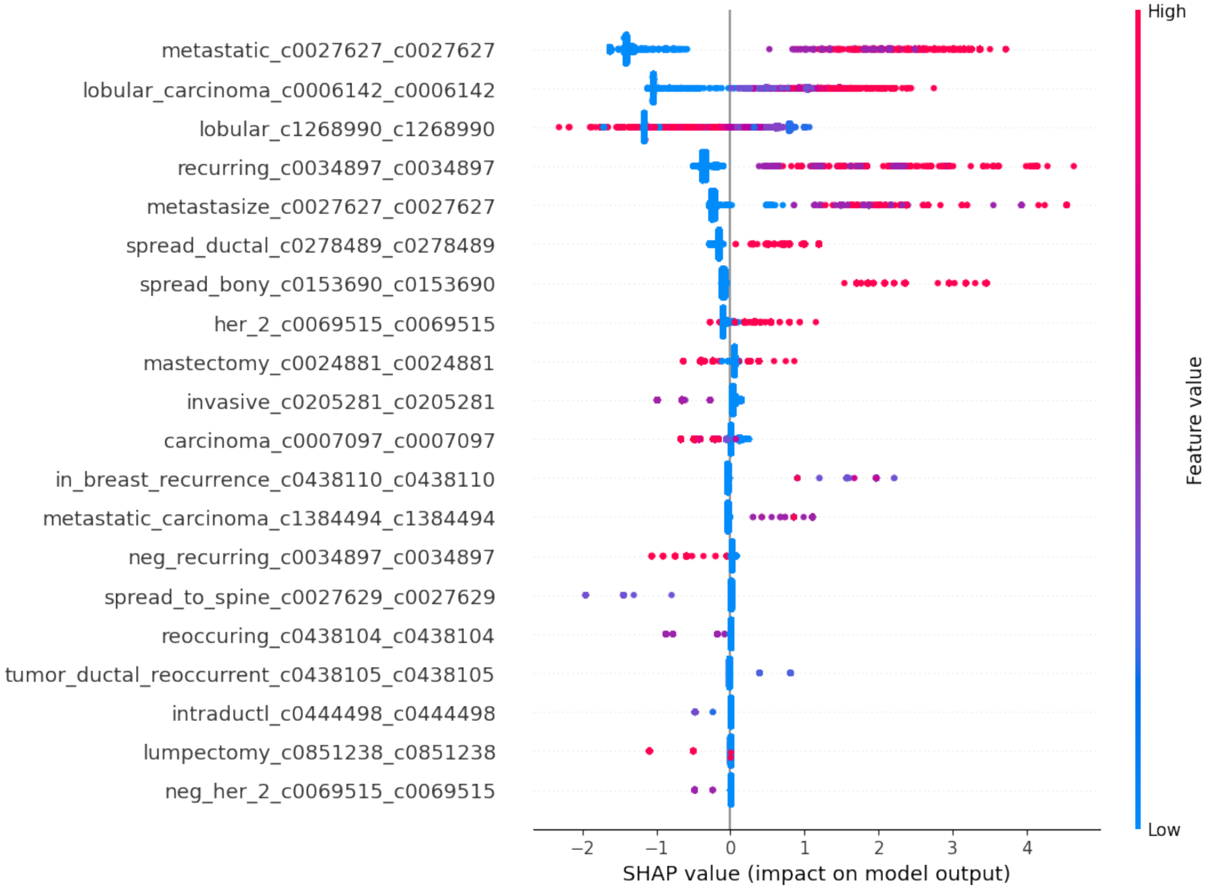

Figure 6.6: The SHAP feature effects plot that indicate contribution to positive (red), negative (blue) outcome based on each feature.

The overall feature importance based on the SHAP framework is presented in Figure

6.7. The overall importance is computed by using the mean absolute average of Shapely values. This plot provides an idea of the relative feature importance among the ten most significant features.

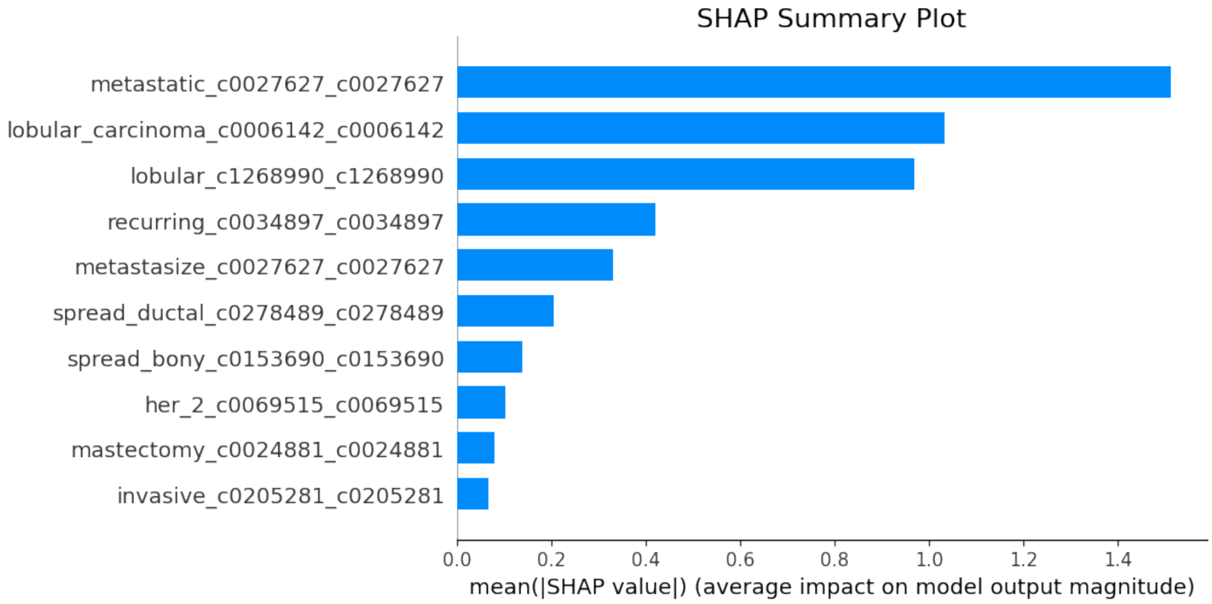

Figure 6.7: The SHAP feature importance summary plot computed by using the mean absolute Shapely values.

Permutation Feature Importance (PFI) [49] is the importance of a specific feature measured by calculating the increase in the prediction error of the model after the feature's values are permuted. First, the dataset is split into training-testing split of 80%/20%. The training set is used to train the model and find the best hyperparameters using stratified 10-fold cross validation. Next, by using the trained model with all the features, we calculate the baseline accuracy as the average prediction score in the specified test set. Afterwards, iteratively, the values of different chosen features are shuffled on the test set and the average prediction score of the previously trained model on the modified dataset is calculated. Finally, the importance of each feature is calculated as the reduction of their score to the baseline accuracy. The results of the analysis of the PFI algorithm on the features of the dataset are reported in the Table 6.2.

Feature	PFI Weight
metastatic_c0027627	0.0370 ± 0.0023
lobular_c1268990	0.0177 ± 0.0029
metastasize_c0027627	0.0152 ± 0.0012
recurring_c0034897	0.0152 ± 0.0012
lobular_carcinoma_c0006142	0.0119 ± 0.0017
metastatic_carcinoma_c1384494	0.0033 ± 0.0004
spread_bony_c0153690	0.0019 ± 0.0008
spread_ductal_c0278489	0.0015 ± 0.0006

Table 6.2: The Permutation Feature Importance (PFI) ranking of the ten most significant features in descending order.

The various interpretability techniques used were combined and the features ranking of each of these methods were compared to understand the similarities and differences. The feature rank table is shown in Table 6.3. The ten most important features for each method are displayed. The most important feature that was consistent across all methods was metastatic_c0027627 (except for logistic regression). The importance of this feature is known intuitively as the mentioning of metastatic cancer is an indication of breast cancer recurrence and therefore is a very strong predictor if this tag is identified in the note. Other features that were found to be strong predictors include recurring_c0034897 and lobular_carcinoma_c0006142. These also make intuitive sense because the presence or absence of these terms indicate if the notes are discussing the patients conditions relating to breast cancer recurrence.

Features	DT	RF	XGB	LR	AdaBoost	SHAP	PFI
metastatic_c0027627	**1**	**1**	**1**	**6**	**1**	**1**	**1**
lobular_c1268990	5	5	7	-	2	3	2
recurring_c0034897	3	8	4	1	4	4	4
lobular_carcinoma_c0006142	2	3	2	-	3	2	5
metastatic_carcinoma_c1384494	-	-	-	2	-	-	6
spread_bony_c0153690	-	9	8	4	7	7	7
spread_ductal_c0278489	6	2	3	5	6	6	8
her_2_c0069515	-	-	-	-	-	8	-
metastatic_lobular_carcinoma_c0278488	-	4	10	-	9	-	9
mastectomy_c0024881	-	10	-	-	10	9	-
invasive_c0205281	-	-	-	7	8	10	10
in_breast_recurrence_c0438110	-	-	6	8	-	-	-
neg_carcinoma_c0007097	-	-	9	3	-	-	-
carcinoma_c0007097	-	6	-	-	-	-	-

Table 6.3: Ranking of the most important feature for each model and interpretability methods. The SHAP interpretability method was run using the XGB model. Bolded is the most significant feature across most methods.

6.3 Model Evaluation on Validation Dataset

The model constructed on the original dataset is evaluated on an unseen external validation dataset. This approach is used to ensure that the model constructed on the training set is capable of performing with similar results on unseen data that may be encountered if the model is deployed in a working environment. The main objective is to determine if the model can perform consistently on the external validation set.

In order to determine if the model performs consistently, several techniques are used. The classification metrics computed in Chapter 5 are reported for the external validation set. The confusion matrix is reported to identify proportions of misclassifications. The ROC AUC plot and calibration curves are used to visualize the performance of the models. In particular, the calibration curve is used to get a better understanding of how well the prediction score output from the model translates to a confidence level of its predictions.

6.4 Results on Validation Dataset

The external validation dataset was collected from a different time period from the same institution. There are slight differences to these notes in comparison to the original dataset. Primarily, the formats of these notes vary slightly as the structure of some of these notes have changed. Some of these notes have semi-structured information with headings indicated within the text such as “FAMILY HISTORY:” or “PAST MEDICAL CONDITION:”.

This external validation dataset provides a way to test whether the model constructed on the original dataset is capable of performing well on brand new data that may contain previously unseen concepts. This form of validation is commonly used in the medical research field to ensure that models are capable of performing in new environments without major issues in performance.

The best performing model that was found on the training dataset was used on the validation dataset to observe how well it performs on unseen data. The results are shown in Table 6.4.

		actual outcome		
		P	**N**	**total**
predicted outcome	**P**	2689	1946	4635
	N	436	73389	73825
	total	3125	75335	

Figure 6.8: The confusion matrix for the evaluation of the XGBoost model on the **validation dataset**.

Model	Bal. Acc.	Sens.	Spec.	PPV	NPV	AUC	F-1	F-2
DT	0.915	0.887	0.943	0.392	0.995	0.915	0.544	0.708
RF	0.876	0.788	0.964	0.475	0.991	0.876	0.593	0.697
XGB	0.909	0.843	**0.974**	**0.575**	0.993	0.909	**0.684**	**0.771**
LR	0.919	0.890	0.948	0.416	0.995	0.919	0.567	0.725
AdaBoost	0.921	0.880	0.962	0.489	0.995	0.921	0.628	0.758
Undersampling XGB	**0.927**	**0.919**	0.935	0.368	**0.996**	**0.927**	0.526	0.707
ROSE XGB	0.925	0.897	0.953	0.440	**0.996**	0.925	0.590	0.743

Table 6.4: Balanced Accuracy, Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), ROC AUC, F-1 and F-2 classification measures of the models tested on the entire **validation dataset**. Bolded are the best performances for each measure.

Chapter 7

Discussion

This study developed the model for breast cancer recurrence detection using a hierarchical architecture, combining natural language processing (Apache cTAKES) and machine learning classification techniques. This study's strength partially lies in the training datasets that combine administrative data and expert chart-reviewed clinical notes with large size. Notably, the algorithms consider syntactic and semantic variations of documenting cancer recurrence in a real-world clinical setting. Most deep learning NLP models are difficult for a modeller or a clinician to understand how the classification of recurrence or non-recurrence is determined. On the contrary, our algorithms can identify the most impactful UMLS concepts and controlled vocabulary. We can easily exam the confusion matrix and identify specific language variables or clinical decision logic in the real-world practices that cause the model to perform poorly. Thus the algorithms can be engineered to adopt local variance better.

7.1 HML Approach to Breast Cancer Recurrence Identification

The HML approach proves to be successful in providing higher quality predictions however it does have some drawbacks. The HML approach introduces a two step process to the prediction, firstly the patient based on all their notes is predicted as positive or negative for recurrence. If the patient is positive, then their notes are assessed to identify which of their notes mention information about positive breast cancer recurrence. If the patient is predicted as negative no further predictions are made on that patient. Due to this definitive boundary between positive and negative recurrence patients, it is vital to achieve very high sensitivity in the patient prediction model. If patients are misclassified as negative, it incurs a very high penalty on the overall performance.

One of the benefits of this approach is that it is easier to filter negative patients from the cohort since many of their notes do not typically include keywords for breast cancer recurrence patients. However, in such cases where negative recurrence patients do have some mentioning of recurrence keywords that are out of context of the patient (ex. doctor's explanation of medical terms to patient, past medical history or family history), the patient level model may misclassify them as positive but the note level model may handle these cases better as it is trained to perform better on the individual notes and with greater precision.

7.2 Performances of the Classification Models

Overall many of the decision tree based classification models performed well. The ensemble techniques improved the performances, particularly on the external validation dataset in which the model was tested on unseen data. These results are promising since they indicate potential use of these models in new environments where the data the model was trained with is different from the data used for prediction.

Using the Isotonic Regression technique with the ensemble tree based methods, the classifiers are able to provide prediction scores that are more useful for the end users. The prediction scores can potentially be used as a proxy to judge the classifiers confidence level in its predictions. For example, considering the relative prediction score for a positive recurrence note of 0.8 to a note with score of 0.95. The latter score indicates higher level of confidence in the estimate compared to the former. In both situations the final output would be positive for recurrence however these scores help to understand how confident the model is in the prediction.

7.3 Identifying Inconsistencies

Many datasets have some noise due to errors in data collection or other limitations that prevent perfect data collection. One of the benefit of using the cTAKES and traditional machine learning approach is that it is easier to identify data inconsistencies. The approach used in this thesis has transparent outputs at every stage, aiding in identification of errors and anomalies. For example, the outputs from the cTAKES processing of the clinical notes is human readable and easy to process, this helps to get an idea of underlying data patterns prior to running classification models. This capability is leveraged to identify inconsistent patterns in the dataset collection regarding breast cancer recurrence. Several types of inconsistencies were found with the definition of recurrence and labelling practices. Since these were found and corrected prior to running the machine learning algorithms, it saved time for debugging and improved the quality of the dataset. These inconsistencies are summarized in Table 7.1.

Since multiple coders were involved in the initial data collection phase, there was a misunderstanding of the labeling metastatic cases during the data collection phase. Some of the clinical notes had metastatic breast cancer cases labelled as non-recurrent however these cases should have been labelled recurrent. Using cTAKES the clinical notes were analyzed and those that mentioned phrases related to "metastatic" were filtered and reviewed again to ensure data consistency. Table 7.1 lists the types of inconsistencies found and the counts, using Apache cTAKES.

Dataset Inconsistencies	Number of Notes, n(%)
Training Dataset	117365 (100)
Metastatic Cases	3838 (3.27)
Recurrence Mislabelling	1033 (0.88)

Table 7.1: The types of dataset inconsistencies found using the cTAKES and traditional machine learning approach.

7.4 Comparison with the State of the Art (BERT Model)

The original BERT model [50], based on multi-layer bidirectional transformers [51], can generate contextualized word representations. Incorporating information from bidirectional representations allows the BERT model to capture more accurately the meaning of a word based on its surrounding context (sentence).

In recent years, there have been significant improvements in the field of NLP. In particular the current state of the art is the BERT model which is a transformer-based machine learning model. The BERT model can train on massive amounts of public data to build a pre-trained model that is capable of performing many downstream tasks. Several other advantage of the BERT model includes ease of scalability as the terminology for BERT models are not pre-programmed but rather trained by provided large number of samples such as biomedical research articles (ex. PubMED) or de-identified patient notes from large public dataset (ex. MIMIC dataset).

The classical classification techniques used in this paper do still hold some advantages over the new BERT model architecture. Since the BERT model involves using neural networks, the model is not inherently interpretable. There are current research efforts to building interpretable techniques to better understand outputs from neural networks such as BERT. Also, classical NLP and machine learning methods allow some more flexibility in terms of pre-programing some custom phrases or to handle consistent semi-structure information that may not be easy to train on.

Model	Bal. Acc.	Sens.	Spec.	PPV	NPV	AUC	F-1
BERT (0.02)	0.959	0.962	0.955	0.458	0.998	0.985	0.620
BERT (0.1)	0.949	0.929	0.969	0.543	0.997	0.985	0.686
BERT (0.25)	0.940	0.906	0.974	0.579	0.996	0.985	0.706
BERT (0.5)	0.929	0.879	0.978	0.611	0.995	0.985	0.721
BERT (0.75)	0.912	0.843	0.981	0.643	0.994	0.985	0.729
BERT (0.9)	0.882	0.777	0.986	0.687	0.991	0.985	0.729
XGB	0.909	0.843	0.974	0.575	0.993	0.909	0.684
Undersampling XGB	0.927	0.919	0.935	0.368	0.996	0.927	0.526

Table 7.2: Comparison of the BERT model performances with the traditional ML methods. The number in brackets for the BERT models indicate the cut off values. The table shows the classification metrics: Balanced Accuracy, Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), ROC AUC and F-1 on the **validation dataset**.

Chapter 8

Conclusion

This thesis has demonstrated the ability to use Natural Language Processing and Machine Learning techniques to assist in chart review by 1) excluding a large amount of notes which contain no relevant information, 2) identifying notes most likely contains relevant recurrence information, in order to accurately identify the timing of recurrence. Given the severely imbalanced ratio (1:10) of notes with and without recurrence information, the algorithm is deemed useful for assisting chart review, but not replacing the chart reviewing, nor use as a diagnostic tool for recurrence detection.

Despite high sensitivity and specificity, the model performed at a substantially lower positive predictive value (PPV). This less satisfactory PPV can be attributed to the low prevalence of recurrence in the test data at the note level. One of the future works in this study is to improve the model's PPV. Possible strategies include: 1) continuously enhancing the techniques to identify ambiguous terms and implicit logics when clinicians document their statement or suspicion of a recurrence event in the clinical notes, and 2) incorporating meta-data of the document types to filter out noisy data in the training set. Our clinical collaborators at CCMB observed that certain types of clinical reports/notes, such as those documented by oncology specialists, could contain more comprehensive information with consistent terms when referring to recurrence. This has been identified as the next step to improve the performance and utility of this model in CCMB's practice.

Another limitation of the current approach is the transferability of the model. In addition to the standard UMLS vocabulary, the NLP algorithm utilizes a vocabulary specific to breast-cancer recurrence, with some terms specific for CCMB clinicians. This additional "localized" vocabulary is very helpful to improve the performance of the BCR model, but it also makes the model less performant on a dataset with different diseases, for example, colorectal cancer or lung cancer. Our initial experiment on a BERT-based NLP algorithm has demonstrated superior transferability on different disease datasets. This is a promising new algorithm development venture already started with the CCMB researchers at the time of this writing.

Given the great demand for reliable, population-based measures of cancer outcomes from professional, governmental and researchers, robust recurrence detection algorithms represent a significant advance. To foster adoption of the Breast Cancer Recurrence Detection algorithms as presented in this thesis, we have published the terms specific for breast cancer recurrence identification [1]. This is only one example of many "big data" initiatives for "liberating" information from unstructured data with AI and machine learning techniques to create broader and clinically enhanced real world datasets for research and evidence-based policymaking. These advances are essential because administrative datasets do not encompass all important clinical information about the patient, the context of treatment and disease progression, thus with limited generalizability and lack of relevant cancer outcomes. The machine-learning algorithms and disease-specific vocabularies developed in this study have a potential to become a useful tool in processing real-world data at scale in healthcare. The appropriate application and wider adaptation of machine-learning techniques in healthcare could accelerate system transformation in real-world evidence and personalized healthcare.

[1] The complete dictionary of breast cancer recurrence specific terminology was uploaded to https://github.com/sujan-suzbe/bcr-terminology/.

References

[1] P. A. Jones and S. B. Bylin, “The epigenomics of cancer,” *Cell Press*, vol. 128, no. 4, pp. 683–692, 2007. DOI: 10.1016/j.cell.2007.01.029.

[2] R. W. Carlson, C. Allred, B. O. Anderson, H. J. Burstein, W. B. Carter, S. B. Edge, J. K. Erban, W. B. Farrar, L. J. Goldstein, W. J. Gradishar, D. F. Hayes, C. A. Hudis, M. Jahanzeb, K. Kiel, B.-M. Ljung, P. K. Marcom, I. A. Mayer, B. McCormick, L. M. Nabell, L. J. Pierce, E. C. Reed, M. L. Smith, G. Somlo, R. L. Theriault, N. S. Topham, J. H. Ward, E. P. Winer, and A. C. Wolff, “Breast cancer,” *Official Journal of the National Comprehensive Cancer Network*, vol. 7, no. 2, pp. 122–192, 2009. DOI: 10.6004/jnccn.2009.0012.

[3] M. Colleoni, Z. Sun, K. N. Price, P. Karlsson, J. F. Forbes, B. Thürlimann, L. Gianni, M. Castiglione, R. D. Gelber, A. S. Coates, and A. Goldhirsch, “Annual hazard rates of recurrence for breast cancer during 24 years of follow-up: Results from the international breast cancer study group trials i to v,” *Journal of Clinical Oncology*, vol. 34, no. 9, pp. 927–935, 2016. DOI: 10.1200/JCO.2015.62.3504.

[4] I. Ma, A. Dueck, R. Gray, N. Wasif, M. Giurescu, R. Lorans, V. Pizzitola, and B. Pockaj, “Clinical and self breast examination remain important in the era of modern screening,” *Annals of Surgical Oncology*, vol. 19, no. 3, pp. 1484–1490, 2012. DOI: 10.1245/s10434-011-2162-9.

[5] W. E. Barlow, C. D. Lehman, Y. Zheng, R. Ballard-Barbash, B. C. Yankaskas, G. R. Cutter, P. A. Carney, B. M. Geller, R. Rosenberg, K. Kerlikowske, D. L. Weaver, and S. H. Taplin, “Performance of diagnostic mammography for women with signs or symptoms of breast cancer,” *Journal of the National Cancer Institute*, vol. 94, no. 15, pp. 1151–1159, 2002. DOI: 10.1093/jnci/94.15.1151.

[6] I. K. Maitra, S. N. Samir, and K. Bandyopadhyay, “Technique for preprocessing of digital mammogram,” *Computer Methods and Programs in Biomedicine*, vol. 107, no. 2, pp. 175–188, 2012. DOI: 10.1016/j.cmpb.2011.05.007.

[7] L. L. Visser, E. J. Groen, F. E. van Leeuwen, E. H. Lips, M. K. Schmidt, and J. Wesseling, “Predictors of an invasive breast cancer recurrence after dcis: A systemic review and meta-analysis,” *Cancer Epidemiol Biomarkers Prevention*, vol. 28, no. 5, pp. 835–845, 2019. DOI: 10.1158/1055-9965.EPI-18-0976.

[8] M. V. Dieci, E. Orvieto, M. Dominici, P. Conte, and V. Guarneri, “Rare breast cancer subtypes: Histological, molecular and clinical peculiarities,” *The Oncologist*, vol. 19, no. 8, pp. 805–813, 2014. DOI: 10.1634/theoncologist.2014-0108.

[9] K. D. Miller, L. Nogueira, A. B. Mariotto, J. H. Rowland, K. R. Yabroff, C. M. Alfano, A. Jemal, J. L. Kramer, and R. L. Siegel, “Cancer treatment and sruvivorship statistics, 2019,” *American Cancer Society*, pp. 363–385, 2019. DOI: 10.3322/caac.21565.

[10] S. K. Plevritis, D. Munoz, and A. W. Kurian, “Association of screening and treatment with breast cancer mortality by molecular subtype ub us women, 2000-2012,” *JAMA Network Open*, vol. 319, no. 2, pp. 154–164, 2018. DOI: 10.1001/jama.2017.19130.

[11] C. E. Santis, J. Ma, M. Gaudet, K. Miller, A. G. Sauer, A. Jemal, and R. Siegel, “Breast cancer statistics, 2019,” *A Cancer Journal for Clinicians*, vol. 60, no. 6, pp. 438–451, 2019. DOI: 10.3322/caac.21583.

[12] N. D. Arvold, A. G. Taghian, A. Niemierko, R. F. A. Raad, M. Sreedhara, P. L. Nguyen, J. R. Bellon, J. S. Wong, B. L. Smith, and J. R. Harris, “Age, breast cancer subtype approximation, and local recurrence after breast-conserving therapy,” *Journal of Clinical Oncology*, vol. 29, no. 29, pp. 3885–3891, 2011. DOI: 10.1200/JCO.2011.36.1105.

[13] J. E. Panoff, J Hurley, C Takita, I. M. Reis, V. S. W Zhao, C. R. Gomez, M Jorda, L Koniaris, and J. L. Wright, “Risk of locoregional recurrence by receptor status in breast cancer patients receiving modern systemic therapy and post-mastectomy radiation,” *Breast Cancer Research and Treatment*, vol. 128, no. 3, pp. 899–906, 2011. DOI: 10.1007/s10549-011-1495-1.

[14] C. C. Earle, A. B. Nattinger, A. L. Potosky, K. Lang, R. Mallick, M. Berger, and J. L. Warren, “Identifying cancer relapse using seer-medicare data,” *Med Care*, vol. 40, no. 8, pp. IV–75–81, 2002. DOI: 10.1097/00005650-200208001-00011.

[15] E. B. Lamont, J. E. Herndon, J. C. Weeks, C. Henderson, C. C. Earle, R. L. Schilsky, N. A. Christakis, Cancer, and L. G. B, “Measuring disease-free survival and cancer relapse using medicare claims from calgb breast cancer trial participants (companion to 9344),” *J National Cancer Institution*, vol. 98, no. 18, pp. 1335–1338, 2006. DOI: 10.1093/jnci/djj363.

[16] J. Chubak, O. Yu, G. Pocobelli, L. Lamerato, J. Webster, M. N. Prout, M. U. Yood, W. E. Barlow, and D. S. M. Buist, "Administrative data algorithms to identify second breast cancer events following early-stage invasive breast cancer," *J National Cancer Institution*, vol. 104, no. 12, pp. 931–940, 2012. DOI: 10.1093/jnci/djs233.

[17] M. J. Hassett, D. P. Ritzwoller, N. Taback, N. Carroll, A. M. Cronin, G. V. Ting, D. Schrag, J. L. Warren, M. C. Hornbrook, and J. C. Weeks, "Validating billing/encounter codes as indicators of lung, colorectal, breast, and prostate cancer recurrence using two large contemporary cohorts," *Med Care*, vol. 52, no. 10, e65–e73, 2014. DOI: 10.1097/MLR.0b013e318277eb6f.

[18] D. E. Stewart, A. M. Cheung, S. Duff, F. Wong, M. McQuestion, T. Cheng, L. Purdy, and T. Bunston, "Attributions of cause and recurrence in long-term breast cancer survivors," *Journal of the Psychological, Social and Behavioural Dimension of Cnacer*, vol. 10, no. 2, pp. 179–183, 2001. DOI: 10.1002/pon.497.

[19] R. Kavuluru, I. Hands, E. B. Durbin, and L. Witt, "Automatic extraction of icd-o-3 primary sites from cancer pathology reports," *AMIA Summits on Translational Science Proceedings*, vol. 2013, p. 112, 2013.

[20] Z. Zeng, X. Li, S. Espino, A. Roy, K. Kitsch, S. Clare, S. Khan, and Y. Luo, "Contralateral breast cancer event dtection using natural language processing," *AMIA Annu Symp Proc*, pp. 1885–1892, 2017.

[21] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, Sep. 2010, ISSN: 1067-5027. DOI: 10.1136/jamia.2009.001560.

[22] G. Michalopoulos, H. Qazi, A. Wong, Z. Butt, and H. Chen, "Automatic extraction of risk factors for dialysis patients from clinical notes using natural language processing techniques," *Studies in Health Technologies and Informatics*, vol. 270, no. 1, pp. 53–57, 2020. DOI: 10.3233/SHTI200121.

[23] A. W. Forsyth, R. Barzilay, K. S. Hughes, D. Lui, K. A. Lorenz, A. Enzinger, J. A. Tulsky, and C. Lindvall, "Machine learning methods to extract documentation of breast cancer symptoms from electronic health records," *Journal of Pain and Symptom Managment*, vol. 55, no. 6, pp. 1492–1499, 2018. DOI: 10.1016/j.jpainsymman.2018.02.016.

[24] D. S. Carrell, S. Halgrim, D.-T. Tran, D. S. M. Buist, J. Chubak, W. W. Chapman, and G. Savova, "Using natural language processing to improve efficiency of manual chart abstraction in research: The case of breast cancer recurrence," *American Journal of Epidemiology*, vol. 179, no. 6, pp. 749–758, 2014. DOI: 10.1093/aje/kwt441.

[25] W.-H. Weng, K. B. Wagholikar, A. T. McCray, P. Szolovits, and H. C. Chueh, "Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach," *BMC Medical Informatics and Decision Making*, vol. 17, no. 1, p. 155, DOI: 10.1186/s12911-017-0556-8.

[26] A. R. Aronson and F.-M. Lang, "An overview of metamap: Historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, 2010. DOI: 10.1136/jamia.2009.002733.

[27] R. Reategui and S. Ratte, "Comparison of metamap and ctakes for entitiy extraction in clinical notes," *BMC Medical Informatics and decision Making*, vol. 18, no. 3, 2018. DOI: 10.1186/s12911-018-0654-2.

[28] R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015. DOI: 10.1161/CIRCULATIONAHA.115.001593.

[29] B. J. Mortazavi, N. S. Downing, E. M. Bucholz, K. Dharmarajan, A. Manhapra, S.-X. Li, S. N. Negahban, and H. M. Krumholz, "Analysis of machine learning techniques for heart failure readmissions," *Circulation: Cardiovascular Quality and Outcomes*, vol. 9, no. 6, pp. 629–640, 2016. DOI: 10.1161/CIRCOUTCOMES.116.003039.

[30] L. Breiman, J. Friedman, C. Stone, and R. Olshen, *Classification and Regression Trees*. Taylor & Francis, 1984.

[31] J. P. Bradford, C. Kunz, R. Kohavi, C. Brunk, and C. E. Brodley, "Pruning decision trees with misclassification costs," *Machine Learning: ECML-98, 1998*, vol. 1398, 1998. DOI: 10.1007/BFb0026682.

[32] J. R. Quinlan, *C4.5: Programs For Machine Learning*. Morgan Kaufmann Publishers, Inc., 1993.

[33] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997. DOI: 10.1006/jcss.1997.1504.

[34] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324.

[35] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002. DOI: 10.1016/S0167-9473(01)00065-2.

[36] K. Decker, P. Lambert, M. Pitz, and H. Singh, "Evaluation of algorithms using administrative health and structured electronic medical record data to determine breast and colorectal cancer recurrence in a canadian province," *PREPRINT*, 2020. DOI: 10.21203/rs.3.rs-87456/v1.

[37] F. E. Harrell, K. L. Lee, and D. B. Mark, "Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Stat Med*, vol. 15, no. 4, pp. 361–387, 1996. DOI: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

[38] K. A. Spackman, K. E. Campbell, and R. A. Côté, "Snomed rt: A reference terminology for health care.," *Proceedings of the AMIA annual fall symposium*, p. 640, 1997.

[39] T. S. Morton, "Coreference for nlp applications," in *Proceedings of the 38th annual meeting of the association for computational linguistics*, 2000, pp. 173–180.

[40] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 04, pp. 687–719, 2009. DOI: 10.1142/S0218001409007326.

[41] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Special issue on learning from imbalanced data sets," *ACM SIGKDD explorations newsletter*, vol. 6, no. 1, pp. 1–6, 2004. DOI: 10.1145/1007730.1007733.

[42] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002. DOI: 10.1613/jair.953.

[43] G. M. Weiss and F. Provost, "Learning when training data are costly: The effect of class distribution on tree induction," *Journal of artificial intelligence research*, vol. 19, pp. 315–354, 2003. DOI: 10.1613/jair.1199.

[44] H. Chen, G. Michalopoulos, S. Subendran, R. Yang, R. Quinn, M. Oliver, Z. Butt, and A. Wong, "Interpretability of ml models for health data - a case study," 2019.

[45] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you? explaining the predictions of any classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016. DOI: 10.1145/2939672.2939778.

[46] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.," *Association for Computing Machinery*, vol. 16, no. 3, 31–57, 2018, ISSN: 1542-7730. DOI: 10.1145/3236386.3241340.

[47] M. Du, N. Liu, and X. Hu, “Techniques for interpretable machine learning,” *Communications of the ACM*, vol. 63, no. 1, pp. 68–77, 2019. DOI: 10.1145/3359786.

[48] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, pp. 4765–4774, 2017.

[49] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, “Permutation importance: A corrected feature importance measure,” *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010. DOI: 10.1093/bioinformatics/btq134.

[50] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171–4186, 2019. DOI: 10.18653/v1/N19-1423.

[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010, 2017.